Міністерство освіти і науки України
Національний технічний університет
«Дніпровська політехніка»

Факультет інформаційних технологій
(факультет)

Кафедра системного аналізу та управління
(повна назва)

ПОЯСНЮВАЛЬНА ЗАПИСКА
кваліфікаційної роботи ступеня магістра

Студента Кочерги Вікторії Сергіївни
академічної групи 124м-23-1
спеціальності 124 Системний аналіз

на тему: «Застосування алгоритмів кластеризації для сегментації клієнтів банку»

Керівники	**Прізвище, ініціали**	**Оцінка за шкалою**		**Підпис**
		рейтинговою	**інституційною**	
кваліфікаційної роботи	*к.ф.-м.н., доц. Хом'як Т.В.*			
розділів:				
Інформаційно-аналітичний	*к.ф.-м.н., доц. Хом'як Т.В.*			
Спеціальний розділ	*к.ф.-м.н., доц. Хом'як Т.В.*			
Рецензент	*д.т.н., проф. Алексєєв М.О.*			
Нормоконтролер	*к.ф.-м.н., доц. Хом'як Т.В.*			

Дніпро

2024

ЗАТВЕРДЖЕНО:
завідувач кафедри
Системного аналізу та управління
(повна назва)

_____ *к.т.н., доц. Желдак Т.А.*
(підпис) (прізвище,ініціали)

«____» _____20___року

ЗАВДАННЯ
на кваліфікаційну роботу
ступеня *магістра*

студентці *Кочерзі В. С.* **академічної групи** *124м-23-1*
спеціальності: *124 Системний аналіз*
на тему *«Застосування алгоритмів кластеризації для сегментації клієнтів банку»*
затверджену наказом ректора НТУ «Дніпровська політехніка»
від *16.10.2024 р. №1388 - С*

Розділ	Зміст	Терміни виконання
1. Інформаційно-аналітичний розділ	*Розглянути поняття «кластеризація» та методи, що будуть використовуватися у кваліфікаційній роботі. Зазначити показники для оцінки точності отриманих результатів.*	08.09.2024 – 29.10.2024
2. Спеціальний розділ	*Ознайомитися з предметною областю; провести кластеризцію методом k-середніх; оцінити результати.*	30.10.2024 – 15.12.2024

Завдання видано _____ *доц. Хом'як Т.В.*
(підпис) (прізвище, ініціали)

Дата видачі: *06.09.2024 р.*

Дата подання до екзаменаційної комісії: *26.12.2024 р.*

Прийнято до виконання _____ Кочерга В. С.
(підпис студента) (прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка: 80 с., 37 рис., 6 табл., 4 додатки, 28 джерела.

Об'єктом дослідження виступають аплікаційні дані клієнтів АТ «Банк Кредит Дніпро», що отримані в ході анкетування.

Предметом дослідження є методи ітераційної кластеризації.

Метою даної кваліфікаційної роботи є застосування кластеризації для сегментації клієнтів АТ «Банк Кредит Дніпро» для виявлення неплатоспроможних клієнтів.

Методи дослідження: метод ітераційної кластеризації k-середніх.

В *інформаційно-аналітичному розділі* визначені поняття «кластер» і «кластеризація». Наведено характеристику методів кластерного аналізу. Розглянуто метрики для вимірювання подібності.

У *спеціальному розділі* проведено попередній аналіз предметної області, та наданих аплікаційних даних. Продемонстровано практичне використання методу кластеризації k-середніх, та визначено характеристики, що притаманні клієнтам з поганою кредитною історією.

Практична цінність отриманих результатів полягає в розбитті клієнтів на групи зі схожими характеристиками, з метою оцінки рівня клієнта, а також створення програмної реалізації методу k-середніх безпосередньо в СУБД.

Ключові слова: КЛАСТЕРИЗАЦІЯ, МЕТОД K-MEANS, СЕГМЕНТАЦІЯ, КЛАСТЕР, ІТЕРАТИВНА КЛАСТЕРИЗАЦІЯ, НАВЧАННЯ БЕЗ ВЧИТЕЛЯ.

ABSTRACT

Explanatory note: 80 pages, 37 figures, 8 tables, 4 appendices, 28 sources.

The object of the study is the application data of clients of JSC «Bank Credit Dnipro», obtained during the survey.

The subject of the study is the methods of iterative clustering.

The purpose of this qualification work is to use clustering to segment customers of JSC «Bank Credit Dnipro» to identify insolvent customers.

Research methods: k-means iterative clustering method.

The information-analytical section defines the concepts of «cluster» and «clustering». There are given the characteristic of cluster analysis methods. Metrics for measuring similarity are considered.

In a special section, a previous analysis of the subject area and the provided application data is carried out. The practical use of the k-means clustering method is demonstrated, and the characteristics inherent in customers with bad credit history are identified.

The practical value of the results lies in the division of clients into groups with similar characteristics in order to assess the level of the client, as well as the creation of a software implementation of the k-means method directly in a DBMS.

Keywords: CLUSTERING, K-MEANS METHOD, SEGMENTATION, CLUSTER, ITERATIVE CLUSTERING, UNSUPERVISED LEARNING.

ЗМІСТ

ВСТУП

Класифікація – це основний процес інтелектуальної діяльності людини. Дізнаючись щось нове, чи зіштовхуючись з якимось новим явищем, людина одразу починає зіставляти її з вже відомими їй знаннями. Саме цей процес і є класифікацією.

Кластерний аналіз вперше з'явився наприкінці 30-х років, проте активно досліджувати та вивчати дану дисципліну, почали наприкінці лише 60-х років.

Кластерний аналіз – це сукупність методів, які реалізують класифікацію якихось об'єктів чи явищ на основі вхідних змінних, що описують ці явища. В результаті застосування кластерного аналізу, утворюються групи схожих між собою кластерів.

Кластеризація може застосовуватися в різних сферах та областях. Наприклад, в області медицини, для класифікації захворювань чи симптомів тощо. Тобто кластеризацію можна застосовувати будь-де, де постає задача організувати дані у якусь структуру. За допомогою кластеризації можна вирішити ряд деяких задач: поглибити знання щодо даних, які досліджуються; перевірити вже висунуті припущення щодо наявності деякої структури в даних; або навпаки побудувати таку структуру для малознайомих даних.

Переважна більшість алгоритмів кластеризації працюють, підпорядковуючись деяким етапам: задати початкове розбиття об'єктів на групи; ітераційно переносити об'єкти з одного кластеру в інший, доки критерій якості не перестане поліпшуватися.

Дана кваліфікаційна робота присвячена розробці збереженої процедури на SQL Server для виконання кластеризації методом k-середніх. Предметною областю, яка надає дані для реалізації алгоритму, виступає АТ «Банк Кредит Дніпро».

Суть роботи полягає в групуванні вибірки клієнтів на класи, з метою виявлення клієнтів, що більш схильні до неповернення кредиту. Розроблена процедура на базі SQL, призначена для того аби швидко можна було звертатися

до вибірки даних, та при оновленні таблиці, автоматично сегментувати клієнтів. Отриманий розв’язок слугує гіпотезою, яку в подальшому необхідно детально перевірити, з метою прийняття рішення щодо того, чи дійсно отримані характеристики впливають на платоспроможність клієнта.

РОЗДІЛ 1 ІНФОРМАЦІЙНО-АНАЛІТИЧНИЙ

1.1 Кластерний аналіз

Кластерний аналіз – це процедура, що полягає у розподілі множини об’єктів на підмножини, що не перетинаються та називаються кластерами. За правилом, кластер налічує в собі схожі об’єкти, проте самі кластери мають відрізнятися один від одного. Взагалі, задачу кластеризації можна прирівняти до задачі класифікації, проте вони принципово відрізняються одна від одної. Під час класифікації кожний з об’єктів відноситься до одного із заздалегідь відомих класів, а в кластеризації – до одного з невизначених класів [1, 21].

Вперше термін «кластер» з’явився в 1939 році, що і вважається початком зародження кластерного аналізу, проте активного розвитку та широкого застосування набуває наприкінці 60-х – початку 70-х років. Спочатку методи кластеризації застосовувалися в психології, археології, біології, але з появою та розвитком ЕОМ, особливо персональних комп’ютерів, їх використання поширилося на області соціології, економіки та статистики. Велике значення приділяється розвитку методів кластерного аналізу, бо вони дають можливість побудувати науково обґрунтовану класифікацію, виявити зв’язки між об’єктами, які досліджуються. В області статистики, при наявності великого потоку інформаційних даних, які постійно збільшуються, методи кластерного аналізу застосовують як інструмент для стиснення інформації, що допомагає при аналізі, наприклад, економічної діяльності різних груп підприємств чи країн тощо [24].

Так як основна задача кластерного аналізу – це виділення компактної групи об’єктів. Тому його можна застосовувати в різних сферах, наприклад, біологи застосовують кластерний аналіз для розбиття тварин на різні види задля опису відмінностей між ними, в маркетингу – сегментація клієнтів, в

менеджменті – розбиття працівників на групи тощо, в медицині – для розподілу захворювань [15].

Кластеризація є одним із напрямів класичного машинного навчання без вчителя. Розділяють два види класичних алгоритмів машинного навчання: навчання з та без вчителя. Під час навчання з вчителем, «вчитель» заздалегідь розподіляє всі дані, а обчислювальна машина вчиться на цих прикладах, тобто вчитель допомагає комп'ютерній системі вивчати предмети. За навчання без вчителя, на вхід подаються дані та задача визначити їх схожість за певними ознаками. Задача комп'ютерної системи – самостійно виявити всі закономірності, так як вчитель відсутній. Навчання без вчителя більше несе в собі аналітичний характер, а не використовується як основний алгоритм [19].

Перед проведенням кластерного аналізу, користувач має врахувати деякі його особливості, та на основі цього вирішувати чи дійсно дана процедура підходить та є доречною. По-перше, на даний момент, алгоритми кластерного аналізу досі не мають статистичного обґрунтування, тобто методи кластеризації мають евристичний характер. По-друге, деякі методи розроблені під потреби деякої наукової дисципліни, тому перед використанням методу, необхідно дослідити особливості щодо вимоги до форми подання даних тощо. По-третє, при застосуванні різних методів кластеризації, вони можуть надавати абсолютно різні результати [19].

Незалежно від цілі задачі чи обраного методу, алгоритм застосування кластерного аналізу можна описати наступними кроками:

1) Формування вибірки даних.
2) Формування множини ознак, що будуть використовуватися в процесі кластеризації.
3) Обчислення міри подібності.
4) Оцінка отриманих результатів [26].

1.2.1 Кластеризація. Методи кластеризації

Формально задачу кластеризації можна описати наступним чином: є множина об’єктів $I = \{i_1, i_2, \dots, i_n\}$, кожний з яких має вектор параметрів $x_j = \{x_{j1}, x_{j2}, \dots, x_{jm}\}$. Кластеризація полягає у побудові відображення множини I на множину кластерів C: (1.1)

$$I \to C, C = \{c_1, c_2, \dots, c_k\},$$

де c_k – кластер з подібними об’єктами множини I:

$$c_k = \{i_j, i_p \mid i_j \in I,\ i_p \in I \text{ та } d(i_j, i_p) < \sigma\}, \quad (1.2)$$

де σ – величина, що визначає ступінь подібності для включення об'єктів до кластеру,

$d(i_j, i_p)$ – відстань між кластерами [21].

Якщо $d(i_j, i_p)$ (міра подібності об'єктів) менше за σ, значить об'єкти схожі, тому їх слід розмістити в одному кластері. Кластер – це група об’єктів, які мають спільні ознаки. Кластер має певні властивості, серед них: щільність, дисперсія, розмір, форма та віддільність [8].

Ознака щільності розглядає кластер як групу точок у просторі, яка може бути щільніше в порівнянні з іншими областями, складатися з невеликої кількості точок, або взагалі їх не налічувати. Дисперсія, як і в статистиці, визначає ступінь розкидання точок у просторі навколо центру кластера, проте кластери не завжди являють собою багатовимірну нормальну популяцію, тому дисперсію слід розуміти, як величину, що вимірює наскільки близько один до одного розташовані точки кластеру. З цього твердження, випливає, що кластер вважається щільним, якщо всі точки знаходяться близько до його центру, і не щільним – при розкиданні точок навколо центру. Розмір кластеру напряму пов’язаний з поняттям «дисперсії», бо якщо кластер можна ідентифікувати, отже можна виміряти його радіус. Звичайно, дана характеристика має сенс тільки тоді, коли кластер має круглу форму. Формою являється положення точок у просторі. Зазвичай за формою кластери нагадують гіперсферу чи еліпсоїд, проте мають місце і інші форми, наприклад, подовжені (при такій формі виміряти радіус немає сенсу, проте можна розрахувати зв’язність точок,

тобто відносну відстань між ними). Віддільність означає наскільки кластери перекриваються, або іншими словами, наскільки далеко вони розташовані один від одного. Вони можуть бути відносно близько один до одного, що буде розмивати їх кордони, або навпаки, розділятися широким пустим простором [8, 26].

Мету кластеризації можна описати декількома пунктами:

- Визначення структури множини об’єктів, розділивши її на групи схожих об’єктів.
- Скорочення об’єму даних, залишивши по одному типовому представнику з кожного кластера.
- Виявлення об’єктів, які суттєво відрізняються від типових представників кластеру.
- Статистичний аналіз отриманих кластерів [9].

Розв’язок задачі кластеризації не є однозначним, що пояснюється рядом причин. По-перше, не існує критерію якості кластеризації; багато алгоритмів не мають чітко вираженого критерію, проте проводять якісну кластеризацію, і звичайно кожний з них дає різний результат. По-друге, за замовчуванням залишається невідомим число кластерів, що як правило визначається суб’єктивно, за думкою експерта. По-третє, результат кластеризації залежить від обраної міри подібності, що також є суб’єктивною [13].

Згрупувавши усі методи кластеризації, можна виділити 7 груп:

1) Ієрархічні агломеративні методи;
2) Ієрархічні дивізимні методи;
3) Ітеративні методи групування;
4) Методи пошуку модальних значень щільності;
5) Факторні методи;
6) Методи згущень;
7) Методи на основі теорії графів [26].

Найбільш часто використовуються ієрархічні агломеративні методи, до складу яких входять методи одиничного зв’язку, повного та середнього

зв’язків, і метод Уорда. Агломеративні методи володіють деякими особливостями: матриця подібності має бути розміром $N \times N$, де N – число об’єктів; послідовність застосування методу можна зобразити візуально у вигляді деревоподібної діаграми – дендрограми, де кожний крок, на котрому об’єднувалися об’єкти, зображується у вигляді гілки цього дерева. На рисунку 1.1 зображений приклад дендрограми, яка показує ієрархічну структуру зв’язків між 11-ю точками даних, де на найнижчому рівні всі точки незалежні, на найвищому – об’єднані в одну велику групу [26].

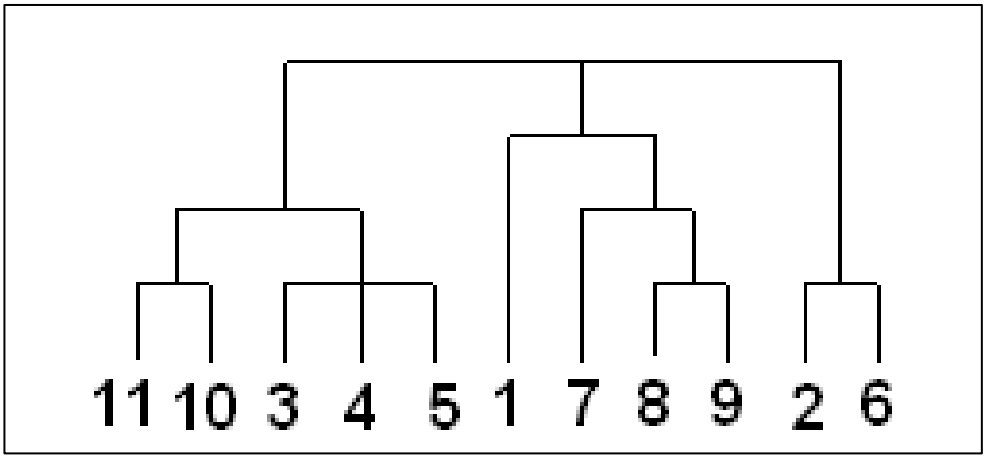

Рисунок. 1.1 – Дендрограма.

Наступна особливість полягає у тому, що для матриці розміром $N \times N$, повна кластеризація виконується за $N - 1$ кроків. Агломеративні методи не потребують глибоких знань та розумінь матричної алгебри, так як всі кроки підпорядковуються правилу, за яким об’єкти об’єднуються у кластери. Результатом застосування цих правил є кластери, які не перекриваються, проте являються вкладеними в тому сенсі, що кожний кластер може бути елементом іншого, більшого кластеру на вищому рівні подібності. До недоліків методів можна віднести те, що їх застосування можливе тільки за умови обчислення матриці подібності, що для великого числа елементів буде потребувати зберігати матрицю великого розміру, що буде витрачати купу часу на її переогляд при розрахунках [26].

Метод одиничного зв’язку будується на пошуку двох найбільш схожих об’єктів в матриці подібності. За правилом, об’єкт записується у кластер, якщо він має найвищу ступінь збіжності з одним із представників кластеру, тобто хоча б один елемент з кластеру знаходиться на одному й тому ж рівні подібності, що і елемент, який розглядається. Звідси, даний метод реалізує об’єднання лише за наявністю одного зв’язку між об’єктом і кластером. До недоліку методу можна віднести формування кластерів подовженої форми (у вигляді ланцюжка) [26].

Метод повного зв’язку являється протилежністю минулому методу, так як за правилом, ступінь збіжності з одним із представників кластеру має бути не менше заданого граничного значення [26].

За методом середнього зв’язку, відбувається розрахунок середнього значення ступеню подібності об’єкта з кожним представником кластеру. Серед отриманих значень, знаходиться те, яке більше або дорівнює заданому граничному рівню, що і буде відповідати за включення об’єкту до кластеру. Зазвичай, розраховується саме середнє арифметичне між об’єктами кластеру та потенційним об’єктом на включення, проте, зустрічаються варіації методу, де розраховується подібність між центрами ваг двох кластерів, які хочуть об’єднати [26].

В методі Уорда лежить в основі оптимізація мінімальної дисперсії в кластері. Цільова функція задачі відома як сума квадратичних відхилень (СКВ). На першому кроці, кожний кластер складається з одного об’єкта, тому СКВ дорівнює 0. Далі об’єднання в групи відбувається для об’єктів, у яких СКВ має мінімальний приріст. За спостереженнями, метод схильний формувати кластери приблизно однакових розмірів, які мають гіперсферичну форму [26].

Не зважаючи на те, що на даний момент достатньо програмних засобів, які будуть здійснювати кластерний аналіз та самостійно виводити дендрограму, важливо також розуміти принцип її побудови. Побудова дендрограми не вимагає сильних зусиль та якихось додаткових розрахунків, так як для її побудови використовуються результати кластеризації, а саме, значення

мінімальних відстаней на кожній ітерації об’єднання. Процедуру побудови дендрограми можна описати наступними кроками:

1) Побудова початково ескізу. Спочатку зображується «основа дерева», що складається з останнього об’єднання, та позначається точкою з його номером.
2) З основного вузла проводяться дві «гілки», що відповідають точкам наступних груп.
3) У разі, якщо гілка закінчена, то її кінець позначається номером об’єкта, в протилежному випадку – вказується відповідний номер ітерації.
4) Коли всі гілки помічені, слід скоригувати масштабування дендрограми. На осі ординат зображуються значення мінімальних відстаней, об’єкти з групами, а довжині гілки відповідає значення мінімальної відстані. По осі абсцис відображаються об’єкти вибірки [19].

На рисунку 1.2 зображено принцип роботи ієрархічних агломеративних методів кластеризації.

