Міністерство освіти і науки України
Національний університет «Києво-Могилянська академія»
Факультет інформатики
Кафедра математики

Кваліфікаційна робота

освітній ступінь – бакалавр

на тему: **«Методи регуляризації в задачах кластеризації»**

Виконав: студент 4-го року навчання
освітньої програми «Прикладна математика»,
спеціальності 113 Прикладна математика

Міщеня Володимир Володимирович

Керівник: Крюкова Г. В.
кандидат фіз.-мат. наук, доцент

Рецензент:

Кваліфікаційна робота захищена
з оцінкою _____

Секретар ЕК_____
(підпис)

«_____» _____ 20___р.

Київ - 2025

Міністерство освіти і науки України
Національний університет «Києво-Могилянська академія»
Факультет інформатики
Кафедра математики

ЗАТВЕРДЖУЮ
Зав.кафедри математики,
доцент, кандидат фіз.-мат. наук
_____ *Чорней Р.К.*
(підпис)
“____”_____ 2025

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ
для кваліфікаційної роботи
студенту 4-го курсу, факультету інформатики
Міщені Володимиру Володимировичу

Тема: «Методи регуляризації в задачах кластеризації»

Зміст кваліфікаційної роботи:

Анотація

1. Вступ
2. Основні поняття та означення кластеризації
3. Методи регуляризації в задачах кластеризації, огляд підходів та способів застосування
4. Аналіз та порівняння результатів кластеризації з використанням різних методів регуляризації

Висновки

Список літератури

Дата видачі “____”_____ 2025 Керівник _____
(підпис)
Завдання отримав _____
(підпис)

Графік підготовки кваліфікаційної роботи до захисту

Графік узгоджено «_____» _____ 2025р.

№ з/п	Перелік робіт	Термін виконання етапу	Підпис наукового керівника	Дата ознайомлення наукового керівника	Примітка
1.	Отримання теми кваліфікаційної роботи.	16.09.2024			
2.	Ознайомлення з темою кваліфікаційної роботи.	27.09.2024			
3.	Розробка плану та структури роботи.	19.10.2024			
4.	Робота з науковою літературою, опис основних означень. Написання вступу та анотації.	18.11.2024			
5.	Дослідження методів регуляризації.	20.12.2024			
6.	Робота над текстовим оформленням теоретичної частини та одержаних результатів.	18.02.2025			
7.	Попередній аналіз кваліфікаційної роботи. Виправлення помилок.	11.04.2025			
8.	Попередній захист кваліфікаційної роботи.	23.05.2025			
9.	Захист кваліфікаційної роботи.	06.06.2025			

Науковий керівник _____
(ПІБ)

Виконавець кваліфікаційної роботи _____
(ПІБ)

Зміст

Анотація **5**

1 Вступ **6**
1.1 Актуальність 6
1.2 Мета, завдання дослідження 6

2 Основні поняття та означення кластеризації **7**
2.1 Основи кластеризації 7
2.2 Кластеризація метод k-середніх 8
2.3 Підходи до оцінки якості кластеризації 10

3 Методи регуляризації в задачах кластеризації **14**
3.1 Постановка задачі регуляризації в методі кластеризації k-середніх 16
3.2 Типи штрафних функцій 17
3.3 Обчислення регуляризованого методу k-середніх 20
3.4 Вибір параметра регуляризації λ 22

4 Практичне застосування регуляризації в задачах кластеризації **25**
4.1 Постановка задачі 25
4.2 Методи оцінювання 27
4.3 Підготовка даних 28
4.4 Застосування алгоритму k-середніх 28
4.5 Регуляризація методом Hard-Thresholding 31
4.6 Регуляризація методом Lasso 34
4.7 Регуляризація методом Ridge 37
4.8 Регуляризація методом Group Lasso 39
4.9 Аналіз ефективності методів та результати задачі 42

Висновки **45**

Список літератури **47**

Анотація

У кваліфікаційній роботі було досліджено методи регуляризації в задачах кластеризації. Детально розглянуто теоретичні аспекти поняття регуляризації, її властивості та застосування.

У другому розділі розглянуто основи кластеризації, описано найпоширеніший метод – k-середніх, та підходи до оцінки якості методу.

У третьому розділі наведені основні концепції регуляризації та розглянуто практичні підходи до використання регуляризації в задачах кластеризації. Проведено огляд методів, які покращують стійкість та точність результатів.

У четвертому розділі проаналізовано та порівняно результати кластеризації з використанням різних методів регуляризації для прикладної задачі сегментації користувачів мобільного застосунку.

1 Вступ

1.1 Актуальність

Кластеризація є одним із ключових інструментів аналізу даних, який знаходить широке застосування у багатьох галузях, таких як: фінанси, біоінженерія, маркетинг, логістика та освіта. На жаль, часто стандартні методи кластеризації мають неточності, пов'язані з перенавчанням, нерівномірністю розподілу даних, чутливістю до інформації, яка не є значущою, та високою варіативністю результатів.

Тому для підвищення стійкості та покращення результатів кластеризації часто використовують різні методи регуляризації, що дозволяють зменшити вплив нерелевантних даних, стабілізувати процес навчання алгоритмів та уникнути перенавчання. Регуляризація допомагає адаптувати моделі до реальних умов, коли дані можуть бути нерівномірно розподіленими або містити надлишкову та нерелевантну інформацію.

Актуальність даної роботи полягає в тому, що при стрімкому зростанні кількості даних та їх складності постає гостра потреба в розробці інструментів, які забезпечать якісний аналіз і обробку інформації. У багатьох галузях точність кластеризації відіграє критичну роль у прийнятті рішень. Проте, традиційні методи кластеризації часто не справляються з викликами, які виникають через високий рівень нерелевантності та нерівномірності даних та наявністю складних зв'язків між об'єктами. Саме тому застосування регуляризації в задачах кластеризації стає актуальним і перспективним напрямком.

1.2 Мета, завдання дослідження

Метою кваліфікаційної роботи є дослідження та впровадження методів регуляризації у задачах кластеризації для підвищення їхньої стійкості, точності та здатності до узагальнення при роботі з великими обсягами даних, що характеризуються нерівномірністю розподілу та наявністю нерелевантних даних. У рамках роботи будуть розглянуті способи впровадження методів регуляризації в задачі кластеризації, зокрема аналіз впливу регуляризаційних параметрів на якість кластеризації, а також дослідження можливостей зменшення впливу нерелевантних даних на результати кластеризації. Окрім цього, буде проведено експериментальне порівняння різних регуляризаційних підходів для прикладної задачі кластеризації.

2 Основні поняття та означення кластеризації

2.1 Основи кластеризації

Кластеризація, або кластерний аналіз [1], є статистичним методом, мета якого полягає у поділі вибірки об'єктів на групи, які називаються кластерами. Важливо, щоб об'єкти в межах одного кластера були максимально схожими, а об'єкти з різних кластерів мали суттєві відмінності. Така задача належить до сфери статистичної обробки даних і є частиною великої групи задач навчання без вчителя.

Формально, нехай X — набір об'єктів, а Y — сукупність міток або індексів, що відповідають кожному кластеру. Для визначення близькості між об'єктами використовується метрика $\rho(x, x')$. Розглядається кінцева вибірка $X_m = \{x_1, \ldots, x_m\} \subset X$, яку потрібно поділити на взаємовиключні групи — кластери. В межах кожного кластеру об'єкти повинні бути схожими за метрикою ρ, тоді як об'єкти з різних кластерів мають значно відрізнятися один від одного. Кожному елементу вибірки $x_i \in X_m$ присвоюється мітка $y_i \in Y$, що вказує на його належність до певного кластера.

У рамках цієї роботи метрика $\rho(x, x')$ буде визначена як Евклідова відстань:

$$\rho(x, x') = \sqrt{\sum_{i=1}^{n} (x_i - x'_i)^2}, \tag{1}$$

де $x = (x_1, x_2, \ldots, x_n)$ та $x' = (x'_1, x'_2, \ldots, x'_n)$ — координати точок у n-вимірному просторі.

Алгоритм кластеризації [1] визначається як функція $f : X \to Y$, яка кожному об'єкту $x \in X$ присвоює індекс кластеру $y \in Y$. У деяких випадках множина Y визначена заздалегідь, однак часто необхідно визначити оптимальну кількість кластерів, виходячи з певного критерію якості кластеризації.

На рисунку 1 показано приклад результату кластеризації точок на координатній площині на два кластери.

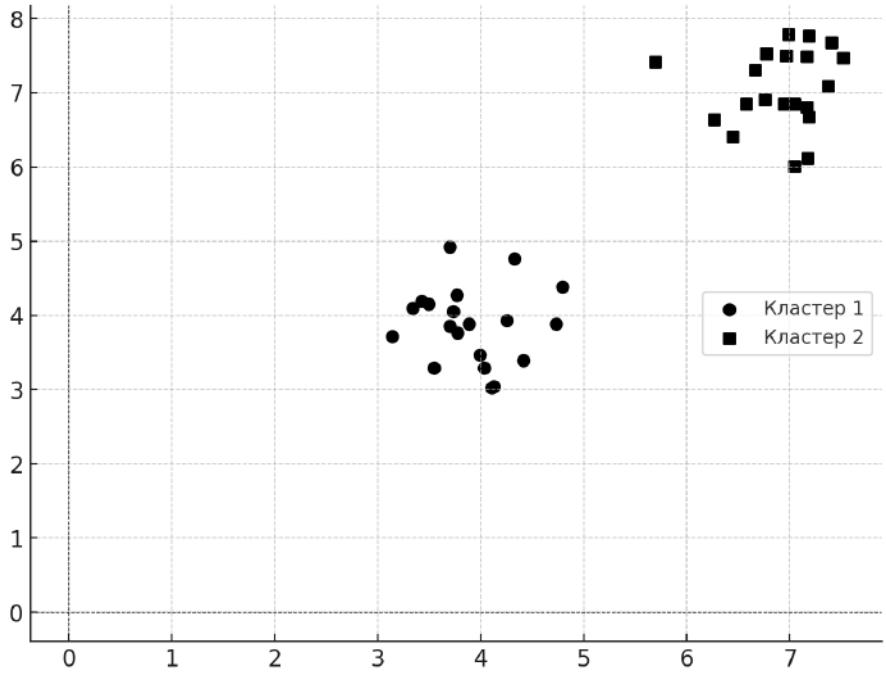

Рис. 1: Кластеризація точок на координатній площині

2.2 Кластеризація метод k-середніх

Серед численних методів кластеризації метод k-середніх [2] є одним із найпопулярніших та найпоширеніших завдяки своїй концептуальній простоті, високій обчислювальній ефективності та здатності масштабуватися для великих наборів даних. Цей метод є особливо корисним у випадках, коли необхідно розділити n об'єктів (спостережень) на K кластерів таким чином, щоб кожна точка належала саме одному кластеру, і була якомога ближчою до свого центру.

Припустимо, що доступні n спостережень $X_1, \ldots, X_n$, де кожне $X_i \in \mathbb{R}^p$ є вектором з p ознаками: $X_i = (X_{i1}, \ldots, X_{ip})^T$. Нехай задано кількість кластерів K, тоді задача кластеризації зводиться до знаходження множини кластерів $A_1, \ldots, A_K$ та їх центрів $C_1, \ldots, C_K$, де $C_k = (C_{k1}, \ldots, C_{kp})^T$, які мінімізують наступну цільову функцію

$$\min_{A_k, C_k} \sum_{k=1}^{K} \sum_{X_i \in A_k} \rho(X_i, C_k)^2. \tag{2}$$

Ця задача є неконвексною через дискретність кластерних призначень A_k,

однак її зазвичай розв'язують за допомогою евристичного алгоритму Ллойда [8], який гарантує збіжність до локального мінімуму цільової функції (2).

Алгоритм кластеризації методом k-середніх (алгоритм Ллойда) [3] включає наступні етапи:

1. **Ініціалізація:**

   - Встановлюється кількість кластерів K.
   - Обираються початкові центроїди кластерів $C_1, \ldots, C_K$. Типовим підходом є випадковий вибір K точок із набору $X = \{X_1, \ldots, X_n\}$, однак для підвищення стабільності можуть застосовуватись вдосконалені методи, наприклад, k-середніх++ [4].

2. **Призначення:**

   - Кожне спостереження X_i призначається до найближчого центру C_k згідно з критерієм мінімізації квадрату Евклідової відстані(1):

   $$k^{(i)} = \arg\min_k \rho(X_i, C_k)^2,$$

   де $k^{(i)}$ — індекс найближчого центру кластера для точки X_i.

