ОСВІТА І НАУКА В СУЧАСНИХ УМОВАХ: ПРОБЛЕМИ Й ПЕРСПЕКТИВИ

УДК 378.016:004.85 **DOI: 10.58407/visnik.253543**

Горошко Юрій

https://orcid.org/0000-0001-9290-7563
Researcher ID GQB-3684-2022
Scopus-Author ID 57952935600

Доктор педагогічних наук, професор,
завідувач кафедри інформатики і обчислювальної техніки
Національного університету «Чернігівський колегіум» імені Т. Г. Шевченка
(Чернігів, Україна) E-mail: horoshko_y@ukr.net

Цибко Ганна

https://orcid.org/0000-0002-1861-3003
Researcher ID AAC-6021-2021
Scopus-Author ID 57952055100

Кандидат педагогічних наук, доцент,
доцент кафедри інформатики і обчислювальної техніки
Національного університету «Чернігівський колегіум» імені Т. Г. Шевченка
(Чернігів, Україна) E-mail: a.tsb@ukr.net

Костюченко Андрій

https://orcid.org/0000-0002-6178-6444
Researcher ID GPX-1175-2022

Кандидат педагогічних наук,
старший викладач кафедри інформатики і обчислювальної техніки
Національного університету «Чернігівський колегіум» імені Т. Г. Шевченка
(Чернігів, Україна) E-mail: kost_andrey@ukr.net

ВИКОРИСТАННЯ КЛАСТЕРНОГО АНАЛІЗУ В ІНТЕЛЕКТУАЛЬНОМУ ОПРАЦЮВАННІ ДАНИХ

У статті розглядаються основні поняття такого розділу науки про великі дані (Data Science), як кластерний аналіз. Висвітлено теоретичні основи та практичні аспекти застосування кластерного аналізу в різних галузях. Здійснено добір вільно поширюваних програмних засобів для кластерного аналізу, доцільних для застосування в освітньому процесі і практичній діяльності. Запропоновано елементи методики навчання основ кластерного аналізу майбутніх учителів інформатики і фахівців з комп'ютерних та соціальних наук.

Мета статті. *Проаналізувати та добрати дидактично доцільні вільно поширювані інструменти для здійснення кластерного аналізу даних, та розробити окремі компоненти методики навчання цієї теми майбутніх учителів інформатики, фахівців з комп'ютерних та соціальних наук.*

Методологія. *Вивчення та аналіз наукових публікацій, навчально-методичних видань, порівняльний аналіз програмного забезпечення, узагальнення досвіду фахівців в галузі освіти, комп'ютерних та соціальних наук, моделювання і синтез компонентів методики навчання, системний підхід до навчання інформатики.*

Наукова новизна. *Відібрано доцільні вільно поширювані інструменти здійснення кластерного аналізу та розроблено окремі компоненти методики навчання майбутніх фахівців.*

Висновки. *У роботі розглянуто основні концепції кластерного аналізу. Висвітлено сутність та призначення кластерного аналізу, зроблено огляд джерел стосовно цієї теми, відібрано вільно поширювані інструменти та наведено методичні підходи до навчання основ кластерного аналізу майбутніх учителів інформатики і фахівців з комп'ютерних та соціальних наук. Зазначені методичні підходи до навчання сучасних методів і засобів кластерного аналізу спрямовані на формування у майбутніх фахівців спеціальних компетентностей, необхідних інтелектуального аналізу даних. Таке формування може бути проведене при вивченні дисципліни «Основи штучного інтелекту та інтелектуального аналізу даних», що актуалізує тематику вказаного курсу.*

Ключові слова: *інтелектуальний аналіз даних, кластерний аналіз, KNIME, Python.*

Актуальність роботи. Питання, пов'язані з машинним навчанням, штучними нейронними мережами, інтелектуальним аналізом даних стали одним з ключових напрямків сучасних інформаційних технологій.

Великим поштовхом у цьому напрямку стала поява великих мовних моделей (LLM). Відповідні зміни відбуваються і у навчанні інформатичних дисциплін, і не тільки майбутніх спеціалістів з інформаційних технологій і комп'ютерних наук, а і майбутніх вчителів інформатики, фахівців соціально-психологічного напрямку. Хоча питання, пов'язані з ШІ, машинним навчанням, інтелектуальним аналізом даних частково розглядалися в інформатичних дисциплінах ОП Середня освіта (Інформатика) НУЧК, визріла нагальна потреба введення в цю ОП окремої дисципліни, присвяченої згаданим вище питанням. Тому з 2025 р. в ОП Середня освіта (Інформатика) введено дисципліну «Основи штучного інтелекту та інтелектуального аналізу даних». Було визначено коло тем, що будуть розглядатися. Однією з таких тем є основи кластерного аналізу.