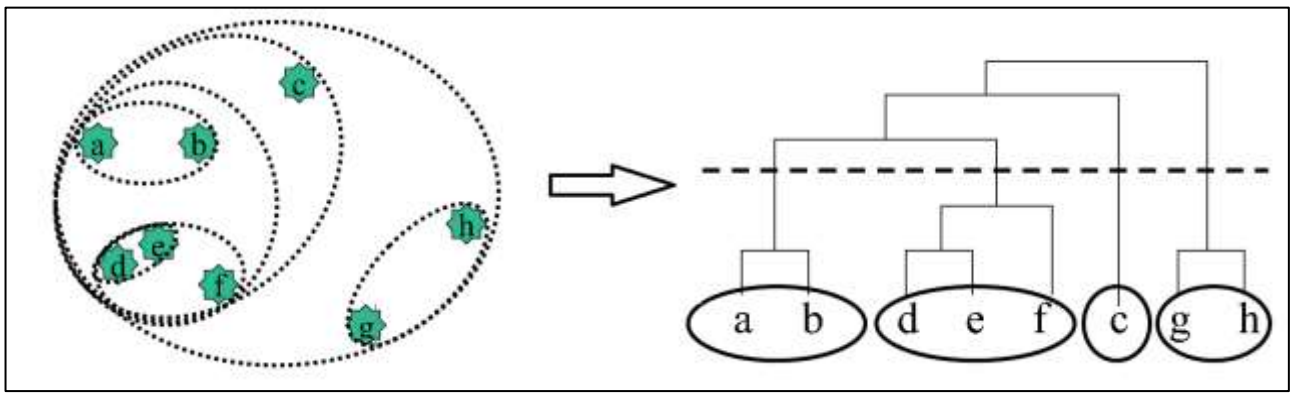

Рисунок. 1.2 – Ієрархічний метод кластеризації.

Алгоритм WaveCluster відноситься до групи ієрархічних методів, та працює через накладання на простір даних багатовимірної решітки. Під час роботи алгоритму аналізуються тільки узагальнені характеристики точок, що знаходяться в одній клітинці решітки. На наступних кроках застосовується хвильове перетворення для визначення кластерів. Проте алгоритм має як недоліки, так і переваги, до переваг відноситься відсутність чутливості методу

до шумів та здатність розпізнавати кластери різних форм. До недоліків відноситься складність реалізації алгоритму та те, що його можна застосовувати для даних невеликої розмірності [18].

Ітеративні методи групування не отримали широкого застосування на відміну від агломеративних методів. Всі вони працюють за однаковим алгоритмом:

1) Початкове розбиття даних на задану кількість кластерів;
2) Обчислення центроїдів отриманих кластерів;
3) Віднесення кожної точки до кластеру з найближчим центроїдом;
4) Обчислення нових центроїдів кластерів. Кластери не замінюються на нові, поки не передивиться весь набір даних;
5) Кроки 3 та 4 повторюються, доки кластери не перестануть змінюватися [26].

На відміну від агломеративних методів, ітеративні не потребують обчислення матриці подібності, а працюють з первинними даними, таким чином ці методи здатні обробляти великі масиви даних. Ще одна перевага цих методів полягає в тому, що вони мають повторний цикл переогляду даних, що дає змогу компенсувати наслідки неякісного первинного розбиття. Також, в результаті, кластери утворюються з одним і тим самим рангом, що не являються вкладеними, і таким чином не утворюють ієрархію та не допускають перекриття [26].

Для ітеративних методів притаманні деякі особливості, які важливо враховувати перед початком кластеризації. Серед них початкове розбиття, тип ітерації та статистичний критерій [26].

Щоб розпочати ітеративний процес необхідно або вказати початкові точки, або задати початкове розбиття. Початковими точками будуть служити центри кластерів (центроїди), при першій ітерації точки об’єктів будуть приписувати до тих центроїдів, до яких ближче всього знаходяться. Якщо вказується початкове розбиття, то точки відносяться до кластерів, і їх центроїди визначаються середніми арифметичними серед віднесених об’єктів. Для

ініціалізації початкового розбиття можна використовувати розбиття, отримане в результаті ієрархічної кластеризації, або задаватися випадковим чином [26].

Тип ітерації визначає, яким чином об’єкт буде потрапляти до кластеру. Існує два види ітерацій: за принципом k-середніх або «сходження на пагорб». Ітерації за k-середнім полягають в переміщенні об’єкта до кластеру з найближчим центроїдом, перерахунок центрів кластерів може виконуватися або після кожного додавання об’єкта, або після перегляду всієї вибірки. В ітераціях «сходження на пагорб» об’єкти не просто записуються до кластеру на основі відстані, а по оцінці чи буде переміщення об’єкта оптимізувати значення заданого статистичного критерію [26].

Статистичними критеріями, що оптимізуються в ході роботи методу, що засновані на «сходженні на пагорб», можуть використовуватися одні із функцій якості кластеризації: $trW, trW^{-1}B, detW$, тут W – об’єднана коваріаційна матриця об’єктів всередині кластера, B – об’єднана коваріаційна матриця об’єктів поза межами кластера. Ці критерії служать для виявлення однорідності кластерів в багатовимірному просторі. Критерій trW буде знаходити кластери гіперсферичних форм, при $detW$ – важливо враховувати, що його використання призведе до утворення кластерів, які мають одну й ту саму форму [26].

Головний недолік ітеративних методів – знаходження субоптимального розв’язку. Існує ймовірність, що в результаті кластеризації буде отримано локальний оптимум, а не глобальний, що пов’язано з тим, що ітеративні методи здатні розпізнати тільки невелику частину серед всіх можливих розбиттів об’єктів. Також головною причиною є погане початкове розбиття даних, наприклад, метод k-середніх дуже чутливий до початкового розбиття, якщо воно обрано довільним шляхом, то це тільки збільшить шанси отримати не глобальний розв’язок. І хоча немає об’єктивної оцінки чи є рішення глобальним оптимумом, слід використовувати методи, який взаємодіє з процедурою перевірки результату [26].

Також метод k-середніх чутливий до шуму, що можуть спотворювати середнє значення, щоб уникнути даної проблеми, можна використовувати модифікований метод k-середніх – метод k-медіан [18].

Вхідними даними методу k-медіан є кількість кластерів та матриця відстаней. Методи k-середніх та k-медіан мають однакові кроки реалізації, тільки різниця полягає в розрахунку відстані до точки. В k-середніх розраховуються відстані між центром кластера та об'єктом на входження, в k-медіан – знаходиться серединна точка кожного кластеру, потім відстань від точки об'єкта до медіани. Обидва методи дають близькі результати для одного й того самого набору даних [19].

Наступний ітеративний метод – алгоритм expectation-maximization (EM). Цей алгоритм застосовують, щоб оцінити максимальну схожість параметрів ймовірних моделей, у тому випадку, якщо вона залежить від не досліджуваних змінних. Алгоритм EM не чутливий до викидів, швидко збігається, проте як і метод k-середніх вимагає задання кількості кластерів, що обумовлюється наявністю апріорних знань про дані. Ітераційний процес складається з двох кроків:

- Крок «expectation». Розрахунок очікуваного значення функції подібності зі змінних, що є прихованими, але для розрахунку вважаються як досліджувані.
- Крок «maximization». Розрахунок оцінки максимальної схожості, що збільшує очікувану схожість, розраховану на кроці 1. Розраховане значення використовується на наступному кроці «expectation» [18].

Алгоритм PAM (partitioning around medoids) – модифікація методу k-середніх на основі k-медіани. Алгоритм не такий чутливий до шумів, на відміну алгоритму k-середніх, так як вплив викидів на медіану мінімальний. PAM слід застосовувати на невеликому наборі даних, та ефективніше замість вихідної множини ознак використовувати узагальнені критерії, що можна отримати за допомогою факторного аналізу, що використовується для стискання даних [18].

Алгоритм FOREL дещо схожий з методом k-середніх, проте вони все-таки відрізняються за принципом роботи. На початку обирається одна точка в просторі, спираючись на інтуїцію чи просто випадковим чином, потім розраховуються відстані від обраної точки до всіх інших точок. Отримані відстані заносяться до матриці в упорядкованому вигляді, що буде використовуватися для встановлення фіксованого радіуса сфери, що буде становити границю кластера. Радіус сфери обирається довільно, наприклад, за допомогою принципу потрапляння в сферу деякої кількості точок, зазвичай це одна третина від загальної кількості об’єктів. Маючи встановлений радіус сфери, всі об’єкти, що знаходяться в ній, формують кластер. На наступному ході обирається нова точка сформованого кластеру, що стає центром іншої сфери фіксованого радіусу. Для вибору координат нового центроїда доцільно використовувати якийсь критерій, наприклад, мінімальну відстань від обраної точки до границі сфери. Таким чином, нова сформована сфера буде включати в себе об’єкти з попередньої сфери, окрім нових. Процедура повторюється, поки в сфери перестають потрапляти нові точки [10].

На рисунку 1.3 зображена графічна інтерпретація алгоритму FOREL.

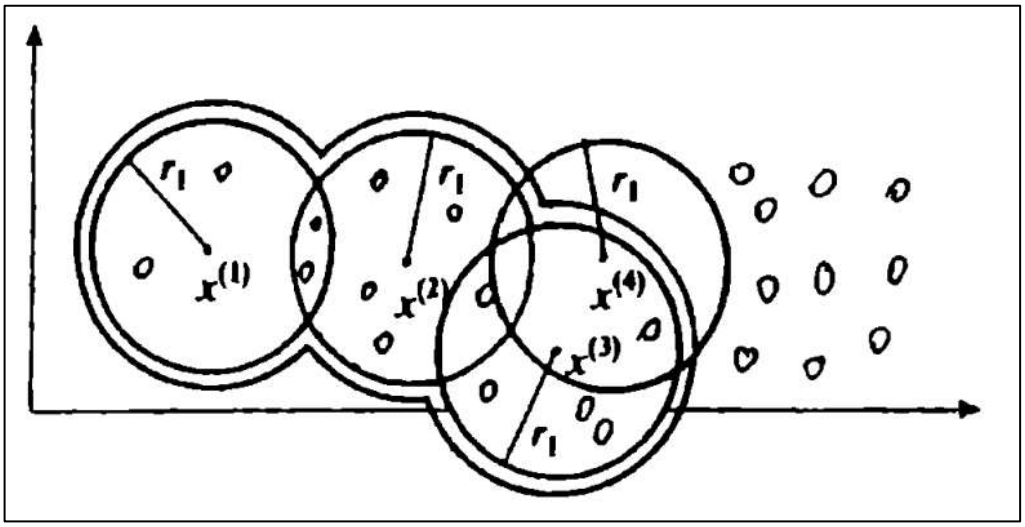

Рисунок. 1.3 – Графічна інтерпретація алгоритму FOREL.

Якість кластеризації алгоритмом FOREL залежить від раціональності обраних центрів сфер та радіусу пошуку. Якщо простір ознак має два чи три

виміри, то оцінити вхідні параметри можна візуально, та за необхідності скоригувати, при багатовимірному просторі вхідні параметри обираються всліпу, і при невдачах необхідно повторювати алгоритм [10].

Алгоритм FOREL має деякі недоліки: відсутність автоматичних критеріїв якості проведеної кластеризації, так як алгоритм припиняє роботу при першому пустому кластері в побудованій сфері; границі кластерів не завжди можуть мати явні функції їх відокремлення [10].

Застосування факторних методів починається з формування кореляційної матриці подібності між об’єктами, за допомогою якої об’єкти відносяться до кластерів в залежності від їх факторних навантажень. Ієрархічні дивізимні методи протилежні агломеративним за логікою, а саме: на початку розрахунків всі об’єкти належать одному кластеру, а потім цей кластер розбивається на послідовно менші групи. На рисунку 1.4 зображено дендрограму для агломеративного та дивізимного методу [26].

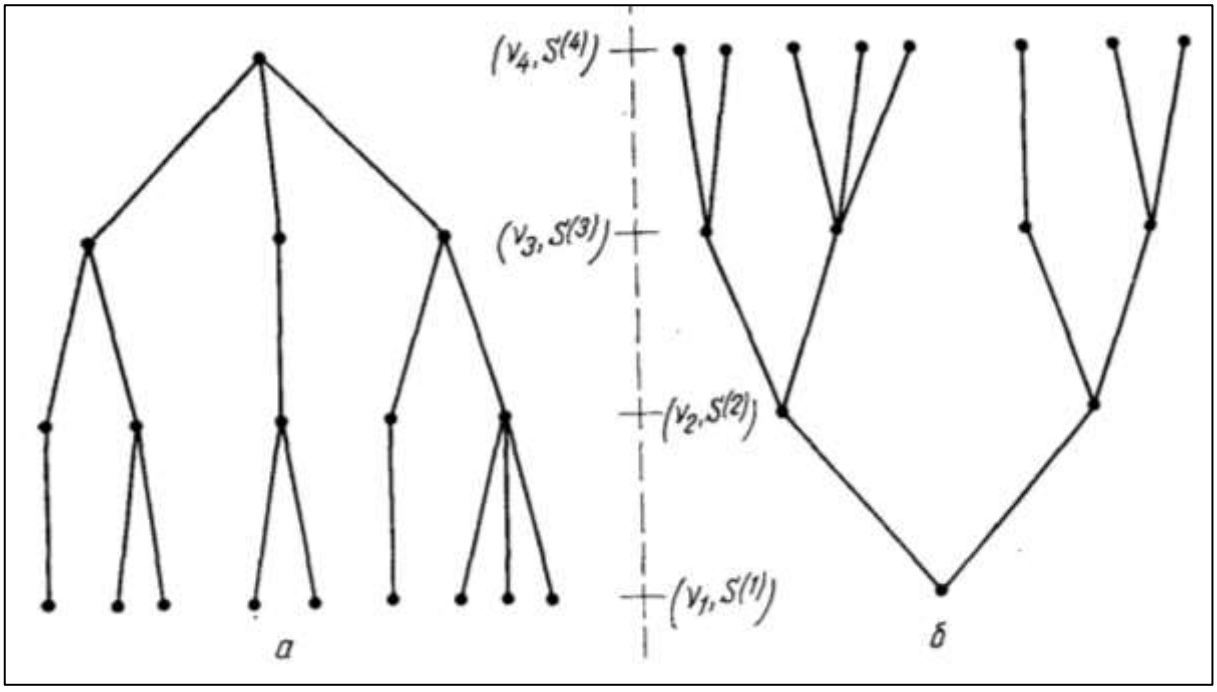

Рисунок. 1.4 – Дендрограма ієрархічного а) агломеративного б) дивізимного методів

Дивізимні методи бувають двох видів: з монотетичним та політетичними кластерами. Монотетичний кластер складається з об’єктів, які мають приблизно одне й те саме значення деякої ознаки, таким чином такі кластери описуються

фіксованими ознаками, значення яких необхідні при віднесенні об’єкта до кластеру. Політетичні кластери являються протилежністю монотетичним, оскільки для віднесення до кластеру достатньо наявності збігу з деякої множини ознак [26].

Методи пошуку модальних значень щільності представляють кластер як область простору з «високою» щільністю точок в порівнянні з навколишніми областями. Вони передивляються простір в пошуках скупчення даних, які являють собою області високої щільності. Методи пошуку модальних значень щільності, що засновані на кластеризації за принципом одиничного зв’язку, перешкоджають утворенню ланцюжків, та працюють за правилом: краще створити новий кластер, ніж приєднати об’єкт до вже створеної групи [26].

Найпоширенішим методом щільності є DBSCAN (density-based spatial clustering of applications with noise), що на вході отримує радіус околиці та мінімальну кількість сусідів. Метод має деякі ознаки та правила: околиця об’єкта (околиця радіуса об’єкту); кореневий об’єкт (якщо його околиця містить не менше заданого мінімального числа об’єктів); об’єкт 1 щільно-досяжний з об’єктом 2, якщо об’єкт 1 знаходиться в околиці об’єкта 2, при чому об’єкт 2 – кореневий; об’єкт 1 щільно-з’єднаний з об’єктом 2, при заданих околицях та мінімального числа точок, якщо є об’єкт 3, з яким інші об’єкти щільно-досяжні [18].

Пошук кластерів за алгоритмом DBSCAN здійснюється шляхом перевірки околиці кожного об’єкта. Якщо деяка околиця має більшу кількість точок, ніж задано, тоді створюється новий кластер з даний кореневим об’єктом. На наступному кроці, ітеративно збираються усі об’єкти серед кореневих, які щільно-досяжні, що можуть спряти об’єднанню щільно-досяжних кластерів. Алгоритм зупиняється, якщо до кластерів вже не можна віднести новий об’єкт [18].

Наступна група методів пошуку модальних значень – методи з виявлення параметрів суміші розподілів. Суміш являє собою сукупність вибірок, що представляють різні популяції об’єктів. Даний підхід до кластеризації

заснований на статистичній моделі, елементи якої з різних груп повинні мати різні ймовірнісні розподіли. Ціль такої кластеризації полягає у виявленні параметрі, що описують розподіл популяції [26].

Методи згущення, на відміну від ієрархічних методів, не породжують ієрархічну структуру, хоча і дозволяють створювати перекриваючі кластери, і об’єкти можуть одночасно знаходитися в декількох кластерах. Застосування методу згущення потребує розрахунку матриці подібності між об’єктами, та оптимального значення статистичного критерію – функції згуртованості. Потім об’єкти переміщуються, доки функція не досягне оптимального значення. Недоліком таких методів є те, що час від часу може виявлятися один і той самий кластер, що не буде надавати нову інформацію [26].

В залежності від методу кластеризації, можуть утворюватися кластери різних структур. На рисунку 1.5 наведені приклади різних кластерних структур [8].

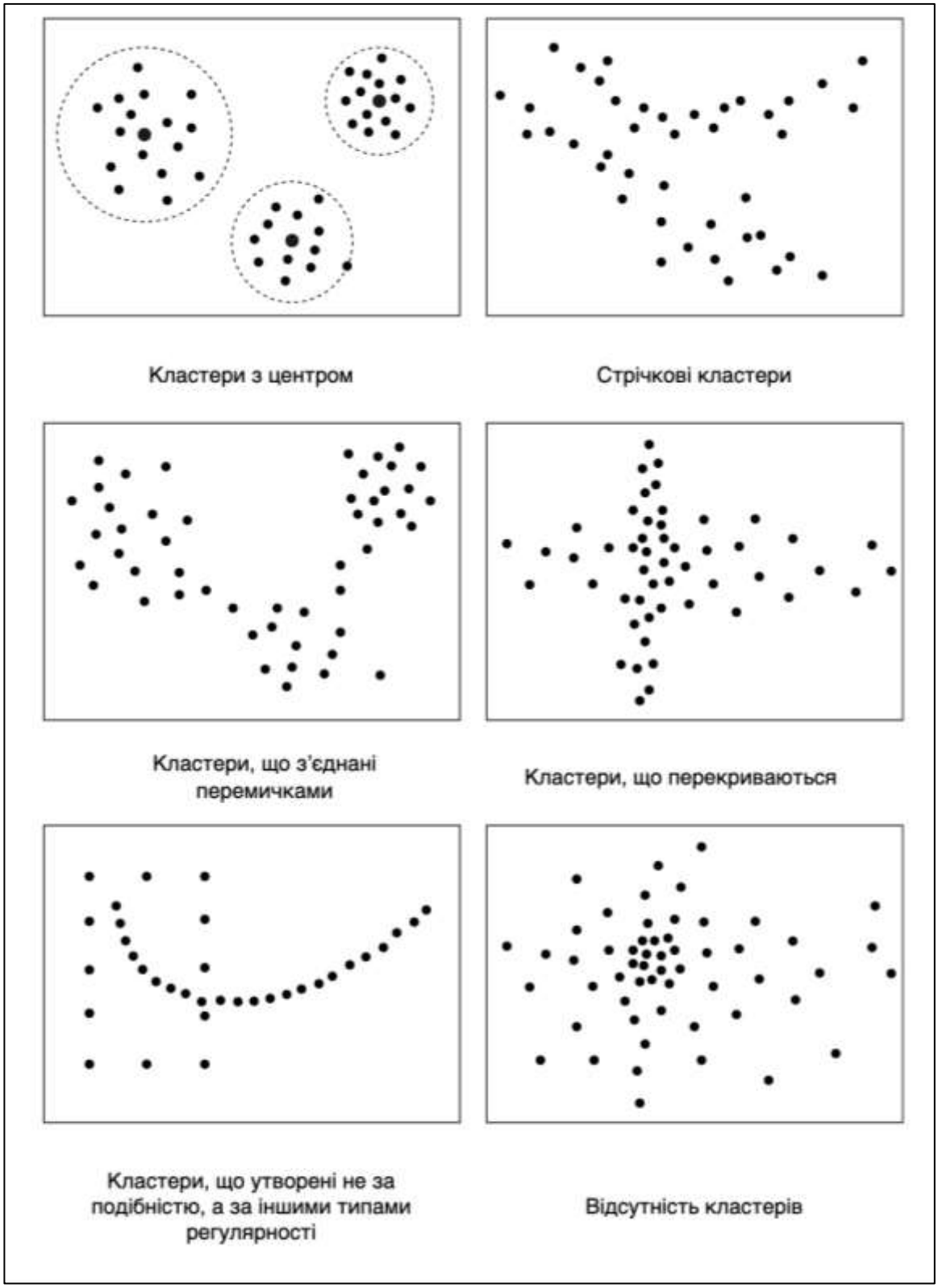

Рисунок. 1.5 – Типи кластерних структур.

На роботу методів кластеризації впливають деякі фактори: характеристики кластерної структури, ступінь перекриття кластерів, обрана міра подібності та наявність шуму [26].