3. **Оновлення центрів:**

   - Для кожного кластера A_k обчислюється новий центр як середнє значення усіх точок, що були до нього призначені:

   $$C_k = \frac{1}{|A_k|} \sum_{X_i \in A_k} X_i,$$

   де $|A_k|$ — кількість точок у кластері A_k.
   - Якщо кластер A_k виявився порожнім, його центр може залишитися незмінним або ініціалізуватися заново.

4. **Критерій зупинки:**

   - Процес повторюється до тих пір, поки:
     - кластерні призначення не змінюються між ітераціями;
     - або зміна цільової функції (2) стає меншою за наперед визначений поріг;

– або досягнуто максимальну кількість ітерацій.

Таким чином, метод k-середніх та алгоритм Ллойда дозволяють здійснити ефективне розбиття на K кластерів, яке мінімізує внутрішньокластерну дисперсію

2.3 Підходи до оцінки якості кластеризації

Існує два основні підходи до оцінювання якості результату кластеризування: внутрішня та зовнішня оцінка.

Внутрішня оцінка аналізує структуру кластерів, використовуючи лише властивості даних без порівняння з еталонною розміткою. Основна мета внутрішньої оцінки — визначити наскільки точки в межах одного кластера подібні між собою та наскільки кластери відрізняються один від одного.

Зовнішня оцінка передбачає аналіз якості кластеризації шляхом порівняння отриманого розбиття з наявною еталонною розміткою. Такі методи застосовуються, якщо доступна інформація про правильне групування даних.

У даній роботі зовнішні оцінки не використовуються, оскільки розглядаються лише випадки без еталонного маркування.

Далі будуть наведені приклади внутрішніх оцінок які застосовуватимуться для аналізу якості кластеризації.

Внутрішньокластерна відстань [2] обчислює середню відстань між кожною точкою кластера та центром цього кластера. Чим менше значення внутрішньокластерної відстані, тим щільніше та компактніше зосереджені точки всередині кластерів. Обчислюється метрика за наступною формулою

$$\text{Intra-cluster distance} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|A_k|}\sum_{X_i \in A_k}\rho(X_i, C_k)^2. \tag{3}$$

Для наочного порівняння значень внутрішньокластерної відстані на рисунку 2, зліва показаний кластер із більшою внутрішньокластерною відстанню, а справа — із меншою.

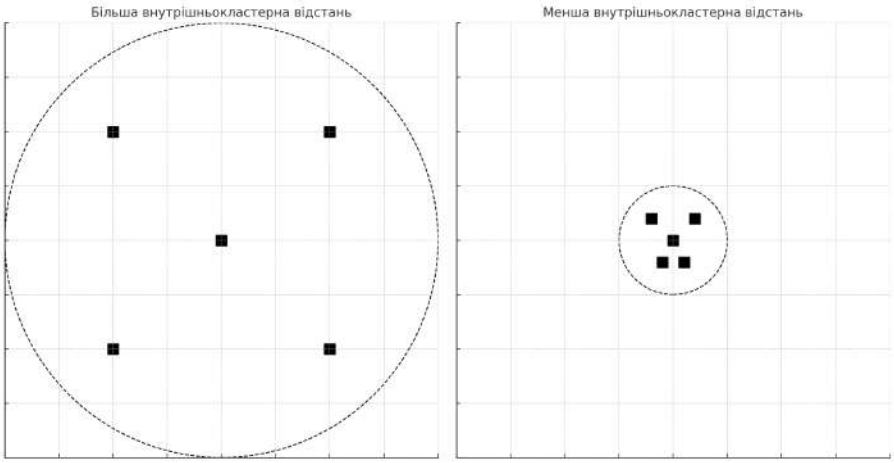

Рис. 2: Порівняння внутрішньокластерних відстаней

Сумарна внутрішньокластерна дисперсія [2]. Для оцінки якості кластеризації також широко використовується показник сумарної внутрішньокластерної дисперсії.

WCSS (Within-Cluster Sum of Squares) — це сума квадратів відстаней усіх спостережень до центроїдів. Формально вона визначається так:

$$\text{WCSS} = \sum_{k=1}^{K} \sum_{X_i \in A_k} \rho(X_i, C_k)^2, \tag{4}$$

Зменшення значення WCSS свідчить про зростання щільності кластерів, тобто об'єкти всередині кластерів розташовуються ближче до своїх центрів. У класичному методі k-середніх мінімізація WCSS є основною ціллю кластеризації.

Міжкластерна відстань [2] оцінює ступінь віддаленості кластерів один від одного. Чим більшими є відстані між кластерами, тим кращим вважається розділення кластерної структури. Існує два поширених способів визначення міжкластерної відстані:

- *Мінімальна міжкластерна відстань* (найменша відстань між центрами будь-яких двох кластерів):

$$\text{Inter-cluster distance}_{\min} = \min_{k \neq j} \rho(C_k, C_j), \tag{5}$$

де C_k і C_j — центри кластерів k та j відповідно.

- *Середня міжкластерна відстань* (усереднена відстань між усіма парами центрів кластерів):

$$\text{Inter-cluster distance}_{\text{avg}} = \frac{2}{K(K-1)} \sum_{1 \le k < j \le K} \rho(C_k, C_j), \tag{6}$$

Індекс Давіса–Болдіна [5] є одним з класичних показників для оцінювання якості кластеризації. Він відображає по кластерах середнє відношення внутрішньої дисперсії до міжкластерної відстані. Чим менше значення DB-індексу, тим компактніше та чіткіше розділені кластери.

Формально індекс Давіса–Болдіна визначається наступним чином:

$$DB = \frac{1}{K} \sum_{k=1}^{K} \max_{j \neq k} \left(\frac{S_k + S_j}{\rho(C_k, C_j)} \right), \tag{7}$$

де S_k — середня відстань (розсіювання) точок кластеру A_k до його центру C_k, яка обчислюється як:

$$S_k = \frac{1}{|A_k|} \sum_{X_i \in A_k} \rho(X_i, C_k).$$

Тобто для кожного кластеру k знаходиться кластер j, який є "найближчим сусідом" з погляду несприятливого співвідношення між розсіюванням та відстанню між центрами. Потім середнє таких максимумів для всіх кластерів дає фінальну метрику.

Силуетний коефіцієнт [6] є популярною внутрішньою метрикою для оцінки якості кластеризації, яка дозволяє оцінити, наскільки добре кожна точка X_i "вписується" у свій кластер, порівняно з найближчим альтернативним кластером.

Значення коефіцієнта силуету знаходиться в межах від -1 до 1:

- значення, близькі до 1, свідчать про добре кластеризовану точку;
- значення, близькі до 0, вказують на майже рівномірне розміщення точки між двома кластерами;
- значення менше нуля означають, що точка, ймовірно, була віднесена до неправильного кластера.

Формально силуетний коефіцієнт обчислюється як:

$$S = \frac{1}{n} \sum_{i=1}^{n} \frac{b_i - a_i}{\max(a_i, b_i)}, \tag{8}$$

де:

- n — загальна кількість точок;
- a_i — середня відстань від точки X_i до всіх інших точок свого кластеру A_k, тобто:

  $$a_i = \frac{1}{|A_k| - 1} \sum_{\substack{X_j \in A_k \\ j \neq i}} \rho(X_i, X_j),$$

  де $X_i \in A_k$;
- b_i — найменша середня відстань від точки X_i до всіх точок іншого кластеру $A_{k'}$, тобто:

  $$b_i = \min_{k' \neq k} \left(\frac{1}{|A_{k'}|} \sum_{X_j \in A_{k'}} \rho(X_i, X_j) \right).$$

Цей показник забезпечує точкову оцінку якості кластеризації і дозволяє візуалізувати ефективність розбиття через графік силуетів, який особливо корисний для вибору кількості кластерів K.

3 Методи регуляризації в задачах кластеризації

У попередньому розділі було розглянуто основи кластеризації, метод k-середніх та підходи до оцінки якості кластеризації. Однак, при роботі з реальними даними, особливо великої розмірності, стандартний алгоритм k-середніх може мати неточності, пов'язані з перенавчанням, нерівномірністю розподілу даних, чутливістю до інформації, яка не є значущою, та високою варіативністю результатів. Наприклад, не всі ознаки можуть бути однаково важливими для формування кластерів, а деякі можуть бути просто шумом, що ускладнює виявлення справжньої структури даних. У таких випадках доцільно використовувати методи, які дозволяють відбирати лише найінформативніші ознаки для кластеризації. Це призведе до більш інтерпретованого поділу даних та точнішого відновлення "справжніх" кластерів. Навіть у випадках, коли кількість ознак невелика, відбір ознак може бути дуже корисним.

Один зі способів покращення кластеризації при великій кількості ознак — це регуляризація, яка модифікує цільову функцію методу k-середніх так, щоб сприяти формуванню центрів кластерів з потрібними властивостями, зокрема здійснювати відбір найбільш релевантних ознак.

У загальному сенсі регуляризація — це сукупність методів, метою яких є стабілізація розв'язків, усунення неоднозначності в некоректно поставлених задачах та покращення здатності моделей до узагальнення. Це досягається шляхом включення до задачі додаткових припущень, обмежень або штрафів.

Регуляризаційні підходи умовно поділяють на два типи:

- *Явна регуляризація* полягає у додаванні до цільової функції спеціального доданка (штрафу, апріорного припущення або обмеження), що змінює цільову функцію. Такий підхід особливо корисний для задач, де існує нескінченно багато розв'язків або рішення нестабільні.

- *Неявна регуляризація* включає всі інші способи контролю складності моделі або стабілізації розв'язку без явного введення додаткового члена у функцію втрат. Прикладами є рання зупинка, використання робастних функцій втрат, або ігнорування викидів

У контексті задач кластеризації в даній роботі буде застосовано саме явну регуляризацію. Основна ідея якої полягає у модифікації цільової функції кла-

сичного алгоритму k-середніх шляхом додавання до неї спеціального регуляризаційного члена. Цей доданок накладає певні обмеження або пріоритети на структуру центрів кластерів, зокрема сприяє вибору найбільш інформативних ознак. Для задачі кластеризації k-середніх рідко застосовується неявна регуляризація, оскільки вона не модифікує цільову функцію і погано узгоджується з геометричною природою методу k-середніх. Наприклад, техніки ранньої зупинки чи робастні функції втрат не мають прямого застосування в задачі кластеризації без учителя, де немає процесу навчання у звичному сенсі. Натомість явна регуляризація дозволяє чітко керувати вибором ознак та формою центрів кластерів.

Загальна форма явної регуляризованої задачі оптимізації має вигляд:

$$\hat{\theta} = \arg\min_{\theta} \left[\mathcal{L}(\theta; X, y) + \lambda \cdot \Omega(\theta) \right],$$

де:

- $\mathcal{L}(\theta; X, y)$ — основна функція втрат, що оцінює якість моделі θ на даних X з мітками y,
- $\Omega(\theta)$ — регуляризаційний штраф, що накладає бажані властивості на модель, наприклад, розрідженість або згладженість,
- $\lambda > 0$ — параметр, що визначає силу регуляризації.

3.1 Постановка задачі регуляризації в методі кластеризації k-середніх

Нехай X — це матриця даних розміром $n \times p$, де n — кількість спостережень $X_1, \ldots, X_n$, де $X_i = (X_{i1}, \ldots, X_{ip})^T$, а p — кількість ознак. Припускаємо, що дані стандартизовані, тобто для кожної j-ої ознаки $\frac{1}{n}\sum_{i=1}^{n} X_{i,j} = 0$ та $\frac{1}{n}\sum_{i=1}^{n} X_{i,j}^2 = 1$. Кластери позначаються як $A_1, \ldots, A_K$, а їхні центри — $C_1, \ldots, C_K$, де $C_k = (C_{k1}, \ldots, C_{kp})^T$.

Тоді загальну форму цільової функції регуляризованого методу k-середніх для заданої кількості кластерів K можна записати наступним чином:

$$\min_{A_k, C_k} \sum_{k=1}^{K} \left(\sum_{X_i \in A_k} \rho(X_i, C_k)^2 \right) + \lambda P(C), \tag{9}$$

де:

- $P(C)$ — штрафна функція, що залежить від матриці центроїдів C,
- $\lambda \geq 0$ — параметр регуляризації, що контролює силу штрафу.

Ідея застосування штрафної функції для координат центроїдів базується на припущенні, що ознаки, які не мають впливу на якість кластеризації або навпаки погіршують її, повинні мати середнє значення центру близьке до нуля. Тобто, якщо змінна не містить інформації для розділення об'єктів на кластери, її значення у центрах мають бути нульовими.

Аналогічний висновок можна зробити за допомогою асимптотичного аналізу [7]. Розглянемо випадкову величину X, що розподілена за деяким розподілом Q у просторі $\mathbb{R}^m$, де $m < p$. Тоді оптимальне значення цільової функції k-середніх задається виразом:

$$\text{opt} = \int \min_{k \in \{1,\ldots,K\}} \rho(x, C_k)^2 \, Q(dx),$$

де кожному кластеру відповідає певна область R_k простору.

