МІНІСТЕРСТВО НАУКИ І ОСВІТИ УКРАЇНИ

КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

ІМЕНІ ТАРАСА ШЕВЧЕНКА

ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

ЗВІТ З ЛАБОРАТОРНОЇ РОБОТИ № 3 ЗА ТЕМОЮ:

**Задача кластеризації в середовищі аналізу даних**

Група \_\_\_\_\_11\_\_\_\_\_\_

Курс \_\_\_\_\_1\_\_\_\_\_\_\_\_\_

Студент (ка) \_\_\_\_Супруненко М. І. \_\_\_\_

Дата оформлення \_\_\_\_\_\_26.02.2025\_\_\_\_\_\_\_

Перевірив \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дата \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

МЕТА РОБОТИ: Метою лабораторної роботи є формування професійних вмінь та навичок щодо використання інструментів та методів кластеризації даних з використанням методів кластеризації K-Means та ієрархічної кластеризації, вміння застосовувати отримані знання на практиці в практичних задачах аналізу даних клієнтів.

ТЕОРІЯ: K-Means — це ітераційна техніка жорсткої кластеризації, яка використовує алгоритм неконтрольованого навчання. У цьому випадку загальна кількість кластерів попередньо визначається користувачем, і на основі подібності кожної точки даних точки даних групуються. Цей алгоритм також визначає центроїд кластера.

Алгоритм для K-Means:

1. Вкажіть кількість кластерів (K).
2. Випадково призначте кожну точку даних кластеру.
3. Обчислити центроїд кластера.
4. Перерозподіліть кожну точку даних до найближчого центроїда кластера.
5. Змінити центроїд кластера.

В ієрархічній кластеризації об’єкти класифікуються в ієрархію, подібну до деревоподібної структури, яка використовується для інтерпретації моделей ієрархічної кластеризації. Алгоритм наступний:

1. Зробіть кожну точку даних одним кластером, який утворює N кластерів.
2. Візьміть дві найближчі точки даних і зробіть з них один кластер, який утворює N-1 кластерів.
3. Візьміть два найближчі кластери та зробіть з них один кластер, який утворює N-2 кластери.
4. Повторюйте кроки 3, доки не залишиться лише один кластер.

ХІД РОБОТИ:

1. Завантажте навчальний набір даних (бажано обирати невеликі набори) та опишіть його параметри та змінні.
2. У разі необхідності здійсніть первинну обробку набору (знайдіть викиди, пропущені дані та очистіть набір). Можливо знадобиться нормалізація набору даних, скористайтеся відомими вам функціями нормалізації, наприклад Z-score.
3. Створіть модель kmeans(), опишіть параметри моделі, здійсніть дослідження якості кластеризації з різною кількістю кластерів.
4. Здійсніть візуалізацію отриманих результатів з використанням функції fviz\_cluster з пакету factoextra.
5. Здійсніть кластеризцію обраного набору за допомогою ієрархічної кластеризації. Опишіть параметри налаштування функції hclust().
6. Побудуйте матрицю відстаней, зробіть аналіз.
7. Побудуйте дендрограму кластеризації (дерево).

ВХІДНІ ДАНІ:

1. В ході лабораторної роботи для вивчення параметрів функцій використовується навчальний набір mtcars.
2. Кожен студент самостійно обирає навчальний набір для проведення самостійних досліджень та оформлює лабораторну роботу.

Необхідні бібліотеки:

# Installing Packages

install.packages("fpc")

# Loading package

library(fpc)

# Installing the package

install.packages("dplyr")

# Loading package

library(dplyr)

**РЕЗУЛЬТАТ ВИКОНАННЯ:**

1. Встановлення необхідних бібліотек (тільки один раз)

|  |
| --- |
| install.packages("factoextra")  # Для візуалізації кластерів  install.packages("lubridate")   # Для роботи з датами  install.packages("ggplot2")     # Для побудови графіків  install.packages("dplyr")       # Для роботи з таблицями |

1. Завантаження бібліотек

|  |
| --- |
| library(dplyr)  library(factoextra)  library(lubridate)  library(ggplot2) |

1. Завантаження і підготвока даних до кластеризації

|  |
| --- |
| # Завантаження даних  df <- read.csv("expenses1.csv")  # Перетворення дати у формат Date  df$Transaction.Date <- as.Date(df$Transaction.Date, format="%Y-%m-%d")  # Отримання першого числа місяця  df$First.Day.Of.Month <- floor\_date(df$Transaction.Date, unit = "month")  # Обчислення кількості днів від початку місяця  df$Days.Since.First <- as.numeric(df$Transaction.Date - df$First.Day.Of.Month)  # Видалення зайвих колонок  df\_cleaned <- df %>% select(Income.Spendings, Days.Since.First) |

1. Масштабування даних (Z-score normalization)

|  |
| --- |
| # Масштабування даних (Z-score normalization)  df\_scaled <- scale(df\_cleaned) |

1. Визначення оптимальної кількості кластерів методом "Elbow"

|  |
| --- |
| # Визначення оптимальної кількості кластерів методом "Elbow"  fviz\_nbclust(df\_scaled, kmeans, method = "wss") |
|  |

1. Побудова K-Means моделі з 3 кластерами (можна змінити кількість)

|  |
| --- |
| # Побудова K-Means моделі з 3 кластерами (можна змінити кількість)  set.seed(42)  kmeans\_model <- kmeans(df\_scaled, centers = 3, nstart = 25)  # Візуалізація кластерів  fviz\_cluster(kmeans\_model, data = df\_scaled) |

1. Ієрархічна кластеризація + дендограма

|  |
| --- |
| # Ієрархічна кластеризація  dist\_matrix <- dist(df\_scaled, method = "euclidean")  # Обчислення відстаней  hclust\_model <- hclust(dist\_matrix, method = "ward.D2")  # Побудова дендрограми  plot(hclust\_model, main = "Дендрограма ієрархічної кластеризації", sub = "", xlab = "")  rect.hclust(hclust\_model, k = 3, border = "red")  # Відображення 3 кластерів |

1. Кластери

|  |
| --- |
| # Додавання кластерів до основного датасету  df$Cluster <- kmeans\_model$cluster  # Відкриття кожного кластеру в окремому вікні для перегляду  View(df\_cluster\_1)  View(df\_cluster\_2)  View(df\_cluster\_3) |
|  |
|  |
|  |

**Висновок:**

У результаті виконаної кластеризації даних було застосовано два методи: K-means та ієрархічну кластеризацію. Перший метод дозволив розподілити дані на три основні групи за допомогою алгоритму K-means, який був налаштований на три центри (кластерів). Визначення оптимальної кількості кластерів методом "Elbow" підтвердило правильність вибору кількості кластерів. Для кожного з кластерів було здійснено візуалізацію, що дозволило зрозуміти структуру розподілу даних та їх характерні ознаки.

Ієрархічна кластеризація дозволила доповнити результати, надавши більш гнучке бачення структури кластерів через побудову дендрограми, що показує відстані між об'єктами і дає змогу візуально оцінити оптимальну кількість груп. Обидва методи показали подібні результати, що підкріплює коректність проведеного аналізу.