До характеристик кластерної структури, що впливає на роботу методів, відноситься форма та розмір кластера, та їх кількість. Метод Уорда та ітеративні методи в результаті утворюють кластери гіперсферичної форми, тому такі методи ще відносять до таких, що розширюють простір. Окрім цього, дані методи утворюють кластери однакових розмірів. Якщо необхідна повна класифікація даних, тобто кожен об’єкт має бути віднесений до кластера, і при цьому дані не мають багато викидів, то метод Уорда ідеально буде підходити для відновлення кластерної структури. Перекриття кластерів аналогічна проблемі наявності викидів. Перекриття означає ступінь, з якою кластери займають один й той самий простір. Кластери можуть бути або добре розподілені в просторі, або навпаки, знаходитися близько один до одного, а також має сенс поняття шумових точок, тобто таких, що знаходяться між границями кластерів [26].

Перед початком кластеризації необхідно підготувати та дослідити інформативні дані. Це можна виконати за допомогою формування таблиці «об’єкт-ознака» (для агломеративних методів) та процедури нормування. Таблиця «об’єкт-ознака» являється типовою формою представлення даних, вона вміщує в себе кількісні та якісні значення властивостей (довжина, колір, ціна тощо), які описують об’єкт (люди, райони тощо). Кожен такий об’єкт може мати нескінченне число ознак, якими можна його описати, проте, в залежності від мети дослідження, не всі характеристики є істотними, тому слід обирати найбільш важливі для дослідження ознаки. Саме тому для спрощення інформація про об’єкти подається у формі таблиці, де в першому стовпчику перелічено усі об’єкти, а в наступних – ознаки та їх значення для кожного об’єкта [19].

Вибір змінних, по яким буде проводитися кластеризація, являється важливим етапом, так як це напряму залежить від того, чи можна буде оцінити подібність об’єктів, чи буде утворюватися структура. Бо кластерний аналіз саме для того і використовується, щоб отримати об’єктивну розбивку об’єктів, тому при виборі всіх змінних із вхідної вибірки, зростає ризик того, що дані не

зможуть утворювати об’єднуватися в структури незалежно від кількості об’єктів [26].

Повнота опису досліджуваних об’єктів може складатися з максимальної кількості об’єктів. І зрозуміло, що чим більше характеристик, тим точніше об’єкт описується в деякому просторі параметрів, проте, така точність може шкодити при класифікації. По-перше, через багатовимірний опис об’єкту може ускладнюватися обчислювальний процес, та робити об’єкти незіставними. По-друге, кожна задача потребує обмежень числа характеристик, бо багато характеристик можуть бути несуттєвими для розуміння всієї суті, тому такі зайві змінні будуть тільки спотворювати загальну картину дослідження та розмивати границі між кластерами. По-третє, динаміка зміни різних показників може бути корельована між собою, що не допускається при статистичному дослідженні. Таким чином, перед початком дослідження, всі характеристики об’єкту необхідно перевірити на рівень мінімальної статистичної залежності. Якщо об’єкт налічує i ознак, то розрахувати коефіцієнт попарної кореляції необхідно для $i * (i - 1)$ пари ознак. Якщо значення отриманої кореляції більше 0.7, то змінні мають високий зв’язок, тому одну з них необхідно виключити з дослідження. По-четверте, треба враховувати причинно-наслідковий зв’язок між ознаками, що являються причинами деякий явищ, тоді їх виключення з дослідження недопустиме [10].

Дотримування перелічених правил щодо кількості ознак для дослідження, сприяють збільшенню точності кластеризації. Процес відбору характеристик поділяється на два етапи: змістовний відбір та формальний аналіз відібраних змінних, іншими словами застосування математичного апарату відповідно до заданої задачі [10].

Простір з ознаками називається гіперпростіром, якщо їх кількість більше трьох, та системою координат – при двох ознаках. Кожний показник являє собою віссю координат, до яких можна віднести не тільки характеристики, що відповідають аксіомі Евкліда, але і якісні, які не виміряються числами. При високій розмірності простору, стає неможливим зобразити їх графічно, та і

взагалі заважають зосередитися на головних характеристиках, так як вся увага буде розсіюватися серед дрібних деталей. Але одночасно і дуже мала розмірність не дає змістовних результатів, так як результат кластеризації буде тривіальним [10].

Щоб досягти компромісу при процедурі вибору кількості характеристик, можна вводити інтегральні змінні. Узагальнити змінні можна по об'єктивним ознакам: подібність елементів управління, ідентичність внутрішньої структури об'єктів, однаковість природи виникнення тощо. Інтегральні змінні допомагають зменшити розмірність простору ознак, при цьому зберігаючи різноманіття характеристик [10].

Розрахунок інтегральних змінних становить окрему частину дослідження, успішний розв'язок якої залежить від інтуїції спеціаліста, розумінні змісту дослідження. Для кластерного аналізу ефективним вважається розподіл характеристик на дві інтегральні групи. Таким чином можна буде графічно інтерпретувати розв'язок та скоротити час та зусилля на обчислення. Розподіл на групи може базуватися на точних та наближених вимірювань, об'єктивних та оціночних характеристик тощо. Для кожної інтегральної характеристики ставиться у відповідність вісь координат, а границі кластерів розраховуються геометричними методами. Для більш громіздких задач, можна використовувати тривимірні інтегральні оцінки. В таких випадках кластер необхідно зображувати об'ємно [10].

Для виявлення оптимальної розмірності простору ознак, можна застосовувати логічні процедури необхідності та достатності. Процедури слід починати з аналізу кожного параметру на достатність в контексті розв'язувальної задачі. Логіка дослідження допомагає спеціалісту з'ясувати чи існує зв'язок між параметром та цільовою функцією задачі. На наступному кроці виключаються зайві ознаки, що можуть являтися або малозначущими параметрами, або взаємно корельовані ознаки, або ти, що не змінюються в ході експерименту. Якщо не проводити дослідження для виявлення необхідність ознак, то в результаті, можна отримати спотворену картину класифікації [10].

Існує чотири системи шкал, які можуть характеризувати об’єкт відповідно до їх числових характеристик. Шкала найменувань може містити арифметичні операції та порівняння. Шкала порядку не допускає арифметичні операції без їх попереднього погодження, вона використовує числові порівняння, впорядковування по ознаці чи ранжирування по числовим характеристикам. Числа використовуються для відміни одного об’єкта від іншого, та їх ранжирування по обраній ознаці. Шкала інтервалів розширює діапазон характеристик об’єктів. Вона дозволяє створювати нові одиниці вимірювання. Робота з інтервалами передбачає встановлення границі, задання точності вимірювання. Шкала відношень досліджує взаємозв’язок між об’єктами, вона підпорядковується аксіомі Евкліда, і являється найбільш поширеною в задачах класифікації. Шкала найменувань та порядку слугують для опису якісних ознак, інтервалів та відношень - кількісних [19].

Ознаки об’єкта можна описати різними типами шкал, найчастіше ознака описується якісно та кількісно. З кількісними ознаками і так зрозуміло, що вони задаються числовим значенням відповідно мірі тією величини, якій вони відповідають. Якісні ознаки подаються вербально, тобто якщо це вага, то велика чи мала, довжина – довга або коротка тощо. Але якісні ознаки необхідно переводити в кількісну шкалу, зазвичай використовується якійсь присвоєний експертом бал від 0 до максимального значення N, або задаватися бінарними значеннями, тобто через 0 або 1 [19].

Після визначення, необхідних для дослідження, ознак необхідно їх привести до однієї шкали вимірювання, тобто перевести значення показників до однієї безрозмірної величини, яка буде надавати можливість зіставляти ознаки. Нормалізацію можна провести, використовуючи методи лінійного перетворення (формула 1.3), стандартизації (1.4) та центрування (1.5). В результаті лінійного перетворення, дані приводяться до одиничного масштабу. Оптимально застосовувати даний метод за умови, що значення величин знаходяться в конкретному інтервалі, бо при великих викидах не є виключенням зосередження основної маси даних в області нуля [19].

(1.3)

$$r_i = \frac{x_i - x_{imin}}{x_{imax} - x_{imin}}$$

Стандартизація сприяє зменшенню впливу структур на показник за допомогою лінійного перетворення початкових даних. Для стандартизації необхідно заздалегідь обчислити середнє арифметичне та середньоквадратичне відхилення. За результатом обчислень, дані будуть знаходитися в межі одиничного масштабу, що буде давати змогу в подальшому їх порівнювати між собою. Але не завжди дані будуть в межі одиничного інтервалу, таким чином не можна заздалегідь знати їх максимальний розкид [19].

$$s_i = \frac{x_i - \bar{x}}{\sigma_i} \tag{1.4}$$

Центрування подає початкові дані у вигляді, де сума величин ознак показників буде дорівнювати 0, це можна досягти, якщо від елемента вибірки відняти середнє арифметичне. Вихідна вибірка буде вважатися центрованою [19].

$$w_i = x_i - \bar{x} \tag{1.5}$$

Загалом при статистичному аналізі даних, якщо змінна підпорядковується нормальному розподілу, то виконується логарифмічне перетворення. Але якщо дані виміряні в різних масштабах, то нормування зводиться до того, щоб досягти нульове середнє арифметичне та одиничну дисперсію. Проте, процедура нормування досі не прийнята як основний етап при кластеризації. Як стверджував Еверітт, то нормування до нульового середнього та одиничної дисперсії може зменшити відмінності між об’єктами по тим ознакам, по яким відображалось найбільше відмінностей. Він стверджував, що нормування краще проводити всередині кластерів, але цього неможливо досягти, поки об’єкти не будуть розподілені по кластерам [26].

Едельброк зазначав, що змінні багатовимірних даних можуть змінювати значення параметрів розподілу в залежності від потрапляння до кластеру, таким чином нормування не зможе правильно перетворити ці змінні. Проте, після дослідження методом Монте-Карло впливу нормування на подальший процес

кластеризації з використанням коефіцієнта кореляції, він не знайшов значних відмінностей між результатами кластеризації з та без нормуванням [26].

Окрім нормування, використовуються і інші методи перетворення даних, наприклад, факторний аналіз та метод головних компонент. Ці методи доцільно використовувати, якщо досліджувані змінні мають високу кореляцію. Оскільки якщо обрати декілька таких змінних, то їх сумісний вплив буде по факту рівносильний одній змінній, яка матиме вагу в рази більше за обрані змінні. Таким чином, метод головних компонент та факторний аналіз можна застосовувати, якщо необхідно скоротити вибірку з некорельованих змінних [26].

Під головними компонентами розуміється нова множина змінних, що досліджуються, котра отримана в ході лінійних комбінацій досліджуваних змінних. Отримана множина має ряд статистичних властивостей. По-перше, вони впорядковані по ступеню розкиду (перша змінна має найвищий ступінь, тобто найбільшу дисперсію). Таким чином, це спрощує роботу для спеціаліста, так як їх в першу чергу цікавлять ознаки, що мають найбільший розкид. По-друге, для опису об’єкту не обов’язково використовувати якісь вихідні його ознаки. Бо якщо об’єкт описується, наприклад, десятьма ознаками, то як показує практика, згрупувавши ці ознаки, вийде, що об’єкт можна описати умовно трьома ознаками, які являються комбінаціями від початкової кількості ознак [2].

Ідея факторного аналізу полягає в бажанні пов’язати кореляцію між досліджуваними змінними з тим, що ці змінні залежать від меншої кількості інших ознак, які не досліджуються. Такі ознаки називають загальними факторами, та намагаються розрахувати таким чином, щоб вона були взаємно некорельовані. Так як, не можна вважати, що кожна ознака залежить тільки від деякої кількості загальних факторів, то прийнято, що досліджувана змінна залежить від деякої своєї випадкової компоненти (шум). Дослідження вважається вдалим, якщо якомога більше ознак вдалося описати якомога

меншою кількістю головних факторів, таким чином факторний аналіз стискає інформацію [2].

При побудові факторної моделі, слід враховувати деякі аспекти: не для кожної множини ознак можна побудувати модель факторного аналізу, тобто відшукати загальні фактори, що будуть пояснювати кореляцію між довільними парами ознак; якщо вдалося побудувати модель факторного аналізу, то така модель не єдина [2].

По завершенню кластеризації, стоїть питання оцінки отриманих результатів. Формально не існує єдиного критерію, який буде показувати наскільки успішно виконана кластеризація, проте, часто використовується функціонал якості розбиття. Значення функціоналу залежить від об'єму кластеру та відстані між об'єктами в кластері. Каноном вважається таке розбиття, де досягається екстремум обраного функціоналу. Обирається функціонал довільно, виходячи з експертної думки чи інтуїції. До таких функціоналів можна віднести [11]:

- Сума квадратів відстаней до центроїдів:

$$Q(S) = \sum_{l=1}^{k} \sum d^2(X_i, \overline{X}_l), \tag{1.6}$$

де l – номер кластеру,

$\overline{X}_l$ – центр i-го кластеру,

X_i – значення змінних для i-го об'єкту,

$d(X_i, \overline{X}_l)$ – відстань між i-тим об'єктом та центром l-го кластеру.

При використанні цього функціоналу обирається його мінімальне значення.

- Сума відстаней між об'єктами в кластері:

$$Q(S) = \sum_{l=1}^{k} \sum d_{ij}^2, \tag{1.7}$$

Даний функціонал обирає кластер з мінімальним $Q(S)$, що буде свідчити про те, що кластери мають велику щільність, тобто об'єкти, що знаходяться у кластері, знаходяться поруч одне з одним по тим ознакам, які не використовувалися для кластеризації [11].

- Сума дисперсії всередині кластера:

(1.8)

$$Q(S) = \sum_{l=1}^{k} \sum_{j=1}^{p} \sigma_{lj}^2,$$

де σ_{lj}^2 – дисперсія в кластері S_l.

За цим функціоналом, якщо $Q(S)$ мінімальна, то розбиття можна вважати оптимальним.

Окрім перелічених функціоналів, часто використовуються їх відношення, наприклад, відношення дисперсії даних до суми дисперсій всередині кластерів тощо [11].

Окрім цього, існують методи саме адекватності отриманого рішення, серед них кофенетична кореляція, тест значущості для змінних, що використовувалися в кластеризації, методи Монте-Карло, тест значущості для незалежних змінних та повторна вибірка [26].

Кофенетична кореляція вперше була пропонована в 1962 році, Рольфом та Сокелом, на їх думку це єдина правильна міра для обґрунтування отриманого розв'язку. Даний коефіцієнт можна застосовувати лише для агломеративних методів. Він розраховується, щоб оцінити наскільки добре відображається взаємозв'язок між об'єктами на отриманій дендрограмі. Елементу матриці відповідає значення подібності для того рівня, на котрому були об'єднані об'єкти у кластер. Кофенетична кореляція відображає кореляцію між значеннями початкової матриці подібності та вторинної. Проте, даний коефіцієнт має недоліки, не дивлячись на часте використання. По-перше, для обох значень матриць подібності, значень мають бути корельовані, для початкової матриці дана умова задовольняється, для вторинної – ні. По-друге, матриці мають не рівну кількість значень, таким чином вони налічують різну інформацію [26].

Наступний метод оцінки результатів – тести значущості змінних, що використовувалися в кластеризації, що відносяться до дисперсійного аналізу. Даний метод за допомогою перевірки гіпотези однорідності, оцінює значущість кожного розбиття на кластери. Ці тести можна застосовувати для будь-якого методу кластеризації, що вже розрізняє цей метод від кофенетичної кореляції [26].

Наступний метод – повторна вибірка. В його основі лежить перевірка повторюваності розв’язку серед множини наборів даних. Якщо для різних підмножин даних, які є частиною однієї множини, отримується одне й те саме рішення, то зрозуміло, що такий розв’язок буде притаманний і всій вибірці [26].

Тести значущості для незалежних змінних порівнюють отримані кластери зі змінними, які не використовувалися для кластеризації.

Методи Монте-Карло полягають у генерації випадкових чисел, що будуть утворювати вибірку з тими характеристиками, якими описуються реальні дані, але ці вибірки не мають кластерів. Принцип роботи алгоритму можна описати наступними кроками:

1) Згенерувати випадковий набір даних. За допомогою генератора випадкових чисел, створюється вибірка, що немає кластерів, проте має ті самі характеристики, що і вхідні дані. За допомогою генератора можна створити дані на основі основної вибірки з багатовимірним нормальним розподілом, із заданим вектором середніх та коваріаційною матрицею.
2) Застосувати один і той самий метод для обох наборів даних.
3) Порівняти отримані розв’язки. Розраховуються та порівнюються значення статистичних критеріїв [26].

При успішній кластеризації, отриману модель можна використовувати навіть якщо до вибірки додалися нові елементи, тобто одразу відносити їх до кластерів без повторної кластеризації. Дану процедуру можна виконати за умов, що новий об’єкт описується таким ж ознаками, які і ті, на основі яких проводилась кластеризація; значення ознак нормовані за таким же принципом, що і нормовані об’єкти моделі. Якщо всі умови задовольняються, то віднесення об’єкту до існуючого кластеру можна описати наступними кроками:

1) Розрахунок середньозваженої довжини вектору центру для кожного кластера:

(1.9)

$$M_k = \sqrt{\frac{1}{N}\sum_{j=1}^{N} M_{kj}^2}$$

2) Розрахунок середньозваженого розсіяння даної довжини:

$$\sigma_k = \sqrt{\frac{1}{N}\sum_{j=1}^{N} \sigma_{kj}^2} \quad (1.10)$$

3) За формулою (1.9) розрахувати довжину M_l вектору доданого об’єкта.
4) Розрахунок різниці між відстанями:

$$\Delta_k = M_k - M_l \quad (1.11)$$

5) Якщо всі різниці мають один і той самий знак, то об’єкт записується до кластеру до якого різниця мінімальна, на цьому алгоритм запиняється.
6) В іншому випадку, серед всіх різниць обираються дві з протилежними знаками, для яких модуль є найменшим. Потім до негативної різниці додається значення формули (1.10), а від позитивної – воно віднімається.
7) Якщо перша змінна має знак «+», а друга – «-», або вони змінять знак на протилежний, то об’єкт записується до кластеру, для якого модуль цих параметрів найменший. В іншому випадку, якщо всі змінні зі знаком «+», то об’єкт відноситься до кластеру з негативним значенням (1.11), при всіх від’ємних – до позитивного [24].

1.2.2 Міри подібності

Методи кластеризації мають забезпечувати максимальну подібність об’єктів, які потрапляють до кластеру. Міри подібності можна поділити на види: за типом відстані, зв’язку та інформаційної статистики. При використанні мір подібності типу «відстань», об’єкти вважаються подібними, якщо відстань між ними менша; при типі «зв’язок» - подібність оцінюється силою зв’язку між об’єктами, тобто чим сильніший зв’язок, тим більше вони схожі [20].

Міра подібності має бути метрикою та забезпечувати виконання наступних умов:

- Симетричність, тобто $d(x, y) = d(y, x)$;
- Нерівність трикутника, тобто $d(x, y) \leq d(x, z) + d(z, y)$;
- Невід'ємність, тобто $d(x, y) \geq 0, d(x, y) = 0 \Leftrightarrow x = y$ [12];

Майже завжди використовуються міри відстані, найрозповсюдженіші серед них – евклідова (традиційна міра відстані), манхеттенська (найбільш відома з класу метрик Мінковського) та відстань Махаланобиса (не є метрикою, пов'язується з кореляційними змінними за допомогою дисперсійно-коваріаційної матриці). Серед мір типу «зв'язок» для кількісних ознак використовується коефіцієнт кореляції Пірсона (тільки за умови, що зв'язок між ознаками лінійний), кореляційне відношення та дисперсія-коваріація. При наявності порядкових ознак, доцільно брати в якості міри подібності коефіцієнти рангової кореляції Спірмена та Кендалла, перетворивши їх на міру подібності типу «відстань». Для дихотомічних ознак та таких, що представляються у вигляді таблиць спряженості, рекомендоване використання хеммінгової відстані, показника Жаккара, Рассела й Рао, простого коефіцієнта зустрічальності, та коефіцієнтів асоціації Юла і спряженості Бравайса. Щоб перетворити перелічені показники у відстань, слід відняти обчислені значення від одиниці; не розповсюджується на хеммінгову відстань. При змішаних ознаках застосовується коефіцієнт Гауера [20].

Евклідова відстань розраховується за формулою (1.12). Отримане значення може співпадати з відстанню Махаланобиса за умови, що незалежні змінні некорельовані. Якщо слід надати більшу вагу віддаленим один від одного об'єктам, можна піднести у квадрат формулу (1.12). Метрика Евкліда буде пропорційно збільшувати відстань між об'єктами, якщо абсолютні значення показників різні, що свідчить про те, що дана метрика не враховує знакові розходження. Як результат, розмірність кластерного поля збільшується, а об'єкти штучно віддаляються один від одного, роблячи границі між кластерами більш чіткими та точними [20, 24].