Тепер припустимо, що до даних додається нова ознака, яка є повністю незалежною від існуючих змінних. Нехай X^* — новий вектор розмірності $m+1$, а його розподіл позначимо через Q^*, opt^* — нове значення цільової функції. Тоді справедливі такі нерівності:

$$\text{opt} + 1 \geq \text{opt}^* \geq \text{opt} + \sum_{k=1}^{K} \min_{a \in \mathbb{R}} \int (x_{m+1} - a)^2 I\{x_{1:m} \in R_k\} Q^*(dx),$$

де $x_{1:m}$ — перші m координат вектора x, а x_{m+1} — додана компонента, $I(\cdot)$ — індикаторна функція.

Оскільки нова ознака є незалежною від початкової структури, оптимальне значення a дорівнює середньому значенню доданої змінної.З початкової умови X є стандартизованим, тоді середнє значення дорівнює нулю. Таким чином, координата центру кластеру за доданою змінною також дорівнюватиме нулю для всіх k. У результаті нове значення цільової функції зростає на одиницю:

$$\text{opt}^* = \text{opt} + 1.$$

Цей висновок обґрунтовує необхідність накладання штрафу на величини центрів кластерів для виявлення нерелевантних ознак. Регуляризація, що провокує нульові координати центрів для неінформативних змінних, підвищує точність та інтерпретованість кластеризації.

3.2 Типи штрафних функцій

Вибір штрафної функції $P(C)$ визначає конкретний тип регуляризації та її вплив на центри кластерів.У даній роботі буде розглянуто та застосовано чотири основні види штрафної функції.

L0-регуляризація (Hard-Thresholding): [7]

$$P_0(C) = \sum_{j=1}^{p} I\left(\|C_{\cdot,j}\|_2 > 0\right), \tag{10}$$

де $I(\cdot)$ — індикаторна функція, а $\|C_{\cdot,j}\|_2$ — $L2$-норма j-го стовпця матриці координат центрів C.

Цей штраф підраховує кількість ознак, для яких хоча б один центроїд має ненульове значення. Таким чином, $L0$-регуляризація реалізує жорсткий відбір ознак: кожна ознака або повністю залишається в моделі, або повністю виключається з неї. Якщо для певної ознаки цей критерій не виконується, то всі її компоненти у векторах центрів кластерів занулюються.

Метод, що використовує дану регуляризацію, часто називають HT k-means (Hard-Thresholding k-means). Застосування $L0$-регуляризації дозволяє ефективно повністю усувати нерелевантні або шумові ознаки, що є особливо цінним у задачах високої розмірності, де велика кількість змінних може негативно впливати на якість кластеризації.

L1-регуляризація (Lasso) [7]:

$$P_1(C) = \sum_{j=1}^{p} \|C_{\cdot,j}\|_1, \tag{11}$$

де $\|C_{\cdot,j}\|_1$ — $L1$-норма j-го стовпця матриці центрів кластерів C.

Lasso-регуляризація стимулює розрідженість на рівні окремих елементів матриці центрів кластерів. Це означає, що вона сприяє обнуленню деяких компонент $C_{k,j}$, водночас залишаючи інші ненульовими або зменшеними за абсолютною величиною. На практиці оновлення центрів при використанні $L1$-регуляризації реалізується через оператор м'якого порогування [7], який плавно зсуває координати центрів у напрямку нуля, зберігаючи їхню знакову структуру.

На відміну від $L0$-регуляризації, $L1$-штраф не обов'язково обнуляє всю ознаку одночасно — він діє локально на кожну координату центроїда окремо. Таким чином, відбір змінних у Lasso є м'якшим і поступовішим, що дозволяє зберігати часткову інформацію навіть у слабоінформативних ознаках. Цей компроміс між точністю та розрідженістю робить Lasso особливо ефективним у задачах високої розмірності, де повне усунення ознак може бути небажаним.

L2-регуляризація (Ridge) [7]:

$$P_2(C) = \sum_{j=1}^{p} \|C_{\cdot,j}\|_2^2 \tag{12}$$

Ridge-регуляризація вводить штраф у вигляді суми квадратів $L2$-норм стовпців матриці центрів кластерів. Такий підхід сприяє гладкому зменшенню всіх компонент центрів у напрямку до нуля, не занулюючи їх повністю. Це означає, що ознаки в окремих центрах та загалом не виключаються з моделі, але їхній вплив зменшується, пропорційно до їхньої інформативності.

На відміну від $L0$- або $L1$-регуляризації, Ridge не виконує явного відбору змінних, проте дозволяє зменшити вплив нерелевантних ознак, стабілізуючи процес кластеризації. Це особливо корисно у високовимірних задачах або в ситуаціях, коли ознаки мають сильну кореляцію, оскільки регуляризація зменшує дисперсію без радикального виключення важливої інформації.

Group Lasso-регуляризація [7]:

$$P_3(C) = \sum_{j=1}^{p} \|C_{\cdot,j}\|_2 = \sum_{j=1}^{p} \sqrt{\sum_{k=1}^{K} C_{k,j}^2}. \tag{13}$$

Регуляризація типу Group Lasso застосовує штраф до кожної ознаки як до єдиної групи, що представлена усіма її компонентами у матриці центрів кластерів C. Зокрема, для кожної ознаки j враховується $L2$-норма відповідного стовпця $C_{\cdot,j}$, тобто усіх значень цієї ознаки в кожному кластері.

Цей підхід стимулює групову розрідженість: якщо вплив певної ознаки на формування кластерів незначний (тобто $L2$-норма її стовпця мала), то всі значення цієї ознаки у центрах кластерів одночасно обнуляються. Таким чином, ознака або повністю зберігається у моделі, або повністю виключається — що забезпечує цілісний (груповий) відбір ознак.

Group Lasso є проміжною формою між $L1$- та $L2$-регуляризацією: як і Lasso, вона може занулювати ознаки, але робить це на рівні всіх кластерів одночасно. Це дозволяє ефективно керувати складністю моделі та покращує інтерпретованість кластеризаційного розбиття, особливо при роботі з блоками пов'язаних ознак.

На відміну від $L0$-регуляризації, яка жорстко виключає ознаку лише за умови недосягнення порогу впливу, Group Lasso реалізує поступовий підхід. Проте, якщо норма вектора по ознаці є недостатньою, то, подібно до L0-регуляризації, усі її значення у центрах кластерів обнуляються одночасно:

Таким чином, Group Lasso також може виключити всі ознаки, але робить це у більш стабільний та диференційовний спосіб, що дозволяє ефективніше оптимізувати модель у випадках великої кількості ознак або наявністю шуму.

Підсумовуючи, можна сказати, що використання різних типів регуляризаційних штрафів дозволяє гнучко керувати процесом кластеризації залежно від завдань і характеристик даних. Зокрема, L0 та Group Lasso регуляризації забезпечують чіткий відбір змінних на рівні ознак, тоді як $L1$- та $L2$-регуляризації сприяють поступовому зменшенню впливу менш значущих координат без їх обов'язкового обнулення. Правильний вибір форми штрафу дає змогу підвищити точність, стійкість і інтерпретованість кластеризаційних моделей, що особливо важливо при роботі з даними високої розмірності або при наявності шумових змінних.

Для зручнішого застосування різних видів регуляризації у задачах кластеризації у таблиці 1 наведено коротке порівняння основних типів штрафних функцій.

Регуляризація	Формула штрафу	Механізм впливу	Особливості
L0 (Hard-Thresholding)	$\sum_{j=1}^{p} I(\|C_{\cdot j}\|_2 > 0)$	Повне включення або виключення ознаки	Чіткий відбір ознак, висока інтерпретованість
L1 (Lasso)	$\sum_{j=1}^{p} \|C_{\cdot j}\|_1$	М'яке обнулення окремих координат центрів	Розрідженість без повного виключення ознак
L2 (Ridge)	$\sum_{j=1}^{p} \|C_{\cdot j}\|_2^2$	Гладке зменшення центрів до нуля	Не здійснює відбору ознак, стабілізує модель
Group Lasso	$\sum_{j=1}^{p} \|C_{\cdot j}\|_2$	Стискає координати ознаки	Може обнулити всю ознаку

Табл. 1: Порівняння методів регуляризації

3.3 Обчислення регуляризованого методу k-середніх

Після розгляду основних типів регуляризацій перейдемо до опису алгоритму обчислення регуляризованого методу k-середніх. Мінімізація функції (9) здійснюється за допомогою модифікованого алгоритму Ллойда [8], що поєднує кластеризацію та відбір ознак.

Алгоритм складається з наступних етапів:

1. **Ініціалізація:** Вибираються початкові центри кластерів $C = (C_1, \ldots, C_K)$.

2. **Крок призначення кластерів:** Кожне спостереження X_i призначається до найближчого центру C_k за квадратом Евклідової відстані (1):

$$k = \arg \min_{1 \le k' \le K} \rho(X_i, C_{k'})^2.$$

3. **Крок оновлення центрів кластерів:** При фіксованих призначеннях кластерів матриця центрів оновлюється шляхом розв'язання регуляризованої оптимізаційної задачі для кожного кластера k та ознаки j. Позначимо:

$$C^*_{k,j} = \frac{1}{|A_k|} \sum_{X_i \in A_k} X_{i,j},$$

де $|A_k|$ — кількість точок у кластері A_k.

Формули оновлення центрів для різних штрафних функцій мають вигляд:

- **Hard-Thresholding (P_0):**

$$C_{k,j} = \begin{cases} C^*_{k,j}, & \text{якщо} \quad \|X\|_2^2 > \rho(X, MC^*_{\cdot,j})^2 + n\lambda, \\ 0, & \text{інакше.} \end{cases}$$

Тобто ознака j зберігається у моделі лише в тому випадку, якщо її внесок у зменшення внутрішньокластерної дисперсії є достатньо великим і перевищує регуляризаційний штраф λ.

Тут матриця $M \in \{0,1\}^{n \times K}$ — це матриця, де $M_{i,k} = 1$, якщо спостереження i належить кластеру k, і 0 — інакше. Таким чином, добуток $MC^*_{\cdot,j}$ повертає вектор, у якому кожному об'єкту i призначено середнє значення ознаки j відповідно центру його кластера. Це дозволяє обчислити залишки від зміни ознаки через центри кластерів.

- **Lasso (P_1):**

$$C_{k,j} = \max\left(0, 1 - \frac{n\lambda}{2|A_k|\left|C^*_{k,j}\right|}\right) C^*_{k,j}.$$

Формула реалізує м'яке скорочення модулів координат центрів кластерів.

- **Ridge (P_2):**

$$C_{k,j} = \frac{1}{1 + \frac{n\lambda}{|A_k|}} C^*_{k,j},$$

що відповідає регулярному "стисканню" центрів кластерів до нуля без обнулення.

- **Group Lasso (P_3):**
Якщо $\|C^*_{\cdot,j}\|_2 \neq 0$, тоді:

$$C_{k,j} = \frac{1}{1 + \frac{n\lambda}{2|A_k|\|C^*_{\cdot,j}\|_2}} C^*_{k,j}.$$

В іншому випадку центри за ознакою j обнуляються. Даний механізм проводить груповий відбір ознак на рівні усієї змінної.

4. **Критерій зупинки:** Алгоритм повторюється доти, доки:

- кластерні призначення перестають змінюватися,
- зменшення цільової функції стає меншим за наперед заданий поріг,
- досягнуто максимальну кількість ітерацій.

Через наявність локальних мінімумів рекомендується запускати алгоритм з кількома різними початковими ініціалізаціями та обирати найкраще рішення за значенням цільової функції.

Таким чином, модифікований алгоритм Ллойда для регуляризації дозволяє ефективно здійснювати як кластеризацію, так і автоматичний відбір інформативних ознак.

3.4 Вибір параметра регуляризації λ

Важливою складовою побудови регуляризованої кластеризаційної моделі є правильний вибір значення параметра регуляризації λ. Значення λ визначає баланс між якістю кластеризації та кількістю вибраних ознак, впливаючи на ступінь розрідженості рішення.

У даній роботі розглядаються наступні підходи до вибору оптимального значення λ.

1. Інформаційні критерії: AIC та BIC

Один із базових методів підбору λ полягає у використанні інформаційних критеріїв, таких як AIC (Akaike Information Criterion) та BIC (Bayesian Information Criterion), які поєднують оцінку якості кластеризації з урахуванням складності моделі.