Мета написання статті – проаналізувати та добрати дидактично доцільні вільно поширювані інструменти для здійснення кластерного аналізу даних, та розробити окремі компоненти методики навчання майбутніх учителів інформатики, фахівців з комп'ютерних та соціальних наук цієї теми.

Аналіз останніх досліджень і публікацій. У статті [1] розглядаються основні поняття науки про великі дані (Data Science). Проаналізовано технології машинного навчання (Machine Learning) та інтелектуального аналізу даних (Data Mining) та відповідні вільно поширювані програмні засоби і джерела даних, доцільні для застосування у навчальному процесі. Запропоновано елементи методики навчання основ Data Science майбутніх учителів інформатики і фахівців з комп'ютерних наук. Проте питання, присвячені кластерному аналізу та методики його навчання в ній не розглянуті.

У роботі [2] розкрито сутність кластерного аналізу, обговорюються особливості застосування методів кластерного аналізу саме в прикладних дослідженнях, не лише теорія, а і методика застосування для реальних даних. В ній описано ключові кроки застосування кластерного аналізу у прикладних дослідженнях, а саме вибір змінних, визначення метрики схожості, вибір алгоритму кластеризації, інтерпретація результатів.

Є ряд статей, присвячених застосуванню кластерного аналізу в різних сферах діяльності. Так у роботі [3] розглядається використання кластерного аналізу в бібліотечній і інформаційній сфері.

У дослідженні [4] Досліджено поведінку студентів в цифровому середовищі. У статті показано, що методи кластерного аналізу можна застосовувати до вивчення цифрових слідів (ЦС) студентів у цифровому освітньому середовищі (ЦОС) закладу освіти. Автори встановили, що за допомогою кластеризації можна виділити шість типів користувачів-студентів, які характеризуються подібними патернами активності в ЦОС та з точки зору інформаційної безпеки (ІБ). Зазначено, що проведений аналіз допомагає покращити персоналізацію навчання, підвищити ефективність освітніх програм, виявляти аномальну поведінку або потенційні загрози ІБ в цифровому середовищі.

У роботі [5] розглянуто важливий алгоритм k-means з царини кластерного аналізу. Розглянуто його слабкі місця і порівняно продуктивність кількох підходів на наборах даних. Автори також описують алгоритми-альтернативи/поліпшення (ініціалізація, робота з категоріями, масштабування, прискорення) і наводять рекомендації для практичного застосування.

У роботах В. О. Климчука [6; 7] продемонстровано підходи до застосування кластерного аналізу у психологічних дослідженнях, а у роботі [8] ці підходи висвітлюються для дослідження економічних проблем.

Результати дослідження. У багатьох дослідженнях важливо організувати отримані дані у вигляді наочної і зрозумілої структури, придатної для подальшого аналізу і досліджень. Наприклад в біології це розбиття сукупності тварин на види і підвиди, у психології – класифікація видів поведінки, у медицині такий організації піддаються симптоми захворювання чи види лікування. Це можна зробити за допомогою *кластерного аналізу*.

Загалом, коли необхідно розбити великі масиви даних на групи, які придатні для подальшого аналізу, застосовують кластерний аналіз.

Таким чином кластерний аналіз – це метод машинного навчання, який використовується для групування об'єктів таким чином, щоб об'єкти в межах однієї групи (кластера) були максимально схожими між собою та максимально відрізнялися від об'єктів інших груп. Кластеризація відноситься до некерованого навчання, коли алгоритм не має заздалегідь визначених міток класів і самостійно знаходить структуру даних.

Важливими напрямками застосування кластерного аналізу є виявлення природних груп (кластерів) у даних, зменшення розмірності даних та візуалізація, виявлення аномалій в даних та підготовка їх даних для подальшого аналізу (класифікації).

Виділяють два основних методи кластерного аналізу: деревоподібна кластеризація та метод К-середніх.

Метод деревоподібної кластеризації (ієрархічна кластеризація, tree clustering) дозволяє побудувати ієрархічне кластерне дерево (дендрограму).

Ієрархічна кластеризація послідовно об'єднує або розділяє об'єкти, створюючи систему кластерів на різних рівнях. Існують два основних підходи:

1. Агломеративний (знизу вгору) – поступове об'єднання кластерів доти, доки не залишиться один великий кластер.

2. Дивізивний (зверху вниз) – починаючи з одного великого кластера, послідовно розбивають на підкластери, доки не отримають окремі об'єкти.