(1.12)

$$d_{ij} = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2},$$

де d_{ij} – відстань між об’єктами i та j,

x_{ik} – значення k-ої змінної для i-го об’єкта,

x_{jk} – значення k-ої змінної для j-го об’єкта [10].

Застосовувати евклідову відстань доцільно, якщо: спостереження належать до генеральних сукупностей, які описуються багатовимірним нормальним законам, а компоненти вектору спостережень незалежні та мають однакову дисперсію; компоненти вектору спостережень мають бути однорідними та однаково важливими для класифікації; розмірність простору ознак складає 1, 2 або 3. Недолік даної метрики проявляється тоді, коли ознаки мають різні одиниці вимірювання, і зміна масштабу одиниць вимірювання приводить до суттєвої зміни результатів класифікації. Щоб цьому запобігти, слід проводити нормування даних, проте, застосування методів нормування також може вплинути на кінцеві результати, наприклад, коли кластери сильно розділяються за деякими ознаками, і при цьому слабко за іншими, то нормалізація може привести до збільшення шумового ефекту тих ознак, які зменшують дискримінуючи можливості першої групи ознак. Також обчислення евклідової відстані можна вважати безглуздим, якщо ознаки вимірюються у якісно різних одиницях [20].

Модифікацією метрики Евкліда є зважена евклідова відстань, що розраховується за формулою (1.13). Процес визначення вагових коефіцієнтів за аналізованою вибіркою, зазвичай, вважається недоцільним, так як може призвести до помилок. Обґрунтованим можна вважати, в залежності від певних незначних варіацій змістової та статистичної природи вихідних даних, надання значенням вагових коефіцієнтів значення пропорційні середньоквадратичній похибці відповідної ознаки або оберненій до цієї похибки величини, проте, рекомендовано обирати ваги за результатами експертних опитувань [20].

$$d_{ij} = \sqrt{\sum_{k=1}^{m} w_k(x_{ik} - x_{jk})^2}, \qquad (1.13)$$

де w_k – вага k-ої ознаки, яка є пропорційною ступеню важливості критерію, $0 \leq w \leq 1$ [8].

Узагальненням евклідової відстані служить метрика Мінковського, загальна формула якої (1.14). При $r = 2$, формула (1.14) буде ідентична формулі (1.12), що відповідає евклідовій метриці; при $r = 1$ утворюється манхеттенська відстань, що обчислюється за формулою (1.15); при $r \to \infty$ – відстань Чебишева (супремум-норма), що обчислюється за формулою (1.16) [20].

$$d_{ij} = \sqrt[r]{\sum_{k=1}^{m} \left|x_{ik} - x_{jk}\right|^r}, \tag{1.14}$$

де r задається користувачем [10].

У випадку з манхеттенською відстанню (відстань міських кварталів або метрика L-норми), вплив окремих викидів менший, оскільки координати не підносяться до квадрату, як це робиться при розрахунку евклідової відстані. Різниця метрик в евклідовому просторі та L-норми залежить лише від абсолютних числових значень та кількості розглянутих показників [24-25].

$$d_{ij} = \sum_{k=1}^{m} \left|x_{ik} - x_{jk}\right| \tag{1.15}$$

Відстань Чебишева (норма «верхня границя») дослідно використовувати коли розділити кластери на підмножини досить складно через їх велику компактність. Простими словами, з усіх різниць значень факторів, взятих по модулю, обирається одна – найбільша, що і буде вважатися відстанню між об’єктами [24].

$$d_{ij} = \sup\left\{\left|x_{ik} - x_{jk}\right|\right\}, k = 1,2,\dots,m \tag{1.16}$$

Відстань Махаланобиса – це відстань від точки спостереження до центра ваги в багатовимірному просторі ознак, що задається у матричній формі, розраховується за формулою (1.17) [20].

$$d_{ij} = \sqrt{\left(X_i - X_j\right)^T \Delta^T \Sigma^{-1} \Delta \left(X_i - X_j\right)}, \tag{1.17}$$

де Δ – симетрична векторна невід'ємна матриця вагових коефіцієнтів (зазвичай діагональна),

Σ – коваріаційна матриця генеральної сукупності, до якої належить спостереження.

Окрім мір відстані, користуються популярністю коефіцієнти кореляції (формула (1.18)), які часто називають кутовими мірами через їх геометричну інтерпретацію. Вперше даний коефіцієнт був застосований для кількісної класифікації в якості методу для визначення залежності між змінними [26]. (1.18)

$$r_{ij} = \frac{\sum_{k=1}^{n}(x_{ki}-\overline{x_i})(x_{kj}-\overline{x_j})}{\sqrt{\sum_{k=1}^{n}(x_{ki}-\overline{x_i})^2 \sum_{k=1}^{n}\left(x_{kj}-\overline{x_j}\right)^2}},$$

де r_{ij} – коефіцієнт кореляції між об'єктами i та j,

x_{ki} – значення k-ої змінної для i-го об'єкта,

x_{kj} – значення k-ої змінної для j-го об'єкта,

$\overline{x_i}$ – середнє значення всіх змінних i-го об'єкта,

$\overline{x_j}$ – середнє значення всіх змінних j-го об'єкта.

Використовувати даний коефіцієнт доцільно для змінних, які описані шкалою відношень чи шкалою інтервалів, у випадку бінарних даних, коефіцієнт кореляції переводиться у φ-коефіцієнт. Діапазон зміни коефіцієнта кореляції варіюється від 0 до 1 включно, при чому 0 свідчить про відсутність зв'язку між об'єктами. Недоліком даного коефіцієнта як для міри подібності виступає чутливість до форми за рахунок зниження чутливості до величини розбіжностей між змінними. Це означає, що два об'єкти можуть мати кореляцію 1, але при цьому бути різними, тобто не проходити через одні й ті ж самі точки. Ще один важливий недолік, коефіцієнт кореляції дуже часто може не задовольняти нерівність трикутника [26].

Коефіцієнт асоціативності використовується для бінарних даних, які описуються таблицею асоціативності, де 1 вказує на наявність змінної, а 0, відповідно, на відсутність. Прикладом такої таблиці є таблиця 1.1. Таких коефіцієнтів існує більше 30 видів, проте, використовуються лише 3 міри, які

були піддані перевірці, серед них: простий коефіцієнт зустрічальності, коефіцієнт Жаккара та Гауера [26].

Таблиця 1.1

Таблиця асоціативності

	1	0
-1-	**-2-**	**-3-**
1	a	b
0	c	d

Простий коефіцієнт зустрічальності розраховується за формулою (1.19), він приймає значення від 0 до 1, а також здатен враховувати відсутність тієї чи іншої ознаки в обох об'єктах, в даному прикладі це комірка d таблиці 1.2. Варто відмітити, що даний коефіцієнт важко перетворити на метрику [26].

$$S = \frac{a + d}{a + b + c + d} \tag{1.19}$$

Коефіцієнт Жаккара визначається формулою (1.20), на відміну від попереднього коефіцієнта, не враховує одночасної відсутності ознаки в об'єктах. Широке застосування даного коефіцієнта відмічається в біології, і як помітили самі біологи, то після застосування простого коефіцієнту зустрічальності, деякі об'єкти виявлялися схожими за рахунок тільки за рахунок того, що вони якраз не мали якоїсь ознаки, а не тому що мали спільні характеристики. Якраз коефіцієнт Жаккара приймає в розрахунок тільки ті ознаки, які має хоча б один з об'єктів [26].

$$J = \frac{a}{a + b + c} \tag{1.20}$$

Коефіцієнт Гауера (формула (1.21)) відрізняється тим, що він єдиний, хто допускає одночасне використання змінних, які описані різними шкалами [18].

$$S_{ij} = \frac{\sum_{k=1}^{p} S_{ijk}}{\sum_{k=1}^{p} W_{ijk}}, \tag{1.21}$$

де S_{ijk} – значущість ознаки k порівнянні об'єктів i та j,

W_{ijk} – ваговий коефіцієнт (1 – якщо порівняння об'єктів за ознакою k варто враховувати, 0 – в протилежному випадку) [8].

Останній тип міри подібності – імовірнісні коефіцієнти подібності (формула 1.22), які застосовуються лише для бінарних даних. Особлива відмінність цих коефіцієнтів від всіх інших в тому, що по факту не відбувається розрахунку подібності між об’єктами. Замість цього дана міра застосовується до даних ще до їх обробки, при формуванні кластерів обчислюється інформаційний виграш від об’єднання двох об’єктів, потім об’єднання, які дають мінімальний виграш, вважаються одним об’єктом [26].

$$I_{ij} = \sum_{x,y} p_{xy}\, log \frac{p_{xy}}{p_x^i p_y^j}, \tag{1.22}$$

де p_{xy} – ймовірність спільної появи ознак x та y,

p_x^i – ймовірність появи ознаки x в об’єкті i,

p_y^j – ймовірність появи ознаки y в об’єкті j [8].

Окрім метрик, в якості міри подібності можуть виступати інші способи оцінювання схожості між об’єктами. Наприклад, вимірювати взаємозв’язок між об’єктами. Часто може бути таке, що об’єкт 1 може відповідати об’єкту 2, проте об’єкт 2 може не відповідати об’єкту 1, таким чином випливає асиметрія відношення подібності. Така асиметрія може спостерігатися в економіці, наприклад, якщо витрати більші, ніж дохід тощо. Через асиметрію можуть виникати труднощі при розрахунку коефіцієнтів подібності [26].

1.2.3 Метод k-середніх

Метод k-середніх відноситься до групи ітераційних алгоритмів, та, мабуть, найбільш поширений серед них. Цільовою функцією виступає мінімізація середньої квадратичної відстані між точками в кластері. На відміну від ієрархічних методів, необхідно заздалегідь знати оптимальну кількість кластерів. Це можна визначити на основі попередніх досліджень або суто інтуїції [1, 13].

До переваг методу можна віднести його швидкість при обробці великого масиву даних, проте він не гарантує абсолютної точності, та може бути

чутливим до шуму, що впливає на середнє арифметичне. Для оцінки результатів кластеризації може використовуватися параметр, що вимірює наскільки кластери різні, за цим параметром розраховуються середнє арифметичне для кожного кластеру. Якщо усі отримані значення істотно відрізняються один від одного, або хоча б переважна частина, то це свідчить про якісно виконану кластеризацію [1, 13].

Покрокову роботу методу можна описати наступним чином:

1) Початковий вибір центроїдів. Для заданого числа k кластерів призначити їх первинні центри (обрати точки центрів можна випадковим чином).

   На рисунку 1.6 продемонстровано розподіл точок в просторі, та відмічено початкові точки центроїдів.

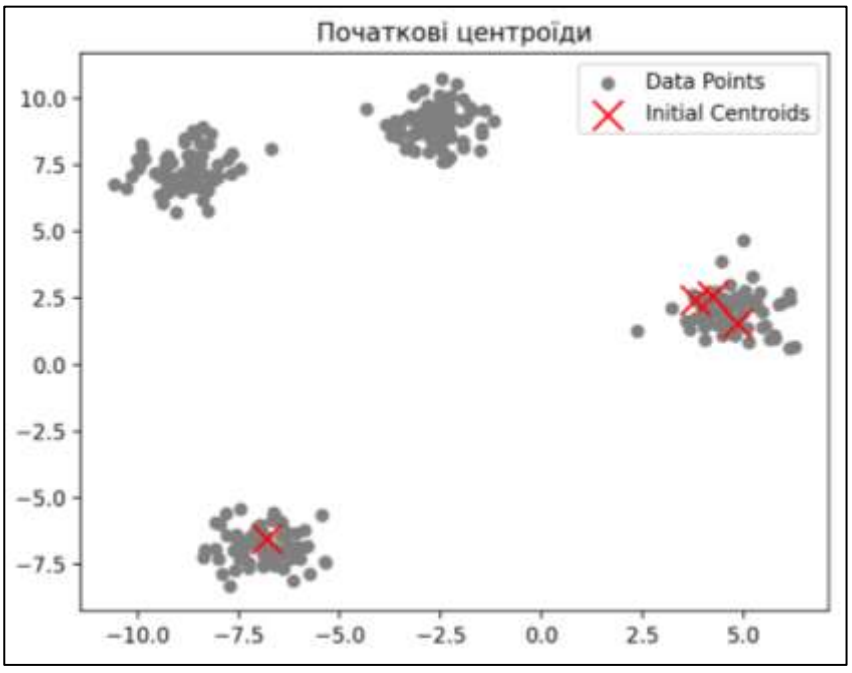

Рисунок. 1.6 – Початковий вибір центроїдів.

2) Призначення кластеру. На основі розрахованої евклідової відстані між центроїдом та об’єктом, віднести його до найближчого кластеру.

   На рисунку 1.7 зображено початкове віднесення об’єктів до кластеру.

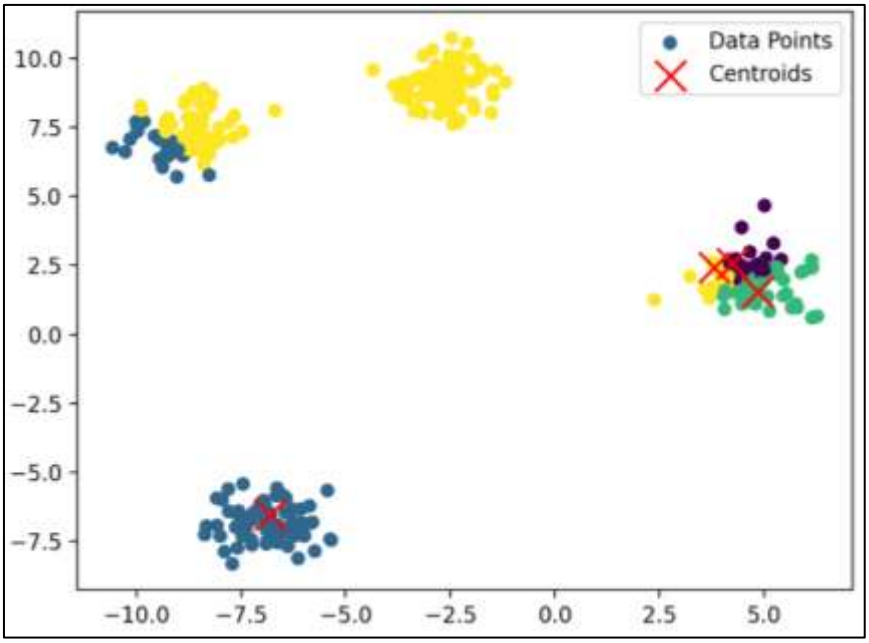

Рисунок. 1.7 –Початковий віднесення об’єктів до кластерів.

3) Перерахунок центроїдів. Для кожного кластеру перерахувати центроїд, шляхом розрахунку середнього арифметичного для усіх точок в кластері.

На рисунку 1.8 зображено перехóвані центри кластерів.

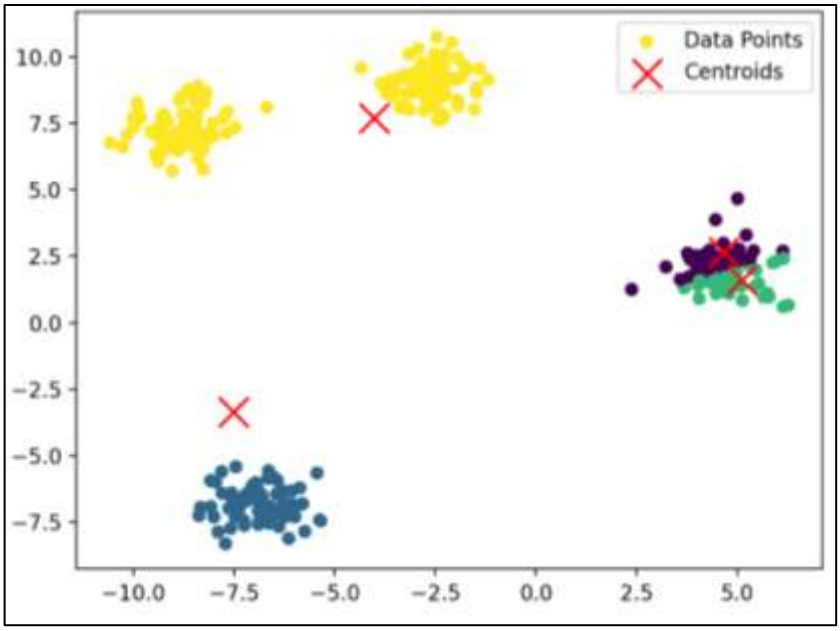

Рисунок. 1.8 – Перерахування центроїдів.

4) Ітераційний процес. Повторювати кроки 2-3, доки центроїди не зміняться або не буде досягнуто задану кількість ітерацій, чи іншого параметру для зупинки алгоритму [1].

На рисунку 1.9 зображено кінець роботи алгоритму, що включає розраховані центри кластерів, та розподілені об’єкти.

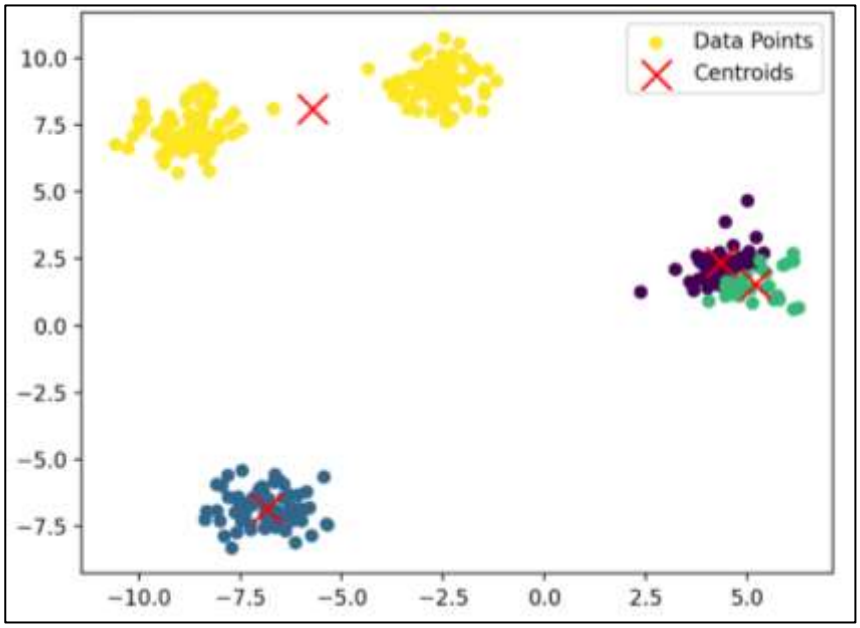

Рисунок. 1.9 – Кінець роботи алгоритму.

За результатами роботи алгоритму, може виявитися таке, що деякі об’єкти не потрапили до жодного кластеру, якщо це число перевищує 25% всієї вибірки, то проведену кластеризацію не можна назвати якісною. Якщо немає експертної підстави для задання кількості кластерів, можна скористатися «ліктьовим» методом [18].

Суть методу полягає в розрахунку метрики за формулою (1.12), що представляє відстань від точок до центрів кластеру. Зі збільшенням кількості кластерів, значення метрики буде наближатися до 0, проте, починаючи з деякого кластеру, наближення буде гальмуватися, та майже не змінюватися. Останній кластер, на якому був пік, і буде вважатися оптимальною кількістю для заданого набору даних. Відображення результатів обчислення метрики продемонстровано на рисунку 1.10. З рисунку видно, що після 3-го кластеру,

метрика не так швидко наближається до 0, отже, число 3 – буде оптимальним для кількості кластерів [18].

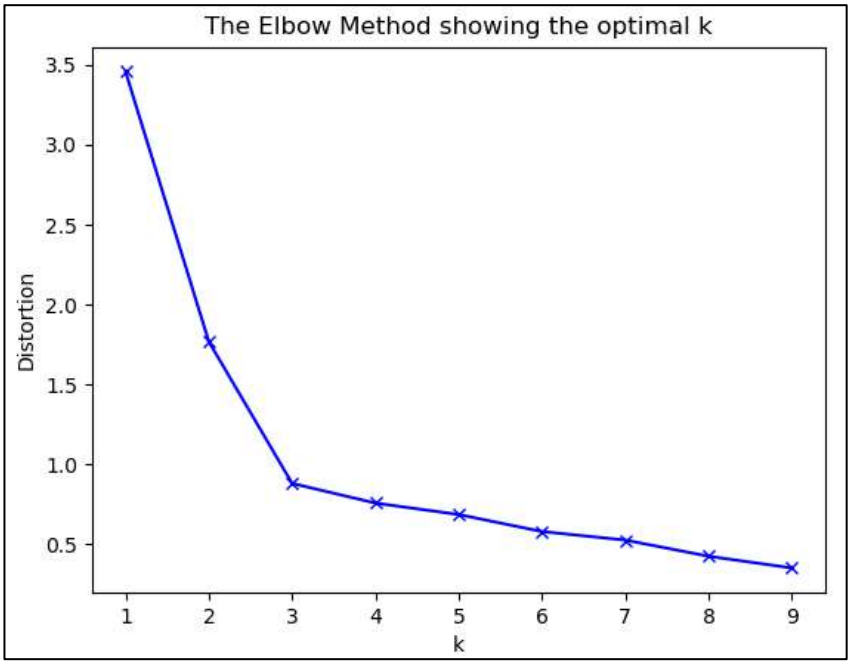

Рисунок. 1.10 – Демонстрація ліктьового методу.