В обчисленнях використовуються наступні формули:

$$\mathrm{AIC} = \mathrm{WCSS} + 2Kp_{\mathrm{sel}},$$

$$\mathrm{BIC} = \mathrm{WCSS} + Kp_{\mathrm{sel}} \ln(n),$$

де:

- WCSS — сумарна внутрішньокластерна дисперсія,
- K — кількість кластерів,
- p_{sel} — кількість активних (вибраних) ознак,
- n — кількість спостережень.

2. Підхід на основі стабільності кластеризації

Альтернативною стратегією є вибір λ на основі оцінки стабільності результатів кластеризації [7] при незначних варіаціях у даних.

Інтуїтивно, якщо кластеризаційний алгоритм є стабільним, невеликі зміни у вибірці не повинні призводити до суттєвої зміни кластерних призначень. Формально відстань між двома кластеризаціями ψ_1 та ψ_2 визначається як:

$$d(\psi_1, \psi_2) = \mathbb{P}\left[I\{\psi_1(X) = \psi_1(Y)\} + I\{\psi_2(X) = \psi_2(Y)\}\right],$$

де $I(\cdot)$ — індикаторна функція, а X та Y — незалежні вибірки з однакового розподілу.

На основі цієї відстані визначається кластеризаційна нестабільність:

$$s(\psi; \lambda) = \mathbb{E}[d\{\psi(X_1; \lambda), \psi(X_2; \lambda)\}],$$

де очікування береться за незалежними вибірками X_1 та X_2.

Для оцінки нестабільності пропонуються різні варіанти процедур:

- Використання бутстреп-зразків: на кожній ітерації формуються дві бутстреп-вибірки розміру n, а валідаційним набором виступає оригінальна вибірка.
- Використання перетину унікальних об'єктів з двох бутстреп-вибірок як валідаційного набору.
- Використання третьої незалежної бутстреп-вибірки як валідаційного набору.

Хоча методи, засновані на оцінці стабільності, теоретично мають сильну мотивацію, на практиці вони можуть бути обчислювально складними та менш ефективними у порівнянні з простішими підходами (такими як AIC/BIC).

3. Gap-метод для вибору λ

Ще одним підходом до вибору оптимального значення λ є адаптація ідеї Gap-методу до задачі регуляризованої кластеризації [7].

Підхід полягає в аналізі зміни WCSS при послідовному включенні нових ознак у модель:

$$\Delta_{q+1} = \frac{1}{n}\mathrm{WCSS}_{q+1} - \frac{1}{n}\mathrm{WCSS}_q,$$

де q — кількість вже включених ознак.

Інтерпретація:

- Якщо Δ_{q+1} близьке до 0, додана ознака добре узгоджується з поточним розбиттям.

- Якщо Δ_{q+1} близьке до 1, ознака випадкова і не покращує кластеризацію.

Для об'єктивної оцінки величини Δ_{q+1} використовується порівняння з очікуваним зростанням WCSS при додаванні випадкової (перемішаної) змінної. Для цього:

- Перемішується одна ознака у даних.
- Знову виконується кластеризація і обчислюється Δ^*_{q+1}.
- Процедура повторюється S разів (наприклад, $S = 50$) для оцінки середнього та дисперсії Δ^*_{q+1}.

Далі обчислюється стандартизована відстань:

$$D_{q+1} = \frac{m - \Delta_{q+1}}{s},$$

де m і s — середнє та стандартне відхилення отриманих Δ^*_{q+1}.

Оптимальне значення λ вибирається таким чином, щоб відповідати максимуму D_{q+1}, або мінімальному значенню λ, для якого D_{q+1} знаходиться в межах заданої кількості стандартних відхилень від максимуму.

Gap-метод ефективно враховує як якість пояснення даних новими ознаками, так і ймовірність випадкового покращення кластеризації, що робить його особливо корисним у задачах з великою кількістю шумових змінних.

4 Практичне застосування регуляризації в задачах кластеризації

4.1 Постановка задачі

У сучасному світі безконтактних платежів фрод (шахрайство) є однією з найбільших загроз для фінансових установ, платіжних систем та продавців. Це не лише знижує прибутковість компаній, але й підриває довіру клієнтів, що є ключовим фактором стабільності сучасної цифрової економіки. Фрод може набувати різних форм залежно від методів його здійснення, зокрема:

- **Класичний (зловмисний) фрод** — ситуації, коли шахрай отримує доступ до платіжних реквізитів користувача без його відома і здійснює несанкціоновані транзакції. До таких випадків належать використання вкрадених карткових даних, скімінг, фішинг, крадіжка особистої інформації тощо.

- **Дружній фрод** — це тип шахрайства, який відбувається зі сторони клієнта, коли власник картки ініціює повернення коштів за транзакцію, яку насправді здійснив свідомо. Основні причини такого фроду можуть бути:

  – Забуття про покупку.

  – Незадоволеність продуктом або послугою.

  – Свідоме введення банку в оману з метою отримання безкоштовного продукту або послуги.

  – Використання картки дітьми або членами сім'ї без дозволу власника.

Платіжні системи встановлюють суворі обмеження на рівень допустимого фроду, оскільки високі показники шахрайства можуть негативно вплинути на їхню репутацію та стабільність фінансових потоків. Невиконання цих вимог може призвести до значних фінансових санкцій або навіть втрати права на обробку карткових платежів, що є серйозним ризиком для бізнесу.

Тому в IT-компанії, що розробляє підписочні додатки, виникає необхідність сегментувати користувачів за рівнем ризику створення дружнього фроду, щоб покращити управління підписками та мінімізувати фінансові ризики.

Метою цієї задачі є кластеризація користувачів на кілька груп із різними рівнями ризику здійснення дружнього фроду. Оскільки дані користувачів можуть містити велику кількість нерелевантних або шумових ознак, які не лише ускладнюють процес кластеризації, але й можуть призводити до неправильної інтерпретації результатів, будуть використані методи регуляризації, що були описані в попередньому розділі.

Задача кластеризації буде виконуватися на основі великого набору історичних даних, що відображають різноманітні аспекти поведінки користувачів і включають такі змінні:

- user uuid: Унікальний ідентифікатор користувача;
- payment method: Метод оплати користувача;
- trial period: Тривалість пробного періоду підписки користувача;
- group country: Група країни користувача;
- gender: Стать користувача;
- age group: Вікова група користувача;
- login: Факт логіну користувача в додаток;
- source: Джерело реєстрації користувача (наприклад, реклама в соцмережах);
- has cons: Факт придбання додаткової консультанційної послуги користувачем;
- has otp upsell: Факт придбання додаткового матеріалу користувачем ;
- rebill number: Кількість повторних платежів по підписці;
- card type: Тип картки користувача (наприклад, дебетова або кредитна);
- issuing bank: Банк-емітент, що випустив картку користувача;
- card brand: Бренд картки;
- fraud count: Кількість повідомлень про шахрайські операції, пов'язані з користувачем;
- income count: Кількість здійснених платежів користувачем.

4.2 Методи оцінювання

У рамках цієї задачі для порівняння якості кластеризації, здійсненої за допомогою різних методів регуляризації, будуть використані оцінки, розглянуті в розділі 2.3. Однак, для того, щоб оцінити не лише якість самого процесу кластеризації, а й ефективність сегментації користувачів, буде введена додаткова специфічна метрика FDS(Fraud Disproportion Score) - оцінка диспропорції частки фроду до частки доходу для кластеру.

Метрика FDS буде використовуватися для порівняння результатів кластеризації, що дозволяє оцінити не тільки внутрішню однорідність кластерів, а й їх здатність адекватно розділяти користувачів на групи з високим і низьким ризиком здійснення фроду. Це дозволить глибше проаналізувати ефективність кожного методу регуляризації у контексті виявлення ризикових груп користувачів. FDS для даної задачі обчислюється за наступною формулою:

$$\text{FDS} = \sum_{k=1}^{C} \left| \frac{p_k}{TP} - \frac{f_k}{TF} \right|, \tag{14}$$

де:

- TP — загальна кількість платежів усіх користувачів.
- TF — загальна кількість випадків фроду серед усіх користувачів.
- p_k — кількість платежів в кластері k.
- f_k — кількість випадків фроду в кластері k.
- C — кількість кластерів.

FDS набуває значень у діапазоні від 0 до 1. Чим ближчий результат метрики до 1 , тим краща кластеризація: якщо метрика для кластеру k велика, це означає, що в кластері k значна диспропорція між часткою платежів та часткою фроду. Такий кластер є більш однорідним і відповідає більш чітким характеристикам задачі, що підвищує ефективність виявлення фроду.

Мала метрика, яка наближається до 0, вказує на гармонійне співвідношення платежів та фроду в кластері, що може свідчити про погану кластеризацію, де різні типи користувачів змішані, і важко визначити, які з них є ризикованими.

4.3 Підготовка даних

Перед початком кластеризації було проведено підготовку даних з метою приведення їх до єдиного числового формату та зменшення потенційного впливу нерелевантних ознак. Спочатку з набору даних було виключено ідентифікатор користувача та цільові змінні, що безпосередньо пов'язані з кількістю доходу і випадків фроду, оскільки вони використовуються виключно для подальшої валідації результатів кластеризації - обрахунку метрики FDS.

Бінарні логічні змінні, що містять інформацію про активність користувача та придбання додаткових послуг, були приведені до числового типу шляхом заміни логічних значень на числові (0 або 1). Такий підхід дозволив включити їх до подальшої обробки як числові ознаки.

Категоріальні змінні з невеликою кількістю унікальних значень були закодовані методом One-Hot кодування, що дало змогу перетворити кожне значення у відповідну бінарну ознаку, уникаючи штучної ієрархії між категоріями.

Для змінної з високою кількістю можливих значень (банк-емітент картки) було застосовано частотне кодування, тобто кожному значенню ознаки було призначено число, що відповідає частоті його появи у вибірці. Такий підхід не збільшує розмірність та дозволяє уникнути надлишковості ознак, що може виникати при використанні One-Hot кодування для змінних із великою кількістю унікальних значень.

Далі всі числові ознаки, включно з закодованими логічними та частотно закодованими, було стандартизовано з використанням масштабування до середнього нуль та одиничного стандартного відхилення. Це дозволило привести всі ознаки до однакового масштабу, що є критично важливим при застосуванні методів кластеризації, які базуються на метриках відстані в просторі.

У результаті даної обробки було сформовано матрицю ознак, придатну для подальшої роботи. Також було сформовано повний список назв усіх ознак, включаючи ті, що виникли внаслідок One-Hot кодування. Це дало змогу інтерпретувати результати кластеризації з урахуванням усіх трансформацій.

4.4 Застосування алгоритму k-середніх

На початковому етапі було реалізовано класичний алгоритм кластеризації методом k-середніх. На відміну від готових вбудованих рішень, реалізація була виконана вручну з метою повного контролю над логікою алгоритму та можливості подальшої адаптації до застосування методів регуляризації.

Алгоритм реалізовано за класичною схемою Ллойда. Початково центри кластерів ініціалізуються випадковим чином із вибірки. Далі, на кожній ітерації, кожна точка призначається до найближчого центру кластеру за евклідовою відстанню. Після цього центри кластерів оновлюються як середнє значення точок, що до них належать. Процес повторюється доти, поки зміна координат центрів не стане меншою за наперед заданий поріг - 0.0001, або не буде досягнуто максимальної кількості ітерацій - 50.

Окрім імплементації базового алгоритму кластеризації, реалізація також включає обчислення основних якісних метрик: силуетного коефіцієнта, індексу Девіса-Болдіна, середньої внутрішньо- та міжкластерної відстані, а також специфічної метрики FDS. Це дозволяє не лише кластеризувати дані, а й якісно оцінити отримані результати.

Для вибору оптимальної кількості кластерів було побудовано графіки основних метрик в залежності від кількості кластерів k(Рис. 3). За їх візуальним аналізом можна твердити, що найкращою кількістю кластерів є $k = 6$, оскільки FDS досягає в цій точці максимума, а значення інших метрик є близькими до найкращих.