Підходи (стратегії) до обчислення відстані між кластерами наступні [7]:

– Single linkage – мінімальна відстань між елементами двох кластерів. Тут відстань між двома кластерами визначається як відстань між двома найближчими об'єктами (найближчими сусідами). Результуючі кластери представляються у вигляді довгих «ланцюжків». Стратегія пов'язує два кластери разом, коли будь-які два об'єкти в цих кластерах ближче один до одного, ніж усі інші.

– Complete linkage – максимальна відстань між елементами двох кластерів. При використанні цієї стратегії відстань між кластерами визначається найбільшою відстанню між двома об'єктами з різних кластерів (між найвіддаленішими сусідами). Якщо є природним типом кластерів в отриманих даних є ланцюжки, то ця стратегія є непридатною. Стратегія утворює в основному «кущі» об'єктів.

– Average linkage – середнє арифметичне від усіх пар відстаней. Відстань між двома кластерами визначається як середня відстань між всіма парами об'єктів у них. Метод ефективний випадку реального об'єднання об'єктів як у «кущі», так і в «ланцюжки».

– Ward's метод – мінімізує збільшення внутрішньокластерної дисперсії. Ця стратегія мінімізує суму квадратів для двох гіпотетичних кластерів, які можуть бути сформовані на кожному кроці процесу кластеризації. Метод вважається ефективним, але намагається створювати кластери малого розміру.

Також необхідно обрати міру відстані між об'єктами. Розглянемо, які бувають міри відстаней між об'єктами.

Евклідова відстань (Euclidian distances). Це найуживаніша міра відстані між об'єктами, яка являє собою геометричну відстань між об'єктами у багатомірному просторі.

Манхеттенівська відстань (City-block (Manhattan) distances). Ця міра у більшості випадків призводить до таких же результатів, як і Евклідова відстань, але зменшується вплив окремих великих різниць (викидів) через те, що відстань обчислюється як сума модулів різниць координат.

Є і інші міри відстаней між об'єктами.

Розглянемо алгоритм агломеративної кластеризації.

– На 1-му кроці потрібно обчислити матрицю відстаней між усіма об'єктами (наприклад, за евклідовою метрикою).

– На 2-му етапі відбувається пошук двох найближчих кластерів за вибраним критерієм близькості.

– На 3-му етапі відбувається об'єднання знайдених кластерів у новий.

– На 4-му етапі оновлюється матриця відстаней, враховуючи новий кластер.

– Кроки 2-4 повторюються, поки не лишиться один кластер.

Даний алгоритм дозволяє отримати наочну ієрархічну структуру даних, підходить для візуального аналізу невеликих вхідних даних.

Наступним інструментом кластерного аналізу доцільно розглянути метод K-means (К-середніх). Він використовується тоді, коли є певна гіпотеза стосовно кількості кластерів, на які будуть розбиті вхідні дані. Тому для застосування цього методу потрібно наперед задати кількість кластерів, і алгоритм кластеризації дозволить знати ці кластери так, щоб вони максимально різнилися один від одного. Перевагою цього методу є можливість перевірки статистичної значимості відмінностей між виділеними кластерами.

Розглянемо більш детально суть алгоритму, мета якого мінімізувати суму квадратів відстаней між кожним об'єктом і центром кластера, до якого він належить.

На вході алгоритму маємо множину даних x_i $(i=1..n)$, кожен елемент якої має m ознак, а також k – кількість кластерів.

Потрібно визначити множину c_i *(i=1..k)* – центрів кластерів, щоб сума квадратів відстаней між кожним об'єктом і центром кластера була мінімальною.

– На 1-му кроці алгоритму випадково обирають K початкових центрів кластерів.

– На 2-му кроці для кожного x_i визначають найближчий центр за обраною метрикою.

– На 3-му кроці для кожного кластера визначають новий центр, як середнє всіх точок у кластері.

– Якщо центри не змінилися, або зміни менші за якесь наперед визначене мале eps, алгоритм зупиняється, а інакше переходить до кроку 2.

Максимальна кількість ітерацій, як правило, обмежується.

До переваг даного алгоритму можна віднести його відносно малу часову складність (*O(n × K × кількість ітерацій)*) і простоту, а до недоліків – необхідність задавати кількість кластерів, чутливість до початкової ініціалізації, погана робота з кластерами складної форми.

В якості одного з інструментів для проведення кластерного аналізу доцільно обрати платформу KNIME, про яку вже згадувалося в [1]. Нагадаємо, що KNIME (Konstanz Information Miner) – це потужна *no-code/low-code* платформа для аналітики даних, машинного навчання, а також кластерного аналізу.