У деяких випадках, слід застосовувати алгоритм k-середніх++, що допоможе ініціалізувати початкові центроїди кластерів перед тим, як продовжити кластеризацію k-середнім. Даний метод оптимізує етап обрання початкових центроїдів випадковим чином. Алгоритм виконується в послідовності наступних кроків:

1) Незалежно від заданої кількості кластерів, обрати один кластер випадковим чином (рисунок 1.11).

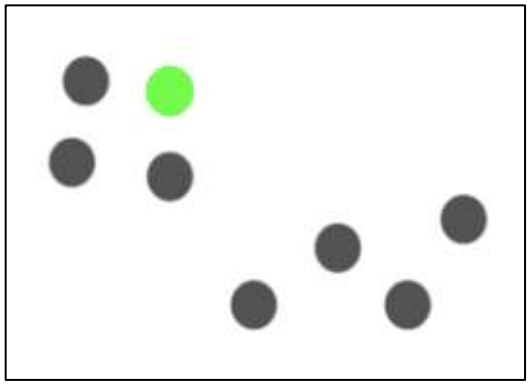

Рисунок. 1.11 – Вибір початкового центроїда.

2) Обчислити відстань між кожною точкою до центроїду обраного кластера (рисунок 1.12).

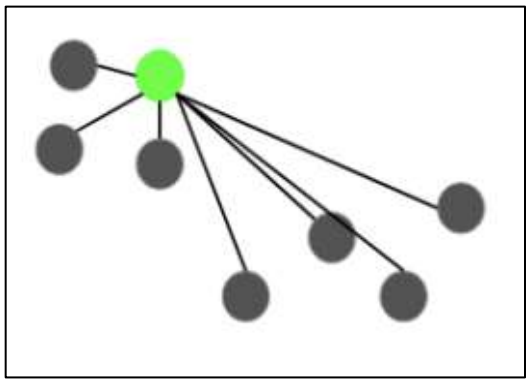

Рисунок. 1.12 – Розрахунок відстані між центроїдом та точками.

3) Обрати новий центроїд з точок з найбільшим квадратом відстані (рисунок 1.13).

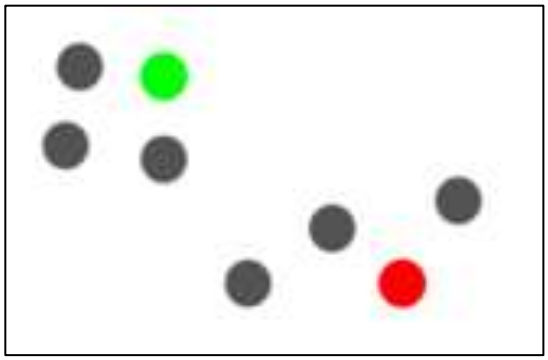

Рисунок. 1.13 – Визначення центроїда іншого кластеру.

4) Повторити кроки 2 та 3 для ініціалізації заданої кількості k кластерів [18].

1.3 Висновок до Розділу 1

Кластеризацію можна назвати класифікацією, але без заздалегідь визначених кластерів. Вона займається розбиттям вибірки на підмножини, які не мають перетинатися та складатися з схожих об’єктів, але самі кластери

мають відрізнятися один від одного. Кожний кластер має такі властивості: щільність, дисперсія, розмір, форма та віддільність.

Щоб оцінити подібність між об’єктами, використовується спеціальна метрика, яка має задовольняти умовам симетрії, нерівності трикутника та невід’ємності. У ролі міри подібності можуть виступати коефіцієнти кореляції, міри відстані, коефіцієнти асоціативності та імовірнісні коефіцієнти подібності. Широке поширення отримали міри відстані.

Методи кластеризації поділяються на ієрархічні (агломеративні, дивізимні), ітераційні, факторні методи, на основі теорії графів тощо. Серед агломеративних методів найпоширеніші: метод одиничного зв’язку, повного та середнього зв’язку, метод Уорда. Серед ітеративних, часто застосовується метод k-середніх.

Суть методу k-середніх полягає в ітераційному обчисленні центроїдів кластеру, та віднесенню об’єкту до кластеру з мінімальною відстані до його центру. Критерії для зупинки алгоритму можна задати самостійно, наприклад, максимальну кількість ітерацій тощо.

РОЗДІЛ 2 СПЕЦІАЛЬНИЙ

2.1 АТ «Банк Кредит Дніпро»

Акціонерне товариство «Банк Кредит Дніпро» (БКД) засновано 7 липня 1993 року відповідно рішенню Загальних зборів акціонерів Банку та згідно законодавства України. 16 липня 2009 року змінено назву та організаційну форму Банку із закритого акціонерного товариства Акціонерний Банк «Муніципальний Банк» на публічне акціонерне товариство АБ «Банк Кредит Дніпро». Проте, згідно змін законодавства України, у квітні 2018 р. відбулися повторні зміни організаційної форми, що перетворило публічне акціонерне товариство на Акціонерне Товариство «Банк Кредит Дніпро» [4, 7].

Згідно класифікації Національного банку України, БКД відноситься до групи банків з приватним капіталом, та входить до ТОП-17 фінансових установ України за розмірами активів [3].

2.1.1 Відомості про банк

Діяльність банку реалізовується у прийманні вкладів від населення та видачі кредитів, а також позики Рефінанс (тобто рефінансування кредитів інших банків чи фінансових компаній); здійсненню переказів грошових коштів та валютно-обмінних операцій не тільки в межах України, а й поза її територією. Також банк розвиває програми кредитування фізичних та юридичних осіб, малого та середнього бізнесу, та надає широкий спектр послуг

клієнтам, що ведуть зовнішньо-економічну діяльність. Бажаючі отримати кредит мають відповідати простим умовам: вік (від 22 до 64 років), працевлаштування (офіційне, ФОП, найманий на ФОП, пенсіонер чи моряк), та додаткова умова для найманих працівників – стаж роботи на нинішньому підприємстві мінімум 3 місяці [4, 6, 23].

Станом на 01.01.2024 р. банк налічує 33 відкритих відділення по Україні, 5 з яких відкрито протягом 2023 року. На рисунку 2.1 зображена динаміка кількості відділень БКД [23].

Рисунок. 2.1 – Динаміка кількості відділень БКД.

Бізнес-модель банку зображено на рисунку 2.2. З нього видно, що основну частку (65%) складаються кошти юридичних осіб, на другому місці – цінні папери (32%), депозити фізичних осіб та кредити юридичним особам займають майже однакову частку – 17% та 19% відповідно.

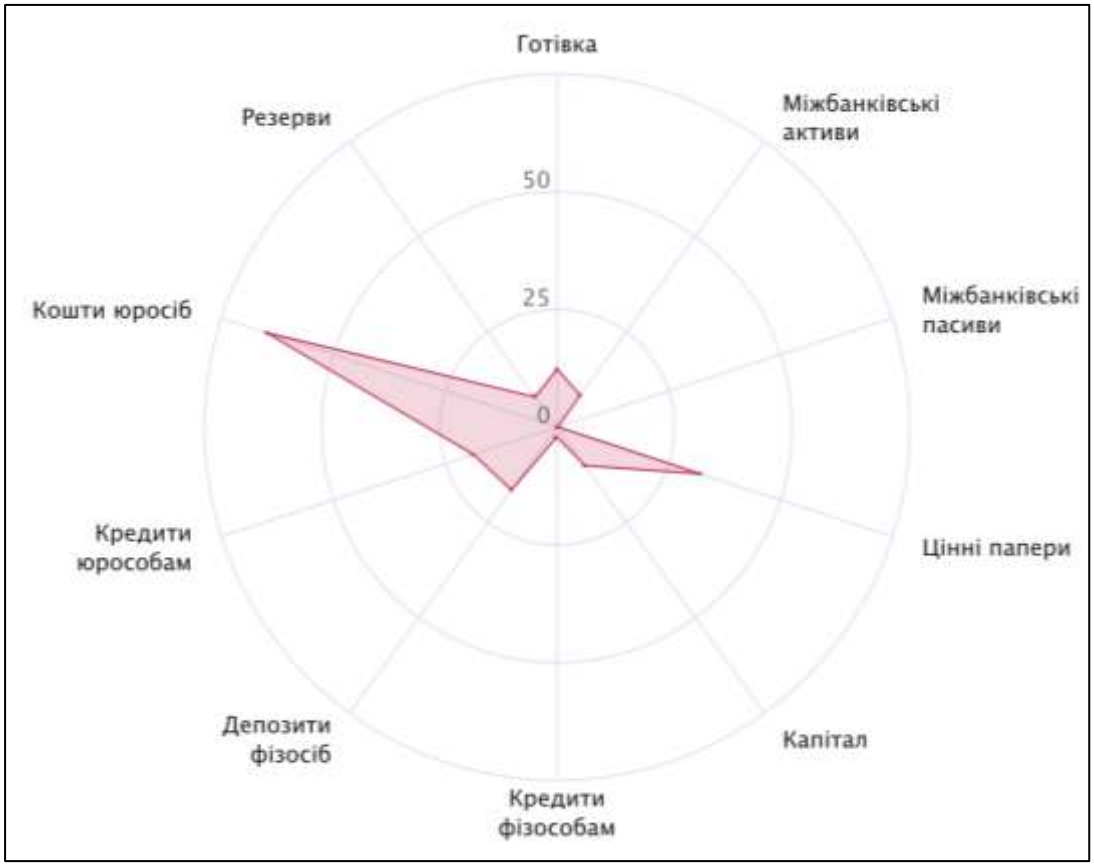

Рисунок. 2.2 – Бізнес-модель БКД.

2.1.2 Фінансовий стан банку

23 січня 2020 р. БКД надано свідоцтво про реєстрацію випуску акцій, загальна сума яких 3,58 млрд. грн., після чого зазначеній сумі відповідає зареєстрований та сплачений розмір статутного капіталу Банку [4].

Фінансовий стан банку, станом на 30.06.2024 р., можна оцінити за допомогою показників, що наведені на рисунку 2.3. Загалом видно, що банк розвивається, що підтверджується приростом власного капіталу [4].

ПРОМІЖНИЙ СКОРОЧЕНИЙ ЗВІТ ПРО ФІНАНСОВИЙ СТАН

Станом на 30 червня 2024 р.

(у тисячах гривень)

	Прим.	*30 червня 2024 р.*	*31 грудня 2023 р.*
Активи			
Грошові кошти та їх еквіваленти	6	5 648 046	8 152 323
Кредити та аванси банкам	7	515 237	267 893
Кредити та аванси клієнтам	8	6 623 457	4 335 277
Інвестиції в цінні папери	9	9 277 647	6 802 864
Похідні фінансові активи		133 976	244 320
Інвестиційна нерухомість	10	96 531	121 807
Відстрочені податкові активи		184 764	292 594
Нематеріальні активи	12	53 868	58 560
Основні засоби	13	333 138	327 772
Інші фінансові активи	11	60 330	20 548
Інші нефінансові активи	11	250 451	219 073
Необоротні активи, утримувані для продажу		38 359	52 913
Усього активів		**23 215 804**	**20 895 944**
Зобов'язання			
Кошти банків		78	40 119
Кошти клієнтів	14	20 432 243	18 476 538
Похідні фінансові зобов'язання		259	-
Інші залучені кошти		91 781	-
Забезпечення	15	53 580	46 792
Інші фінансові зобов'язання	16	152 577	135 502
Інші нефінансові зобов'язання	16	66 764	33 861
Податок на прибуток		-	67 914
Усього зобов'язань		**20 797 282**	**18 800 726**
Власний капітал	17		
Статутний капітал		3 586 561	3 586 561
Непокритий збиток		(1 639 587)	(1 810 628)
Емісійний дохід		17 469	17 469
Резервні та інші фонди банку		75 711	61 430
Інші резерви		378 368	240 386
Усього власного капіталу		**2 418 522**	**2 095 218**
Усього зобов'язань та власного капіталу		**23 215 804**	**20 895 944**

Підписано від імені Правління 30 липня 2024 р.

Голова Правління Сергій ПЛАНОВ

Головний бухгалтер Руслан ЧУДАКІВСЬКИЙ

Рисунок. 2.3 – Проміжний скорочений звіт про фінансовий стан АТ «БКД».

Для розуміння, яке місце посідає банк серед банків України, проводиться аналіз конкурентів за різними фінансовими показниками. «Банк Кредит Дніпро» здійснює моніторинг обсягу робочого портфелю та рівня NPL.

Робочий портфель – це баланс виданих кредитів, заборгованість по яким не перевищує 90 днів. NPL (Non-Performing loans) – це кредити, які на звітну дату мають заборгованість у 90+ днів; такі кредити вважаються сумнівними та безнадійними до повернення. У ролі звітної дати виступає останній робочий день місяця [12].

Всю необхідну статистику, що оновлюється кожного місяця, можна отримати на офіційному сайті НБУ. Станом на 11.11.2024 р. на сайті представлена інформація на 01.10.2024 р. Зібравши всі необхідні дані в діапазоні жовтень 2023 – жовтень 2024 рр., можна представити їх у вигляді графіка 2.4 та таблиці 2.1 [22].

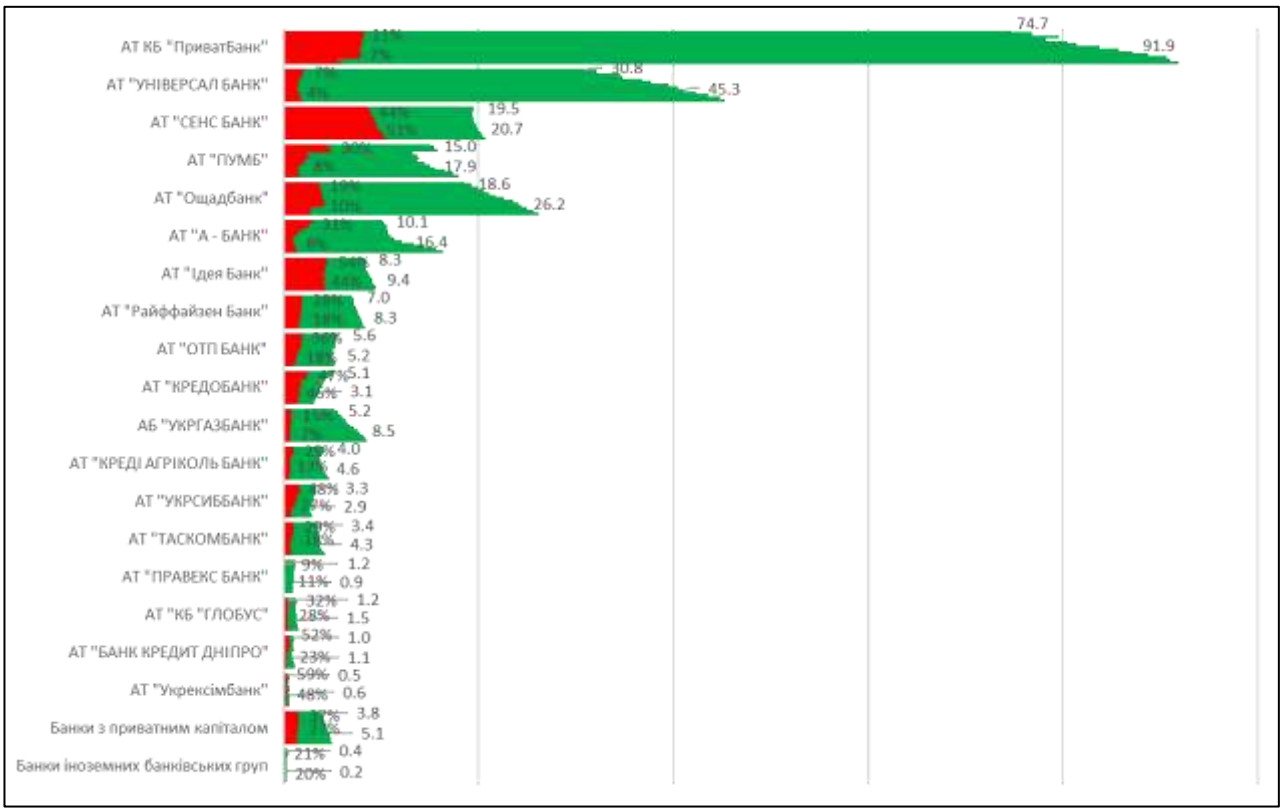

Рисунок. 2.4 – Динаміка обсягу робочого портфелю та рівня NPL серед банків України.

Проаналізувавши таблицю 2.1, видно, що робочий портфель АТ «Банк Кредит Дніпро» з жовтня 2023 р. приріс на 81%, а рівень NPL зменшився на

28,8%. За останній квартал також спостерігається конкурентоспроможний приріст обсягу робочого портфелю.

Таблиця 2.1

Порівняння обсягу робочого портфелю та рівня NPL серед банків України

Банк	Робочий портфель 2024-10, млн. грн.	Приріст робочого портфелю з 2023-10, %	Приріст NPL з 2023-10, %	Приріст робочого портфелю за квартал, %	Приріст NPL за квартал, %
-1-	**-2-**	**-3-**	**-4-**	**-5-**	**-6-**
АТ КБ «ПриватБанк»	85 827	29%	-4.6%	6.2%	-2.4%
АТ «УНІВЕРСАЛ БАНК»	43 564	52%	-3.1%	6.7%	0.3%
АТ «СЕНС БАНК»	10 241	-5%	6.1%	0.7%	1.2%
АТ «ПУМБ»	16 420	56%	-21.4%	16.4%	-0.7%
АТ «Ощадбанк»	23 516	57%	-8.8%	14.3%	-5.4%
АТ «А – БАНК»	15 038	115%	-22.7%	22.7%	0.1%
АТ «Ідея Банк»	5 285	38%	-9.8%	5.8%	-1.0%
АТ «Райффайзен Банк»	6 836	35%	-10.0%	8.7%	-3.2%
АТ «ОТП БАНК»	4 219	17%	-17.8%	13.3%	-6.4%
АТ «КРЕДОБАНК»	1 669	-38%	-1.0%	-11.0%	0.7%
АБ «УКРГАЗБАНК»	7 941	80%	-8.0%	11.1%	-1.5%
АТ «КРЕДІ АГРІКОЛЬ БАНК»	3 989	32%	-11.7%	16.8%	-2.8%
АТ «УКРСИББАНК»	2 113	24%	-21.2%	10.0%	-2.5%
АТ «ТАСКОМБАНК»	3 505	45%	-10.3%	14.0%	1.1%
АТ «ПРАВЕКС БАНК»	846	-20%	1.5%	-2.1%	-2.1%

Продовження табл. 2.1

-1-	-2-	-3-	-4-	-5-	-6-
АТ «ПРАВЕКС БАНК»	846	-20%	1.5%	-2.1%	-2.1%
АТ КБ «ГЛОБУС»	1 110	34%	-4.5%	15.5%	-1.6%
АТ «БАНК КРЕДИТ ДНІПРО»	867	81%	-28.8%	31.3%	-8.1%
АТ «Укрексімбанк»	299	51%	-11.3%	-3.0%	-0.6%
Банки з приватним капіталом	3 688	53%	-10.0%	10.1%	-1.5%
Банки іноземних банківських груп	196	-30%	-0.1%	27.8%	-19.9%

2.2 Кластеризація в середовищі SQL Server Management Studio

SQL Server Management Studio (SSMS) – це інструмент для управління будь-якими компонентами SQL. За допомогою SSMS можна запитувати дані в БД, проектувати та управляти ними [28].

SQL – це мова запитів, що передбачає формулювання користувачем запитів у вигляді директив, що задані за допомогою набору формальний конструкцій, що схожі на англійську мову. SQL надає можливість сформулювати запит на пошук чи вибір даних, їх оновлення тощо [5].