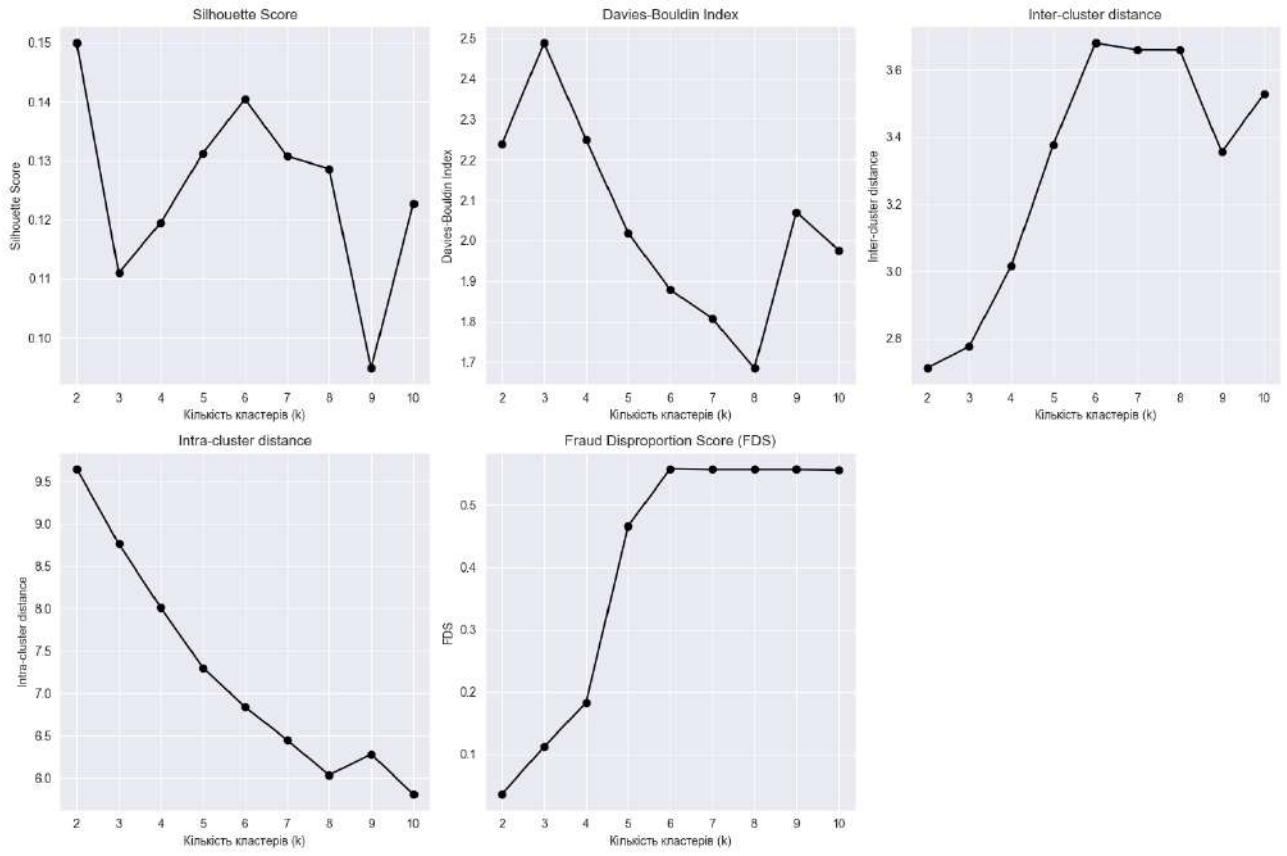

Рис. 3

Для вибраного значення k результати кластеризації було оцінено за відповідними метриками. Їх значення наведено в таблиці нижче:

Метрика	Значення
Силуетний коефіцієнт (Silhouette Score)	0.1404
Індекс Давіса-Болдіні (Davies-Bouldin Index)	1.8793
Середня внутрішньокластерна відстань	6.8393
Середня міжкластерна відстань	3.6802
Fraud Disproportion Score (FDS)	0.5579

Табл. 2: Оцінка якості кластеризації методом k-середніх($k = 6$)

4.5 Регуляризація методом Hard-Thresholding

З метою відбору інформативних ознак і підвищення інтерпретованості кластеризації було реалізовано модифікацію алгоритму k-середніх із застосуванням L0-регуляризації. Відповідно до формули штрафної функції 10, після кожної ітерації оновлення центрів кластерів виконується аналіз значущості кожної ознаки, і якщо її внесок у зменшення загальної похибки кластеризації є недостатнім, значення цієї ознаки занулюється в усіх центроїдах.

Реалізований алгоритм HT-KMeans передбачає виконання класичних кроків кластеризації: ініціалізація центрів, призначення точок до найближчих центрів, оновлення центрів як середнє по точках кластеру. Після цього застосовується регуляризаційний крок, де для кожної ознаки перевіряється умова:

$$\|X\|_2^2 > \rho(X, MC_{\cdot,j}^*)^2 + n\lambda$$

У разі невиконання цієї умови всі значення ознаки j в центроїдах занулюються. Таким чином, ознаки, що не демонструють суттєвого вкладу у відмінність між кластерами, автоматично виключаються з моделі.

Для вибору оптимального значення регуляризаційного параметра λ було використано кілька підходів: інформаційні критерії AIC і BIC, а також метод Gap-статистики. Окрім того, було враховано сукупність якісних метрик, аналогічних до використаних для звичайного методу k-середніх: силуетний коефіцієнт, індекс Девіса-Болдіна, внутрішньо- та міжкластерні відстані, а також специфічна метрика Fraud Disproportion Score (FDS). На рисунках 4-5 зображено графіки відповідних метрик в залежності від значення λ. З візуального аналізу можемо дійти до висновку, що для даної моделі найоптимальніший параметр $\lambda = 0.2$.

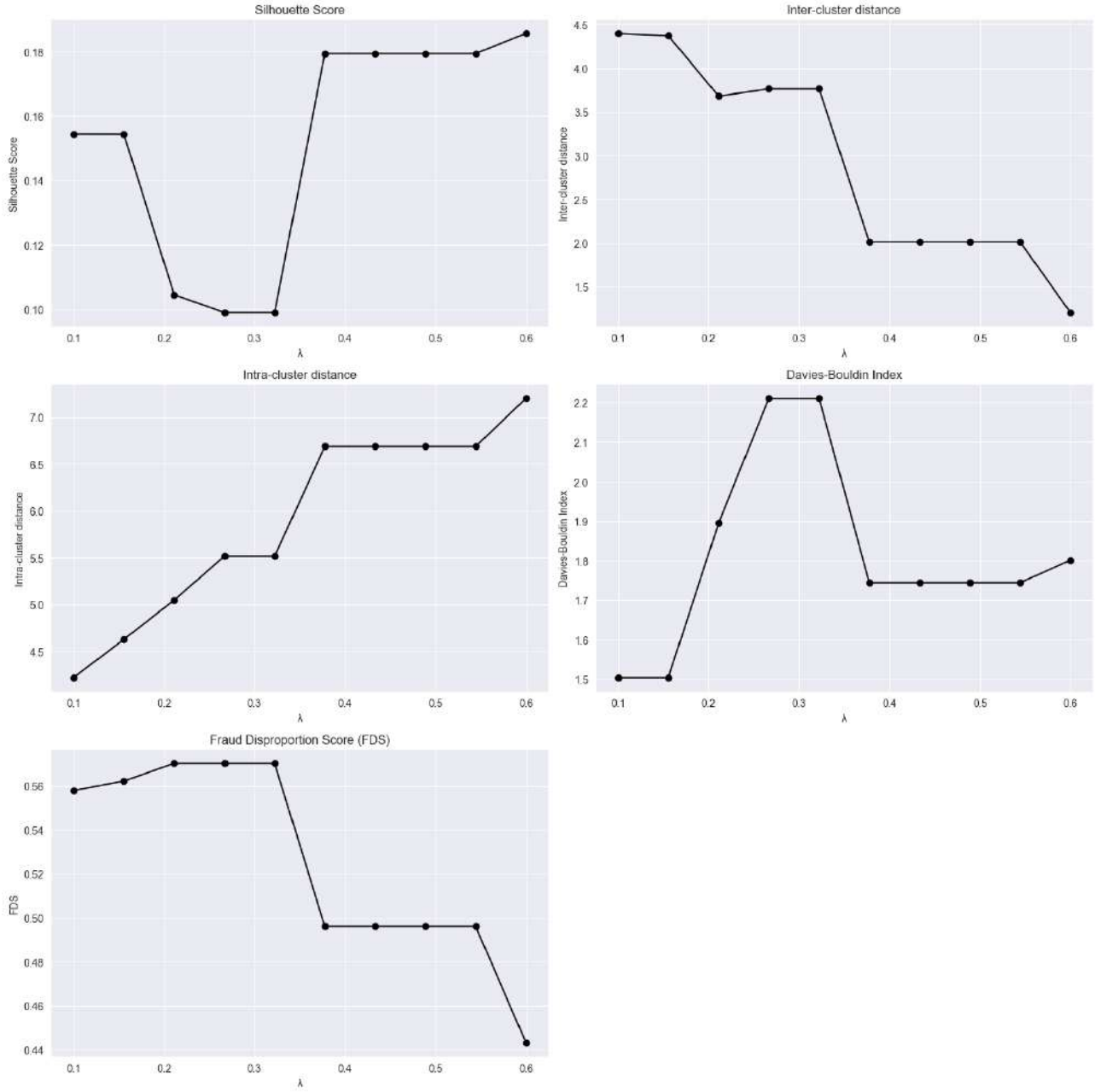
Silhouette Score
Silhouette Score
λ
Inter-cluster distance
Inter-cluster distance
λ
Intra-cluster distance
Intra-cluster distance
λ
Davies-Bouldin Index
Davies-Bouldin Index
λ
Fraud Disproportion Score (FDS)
FDS
λ

Рис. 4

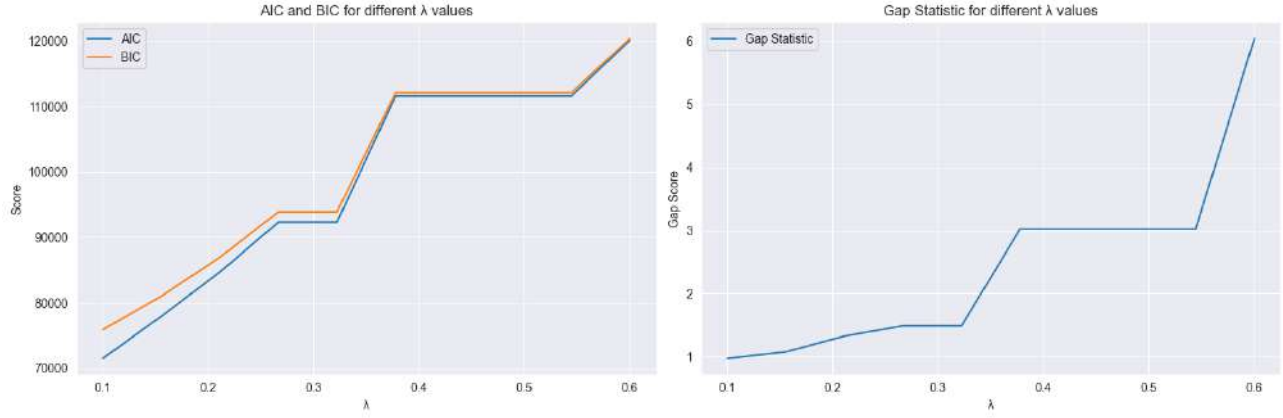
AIC and BIC for different λ values
AIC
BIC
Score
λ
Gap Statistic for different λ values
Gap Statistic
Gap Score
λ

Рис. 5

Для вибраного значення λ результати кластеризації було оцінено за відповідними метриками. Їх значення наведено в таблиці нижче:

Метрика	Значення
Силуетний коефіцієнт (Silhouette Score)	0.1046
Індекс Давіса-Болдіні (Davies-Bouldin Index)	1.8945
Середня внутрішньокластерна відстань	5.0475
Середня міжкластерна відстань	3.6814
Fraud Disproportion Score (FDS)	0.5702

Табл. 3: Оцінка якості кластеризації з використанням L0-регуляризації

Завдяки використанню регуляризації, частина координат центрів кластерів набули нульових значень, що свідчить про виключення відповідних ознак з моделі. Ненульовими залишились лише 6 ознак, координати яких можна побачити в табличці нижче:

Ознака	**Центр 1**	**Центр 2**	**Центр 3**	**Центр 4**	**Центр 5**	**Центр 6**
payment method card	0.512	0.435	0.510	0.478	0.528	0.541
card brand VISA	0.578	0.611	0.590	0.609	0.632	0.594
trial period	-0.760	-0.813	1.234	0.214	-0.046	-0.690
rebill number	-0.182	-0.236	-0.229	-0.134	-0.058	3.375
login	0.458	0.458	0.458	0.458	-2.183	-0.190
has otp upsell	-0.485	-0.485	-0.485	2.063	-0.104	-0.059
card brand MASTERCARD	0.000	0.000	0.000	0.000	0.000	0.000
...	...	...	...	...	...	...
has cons	0.000	0.000	0.000	0.000	0.000	0.000

Табл. 4: Координати центрів для L0-регуляризації

Це демонструє ключову особливість методу — повне занулення ознак, які не сприяють покращенню кластеризації. Таким чином, модель автоматично здійснює відбір ознак, залишаючи лише ті, що є релевантними для кластерного поділу.