Головна її перевага – аналіз даних можна робити візуально, без програмування, а лише вибираючи необхідні вузли (*nodes*), налагоджуючи їх та з'єднуючи ці вузли у робочому процесі (*workflow*). Ця програма є відкритою, підтримує імпорт даних у форматі Excel, CSV та деяких інших форматах, має готові вузли для кластерного аналізу та відповідної візуалізації даних.

Розглянемо наступний приклад [7]. Нехай було проведене дослідження позитивного ставлення студентів психологічного факультету до студентів інших факультетів та до майбутніх професійних ролей, де цікавить об'єднання студентів у групи на основі схожого ставлення. Для цього було створено рольовий перелік факультетів та можливих спеціальностей випускників психологічного факультету. Потім студентам запропонували оцінити своє ставлення до всіх ролей за 10-бальною шкалою. В результаті було отримано такий масив даних.

Таблиця 1

Роль	А.О.	О.В.	М.П	К.П	Р.А.	К.В.	В.Д.	П.Р.	Д.К.	Е.О.	З.А.
Студент психологічного факультету	5	9	9	5	5	10	3	5	1,5	1,5	1,5
Студент педагогічного факультету	2	0,5	3	1	2	3	4	2	9	9	9
Студент фізико-математичного факультету	3	1	2	4	3	4	6	3,5	10	10	10
Студент природничого факультету	9	2	1,5	3	4	2	7	7	9,5	9,5	9,5
Студент історичного факультету	4	3	2,5	3,5	6	5	9	9	7	7	7
Студент філологічного факультету	10	4	1	2	1	6	1,5	1	8	8	8
Психолог	4,5	9,5	10	10	10	9	2,5	5,5	2	2	2
Соціальний педагог	5,5	10	8	9,5	9,5	9,5	3,5	4	1	1	1
Вчитель	8	5	5	7	5	7	5	3	6	6	6
Викладач	6	6	7	9	8	5,5	8	8	5	5	5
Керівник	3,5	7	6	8	7	4,5	10	10	3	3	3
Консультант	7	8	4	6	6	8	2	6	4	4	4

Оберемо метод деревоподібної кластеризації. В якості міри виберемо Евклідову, а в якості стратегії – стратегію найближчого сусіда.

Дану таблицю завантажуємо у вузол *Excel Reader* та налагоджуємо його (*контекстне меню*|*Configure...*), вказавши шлях до файлу за допомогою кнопки *Browse...*, та вибравши пункти *Select first sheet with data*, T*able contains columns names in row 1* та *Read entire data of the sheet*. Запускаємо вузол (*контекстне меню*|*Execute*).

Далі транспонуємо таблицю даних за допомогою вузла Transpose. Наступним кроком є додавання вузла Distance Matrix Calculate для побудови матриці відстаней. Налаштуємо цей вузол, вказавши Евклідову метрику. Потім додаємо вузол Hierarchical Clustering (DistMatrix) для, власне розбиття даних на кластери та налаштовуємо його, вказавши стратегію найближчого сусіда. Для побудови дендрограми використаємо вузол Hierarchical Cluster View.

Налаштований робочий процес виглядає наступним чином (рис. 1)

Якщо в контекстному меню останнього вузла вибрати послугу View: Dendrogram, отримаємо таку дендрограму (рис. 2)