Збережена процедура – це скрипт, що містить в собі SQL-конструкції. Вона зберігається у БД та виконується на стороні сервера. Збережена процедура існує незалежно від таблиці, її можна викликати клієнтською програмою, чи іншою процедурою. Змінювати скрипт процедуру може тільки її власник, або той, кому надано право на зміну. Існує 3 типи процедур: системні, тимчасові та користувальницькі.

Задача даної кваліфікаційної роботи полягає в створенні збереженої процедури SQL, яка буде розподіляти вибірку на кластери, а також мати змогу додавати до кластерів нові об’єкти при їх надходженні.

ТЗ отримано від керівника одного з підрозділів АТ «Банк Кредит Дніпро», саме на основі наданих Банком даних, необхідно провести кластеризацію для

сегментації клієнтів задля виявлення «поганих» клієнтів, саме тих, які ймовірно не будуть сплачувати кредит, задля мінімізації збитків Банку.

Загалом задача автоматизації та спрощування видачі кредиту виникла в 60-х – 70-х роках в США, коли активно почали запроваджуватися кредитні картки. При отриманні заявки на кредит від фізичної особи, усі необхідні характеристики отримуються в ході анкетування позичальника, та через інформації, яку Банк збирає самостійно. Описуватися ознаки клієнта можуть бінарно, номінально, кількісно або порядково. В ролі навчальної вибірки виступають клієнти з відомою кредитною історією, яку має Банк. Таким чином, рішення щодо видачі кредиту базується на тому, до якого класу віднесено клієнта: «поганого» чи «хорошого». Проте, для більш глибокого аналізу, окрім класу клієнта, можна орієнтуватися на його скоринг. Скоринг – це загальна кількість балів, яку отримує клієнт в ході аналізу його ознак. Надійність клієнта буде вважатися високою, якщо він отримав великий скоринговий бал.

Дані для кластеризації подані у вигляді таблиці БД, що зберігає аплікаційну інформацію щодо клієнтів, яка отримується в ході заповнення заявки на отримання кредиту. У таблиці 2.2 подано перелік назв стовпців таблиці.

Таблиця 2.2

Перелік стовпців таблиці з аплікаційними даними

Назва стовпця	Тип даних	Опис
-1-	**-2-**	**-3-**
id_order	рядок	Номер заявки
app_date	дата	Дата створення заявки
inn	рядок	ІПН клієнта
gender	рядок	Стать
Birthday	дата	Дата народження
Age	число	Вік
Family_Status	рядок	Сімейний статус

Industry	рядок	Індустрія, в якій працює клієнт

Продовження табл. 2.2

-1-	-2-	-3-
Education	рядок	Освіта
Empl_type	рядок	Тип працевлаштування
Position	рядок	Позиція, яку займає
Employee_Name	рядок	Назва місця роботи
DateStart_Employee	дата	Дата початку роботи
Num_Empl	число	Кількість найманих працівників
organization_type	рядок	Тип організації, в якій працює
speciality	рядок	Спеціалізація по місцю роботи
total_experience	число	Досвід роботи
income	число	Розмір місячного доходу
Status_last	рядок	Кінцевий статус заявки
Liv_State	рядок	Область проживання
Liv_City	рядок	Місто проживання
addressliv_date_start	дата	Дата початку проживання за фактичним адресом
addressliv_status_value	рядок	Житловий статус за фактичним адресом
Reg_State	рядок	Область реєстрації
Reg_City	рядок	Місто реєстрації
addressreg_dateregistration	дата	Дата реєстрації місця проживання
addressreg_status_value	рядок	Житловий статус за місцем реєстрації
Num_Of_children	число	Кількість дітей

Проаналізувавши таблицю 2.2, видно, що не всі стовпці мають один і той самий тип даних, деякі характеристики вимірюються числом, деякі – описані вербально. Перед початком кластеризації необхідно всі ознаки привести до однієї шкали вимірювання. Для спрощення, таблиця з даними, які будуть

кластеризуватися, буде складатися з 1 та 0. Проте, враховуючи, що стовпці, що мають тип даних «рядок», для кожного клієнта можуть набувати різних значень, наприклад, стовпець «Family_Status» може приймати значення «Одружений», «Вдівець», «Не одружена» і т.д. Зважаючи на це, кожний такий стовпець буде розділятися на декілька інших стовпців, враховуючи всі можливі варіанти відповіді клієнта.

Щоб врахувати всі можливі значення, які може приймати стовпець, необхідно з аплікаційної таблиці за допомогою оператора SELECT вибрати усі унікальні значення стовпців. Виконати запит можна наступним чином:

```
SELECT DISTINCT <column_name>
FROM <table_name>
```

Окрім цього, стопці з числовим типом даних, слід згрупувати з одним кроком, якщо мова йде про великі числа; стовпці, що приймають значення 1/0 залишаються незмінними. Стовпці, що мають велику варіацію ймовірних значень, також слід згрупувати по якомусь критерію, наприклад, стовпець «Reg_State», що містить перелік областей України, можна згрупувати по сторонам світу, тобто на північ, південь, схід та захід.

Перед початком приведення аплікаційних даних до однієї шкали, необхідно проаналізувати, які стовпці необхідні для кластеризації. Дана процедура не має якогось математичного рішення, вибір ознак, необхідних для кластеризації, проводиться експертно. Проаналізувавши таблицю 2.2, для кластеризації залишаються стовпці, що записані в таблиці 2.3.

Таблиця 2.3

Перелік стовпців таблиці для проведення кластеризації

Назва стовпця	Тип даних	Опис
-1-	**-2-**	**-3-**
id_order	рядок	Номер заявки
gender	рядок	Стать
Age	число	Вік
Family_Status	рядок	Сімейний статус

Industry	рядок	Індустрія, в якій працює клієнт

Продовження табл. 2.3

-1-	-2-	-3-
Education	рядок	Освіта
Empl_type	рядок	Тип працевлаштування
total_experience	число	Досвід роботи
income	число	Розмір місячного доходу
Liv_State	рядок	Область проживання
addressliv_status_value	рядок	Житловий статус за фактичним адресом
Num_Of_children	число	Кількість дітей

Таблиця 2.3 налічує в собі 12 параметрів, які на думку експерта важливі для кластеризації. Тепер необхідно для кожного стовпця ознайомитися з його можливими значеннями, та створити для кожного значення окремий стовпець. Якщо стовпець серед значень містить NULL чи пусту клітинку, то для них не створюються окремі стовпці.

Наступний важливий момент, аплікаційна таблиця налічує в собі десятки тисяч рядків, які ще будуть додаватися. Постає питання, яким чином з таблиці вигляду, що зображена на рисунку 2.5, зробити таблицю, що буде містити тільки 1 та 0, при чому виконати це за допомогою SQL запиту, та врахувати, що нова таблиця має заповнюватися по мірі додавання стовпця, тобто динамічно.

id	app_date	gender	Birthday	Age	Family_Status	Industry	Education	Empl_type	Postion	DateStart_Employee
58569	2024-05-14	M	1998-06-04	25	Одружений (заміжня)	Охорона здоров_я/ Фар...	Середня - технічна	Найманий на юрид...	Employee_in...	2017-11-02
88690	2024-06-12	F	2005-08-25	18	Не одружений (не заміжня)	Роздрібна торгівля	Середня - технічна	Найманий на юрид...	Employee_in...	2023-02-01
71364	2024-05-26	M	1998-07-01	25	Не одружений (не заміжня)	Оптова торгівля	Вища	Найманий на юрид...	Senior_man...	2020-08-05
59103	2024-05-14	F	1994-04-03	30	Не одружений (не заміжня)	Охорона здоров_я/ Фар...	Вища	Найманий на юрид...	Employee_in...	2013-08-01
58912	2024-05-14	F	1977-01-27	47	Не одружений (не заміжня)	Сфера обслуговування	Вища	Найманий на юрид...	Senior_man...	2002-08-17
58957	2024-05-14	M	1978-03-09	46	Розлучений (розлучена)	Виробництво (фабрики...	Середня - технічна	Найманий на юрид...	Senior_man...	2020-09-01
58510	2024-05-14	F	1992-01-09	32	Не одружений (не заміжня)	Роздрібна торгівля	Вища	Найманий на юрид...	Senior_man...	2021-09-01
62300	2024-05-17	F	1986-01-09	38	Не одружений (не заміжня)	Фінанси / Страхування	Вища	Найманий на юрид...	Employee_in...	2020-01-01
60533	2024-05-16	F	2002-03-26	22	Не одружений (не заміжня)	Готельний бізнес/ Рест...	Вища	Найманий на юрид...	Employee_in...	2021-09-01
58637	2024-05-14	M	1982-06-10	41	Одружений (заміжня)	Видобуток та переробк...	Дві і більше освіти	Найманий на юрид...	Employee_in...	2005-05-01
58781	2024-05-14	F	1988-11-19	35	Одружений (заміжня)	Сфера обслуговування	Вища	Найманий на юрид...	Employee_in...	2023-10-01
58760	2024-05-14	F	1996-03-05	28	Одружений (заміжня)	Сфера обслуговування	Середня - технічна	Найманий на юрид...	Senior_man...	2022-05-01
56078	2024-05-11	F	1975-11-20	48	Розлучений (розлучена)	Готельний бізнес/ Рест...	Вища	Найманий на юрид...	Senior_man...	2021-03-21
58831	2024-05-14	M	1971-08-24	52	Одружений (заміжня)	Охоронна діяльність	Середня - технічна	Найманий на юрид...	Employee_s...	2015-06-05
59274	2024-05-14	M	1982-11-10	41	Одружений (заміжня)	Органи влади та управл...	Середня - технічна	Найманий на юрид...	Employee_in...	2003-06-09

Рисунок. 2.5 – Фрагмент таблиці з аплікаційними даними клієнтів.

Щоб досягти динамічного додавання стовпця до таблиці, та її одночасного заповнення, слід використовувати динамічні запити SQL. Динамічний запит записується у вигляді рядку, та може містити стандартний запит на вибірку з таблиці, чи більш складний запит, який, наприклад, містить параметр, що змінюється.

Після огляду кожного стовпця аплікаційної таблиці, виявлено кількість унікальних значень, які можуть описувати стовпець. В таблиці 2.4 наведено назву стовпцю та кількість значень, які він може приймати, не враховуючи NULL або пусті клітинки; стовпець «id_order» буде залишатися незмінним, так як це ідентифікатор заявки, за допомогою якого буде визначатися, якому кластеру вона належить.

Таблиця 2.4

Перелік стовпців таблиці та кількість їх унікальних характеристик

Назва стовпця	Кількість значень
-1-	**-2-**
gender	2
Age	58
Family_Status	10
Industry	27
Education	7
Empl_type	4
total_experience	57
income	1 134
Liv_State	27
addressliv_status_value	5
Num_Of_children	13

Переглянувши таблицю 2.4, видно що, якщо дані не згрупувати, то таблиця буде мати 1 344 стовпців, що звичайно не є адекватним. Таким чином,

групуванню підвергнуться стовпці «Age», «total_experience», «income», «Liv_State» і «Num_Of_children». В результаті групування даних, таблиця буде налічувати 99 стовпців.

На рисунку 2.6 зображено фрагмент скрипту SQL, який просто створює майбутні назви для стовпців. Приклад додавання значень стовпця «Liv_State» у нову таблицю з даними наведено в Додатку В.

```
drop table #inserts
create TABLE #inserts (id int identity (1,1), to_insert NVARCHAR(150));

INSERT INTO #inserts (to_insert)
    select distinct
        case when Liv_State is null then 'Liv_State' + '_null'
            when Liv_State in ('Автономна Республіка Крим', 'Севастополь') then 'Liv_State_crimea'
            when Liv_State = 'Київ' then 'Liv_State_kyiv_city'
            when Liv_State in ('Луганська','Донецька') then 'Liv_State_lugansk_donetsk'
            when Liv_State in ('Житомирська','Київська','Чернігівська','Сумська') then 'Liv_State_north'
            when Liv_State in ('Волинська','Рівненська','Тернопільська','Хмельницька','Львівська'
                    ,'Чернівецька','Івано-Франківська','Закарпатська') then 'Liv_State_west'
            when Liv_State in ('Вінницька','Черкаська','Полтавська','Кіровоградська','Дніпропетровська')
                    then 'Liv_State_centr'
            when Liv_State in ('Одеська','Миколаївська') then 'Liv_State_south'
            when Liv_State in ('Запорізька','Херсонська','Харківська') then 'Liv_State_zp_khrsn_khrkv'
             end
    from назва_таблиці
```

Рисунок. 2.6 – Скрипт для групування значень стовпців.

В результаті використання скрипту з Додатку В для кожного стовпця, вийшла таблиця, що містить номер заявки, та 1 і 0 для кожного стовпця. Таким чином, враховуючи з таблиці 2.4, для стовпця «addressliv_status_value» можливі 5 варіантів, яким він може дорівнювати, тому для цього поля буде одна 1, а для інших чотирьох значень – 0.

Фрагмент кінцевої таблиці зображено на рисунку 2.7.

order	gender_M	gender_F	Family_Status_Вдівець (вдова)	Family_Status_Цивільний шлюб	Family_Status_Розлучений/розлучена	Family_Status_Розлучений (розлучена)
758569	1	0	0	0	0	0
788690	0	1	0	0	0	0
771364	1	0	0	0	0	0
759103	0	1	0	0	0	0
758912	0	1	0	0	0	0
758957	1	0	0	0	0	1
758510	0	1	0	0	0	0
762300	0	1	0	0	0	0
760533	0	1	0	0	0	0
758637	1	0	0	0	0	0
758781	0	1	0	0	0	0
758760	0	1	0	0	0	0
756078	0	1	0	0	0	1
758831	1	0	0	0	0	0
759274	1	0	0	0	0	0
763059	0	1	0	0	0	1
774381	1	0	0	0	0	0
758879	0	1	0	0	0	0
758608	0	1	0	0	0	0
758590	0	1	0	0	0	0
758789	0	1	0	0	0	0
759193	1	0	0	0	0	0
758667	1	0	0	0	0	0
758786	1	0	0	0	0	0
758959	0	1	0	0	0	0
759233	0	1	0	0	0	0
788273	1	0	0	0	0	0
761521	1	0	0	0	0	0
759073	0	1	0	0	0	0
759116	1	0	0	0	0	0
759869	0	1	0	0	0	0
773539	1	0	0	0	0	1

Рисунок. 2.7 – Фрагмент таблиці з перетвореними аплікаційними даними клієнтів.

Фінальна таблиця налічує в собі 46 074 рядків, проте, не всі вони будуть використовуватися для кластеризації. Так як, задача полягає в тому, аби знайти ознаки клієнтів, які схильні до несплати кредиту, то доцільно брати таких клієнтів, бо яким відома їх кредитна історія, тобто залишити ті заявки, по яким Банк звертався до УБКІ (Українське бюро кредитних історій) по їх КІ (кредитну історію). В результаті, кількість рядків в таблиці скоротиться до 5 639. Тепер маючи таблицю з нормованими параметрами, можна починати кластеризацію.

2.2.1 Реалізація методу k-середніх

Для визначення відстані між точкою та центроїдом буде використовуватися метрика Евкліда, що розраховується за формулою (1.12). Повний скрипт процедури для проведення кластеризації наведено в Додатку Г.

Перед початком кластеризації, необхідно подбати про те, як будуть проводитися розрахунки відстані, бо при великій кількості стовпців, скрипт

буде громіздкий, якщо в кожній формулі перелічувати всі 99 стовпця. Щоб цього уникнути, знову треба створювати динамічні запити. Фрагмент коду для динамічного обчислення відстані представлено на рисунку 2.8.

```
-- Динамічне обчислення відстані
SET @DistanceCalculation = '';

SELECT @DistanceCalculation = STUFF((
    SELECT ' + POWER(d.[' + c.name + '] - c.[' + c.name + '], 2)'
    FROM tempdb.sys.columns c
        JOIN tempdb.sys.objects o ON c.object_id = o.object_id
    WHERE o.name like '##t1%'
        AND o.type = 'U'
        and c.name <> 'ipn'
        and c.name not in ('id_order','cluster_id') -- зайва колонка
    FOR XML PATH('')), 1, 3, '');
```

Рисунок. 2.8 – Скрипт для обчислення відстані між точкою та центроїдом.

Пояснення щодо даного запиту. В операторі SELECT у вигляді рядку записується формула для обчислення відстані, при чому за допомогою вбудованих таблиць SQL, що зберігають всю інформацію про таблицю, динамічно додаються назви стовпців, на основі яких проводиться кластеризація.

Для призначення нового центроїду кластеру, необхідно створити запит, що буде зіставляти нове значення зі старим. Скрипт для такої процедури продемонстровано на рисунку 2.9. Так як, оновлення центроїду відбувається за мінімальним середнім арифметичним, то для даного розрахунку також необхідно додати динамічний запит, який буде це реалізовувати – рисунок 2.10.

```
SET @update_clusters = '';

SELECT @update_clusters = STUFF((
    SELECT ' , c.[' + c.name + '] = subquery.[' + c.name + '] '
    FROM tempdb.sys.columns c
        JOIN tempdb.sys.objects o ON c.object_id = o.object_id
    WHERE o.name like '##t1%'
        AND o.type = 'U'
        and c.name <> 'ipn'
        and c.name not in ('id_order','cluster_id') -- зайва колонка
    FOR XML PATH('')), 1, 3, '');
```

Рисунок. 2.9 – Скрипт для призначення нового центроїда кластера.

```
SET @avg_clusters = '';

SELECT @avg_clusters = STUFF((
    SELECT ' , avg([' + c.name + ']) as [' + c.name + '] '
    FROM tempdb.sys.columns c
        JOIN tempdb.sys.objects o ON c.object_id = o.object_id
    WHERE o.name like '##t1%'
        AND o.type = 'U'
        and c.name <> 'ipn'
        and c.name not in ('id_order','cluster_id') -- зайва колонка
    FOR XML PATH('')), 1, 3, '');
```

Рисунок. 2.10 – Скрипт для розрахунку середнього арифметичного.

Щоб контролювати зміну центроїда, необхідно задати якесь порогове значення, перевищення якого буде свідчити про зміни центроїда.

Динамічний запит для перевірки зображено на рисунку 2.11. Тут відбувається розрахунок модуля різниці між поточним центроїдом та попереднім.

```
SET @check_changes = '';

SELECT @check_changes = STUFF((
    SELECT ' or abs(c.[' + c.name + '] - p.[' + c.name + ']) > @epsilon '
    FROM tempdb.sys.columns c
        JOIN tempdb.sys.objects o ON c.object_id = o.object_id
    WHERE o.name like '##t1%'
        AND o.type = 'U'
        and c.name <> 'ipn'
        and c.name not in ('id_order','cluster_id') -- зайва колонка
    FOR XML PATH(''), TYPE).value('.', 'NVARCHAR(MAX)'), 1, 4, '');
```

Рисунок. 2.11 – Скрипт для перевірки зміни центроїда.

Продемонструвавши всі основні змінні, реалізацію алгоритму можна описати наступним чином:

1) Задати необхідну кількість кластерів та назви стовпців, по яким проводити кластеризацію.
2) Для коректної роботи динамічних запитів та записування результатів, створити тимчасову таблицю (рисунок 2.12), яка буде зберігати номер кластеру та об’єкти з їх характеристиками.

```
SET @SQL = 'select null cluster_id
    , [id_order]
    ,' + @Columns_name
    + 'into ##t1 from назва_таблиці
EXEC sp_executesql @SQL;
```

Рисунок. 2.12 – Скрипт для створення шаблону кінцевої таблиці.

3) Оголосити змінні для обчислення відстані від об’єктом та центроїдом (рис. 2.8), для оновлення центроїду (рис. 2.9), для розрахунку середнього арифметичного (рис. 2.10) та критерію зупинки алгоритму (рис. 2.11).
4) Створити тимчасову таблицю (рисунок 2.13), що буде зберігати центроїди кластерів. Додати до таблиці початкові центроїди для заданої кількості кластерів.

```
SET @SQL = 'CREATE TABLE ##km_clusters (id INT IDENTITY(1,1) PRIMARY KEY, '
           + @Columns_type + ');
           INSERT INTO ##km_clusters (' + @Columns_name + ')
               SELECT TOP (@NumClusters) ' + @Columns_name + '
               FROM ##t1'
EXEC sp_executesql @SQL, N'@NumClusters INT', @NumClusters;
```

Рисунок. 2.13 – Скрипт для створення тимчасової таблиці, що містить кластери та їх центроїди.