4.6 Регуляризація методом Lasso

Наступним кроком дослідження стало реалізація та застосування L1-регуляризації. На відміну від методу Hard-Thresholding, описаного в попередньому розділі, Lasso передбачає м'яке порогове усічення координат центрів кластерів, що дозволяє поступово знижувати вагу менш значущих ознак у кожному кластері, при цьому зберігаючи їхню часткову інформаційну цінність. Існують випадки, коли значення певної ознаки для конкретного кластера повністю занулюється, проте, на відміну від L0-регуляризації, сама ознака не обнуляється у всіх центроїдах одночасно.

Реалізація алгоритму складається з класичних етапів: початкової випадкової ініціалізації центрів кластерів, призначення кожної точки до найближчого центроїда та оновлення центрів. Після обчислення стандартних центрів здійснюється застосування м'якого порогового усічення до кожної компоненти центру відповідно до формули:

$$C_{k,j} = \max\left(0, 1 - \frac{n\lambda}{2|A_k|\left|C^*_{k,j}\right|}\right) C^*_{k,j}.$$

де $C^*_{k,j}$ — стандартний центр кластера k по ознаці j, n — загальна кількість спостережень, $|A_k|$ — кількість точок у кластері k, λ — параметр регуляризації.

Алгоритм працює до досягнення збіжності або максимального числа ітерацій. Аналогічно до попереднього методу для вибору оптимального значення регуляризаційного параметра λ було використано кілька підходів. На рисунках 6-7 зображено графіки відповідних метрик в залежності від значення λ. Помітно, що найкращим параметром для даної моделі є $\lambda = 0.15$

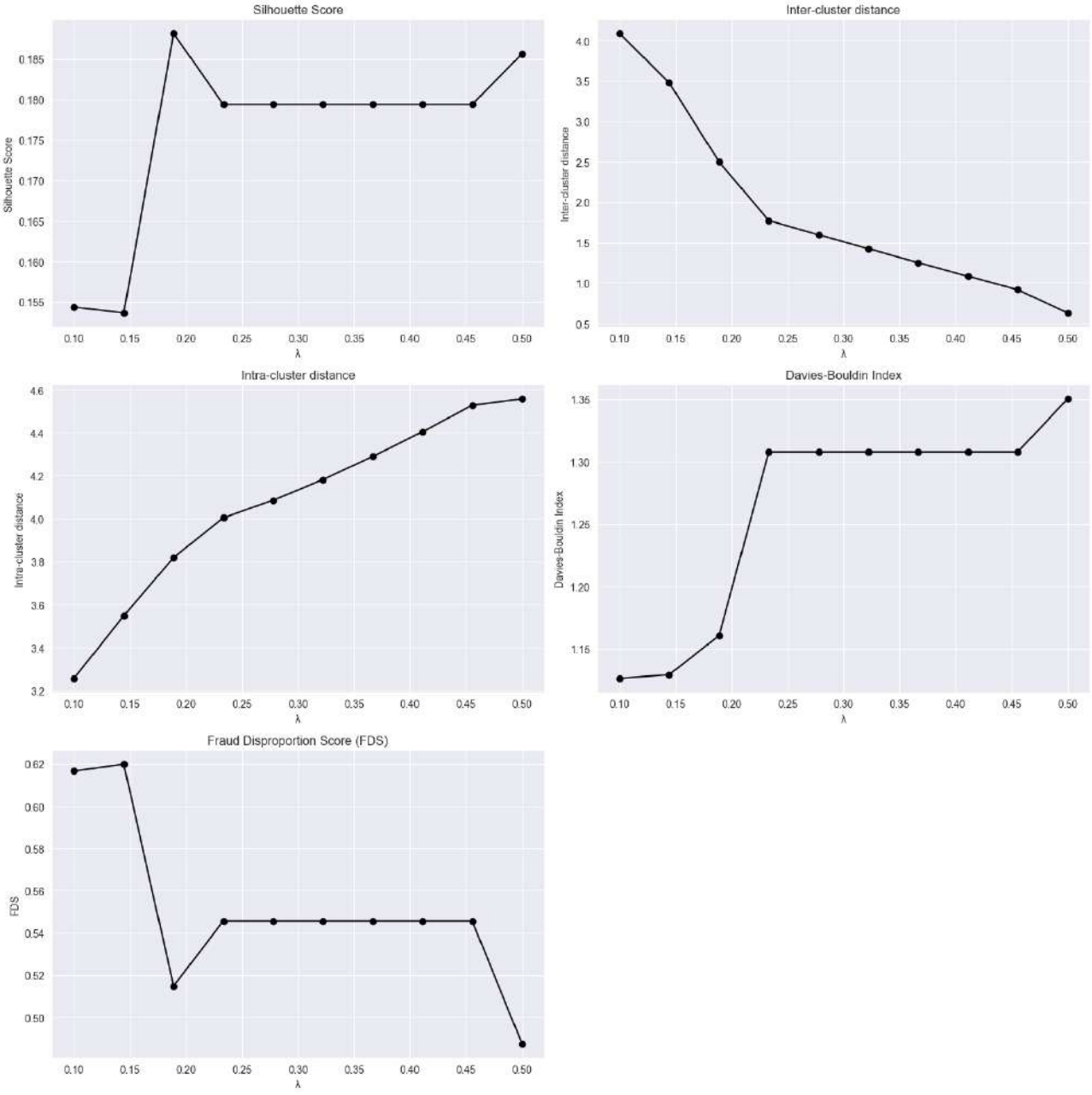
Silhouette Score
Silhouette Score
λ
Inter-cluster distance
Inter-cluster distance
λ
Intra-cluster distance
Intra-cluster distance
λ
Davies-Bouldin Index
Davies-Bouldin Index
λ
Fraud Disproportion Score (FDS)
FDS
λ

Рис. 6

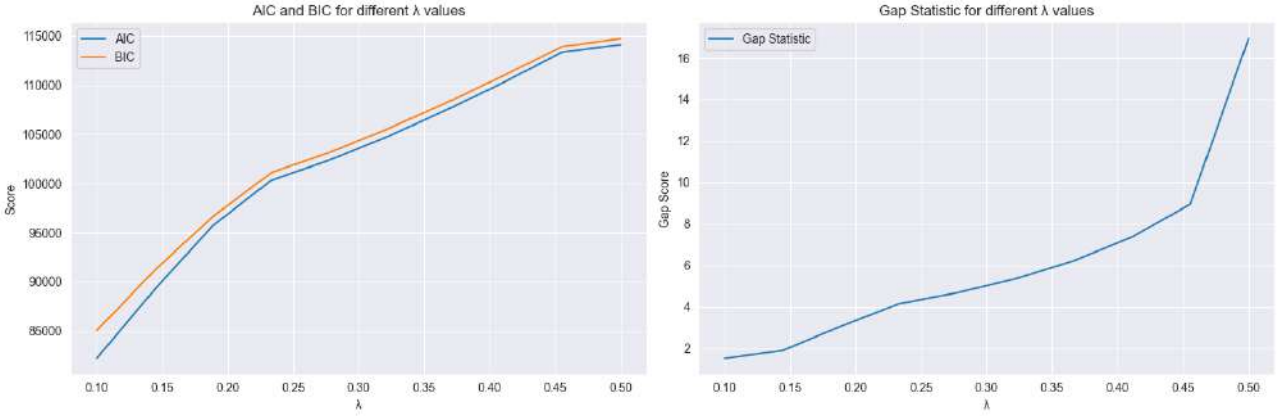
AIC and BIC for different λ values
AIC
BIC
Score
λ
Gap Statistic for different λ values
Gap Statistic
Gap Score
λ

Рис. 7

Для вибраного значення λ результати кластеризації було оцінено за відповідними метриками. Їх значення наведено в таблиці нижче:

Метрика	Значення
Силуетний коефіцієнт (Silhouette Score)	0.1536
Індекс Давіса-Болдіні (Davies-Bouldin Index)	1.1294
Середня внутрішньокластерна відстань	3.5848
Середня міжкластерна відстань	3.4052
Fraud Disproportion Score (FDS)	0.6199

Табл. 5: Оцінка якості кластеризації з використанням L1-регуляризації

Подальший аналіз координат центрів кластерів виявив, що частина ознак набули значень, близьких до нуля, що є характерною властивістю Lasso як методу регуляризації — поступове зменшення впливу менш значущих ознак без їх повного виключення для кожного окремого центроїда. Водночас у деяких ознаках спостерігалося повне занулення значень у окремих координатах, що свідчить про локальне відсіювання менш релевантної інформації.

Таким чином, регуляризаційний метод Lasso дозволяє не лише знаходити структуровані кластери, але й здійснювати відбір релевантних ознак, що є важливим для підвищення якості сегментації та інтерпретованості моделей у задачах з великою кількістю вхідних змінних.

Для підтвердження роботи алгоритму наведено фрагмент таблиці з координатами центрів кластерів для деяких ознак(Табл. 6).

Ознака	**Центр 1**	**Центр 2**	**Центр 3**	**Центр 4**	**Центр 5**	**Центр 6**
rebill number	0.008	0.056	0.000	0.000	0.000	1.852
group country USA	0.074	0.125	0.024	0.000	0.000	0.000
login	0.253	0.000	0.135	0.087	-1.691	0.098
age group 31-45	0.172	0.009	0.047	0.000	0.000	0.000
age group 45+	0.095	0.080	0.077	0.000	0.040	0.000
source Facebook	0.267	0.000	0.172	0.578	0.043	0.020
card brand VISA	0.372	0.000	0.240	0.074	0.121	0.003
card brand MASTERCARD	0.218	0.000	0.114	0.000	0.000	0.190

Табл. 6: Координати центрів для L1-регуляризації

4.7 Регуляризація методом Ridge

На відміну від L0 та L1 методів, регуляризація методом Ridge не призводить до занулення координат. Замість цього вона м'яко зменшує ваги ознак, застосовуючи масштабування центрів кластерів, що забезпечує стабілізацію моделі та зниження впливу менш значущих ознак без їх повного виключення.

У реалізованому алгоритмі k-середніх з Ridge-регуляризацією ітерації складаються з класичних кроків призначення точок до найближчих центрів і оновлення центрів кластерів як середнього значення точок відповідного кластера. Після обчислення стандартних центрів застосовується коефіцієнт стиснення, який залежить від параметра регуляризації λ та розміру кластера, що зменшує амплітуду координат центрів, відповідно до формули оновлення координат центроїдів:

$$C_{k,j} = \frac{1}{1 + \frac{n\lambda}{|A_k|}} C^*_{k,j},$$

Підбір оптимального значення λ для моделі виконувався за допомогою інформаційних критеріїв AIC, BIC та Gap-статистики, силуетного коефіцієнта, індексу Девіса-Болдіні, середніх внутрішньо- та міжкластерних відстаней, а також спеціалізованого показника Fraud Disproportion Score (FDS). На рисунках 8-9 зображені графіки відповідних метрик в залежності від значення λ. З візуального аналізу можна дійти до висновку, що найкращим параметром є $\lambda = 0.13$. Для вибраного значення λ результати кластеризації було оцінено за відповідними метриками. Їх значення наведено в таблиці нижче:

Метрика	Значення
Силуетний коефіцієнт (Silhouette Score)	0.1417
Індекс Девіса-Болдіні (Davies-Bouldin Index)	1.5522
Середня внутрішньокластерна відстань	4.9604
Середня міжкластерна відстань	1.9444
Fraud Disproportion Score (FDS)	0.5781