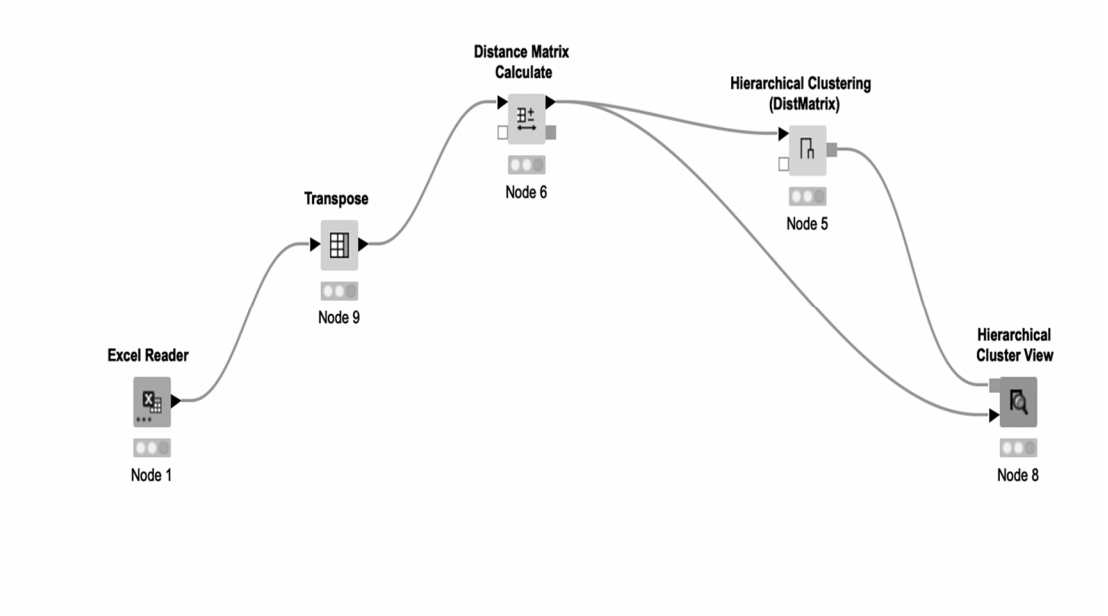

Рис. 1

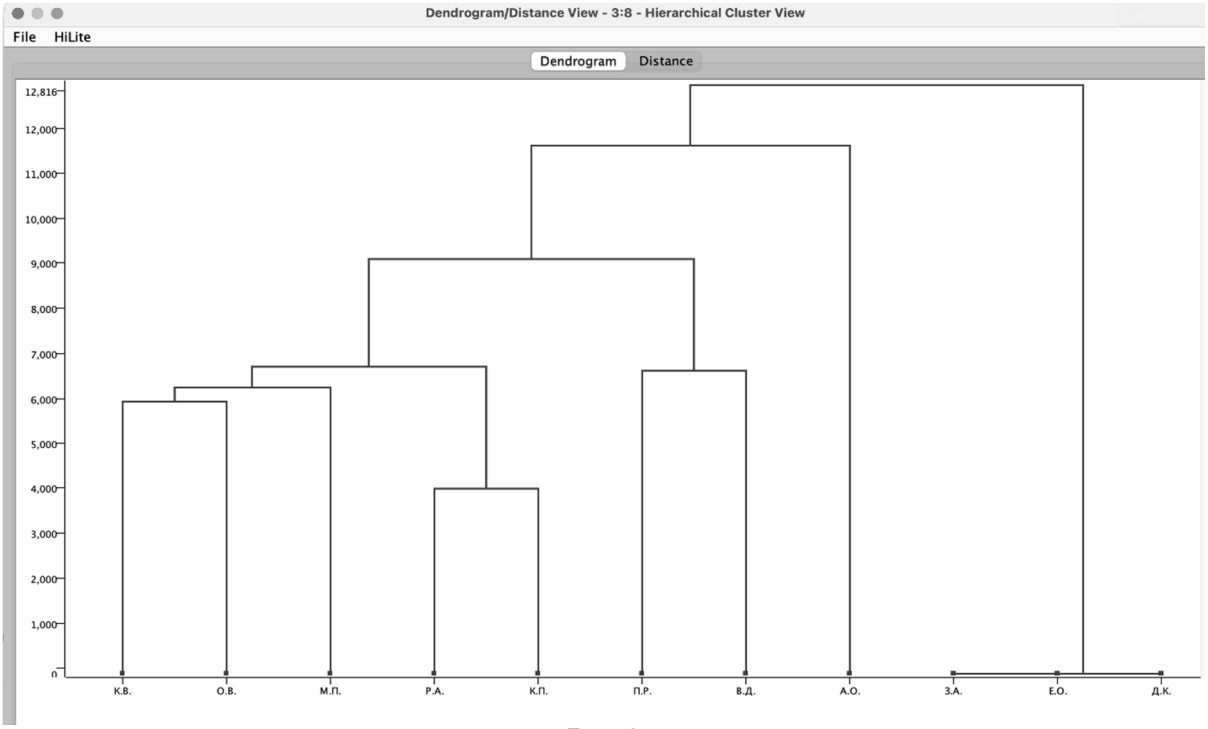

Рис. 2

Тепер чітко видно утворену кластерну структуру – кластери студентів, що мають однакове ставлення до своїх рольових позицій і до студентів інших факультетів. Перша група студентів – Д.К., Е.О., З.А., друга група – К.П. та Р.А., третя група – О.В. та К.В., до якої приєднується М.П., четверта група – В.Д. та П.Р. Далі від усіх знаходиться студент А.О., який не входить у жоден з первинних кластерів, а значить, найбільше відрізняється від усіх інших.

Трактувати це можна наступним чином: наприклад, до першого кластера потрапили студенти Д.К., Е.О., З.А. Ці студенти найпозитивніше ставляться до студентів інших факультетів, і одночасно – не дуже позитивно до майбутніх рольових позицій.