5) Запустити цикл, який триватиме, поки кластери не перестануть змінюватися.
6) Для порівняння центроїдів кластерів, створити таблицю (рисунок 2.14), яка буде містити попередні центроїди кластерів.

```
SET @SQL = 'IF OBJECT_ID(''tempdb..##prev_clusters'')
                IS NOT NULL drop table ##prev_clusters;
            SELECT * INTO ##prev_clusters
            FROM ##km_clusters';
EXEC sp_executesql @SQL;


-- Зберігання проміжних станів
SET @SQL = 'INSERT INTO ##km_steps
            SELECT * FROM ##km_clusters';
EXEC sp_executesql @SQL;
```

Рисунок. 2.14 – Скрипт для створення тимчасової таблиці, що містить інформацію щодо попередніх кластерів.

7) Призначити об’єкту кластер (рисунок 2.15), шляхом обчислення відстані між поточними центроїдами та об’єктом на входження для кластеру.

```
SET @SQL = 'update d
            set cluster_id = (
                select top 1 c.id
                from ##km_clusters c
                order by ' + @DistanceCalculation + ')
            from ##t1 d';
EXEC sp_executesql @SQL;
```

Рисунок. 2.15 – Скрипт для призначення кластера об’єкту.

8) Оновити центроїди кластерів (рисунок 2.16) через розрахунок середнього арифметичного між об’єктами в кластері.

```
SET @SQL = 'update c
            set ' + @update_clusters + '
            from ##km_clusters c
                join ( select cluster_id, ' + @avg_clusters +
                ' from ##t1 group by cluster_id) subquery
                    on c.id = subquery.cluster_id';
EXEC sp_executesql @SQL;
```

Рисунок. 2.16 – Скрипт для оновлення центру кластера.

9) Перевірити зміни кластерів (рисунок 2.17), якщо змін не відбулося, то цикл зупиняється.

```
SET @SQL = 'set @changes = (
            select count(*)
            from ##km_clusters c
                join ##prev_clusters p on p.id = c.id
            where ' + @check_changes + ' );';
EXEC sp_executesql @SQL, N'@epsilon FLOAT, @changes INT OUTPUT'
, @epsilon, @changes OUTPUT;
```

Рисунок. 2.17 – Скрипт для перевірки зупинки алгоритму.

Результатом обчислень будуть вхідна таблиця з додатковим полем з номером кластеру та центроїди кластерів. Вже на основі вихідної таблиці, необхідно згрупувати дані по кластерам, та розрахувати відсоток «поганих» клієнтів в кожному кластері. Де відсоток виявився найбільшим, отже кластер з

тими характеристиками відповідає за неплатоспроможних клієнтів. Повний скрипт процедури знаходиться в Додатку Г.

Так як, перетворена таблиця, що містить вхідні дані, складається з 99 стовпців, то слід додатково проаналізувати, що які характеристики можуть впливати на платоспроможність клієнта. З таблиці 2.4, яка налічує загальні ознаки клієнтів, можна детально проаналізувати кожне можливе значення, яке може призначатися групі.

Стовпці, що відповідають за стать людини, можна вважати не такими важливими, бо судити людину, виходячи з її статі, не є розумним, проте, для загальної статистики ці стовпці можна включити для розрахунків. Стопці, що згруповані по віку можна додати до розрахунків, проте, необхідно додатково проаналізувати можливі результати, якщо виключити якусь групу; стовпці з сімейним статусом можна скоротити до простих значень – «одружений»/«не одружений», за потреби можна буде розширити варіацію значень цього критерію.

Щодо стовпця з індустрією, то його однозначно необхідно скорочувати, залишивши найбільш вагомі індустрії та яким приділяється особлива увага. Стовпець з освітою також можна скоротити до базових значень – «вища»/«середня»/«незакінчена». Тип працевлаштування може виявляти великий вплив на платоспроможність клієнт, тому даний стовпець залишається без змін. Досвід роботи можна обмежити до 10 років, бо немає сенсу окремо дивитися тих, хто мають великий стаж, бо це скоріш за все пенсіонери, які і так окремо показані в групі «тип працевлаштування». Розмір щомісячного доходу також можна обмежити якоюсь сумою, так як очевидно, що людина з великим доходом мало імовірно буде прострочувати платежі по кредиту.

Для стовпця з областю проживання доцільно залишити ті, що налічують в собі області, які на теперішній момент тимчасово окуповані, чи являються зонами активних бойових дій. Тип житла не несе великої користі, але ці стовпці можна залишити, при необхідності – прибрати з аналізу. Стовпці з кількістю дітей також можна обмежити простими відповідями по типу «так» або «ні».

Таким чином, скоротивши кількість стовпців, що необхідні кластеризації, виходить таблиця, що налічує 77 стовпців. В ході експериментів, деякі стовпці будуть видалятися, якщо буде видно, що вони являються шумом.

Маючи вхідну таблицю даних, можна розпочати кластеризацію. Для початку, можна обрати 3 кластери. Результат кластеризації для трьох кластерів наведено на рисунку 2.18.

	cluster_id	count	bad_rate	bad_rate_%
1	2	5639	0.130237	13.023700

Рисунок. 2.18 – Результат кластеризації для 3-х кластерів.

Отримавши результати кластеризації для 3-х кластерів, можна зробити висновок, що задана дуже мала кількість кластерів, так як всі точки потрапили в один кластер. Необхідно повторити обчислення для 5-ти та 10-ти кластерів. Результат обчислень виявився ідентичним для 3-х кластерів.

Повторивши процедуру для 20-ти кластерів, виявлено, що вибірка розбивається на 2 кластери, проте, даний розв'язок не можна вважати оптимальним, оскільки більша частина вибірки згрупована в одному кластері, що не дає можливість оцінити «якість» цієї групи клієнтів. На рисунку 2.19 результат розбиття вибірки при заданих 20-ти кластерах.

	cluster_id	count	bad_rate	bad_rate_%
1	2	356	0.076923	7.692300
2	15	5283	0.134099	13.409900

Рисунок. 2.19 – Результат кластеризації для 20-ти кластерів.

Задавши 50 кластерів, на виході – 4 не пустих кластери, що представлені на рисунку 2.20. Загалом, аналізуючи результати, можна виділити другий кластер, всередині якого знаходиться майже 18% клієнтів, які мають

прострочку по кредиту. Проте, для перевірки можна видалити деякі стовпці з таблиці, та подивитися як працюватиме алгоритм.

	cluster_id	count	bad_rate	bad_rate_%
1	2	229	0.049295	4.929500
2	11	2647	0.179861	17.986100
3	15	2636	0.096178	9.617800
4	48	125	0.136363	13.636300

Рисунок. 2.20 – Результат кластеризації для 50-ти кластерів.

Переглянувши стовпці таблиці, можна виявити ті, які не впливатимуть на платоспроможність клієнта. До таких характеристик відноситься: сімейний статус, що визначений як «спільне проживання», освіта «науковий ступінь» або «2 і більше», вік старше 70 років, стаж роботи більше 50 років, більше 10 дітей, дохід більше 70 тис. грн. Видаливши ці стовпці, таблиця налічуватиме 66 стовпців.

Залишивши 50 кластерів, можна ще раз провести кластеризацію з оновленою таблицею, результат роботи алгоритму зображено на рисунку 2.21. Видно, що виділяються 2 кластери, проте, враховуючи, що вони мають різну кількість елементів, слід повторно провести кластеризацію, зменшивши кількість кластерів до 30-ти.

	cluster_id	count	bad_rate	bad_rate_%
1	2	2007	0.162767	16.276700
2	4	2636	0.096178	9.617800
3	12	204	0.045801	4.580100
4	32	693	0.189265	18.926500
5	35	97	0.414634	41.463400

Рисунок. 2.21 – Результат кластеризації на основі модифікованої таблиці даних для 50-ти кластерів.

На рисунку 2.22 відображено результат кластеризації при заданих 30-ти кластерів. Результат дещо схожий при кластеризації початкової вибірки 50-тьма кластерами. Таким чином, можна вважати, що дійсно видалені стовпці не були суттєво важливими для проведення кластеризації.

	cluster_id	count	bad_rate	bad_rate_%
1	2	2797	0.177068	17.706800
2	4	2636	0.096178	9.617800
3	12	204	0.045801	4.580100

Рисунок. 2.22 – Результат кластеризації на основі модифікованої таблиці даних для 30-ти кластерів

Окрім написаної процедури, що шукає центроїди кластерів, ТЗ гласить, що також необхідна процедура, що буде відносити нові об’єкти до вже існуючих кластерів. Варто наголосити, що нові об’єкти мають описуватися тим самим набором характеристик, що і ті дані, на основі яких проводилася кластеризація.

Для віднесення об’єкту до існуючого кластеру знадобиться таблиця з їх центроїдами та відстань між об’єктом та центроїдом. Для визначення відстані можна скористатися вже прописаним кодом на рисунку 2.8, а скрипт для додавання об’єкту до кластеру продемонстровано на рисунку 2.23.

```
SET @assign_sql = 'SELECT c.id, d.*
    INTO ##res
    FROM ' + @table_name + ' d
        CROSS APPLY (SELECT TOP 1 c.id FROM ' + @centre_name + ' c
            ORDER BY SQRT(' + @DistanceCalculation + ')
        ) AS c;
    ';
```

Рисунок. 2.23 – Скрипт для віднесення нового об’єкта до створеного кластеру.

2.3 Аналіз та оцінка результатів

Отримавши кластери, 18% яких це неплатоспроможні клієнти (рис. 2.21 та 2.22), необхідно проаналізувати, які характеристики притаманні цим клієнтам, та чи схожі кластери між собою. У таблиці 2.5 подано стовпці таблиці, які використовувалися для кластеризації та їх сума.

Таблиця 2.5

Перелік стовпців таблиці та сума значень всіх рядків

Назва стовпців	Сума значень при 30-ти кластерах	Сума значень при 50-ти кластерах
-1-	**-2-**	**-3-**
gender_M	244	234
gender_F	0	0
Family_Status_Вдівець (вдова)	6	1
Family_Status_Цивільний шлюб	6	6
Family_Status_Розлучений/розлучена	0	0
Family_Status_Розлучений (розлучена)	32	30
Family_Status_Не у шлюбі	0	0
Family_Status_Вдівець/вдова	0	0
Family_Status_У шлюбі	0	0
Family_Status_Не одружений (не заміжня)	144	143
Family_Status_Одружений (заміжня)	56	54
Industry_СБУ/ МВС/ Суд/ Прокуратура	4	4
Industry_Роздрібна торгівля	31	31
Industry_Гральний бізнес	1	1
Industry_Сфера обслуговування	5	5
Industry_Збройні сили (Військовослужбовці)	22	22
Industry_Оптова торгівля	5	5
Industry_Органи влади та управління/ Адміністративні органи/ Органи місцевого самоврядування	23	23
Education_Незакінчена вища	23	23
Education_Незакінчена середня	2	2

Продовження табл. 2.5

-1-	-2-	-3-
Education_Середня	25	23
Education_Середня - технічна	125	118
Education_Вища	68	67
Empl_type_Пенсіонер	9	0
Empl_type_Студент	0	0
Empl_type_Самозайнята особа з реєстрацією в ДФС	1	0
Empl_type_Найманий на юридичну особу	234	234
addressliv_status_value_Live_with_relatives	138	136
addressliv_status_value_Apartment_rent	34	34
addressliv_status_value_Personal_apartment	58	55
addressliv_status_value_Hostel	12	7
age_30_39	87	87
age_40_49	43	41
age_50_54	17	11
age_60_69	4	2
age_25_29	27	27
age_less_22	33	33
age_22_24	30	30
age_55_59	3	3
total_experience_0	0	0
total_experience_1_4	67	58
total_experience_15_19	44	44
total_experience_20_29	30	30
total_experience_30_49	10	9
total_experience_10_14	49	49
total_experience_5_9	44	44
Num_Of_children_1_2	60	59
Num_Of_children_0	176	167
Num_Of_children_3_4	6	6
Num_Of_children_5_9	2	2

Продовження табл. 2.5

-1-	-2-	-3-
income_1k_5k	6	1
income_less_1k	0	0
income_5k_10k	2	2
income_50k_70k	20	20
income_10k_20k	69	65
income_30k_50k	70	69
income_20k_30k	67	67
income_70k_100k	6	6
Liv_State_south	41	39
Liv_State_crimea	0	0
Liv_State_zp_khrsn_khrkv	22	22
Liv_State_centr	53	48
Liv_State_lugansk_donetsk	1	1
Liv_State_kyiv_city	31	31
Liv_State_north	36	36
Liv_State_west	60	57

З таблиці 2.5 видно, що утворені кластери подібні, таким чином можна вважати, що видалені стовпці не містили важливої інформації. Проаналізувавши концентрацію по стовпцям, можна виділити такі, по яким не було спрацювання, тобто в кластері немає клієнта, у якого той чи інший стовпець дорівнює 1. Такі стовпці можна прибрати та ще раз здійснити кластеризацію.

В результаті, видаливши декілька стовпців та скоротивши число кластерів до 25-ти, отримано результат, що зображений на рисунку 2.24. На рисунку видно, що результат схожий з рисунком 2.22.

	cluster_id	count	bad_rate	bad_rate_%
1	2	150	0.129870	12.987000
2	3	2840	0.092298	9.229800
3	6	2647	0.179861	17.986100

Рисунок. 2.24 – Результат кластеризації на основі модифікованої таблиці даних для 25-ти кластерів.

Оскільки, провівши декілька експериментів, та модифікувавши таблицю, зафіксовано, що розбиття кластерів не змінюється, тому можна вважати, що кластеризація виконана якісно.

Таким чином, на основі таблиці 2.5, можна визначити, які саме ознаки характеризують клієнта як «поганого».

Перша ознака – це стать. В ході проведення кластеризації було виявлено, що серед тих, хто має на зараз прострочений кредит, більша частина це чоловіки. Звичайно, що Банк не може запровадити таке правило, що буде забороняти кредитувати чоловіків, тому така ознака як «стать» більше несе статистичний характер, а не міру, по якій варто розділяти клієнтів.

Друга ознака – сімейний статус. Найбільша концентрація спостерігається серед неодружених клієнтів, в два рази менше – одружених. Така ознака теж більше несе інформаційний характер, пріоритет в неї не великий. Хоча на ній можна будувати гіпотези, наприклад, ті, хто одружені мають більше шансів заплатити кредит, якщо в сім'ї працюють двоє, таким чином відсоток сплати по кредиту збільшується за рахунок працюючого партнера, який в теорії буде допомагати сплатити кредит.

Третя ознака – індустрія, в якій працює клієнт. Для цієї ознаки спеціально було виділено декілька значень, на які Банк звертає особливу увагу при розгляданні заявки на кредит. До них входять: працівники СБУ, представники грального бізнесу (казино тощо), військовослужбовці (Банк кредитує не всіх військовослужбовців по очевидним причинам, так як під час війни даний сегмент клієнтів стає ненадійним, тому Банк може нести фінансові втрати), та представники органів влади. Найбільша концентрація спостерігається серед

робітників в сфері роздрібної торгівлі, на другому місці – військовослужбовці та представники органів влади. Проте, враховуючи, що кількість клієнтів, які мають прострочку, складає 234, то концентрація в 13% не буде вважатися суттєвою.

Четверта ознака – освіта. Найбільший відсоток припадає на клієнтів, що мають середню технічну освіту, тобто провчились в коледжах, ПТУ тощо. Дана ознак теж вважається додатковою, без великого пріоритету.

П’ята ознака – тип працевлаштування. Тут найбільша концентрація серед тих, хто найманий на юридичну особу. Таким чином, можна зробити додатковий висновок, що юридичні особи схильні до погашення кредитної заборгованості без пропускання щомісячних платежів.

Шоста ознака – тип житлового проживання. Переважна частина відповідає за клієнтів, які проживають з родичами, тобто не мають власного житла або орендованого. Виходячи з цього, можна зробити припущення, що такі клієнти будуть мати труднощі аби самостійно погасити кредит.

Сьома ознака – вік. Тут можна виділити дві категорії: молодші 30-ти років, та люди віком 30-50 років. Концентрація серед молодших 30-ти, розділяється на три рівносильні групи – молодші 22-х років, віком 22-24 роки та відповідно 25-29 років. Тим, хто молодше 22-х років Банк автоматично відмовляє в кредиті, що навіть підтвердила кластеризація. Інші групи займають не таку долю, як клієнти віком 30-39 років, серед них найбільша концентрація по несплати кредиту.

Восьма ознака – досвід роботи. Найбільша концентрація спостерігається серед тих, у кого досвід роботи не перевищує 5-ти років, наступними йдуть люди з досвідом роботи в межах 5-20 років.

Дев’ята ознака – кількість дітей. Можна зробити припущення, що якщо у людини є дитина, то вона витрачає більше коштів, саме тому це може заважати при сплаті кредиту, проте, припущення не підтвердилося, так як найбільше не сплачують кредит ті, хто немає дітей, в три рази менше доля тих, у кого 1-2 дитини.

Десята ознака – щомісячний дохід. Одна з найважливіших ознак, по якій оцінюється платоспроможність клієнта. В утвореному кластері найбільша концентрація припадає клієнтів з доходом – 10-30 тис. грн. на місяць, що в залежності від розміру кредиту може вважатися недостатньою сумою задля вчасного його погашення.

Остання, одинадцята ознака – область проживання. В умовах війни, доречно зробити припущення, що люди з територій, де ведуться активні бої, схильні до несплати кредиту, що є зрозумілим. Проте, за результатами кластеризації, серед таких «небезпечних» областей не було виявлено концентрації, всі області розподілені рівномірно.

Виходячи з проаналізованих ознак, можна зробити висновок, що необхідно провести додаткові дослідження та побудувати аналітику серед клієнтів, які мають середню-технічну освіту, не мають окремого житла, віком 30-39 років, з щомісячним доходом 10-30 тис. грн.

Слід зазначити, що кваліфікаційнв робота несе більш демонстраційний характер, що не означає, що отримані результати обов’язково будуть включені до розгляду, так як кластеризація в даній задачі являється лиш одним із етапів комплексного аналізу потоку клієнтів, для створення скорингової моделі, що буде прискорювати процес прийняття рішень по заявкам на видачу кредиту.

2.4 Висновок до Розділу 2

Під час дослідження, було виявлено, що деякі параметри не є інформативними, тому в ході повторення дослідження їх було виключено з розрахунків. Задля виявлення структур в даних, доводилося задавати велику кількість кластерів, що в результаті розподіляло об’єкти на меншу кількість груп. Багато кластерів виявлялися пустими. Дану проблему можна списати на те, що точки центроїдів задавалися випадковим чином, що впливає на якість кластеризації, проте все ж таки, дані, що розбиті на кластери мають достатній рівень розбиття, тобто можна виявити концентрацію.

Також в результаті довелося написати збережену процедуру, яка успішно та швидко проводить кластеризацію методом k-середніх, а також додаткову процедуру, що класифікує нові об’єкти за попередньо виявленими кластерами, що дозволяє не проводити кожного разу нову кластеризацію при додаванні даних до вибірки.

Після застосування методу кластеризації, було виявлено, що слід звернути увагу на клієнтів, що підпадають під перелічені характеристики: середня-технічна освіта, відсутність окремого житла, вік 30-39 років, щомісячний дохід 10-30 тис. грн. За отриманим результатом буде висунуте припущення, яке потребує додаткових досліджень.

ВИСНОВОК

Кваліфікаційна робота на здобуття ступеня «магістра» була присвячена розробці процедури на SQL Server для проведення кластеризації методом k-середніх, з метою сегментації клієнтів АТ «Банк Кредит Дніпро», щоб виявити характеристики, які притаманні клієнтам, які потенційно являються «поганими», тобто з імовірністю будуть неплатоспроможними. Кластеризація проводилася на основі аплікаційних даних клієнта, що отримуються під час їх анкетування.

В результаті дослідження було виявлено характеристики, які не несуть корисної інформації, такі характеристики було виключено з дослідження, з метою отримання більш якісного результату. Під час проведення кластеризації було використано різну кількість кластерів, під час невеликої кількості кластерів, всі об’єкти потрапляли в один кластер, або розподілялися на два кластери, при чому більша частина об’єктів була зосереджена в одному кластері. Такі отримані результати не вважалися якісними. При визначенні більшої кількості кластерів, об’єкти об’єднувалися в меншу кількість кластерів, проте маючи достатню концентрацію, щоб можна було побудувати висновки.

Виконавши кластеризацію та проаналізувавши результати, було виділено ряд ознак, які мали клієнти, що на даний момент вважаються «поганими». Поганий клієнт чи ні оцінювалося наявністю в нього простроченої заборгованості на момент подачі заявки на отримання кредиту. Таким чином, було встановлено, що клієнти з середньо-технічною освітою, які не мають окремого житла, віком 30-39 років, з щомісячним доходом 10-30 тис. грн., мають заборгованість по кредиту. Для статистики було виявлено, що серед таких клієнтів виключно чоловіки.