Табл. 7: Оцінка якості кластеризації з використанням L2-регуляризації

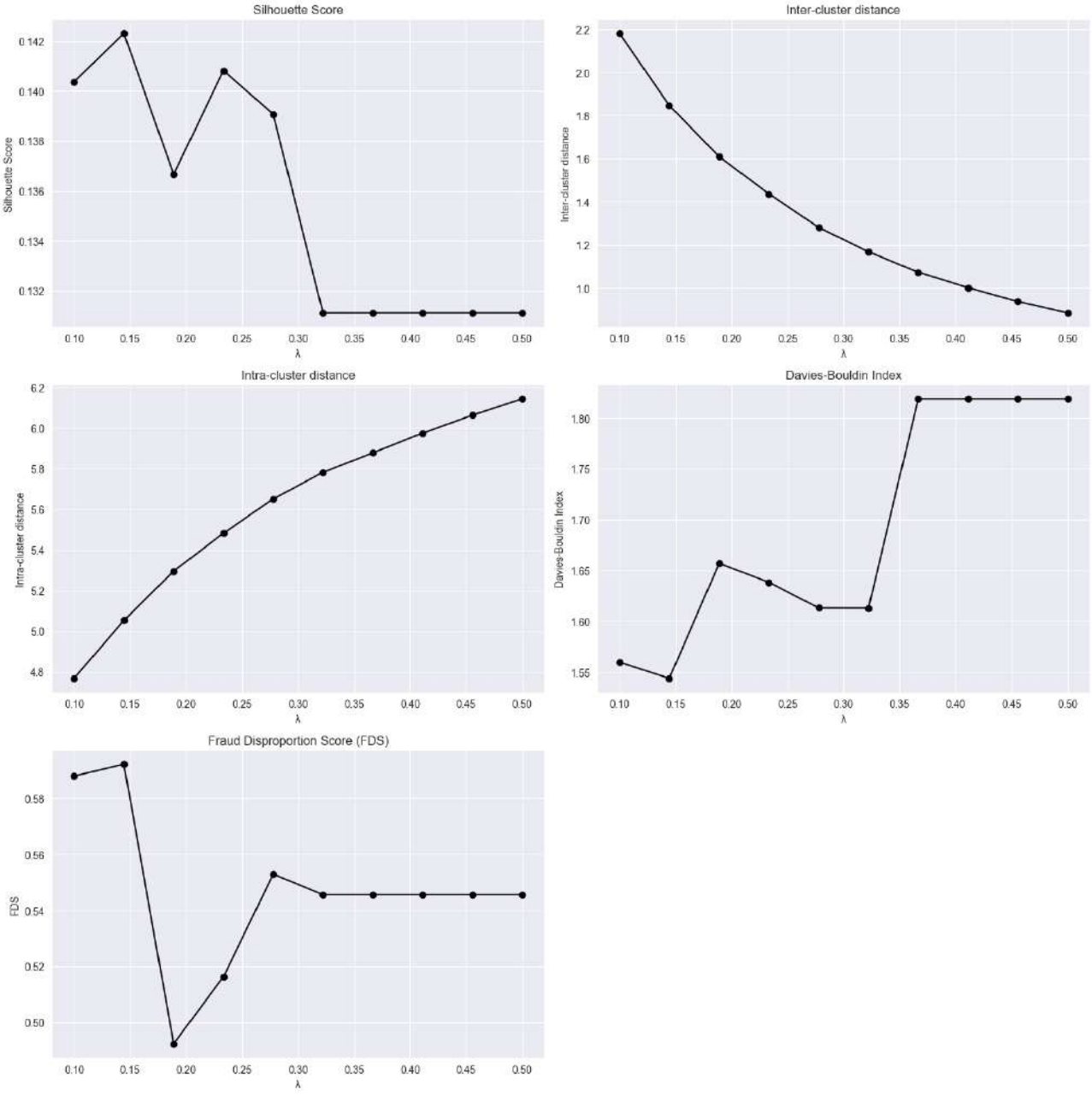
Silhouette Score
Silhouette Score
λ
Inter-cluster distance
Inter-cluster distance
λ
Intra-cluster distance
Intra-cluster distance
λ
Davies-Bouldin Index
Davies-Bouldin Index
λ
Fraud Disproportion Score (FDS)
FDS
λ

Рис. 8

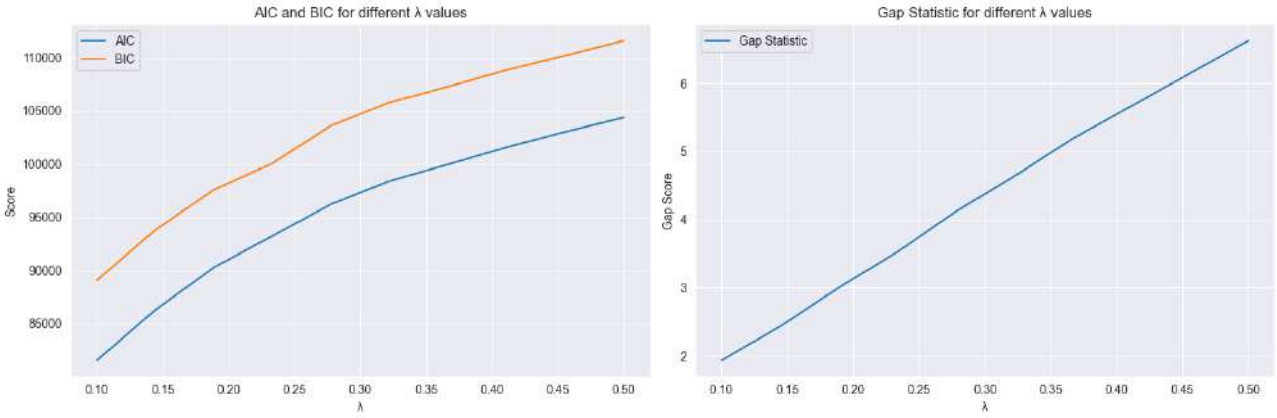
AIC and BIC for different λ values
AIC
BIC
Score
λ
Gap Statistic for different λ values
Gap Statistic
Gap Score
λ

Рис. 9

Для ілюстрації роботи алгоритму Ridge K-Means в таблиці нижче наведено координати центрів кластерів за вибраними деякими ознаками. Як видно з таблиці, L2-регуляризація зменшує значення координат, не занулюючи їх повністю, що дозволяє зберігати внесок усіх ознак у кластеризацію, забезпечуючи при цьому стабільність моделі.

Ознака	**Центр 1**	**Центр 2**	**Центр 3**	**Центр 4**	**Центр 5**	**Центр 6**
payment method card	0.345	0.166	0.327	0.256	0.293	0.165
group country EU	0.345	0.002	0.275	0.230	0.213	0.124
age group 31–45	0.272	0.148	0.230	0.202	0.154	0.113
card brand MASTERCARD	0.304	0.059	0.282	0.223	0.211	0.135
trial period	-0.568	0.050	0.773	0.100	-0.035	-0.200
login	0.328	0.086	0.287	0.244	-1.193	0.024

Табл. 8: Координати центрів для L2-регуляризації

4.8 Регуляризація методом Group Lasso

Метод Group Lasso є розширенням класичних методів регуляризації, який дозволяє виконувати груповий відбір ознак. На відміну від Lasso, що застосовує регуляризацію до окремих координат центрів кластерів, Group Lasso оперує на рівні повних ознак, тобто усіх координат певної ознаки одночасно. Це забезпечує одночасне збереження або виключення ознаки у всіх кластерах.

У реалізації алгоритму для кожної ітерації після обчислення стандартних центрів кластерів визначаються $L2$-норми по колонках матриці центрів (ознак). Для кожної координати центру виконується масштабування з коефіцієнтом стиснення, який залежить від параметра регуляризації λ, кількості елементів у кластері та норми відповідної ознаки. Для ознак з малою нормою коефіцієнт стиснення суттєво знижує їх вагу, фактично зануляючи повністю всі координати цієї ознаки.

Як і в інших методах, для вибору найкращого параметру λ були побудовані графіки якісних метрик, інформаційних критеріїв AIC, BIC та Gap-статистики(Рис. 10-11). З візуального аналізу даних графіків можна дійти до висновку, що найкращим параметром є $\lambda = 0.1$.

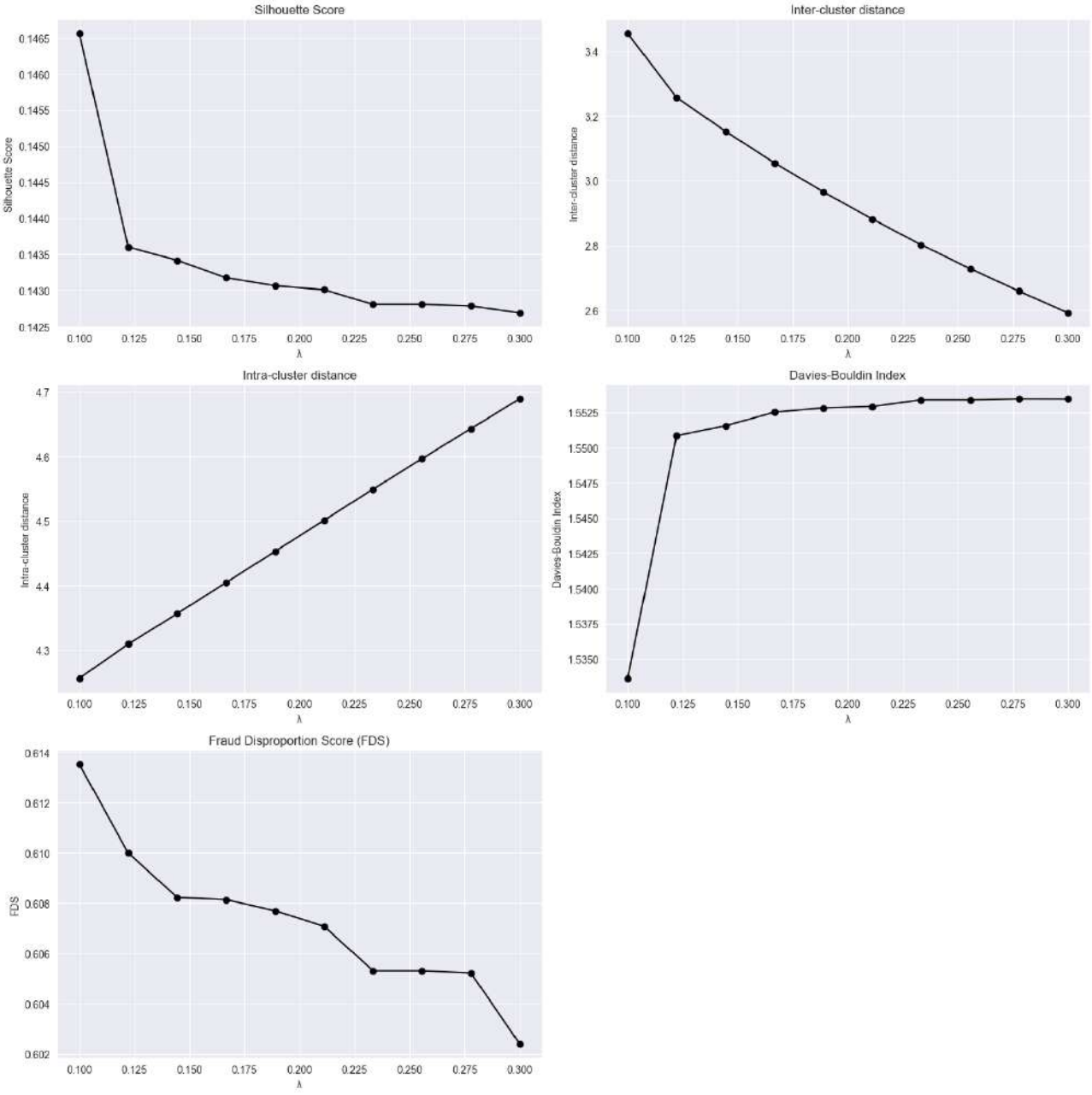
Silhouette Score
Silhouette Score
λ
Inter-cluster distance
Inter-cluster distance
λ
Intra-cluster distance
Intra-cluster distance
λ
Davies-Bouldin Index
Davies-Bouldin Index
λ
Fraud Disproportion Score (FDS)
FDS
λ

Рис. 10

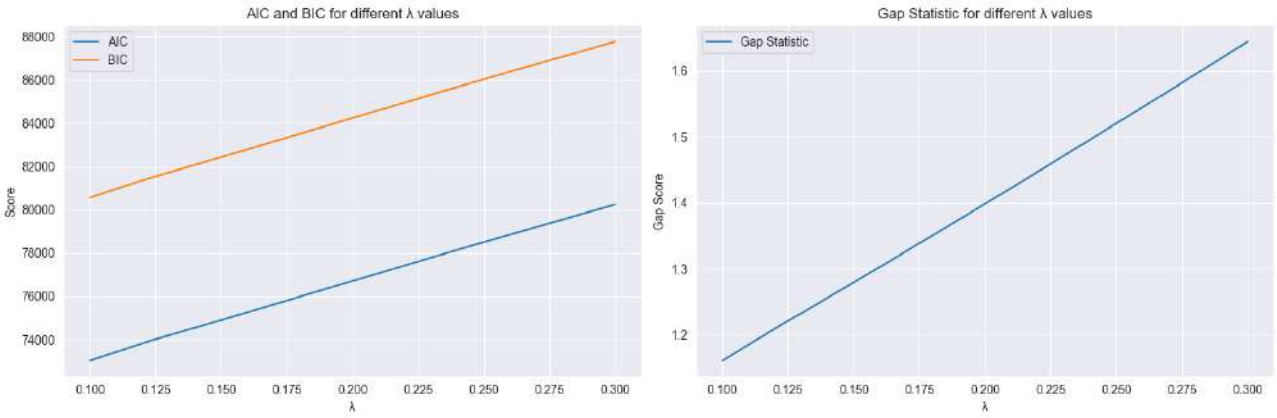
AIC and BIC for different λ values
AIC
BIC
Score
λ
Gap Statistic for different λ values
Gap Statistic
Gap Score
λ

Рис. 11

Для вибраного значення λ результати кластеризації було оцінено за відповідними метриками. Їх значення наведено в таблиці нижче:

Метрика	Значення
Силуетний коефіцієнт (Silhouette Score)	0.1466
Індекс Девіса-Болдіні (Davies-Bouldin Index)	1.5336
Середня внутрішньокластерна відстань	4.2568
Середня міжкластерна відстань	3.4556
Fraud Disproportion Score (FDS)	0.6135

Табл. 9: Оцінка якості кластеризації з використанням Group Lasso

В процесі застосування методу Group Lasso спостерігається очікувана поведінка, що частина ознак набуває значень координат центроїдів близьких до нуля(Табл. 10). Це свідчить про суттєве зменшення їхнього впливу на формування кластерів, хоча повного занулення може і не відбуватися. Такий механізм групового стискання дозволяє ефективно фільтрувати менш інформативні ознаки, підвищуючи стабільність та інтерпретованість моделі кластеризації.