Розглянемо тепер, як можна використовувати спеціалізовані бібліотеки мови python для ієрархічного кластерного аналізу. Для власне кластерного аналізу скористуємося бібліотекою scipy, для отримання діаграми – бібліотекою matplotlib.pyplot, для опрацювання даних – бібліотекою pandas.

У якості навчальних даних для аналізу можна запропонувати таблицю з оцінками студентів, згенерованими випадковим чином.

Таблиця 2

Student	Math	Physics	Informatics	English
Anna	85	78	92	88
Bohdan	60	65	58	70
Olha	90	88	85	80
Petro	50	55	52	60
Iryna	75	80	78	82
Dmytro	65	68	60	65
Kateryna	95	92	96	89
Yulia	70	72	74	77

```
# підключаємо необхідні бібліотеки
from scipy.cluster.hierarchy import linkage, dendrogram
import matplotlib.pyplot as plt
import pandas as pd
# формуємо dataset, включивши в нього тільки числові дані
df=pd.read_csv('students.csv')
inp=df.iloc[0:,1:5]
lb=df['Student'].tolist()
# формуємо легенду для осі X
for i in range(len(lb)):
    lb[i]=str(i)+':'+lb[i]
# застосовуємо метод Ward
Z = linkage(inp, method='ward')
# будуємо дендрограму
plt.figure(figsize=(6, 4))
dendrogram(Z)
plt.title(«Dendrogram»)
plt.xlabel(lb)
plt.show()
```

На дендрограмі (рис. 3) можна побачити, як студенти поєднуються у кластери за успішністю.

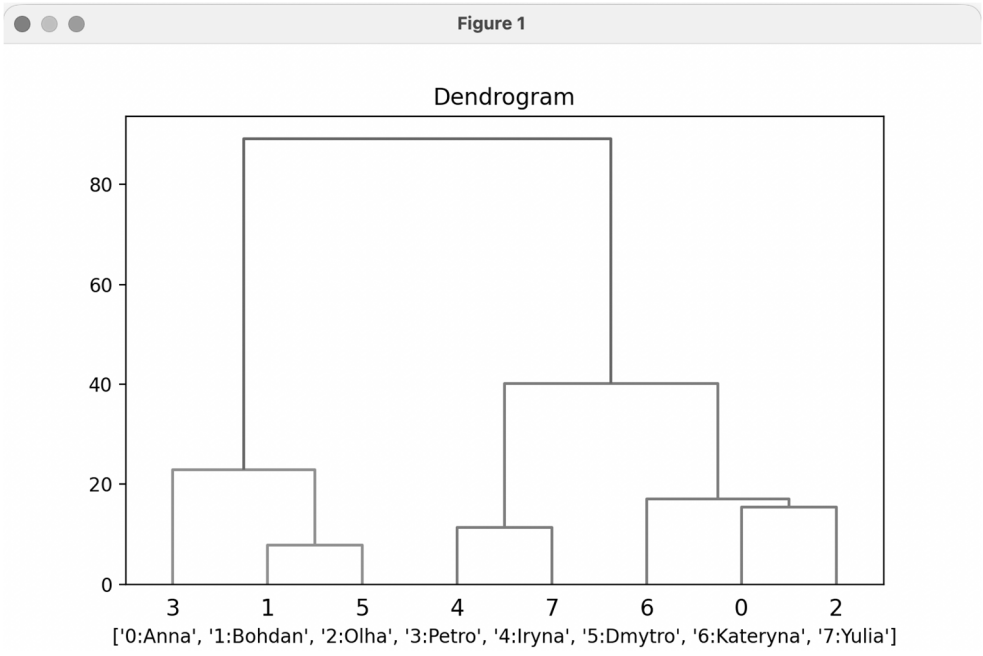

Рис. 3

Перейдемо до методу К-середніх.

Спочатку треба визначитись із набором даних. В бібліотеці Phyton sklearn.datasets є підпрограма make_blobs() генерації багатокласового набору даних, що розподіляє кожен клас на один нормально розподілений кластер точок. Це забезпечує контроль над центрами та стандартними відхиленнями кожного кластера. Цей набір даних використовується для демонстрації кластеризації. Щоб скористатися ним в програмі KNIME, згенеруємо csv-файл за допомогою наступної програми:

```
from sklearn.datasets import make_blobs
import pandas as pd
X, _ = make_blobs(n_samples=300, centers=4, random_state=42)
df=pd.DataFrame(X)
df.to_csv('blobs.csv',index=False)
```

У робочому процесі KNIME розмістимо вузол CSV Reader та налаштуємо його, вказавши шлях до файлу blobs.csv. Приєднаємо вузол k-Means для кластеризації даних методом К-середніх та налаштуємо його, вказавши кількість кластерів 4. Щоб зобразити кожен кластер окремим кольором, скористаємось вузлом Color Manager. Нарешті, для візуалізації результатів аналізу додамо вузол Scatter Plot.
Отримаємо наступний робочий процес (рис. 4) та діаграму (рис. 5), на якій чітко видно 4 кластери точок.