Отримані результати кластеризації необхідно додаткового дослідити по іншим факторам, та вже на основі комплексного аналізу, висунути гіпотезу, яка

підлягає тестуванню. На основі сформованої гіпотези буде будуватися нова скорингова модель Банку.

Варто зазначити, що результати кластеризації не можна вважати ідеальними, що можна пов’язати з тим, що по-перше, дані не можна було представити графічно, а по-друге, обране середовище для кластеризації, не зовсім підходить для даної задачі.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1) Кваліфікаційна робота магістра [Електронний ресурс] : методичні рекомендації для здобувачів ступеня магістра освітньо-професійної програми «Системний аналіз» зі спеціальності 124 Системний аналіз / уклад.: Т.А. Желдак, Т.В. Хом'як, А.В. Малієнко ; М-во освіти і науки України, Нац. техн. ун-т «Дніпровська політехніка». – Дніпро : НТУ «ДП», 2024. – 33 с. https://ir.nmu.org.ua/handle/123456789/167921
2) Основні поняття кластеризації та постановка задачі. URL: https://csc.knu.ua/media/study/asp/mod_probl_inf_tech_sys_analysis_ivohin/lecture/lec11.pdf (дата звернення 12.11.2024 р.)
3) Айвазян, С.А. Классификация многомерных наблюдений / С.А. Айвазян, З.И. Бежаева, О.В. Староверов. – Москва: Статистика, 1974. – 240 с.
4) АТ «Банк Кредит Дніпро». Про Банк. URL: https://creditdnepr.com.ua/pro-bank (дата звернення 30.10.2024 р.)
5) АТ «Банк Кредит Дніпро». Проміжна скорочена фінансова звітність станом на та за період, що закінчився 30 червня 2024 р. URL: https://creditdnepr.com.ua/sites/default/files/ifrs_fs_ukr_iiq2024.pdf (дата звернення 30.10.2024 р.)
6) Бази даних [Текст]: метод. вказівки до виконання комп'ютерного практикуму для студентів спеціальності "Електронні комунікації та радіотехніка» / Уклад.: Суліма С.В., Глоба Л.С., Скулиш М.А.. – К.: КПІ ім. Ігоря Сікорського, 2023. – 54 с.
7) Банк Кредит Дніпро кредит готівкою 2024: умови, калькулятор, онлайн-заявка. URL: https://finsee.com/%D0%B1%D0%B0%D0%BD%D0%BA-%D0%BA%D1%80%D0%B5%D0%B4%D0%B8%D1%82-%D0%B4%D0%BD%D1%96%D0%BF%D1%80%D0%BE/%D0%BA%D1%80%D0%B5%D0%B4%D0%B8%D1%82-

%D0%B3%D0%BE%D1%82%D1%96%D0%B2%D0%BA%D0%BE%D1%8E/ (дата звернення 31.10.2024 р.)

8) Банк Кредит Дніпро. URL: https://minfin.com.ua/ua/company/credit-dnepr/ (дата звернення 30.10.2024 р.)
9) Бізнес-аналітика багатовимірних процесів. URL: http://ebooks.git-elt.hneu.edu.ua/babap/about.html (дата звернення 12.11.2024 р.)
10) Вступ до інтелектуального аналізу даних. Частина 1. Кластерний аналіз та регресійний аналіз: навчальний посібник. / Є.А. Настенко, В.А. Павлов, О.К. Городецька, К.С. Бовсуновська. – КПІ ім. Ігоря Сікорського, 2023. – 131 с.
11) Гитис Л. Х. Статистическая классификация и кластерный анализ. – М.: Издательство Московского государственного горного университета, 2033. – 157 с.
12) Дубровская Л.И., Князев Г. Б. Компьютерная обработка естественно-научных данных методами многомерной прикладной статистики: Учебное пособие. – Томск: ТМЛ-Пресс, 2011, – 120 с.
13) Інструкція користувача з вінтажного аналізу кредитного портфеля розрібного бізнесу АТ "Банк Кредит Дніпро"
14) Інтелектуальний аналіз даних: підручник. / О. І. Черняк, П. В. Захарченко. – Київ, 2010. – 837 с.
15) Кваліфікаційна робота магістра [Електронний ресурс] : методичні рекомендації для здобувачів ступеня магістра освітньо-професійної програми «Системний аналіз» зі спеціальності 124 Системний аналіз / уклад.: Т.А. Желдак, Т.В. Хом’як, А.В. Малієнко ; М-во освіти і науки України, Нац. техн. ун-т «Дніпровська політехніка». – Дніпро : НТУ «ДП», 2024. – 33 с.
16) Кластерный анализ : терминология, методы, задачи / Е. Ю. Леончик. – Одесса: 2011. – 68 с.
17) Купалова Г.І. Теорія економічного аналізу : [навч. посіб.] – К.: Знання, 2008. – 639 с.

18) Курченко О.О., Рабець К.В. Метричні простори у курсі математичного аналізу: навчальний посібник. / О.О.Курченко, К.В.Рабець. – К., 2011. – 146 с.
19) Машинне навчання: комп'ютерний практикум з дисципліни «Машинне навчання» [Електронний ресурс]: навч. посіб. для студ. спеціальності 121 «Інженерія програмного забезпечення» (освітня програма «Інженерія програмного забезпечення мультимедійних та інформаційно-пошукових систем»)/ Л.М. Олещенко; КПІ ім. Ігоря Сікорського. – Електронні текстові дані. – Київ: КПІ ім. Ігоря Сікорського, 2022. – 92 с.
20) Методи аналізу даних : навчальний посібник для студентів / В.Є. Бахрушин. – Запоріжжя : КПУ, 2011. – 268 с.
21) Модели системного анализа в управлении экономическими процессами / Под ред. Докт. Экон. Наук, проф. В.С. Пономаренко, докт. Экон. Наук, проф. Т.С. Клебановой, докт. Экон. Наук, проф. Л.С. Гурьяновой – Братислава-Харьков, ВШЭМ – ХНЭУ им. С. Кузнеца, 2021. – 476 с. Укр. Яз., русск. Яз., англ. Яз.
22) Наглядова статистика. URL: https://bank.gov.ua/ua/statistic/supervision-statist#1 (дата звернення 11.11.2024 р.)
23) Основна інформація про Банк Кредит Дніпро. URL: https://finsee.com/%D0%B1%D0%B0%D0%BD%D0%BA-%D0%BA%D1%80%D0%B5%D0%B4%D0%B8%D1%82-%D0%B4%D0%BD%D1%96%D0%BF%D1%80%D0%BE/#%D1%96%D0%BD%D1%84%D0%BE%D1%80%D0%BC%D0%B0%D1%86%D1%96%D1%8F (дата звернення 31.10.2024 р.)
24) Пістунов І.М., Антонюк О.П., Турчанінова І.Ю. Кластерний аналіз в економіці: Навч. посібник – Дніпропетровськ: Національний гірничий університет, 2008. – 84 с.
25) Тема 10. Методи кластерного аналізу. Ієрархічні методи. URL: https://moodle.znu.edu.ua/pluginfile.php/486140/mod_resource/content/1/%D0%

9B%D0%B5%D0%BA%D1%86%D1%96%D1%8F%2010.pdf (дата звернення 05.12.2024 р.)

26) Факторный, дискриминантный и кластерный анализ: Пер. с англ. / Дж.-О. Ким, Ч. У. Мьюллер, У. Р. Клекка и др.; Под ред. И. С. Енюкова. – М.: Финансы и статистика, 1989. – 215 с.

27) Хом'як Т. В. Бази даних у професійних задачах аналітики [Електронний ресурс] : навч. наочн. посіб. / Т. В. Хом'як, К. С. Хабарлак, Д.М. Гаранжа; М-во освіти і науки України, Нац. техн. ун-т «Дніпровська політехніка». − Дніпро : НТУ «ДП», 2024. − 192с.

28) Что такое SQL Server Management Studio (SSMS)? URL: https://learn.microsoft.com/ru-ru/sql/ssms/sql-server-management-studio-ssms?view=sql-server-ver16 (дата звернення 11.12.2024 р.)

29) Машинне навчання [Електронний ресурс] : методичні рекомендації до виконання практичних робіт для здобувачів ступеня магістра освітньопрофесійної програми «Системний аналіз» зі спеціальності 124 Системний аналіз / уклад.: Т.А. Желдак, О.Б. Владико, А.В. Малієнко, Д.М. Гаранжа ; М-во освіти і науки України, Нац. техн. ун-т «Дніпровська політехніка». – Дніпро : НТУ «ДП», 2024. – 48 с. https://ir.nmu.org.ua/handle/123456789/167920

30) Молоканова, В. М., & Шевченко, Ю. О. (2024). Управління проектною командою. https://ir.nmu.org.ua/handle/123456789/167646

31) Аналіз та обробка великих даних [Електронний ресурс] : методичні рекомендації до виконання практичних робіт для здобувачів ступеня магістра освітньо-професійної програми «Системний аналіз» зі спеціальності 124 Системний аналіз / М-во освіти і науки України, Нац. техн. ун-т «Дніпровська політехніка». – Дніпро : НТУ «ДП», 2024. – 82 с. https://ir.nmu.org.ua/handle/123456789/167968

ДОДАТОК А. Відомість матеріалів кваліфікаційної роботи

№ з/п	Позначення	Найменування	Кількість аркушів	Примітки
1				
2		Документація		
3				
4	124.КР.24.11.ПЗ	Пояснювальна записка	80	Формат А4
5				
6		Демонстраційний матеріал		Презентація на CD-R
7				
8		Копія роботи	1	Диск CD-R
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				

					124.КР.24.07.ДА.ПЗ.
Змін.	Аркуш	№ докум.	Підпис	Дата	

Розроб.		Кочерга В.С.			**Матеріали кваліфікаційної роботи**	Літ.	Аркуш	Аркушів
К. розд.		Хом'як Т.В.						
Керівн.		Хом'як Т.В.				НТУ «ДП», 12; 124м-23-1		
Н.контр.		Хом'як Т.В.						
Зав. каф.		Желдак Т.А.						

ДОДАТОК Б. Відгук керівника кваліфікаційної роботи

Відгук
на кваліфікаційну роботу магістра
студентки групи *124м – 23 – 1*
спеціальності *124 Системний аналіз*

Тема кваліфікаційної роботи: «Застосування алгоритмів кластеризації для сегментації клієнтів банку».

Обсяг кваліфікаційної роботи 80 стор.

Мета кваліфікаційної роботи: виявити ознаки неплатоспроможного клієнта.

Актуальність теми обумовлена початком повномасштабного вторгнення на територію України, що збільшило ризик втрат коштів банківських установ.

Тема кваліфікаційної роботи безпосередньо пов'язана з об'єктом діяльності магістра спеціальності 124 Системний аналіз, оскільки полягає в кластеризації набору даних з його попереднім аналізом, та створення відповідного програмного забезпечення для організації роботи.

Виконані в кваліфікаційній роботі завдання відповідають вимогам другого (магістерського) рівня вищої освіти. Оригінальність наукових рішень полягає в отриманні інформації під час сегментації клієнтів, використовуючи методи кластерного аналізу, та написанні програмного коду, результатом якого є кластерування набору даних на SQL.

Практичне значення результатів кваліфікаційної роботи полягає в їх застосуванні для оцінки клієнтів, та формуванню гіпотези щодо ознак, що характеризують неплатоспроможних клієнтів.

Висновки підтверджують можливість використання результатів роботи в умовах розробки моделі для призначенню клієнту балів, та прийняття стратегічних рішень керівниками Банку.

Оформлення пояснювальної записки та демонстраційного матеріалу до неї виконано згідно з вимогами. Роботу виконано самостійно, відповідно до завдання та у повному обсязі.

Робота має практичну цінність, тому зауважень немає.

Кваліфікаційна робота в цілому заслуговує оцінки: 100б (відмінно).

З урахуванням висловлених зауважень автор заслуговує присвоєння кваліфікації «магістр з системного аналізу».

Керівник кваліфікаційної роботи магістра,

к.ф.-м.н., доц. _____ Хом’як Т.В.

ДОДАТОК В. Скрипт для динамічного додавання стовпця в таблицю БД

```
declare @n int = len('Liv_State_')
DECLARE @i INT = 1; -- Лічильник
DECLARE @maxStatuses INT = (SELECT COUNT(*) FROM #inserts); -- Кількість статусів
DECLARE @statusName NVARCHAR(550); -- Змінна для зберігання статусу
DECLARE @sql NVARCHAR(MAX); -- Змінна для динамічного SQL

WHILE @i <= @maxStatuses
BEGIN
   -- Отримуємо статус з таблиці
   SELECT @statusName = to_insert FROM #inserts WHERE Id = @i;

   -- Формуємо SQL-запит для додавання нового стовпця
   SET @sql = 'ALTER TABLE <table_name>' + ' ADD [' + @statusName + '] varchar(200)';

   EXEC sp_executesql @sql; -- Виконуємо SQL-запит

      SET @SQL = 'UPDATE <table_name> SET [' + @statusName +
                  '] = case when case when pos.Liv_State is null then "null"
                  when Liv_State in ("Автономна Республіка Крим", "Севастополь")
                        then "crimea"
                  when Liv_State = "Київ" then "kyiv_city"
                  when Liv_State in ("Луганська","Донецька") then "lugansk_donetsk"
                  when Liv_State in ("Житомирська","Київська","Чернігівська","Сумська")
                        then "north"
                  when Liv_State in ("Волинська","Рівненська","Тернопільська","Хмельницька"
                        ,"Львівська","Чернівецька","Івано-Франківська","Закарпатська")
                        then "west"
                  when Liv_State in ("Вінницька","Черкаська","Полтавська","Кіровоградська"
                        ,"Дніпропетровська")
                        then "centr"
                  when Liv_State in ("Одеська","Миколаївська") then "south"
                  when Liv_State in ("Запорізька","Херсонська","Харківська")
                        then "zp_khrsn_khrkv"
                   end = '" + right(@statusName, len(@statusName) - @n) + '" then 1 else 0 end '
                  + 'FROM <table_name1> kl
                        left JOIN <table_name2> pos
                              ON kl.id_order = pos.id_order';

      EXEC sp_executesql @SQL;

   SET @i = @i + 1; -- Збільшуємо лічильник
```

END

ДОДАТОК Г. Скрипт процедури для кластеризації методом k-середніх

```
CREATE PROCEDURE [procedure_name]

@NumClusters INT,
@Columns_name varchar(max)

AS
BEGIN
	SET NOCOUNT ON;

declare @Columns_type varchar(max) = REPLACE(@Columns_name, ', ', ' float, ') + ' float';

DECLARE @SQL NVARCHAR(MAX);
DECLARE @changes INT = 1;  -- Для перевірки, чи змінились кластери
DECLARE @epsilon FLOAT = 0.00001;  -- Поріг для зміни в координатах

DECLARE @DistanceCalculation NVARCHAR(MAX);
DECLARE @update_clusters NVARCHAR(MAX);
DECLARE @avg_clusters NVARCHAR(MAX);
DECLARE @check_changes NVARCHAR(MAX);

IF OBJECT_ID('tempdb..##km_clusters') IS NOT NULL DROP TABLE ##km_clusters
IF OBJECT_ID('tempdb..##km_steps') IS NOT NULL DROP TABLE ##km_steps
IF OBJECT_ID('tempdb..##prev_clusters') IS NOT NULL drop table ##prev_clusters
IF OBJECT_ID('tempdb..##t1') IS NOT NULL DROP TABLE ##t1;

SET @SQL = 'select null cluster_id
	, [id_order]
	,' + @Columns_name
	+ 'into ##t1 from <table_name>';
EXEC sp_executesql @SQL;

SET @DistanceCalculation = ''; -- Динамічне обчислення відстані

SELECT @DistanceCalculation = STUFF((
	SELECT ' + POWER(d.[' + c.name + '] - c.[' + c.name + '], 2)'
	FROM tempdb.sys.columns c
		JOIN tempdb.sys.objects o ON c.object_id = o.object_id
	WHERE o.name like '##t1%'
		AND o.type = 'U'
		and c.name <> 'ipn'
		and c.name not in ('id_order','cluster_id') -- зайва колонка
```

```
        FOR XML PATH('')), 1, 3, '');
SET @update_clusters = '';

SELECT @update_clusters = STUFF((
        SELECT ' , c.[' + c.name + '] = subquery.[' + c.name + '] '
        FROM tempdb.sys.columns c
                JOIN tempdb.sys.objects o ON c.object_id = o.object_id
        WHERE o.name like '##t1%'
                AND o.type = 'U'
                and c.name <> 'ipn'
                and c.name not in ('id_order','cluster_id') -- зайва колонка
        FOR XML PATH('')), 1, 3, '');

SET @avg_clusters = '';

SELECT @avg_clusters = STUFF((
        SELECT ' , avg([' + c.name + ']) as [' + c.name + '] '
        FROM tempdb.sys.columns c
                JOIN tempdb.sys.objects o ON c.object_id = o.object_id
        WHERE o.name like '##t1%'
                AND o.type = 'U'
                and c.name <> 'ipn'
                and c.name not in ('id_order','cluster_id') -- зайва колонка
        FOR XML PATH('')), 1, 3, '');

SET @check_changes = '';

SELECT @check_changes = STUFF((
        SELECT ' or abs(c.[' + c.name + '] - p.[' + c.name + ']) > @epsilon '
        FROM tempdb.sys.columns c
                JOIN tempdb.sys.objects o ON c.object_id = o.object_id
        WHERE o.name like '##t1%'
                AND o.type = 'U'
                and c.name <> 'ipn'
                and c.name not in ('id_order','cluster_id') -- зайва колонка
        FOR XML PATH(''), TYPE).value('.', 'NVARCHAR(MAX)'), 1, 4, '');

-- Вибір початкових центрів
SET @SQL = 'CREATE TABLE ##km_clusters (id INT IDENTITY(1,1) PRIMARY KEY, '
                + @Columns_type + ');
                        INSERT INTO ##km_clusters (' + @Columns_name + ')
                                SELECT TOP (@NumClusters) ' + @Columns_name + 'FROM ##t1'
EXEC sp_executesql @SQL, N'@NumClusters INT', @NumClusters;

SET @SQL = 'CREATE TABLE ##km_steps (id INT , ' + @Columns_type + ')';
EXEC sp_executesql @SQL;
```

```
-- Цикл, що триває до моменту, поки кластери не перестануть змінюватись
WHILE @changes > 0
BEGIN
	-- Зберігання поточного стану кластерів у таблиці для порівняння
   SET @SQL = 'IF OBJECT_ID("tempdb..##prev_clusters") IS NOT NULL
                        drop table ##prev_clusters;
                        SELECT * INTO ##prev_clusters
                        FROM ##km_clusters';
   EXEC sp_executesql @SQL;

	-- Зберігання проміжних станів
	SET @SQL = 'INSERT INTO ##km_steps SELECT * FROM ##km_clusters';
	EXEC sp_executesql @SQL;

	-- Призначення кластерів для точок даних
	SET @SQL = 'update d
                        set cluster_id = (
                                select top 1 c.id
                                from ##km_clusters c
                                order by ' + @DistanceCalculation + ')
                        from ##t1 d';
	EXEC sp_executesql @SQL;

	-- Оновлення центрів кластерів
	SET @SQL = 'update c
                        set ' + @update_clusters + '
                        from ##km_clusters c
                                join ( select cluster_id, ' + @avg_clusters + '
                                        from ##t1 group by cluster_id) subquery
                                        on c.id = subquery.cluster_id';
	EXEC sp_executesql @SQL;

	-- Перевірка змін у центрах кластерів
	SET @SQL = 'set @changes = (
                        select count(*)
                        from ##km_clusters c
                                join ##prev_clusters p on p.id = c.id
                        where ' + @check_changes + ' );';
	EXEC sp_executesql @SQL, N'@epsilon FLOAT, @changes INT OUTPUT', @epsilon,
@changes OUTPUT;

end

drop table <result_table>
select * into <result_table> from ##t1
```

```
drop table <result_table1>
select * into <result_table1> ##km_clusters

DROP TABLE ##km_clusters
DROP TABLE ##km_steps
drop table ##prev_clusters
drop table ##t1

select distinct t1.cluster_id
	, t3.count
	, avg(t2.Triger_3*1.0) [bad_rate]
	, avg(t2.Triger_3*1.0) * 100 [bad_rate_%]
from <result_table> t1
	join (select pos.id_order, tr.Triger_1, Triger_3
			from <result_table2> pos
				join <result_table3> tr on tr.inn = pos.inn
					and tr.date1 = dateadd(d,-1,pos.app_date)
		)t2 on t1.id_order = t2.id_order and t2.Triger_1 = 1
	left join (select distinct cluster_id, count(*) [count] f
			rom <result_table> group by cluster_id )t3 on t1.cluster_id = t3.cluster_id
group by t1.cluster_id
	, t3.count
order by 1

END
```