Ознака	**Центр 1**	**Центр 2**	**Центр 3**	**Центр 4**	**Центр 5**	**Центр 6**
payment method google pay	0.082	0.031	0.067	0.054	0.056	0.030
group country OTHERS	0.237	0.013	0.190	0.181	0.179	0.131
source Bing	0.000	0.000	0.000	0.000	0.000	0.000
card type PREPAID	0.040	0.000	0.022	0.007	0.021	0.005

Табл. 10: Координати центрів для Group Lasso

4.9 Аналіз ефективності методів та результати задачі

Для порівняння та вибору найкращого методу кластеризації у межах поставленої задачі було проведено комплексний аналіз ефективності розглянутих методів регуляризації Hard-Thresholding, Lasso, Ridge і Group Lasso, а також класичного методу k-середніх. Цей аналіз включав оцінку ключових метрик, що дозволили всебічно оцінити якість кластеризації з урахуванням її структури, роздільності, компактності та специфічних особливостей самої задачі.

В результаті проведеного аналізу методів було сформовано таблицю (Табл. 11), яка містить ключові метрики для оцінки якості та ефективності кластеризації.

Метод	**Силуетний коеф.**	**Давіса–Болдіна Індекс**	**Внутрішньокластерна відстань**	**Міжкластерна відстань**	**FDS**
k-середніх	0.1404	1.8793	6.8393	3.6802	0.5579
Hard-Thresholding	0.1046	1.8945	5.0475	3.6814	0.5702
Lasso	0.1536	1.1294	3.5848	3.4052	0.6199
Ridge	0.1417	1.5522	4.9604	1.9444	0.5781
Group Lasso	0.1466	1.5336	4.2568	3.4556	0.6135

Табл. 11: Порівняння ефективності методів кластеризації

Провівши аналіз даної таблиці можна з впевненістю сказати, що метод Lasso продемонстрував найкращі результати для даної задачі серед усіх підходів. Завдяки цьому методу вдалося виділити найбільш компактні кластери, що підтверджується найменшим значенням внутрішньокластерної відстані. Крім того, Lasso вирізняється найвищими показниками силуетного коефіцієнта та індексу Девіса-Болдіні, що свідчить про чіткість та роздільність кластерів. Особливо важливо, що найбільш критична бізнесова метрика — Fraud Disproportion Score (FDS) — також досягла найвищого значення при застосуванні саме цього методу,що підкреслює його ефективність у задачі сегментації користувачів.

Переваги Lasso методу для даної задачі зумовлені особливостями механізму м'якого порогового усічення, що дозволяє зменшувати вагу менш інформативних ознак у кожному кластері без їх повного виключення. Такий підхід підтримує баланс між розрідженням моделі та збереженням важливої інформації, що сприяє більш точному та стабільному поділу даних.

Відмінність Lasso від L0-регуляризації полягає в тому, що останнє може дуже радикально вилучати ознаки, що іноді призводить до втрати корисної

інформації. Ridge і Group Lasso, хоча й забезпечують плавне зменшення ваг ознак, не дають такого рівня розрідженості і точного відбору, що відображається у менш виразній кластерній структурі. Таким чином, Lasso забезпечує оптимальний компроміс між якістю кластеризації та інтерпретованістю моделі, що виявилось критичним для ефективної сегментації користувачів.

Водночас інші методи регуляризації також продемонстрували покращення якості кластеризації у порівнянні з класичним методом k-середніх. Це підтверджується покращенням метрик компактності кластерів, таких як внутрішньокластерна відстань, а також підвищенням значення ключової бізнес-метрики — Fraud Disproportion Score (FDS). Загалом можна зробити висновок, що регуляризація позитивно впливає на якість кластеризації у випадках, коли дані містять велику кількість ознак та присутній шум, допомагаючи знизити вплив нерелевантних чи надлишкових змінних.

Результати кластеризації із застосуванням методу регуляризації Lasso зображено на рисунку 12, а саме розподіл користувачів за частками доходу та фроду у кожному кластері, а також метрику рівня фроду, яка відображає відносну частку фроду серед доходу для кожного сегменту.

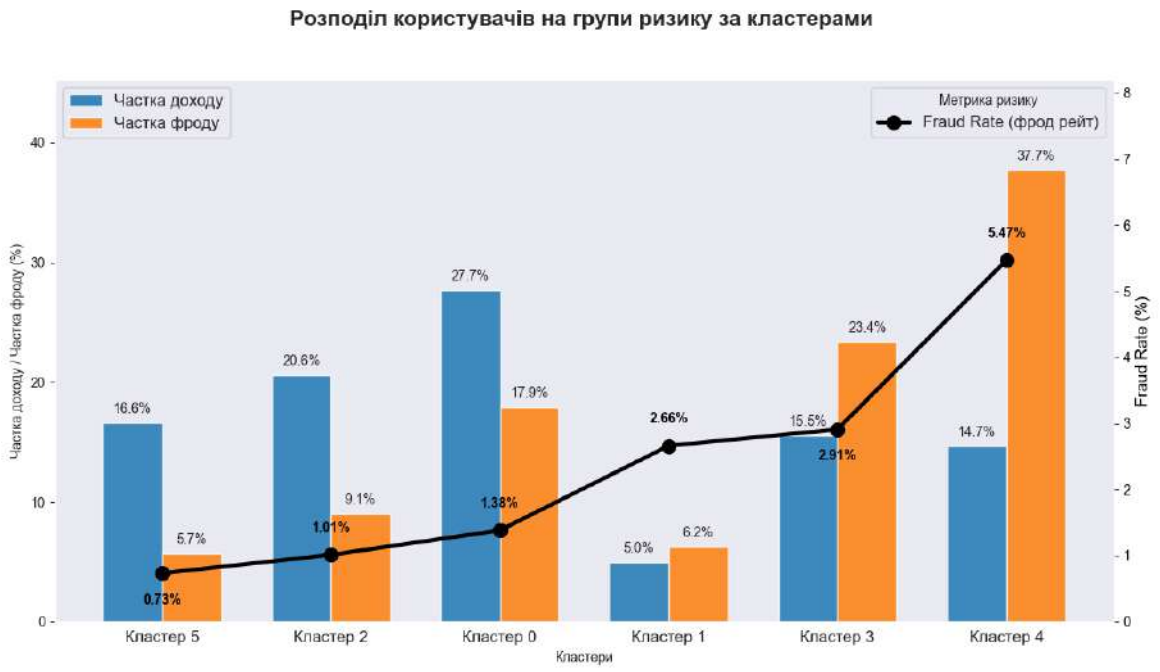

Рис. 12

З графіку чітко видно суттєву диспропорцію між часткою доходу та часткою фроду у різних кластерах, що свідчить про чіткий розподіл користувачів на групи з різним ризиком здійснення фроду.

Такий розподіл дозволяє бізнесу зосередити увагу на найбільш ризикованих сегментах, що допомагає ефективніше управляти ризиками, впроваджувати цільові заходи для запобігання фродових випадків та покращувати користувацький досвід додатоком.

Висновки

У даній роботі було досліджено методи регуляризації в задачах кластеризації, зокрема їх застосування для покращення якості сегментації користувачів мобільного застосунку.

На теоретичному рівні було проаналізовано основи кластеризації, зокрема класичний метод k-середніх. Окрему увагу приділено основним метрикам оцінки якості кластеризації — силуетному коефіцієнту, індексу Девіса-Болдіні, внутрішньокластерній та міжкластерній відстаням.

У рамках дослідження було детально розглянуто постановку задачі регуляризації в методі k-середніх. Вивчено чотири типи штрафних функцій: L0-регуляризацію (Hard-Thresholding), L1-регуляризацію (Lasso), L2-регуляризацію (Ridge) та Group Lasso-регуляризацію. Для кожного з підходів було описано відповідний механізм впливу на модель та реалізовано алгоритм їх обчислення. Також досліджено методи вибору параметра регуляризації λ, включаючи AIC, BIC, Gap-метод та оцінку стабільності кластеризації.

У практичній частині було поставлено задачу сегментації користувачів за ризиком здійснення "дружнього фроду" на основі історичних поведінкових даних. Для оцінки ефективності кластеризації було введено специфічну метрику Fraud Disproportion Score (FDS), що враховує диспропорцію між кількістю фроду та доходом у кожному кластері. Після підготовки даних (кодування категоріальних змінних, стандартизація ознак), було реалізовано базовий метод k-середніх, а також його регуляризовані версії: Hard-Thresholding, Lasso, Ridge та Group Lasso.

Порівняльний аналіз метрик якості кластеризації дозволив встановити, що метод Lasso продемонстрував найкращі результати серед усіх протестованих підходів. Цей метод дозволив виділити найбільш компактні та розділені кластери, що підтверджується найменшим значенням внутрішньокластерної відстані, найвищим значенням силуетного коефіцієнта та найнижчим індексом Девіса-Болдіні. Особливо важливо, що значення метрики FDS також виявилось найвищим саме для Lasso, що свідчить про його ефективність у задачі виявлення ризикованих сегментів.

Переваги методу Lasso для даної задачі пояснюється його здатністю до м'якого порогового усічення координат центроїдів, що дозволяє зменшити вплив менш інформативних ознак без їх повного виключення. Це забезпечує збалансовану модель, здатну до ефективного відбору ознак і водночас до

збереження ключової інформації.

Водночас інші методи регуляризації також продемонстрували покращення якості кластеризації порівняно з класичним k-середніх. Це проявилось у зниженні внутрішньокластерної відстані, збільшені міжкластерної відстані та зростанні значення метрики FDS. Регуляризація виявилася особливо корисною при роботі з даними з великою кількістю ознак та наявністю шуму, дозволяючи знизити вплив нерелевантних змінних.

Отже, результати дослідження підтверджують, що застосування регуляризації, зокрема методу Lasso, значно підвищує ефективність кластеризації, дозволяючи краще виявляти структуру даних, покращувати інтерпретованість моделей та приймати більш обґрунтовані бізнес-рішення.

Список літератури

[1] J. A. Hartigan, M. A. Wong *Algorithm AS 136: A K-Means Clustering Algorithm,* //Journal of the Royal Statistical Society. Series C (Applied Statistics), — 1979. — Vol. 28,— P. 100–108.

[2] Abiodun M. Ikotun,Absalom E. Ezugwu,Laith Abualigah, Belal Abuhaija, Jia Heming *K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data,* //Information Sciences — 2023. — Vol. 622,— P. 178–210.

[3] David MacKay *An Example Inference Task: Clustering,* //Information Theory, Inference, and Learning Algorithms, — 2003. — Vol. 20,— P. 284–292.

[4] David Arthur, Sergei Vassilvitskii, *k-means++: The Advantages of Careful Seeding* //Stanford University — 2020.

[5] Ali Idrus, Nafan Tarihoran, Ucup Supriatna,Ahmad Tohir, Suwarni Suwarni, Robbi Rahim *Distance Analysis Measuring for Clusteringusing K-Means and Davies BouldinIndex Algorithm* //TEM Journal, — 2022. — Vol. 11,— P. 1871–1876.

[6] Peter J. Rousseeuw *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis* //Journal of Computational and Applied Mathematics, — 1987. — Vol. 20,— P. 53–65.

[7] Jakob Raymaekers, Ruben H. Zamar *Regularized K-means through hard-thresholding* // University of British Columbia, Department of Statistics — 2020.

[8] Rafail Ostrovsky, Yuval Rabani, Leonard J. Schulman, Chaitanya Swamy *The Effectiveness of Lloyd-Type Methods for the k-Means Problem* //University of Waterloo — 2020.