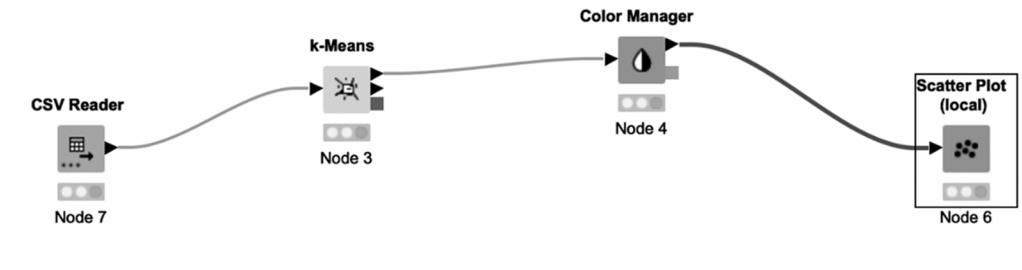

Рис. 4

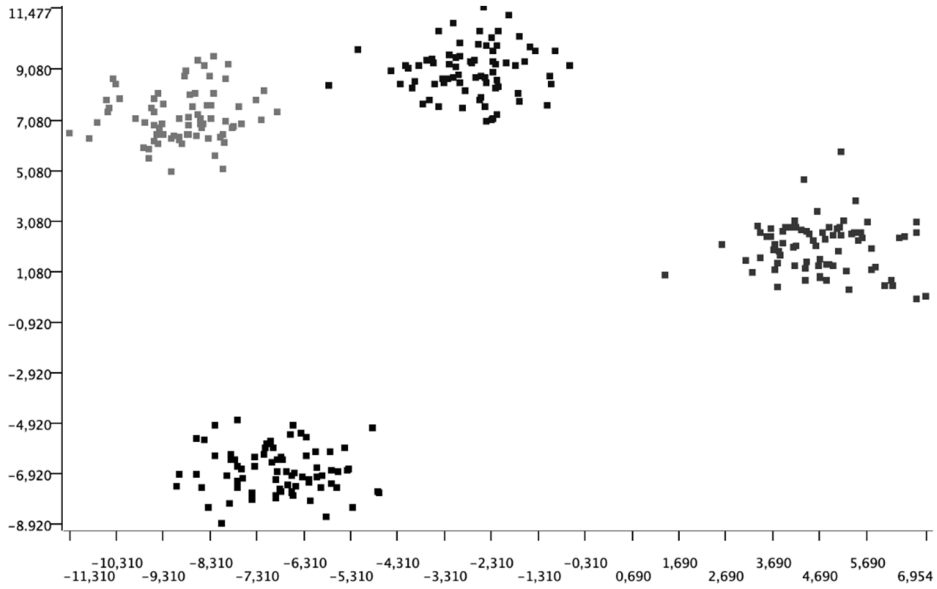

Рис. 5

Щоб провести цей же аналіз засобами бібліотек мови python, можна запропонувати наступну програму

```
# підключаємо необхідні бібліотеки
from sklearn.cluster import KMeans
from sklearn.datasets import make_blobs
import matplotlib.pyplot as plt
# формуємо dataset
```

```
X, _ = make_blobs(n_samples=300, centers=4, random_state=42)
# застосовуємо метод К-середніх
kmeans = KMeans(n_clusters=4)
# будуємо діаграму
labels = kmeans.fit_predict(X)
plt.scatter(X[:, 0], X[:, 1], c=labels)
plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_
[:, 1], marker='X', s=200)
plt.show()
```

Маємо результат, аналогічний попередньому (рис. 6)

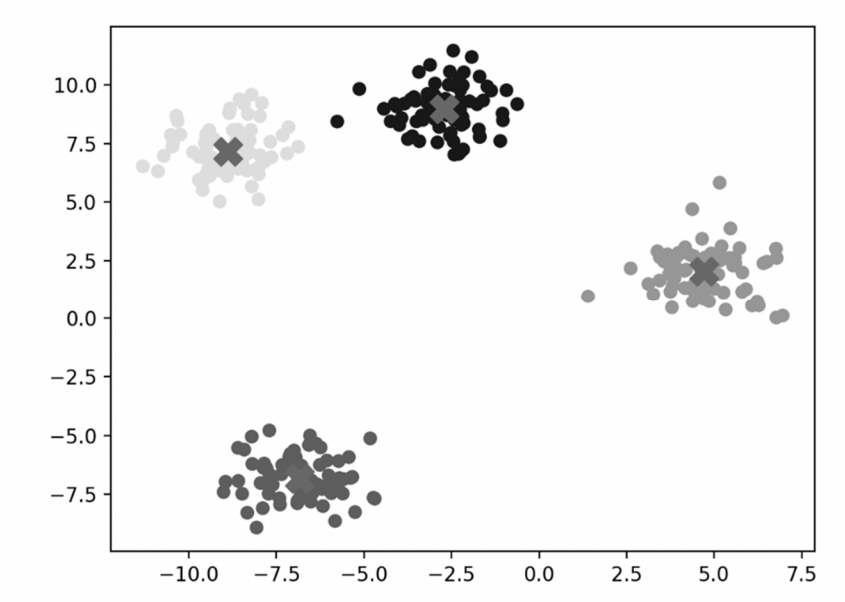

Рис. 6

Розглянемо ще один приклад застосування методу К-середніх. Згенеруємо таблицю з оцінками таким чином, щоб половина студентів мала кращі оцінки з математики та фізики і гірші з історії та англійської мови, а інша половина – навпаки, тобто половина студентів так звані «фізики», а інша – «лірики».

Таблиця 3

Student	Math	Physics	History	English
Anna	85	88	65	63
Bohdan	95	90	58	60
Olha	60	64	85	80
Petro	50	55	90	87
Iryna	62	67	88	94
Dmytro	90	85	60	65
Kateryna	65	62	96	89
Yulia	92	90	62	63

Підключимо цю таблицю до робочого процесу KNIME, налаштовану на метод К-середніх, встановимо кількість кластерів 2. Отримаємо наступний розподіл на два кластери (Рис. 7).

Отже бачимо, що проведення такого кластерного аналізу дозволило чітко відокремити «фізиків» від «ліриків».

Висновки. У роботі розглянуто основні концепції такої технології опрацювання і аналізу великих даних, як кластерний аналіз. Висвітлено сутність та призначення кластерного аналізу, зроблено огляд джерел стосовно цієї теми, відібрано вільно поширювані інструменти та наведено методичні підходи до навчання основ кластерного аналізу майбутніх учителів інформатики і фахівців з комп'ютерних та соціальних наук. Зазначені методичні підходи до навчання сучасних методів і засобів кластерного аналізу спрямовані на формування у майбутніх фахівців спеціальних компетентностей, необхідних для інтелектуального аналізу даних. Таке формування може бути проведене при вивченні дисципліни «Основи штучного інтелекту та інтелектуального аналізу даних», що актуалізує тематику вказаного курсу.

Row ID	S Student	I Math	I Physics	I History	I English	S Cluster
Row0	Anna	85	88	65	63	cluster_1
Row1	Bohdan	95	90	58	60	cluster_1
Row2	Olha	60	64	85	80	cluster_0
Row3	Petro	50	55	90	87	cluster_0
Row4	Iryna	62	67	88	94	cluster_0
Row5	Dmytro	90	85	60	65	cluster_1
Row6	Kateryna	65	62	96	89	cluster_0
Row7	Yulia	92	90	62	63	cluster_1

Рис. 7

References

1. Горошко Ю.В., Цибко Г.Ю., Костюченко А.О. Технології опрацювання великих даних у навчанні інформатичних дисциплін. *Вісник Національного університету «Чернігівський колегіум» імені Т. Г. Шевченка. Вип. 12 (168)*. (Серія: Педагогічні науки). Чернігів : НУЧК, 2021. С. 8-17.
Horoshko Yu.V., Tsybko H.Yu., Kostiuchenko A.O. (2021). Tekhnolohii opratsiuvannia velykykh danykh u navchanni informatychnykh dystsyplin [Big Data processing technologies in Computer Science teaching] Visnyk Natsionalnoho universytetu «Chernihivskyi kolehium» imeni T. H. Shevchenka. Vyp. 12 (168). Chernihiv : NUChK. (Seriia: Pedahohichni nauky). – Bulletin of Taras Shevchenko National University «Chernihiv Colehium». Series: Pedagogical sciences. 12 (168). 8-17. [in Ukrainian].
2. Ваколюк Г. А., Туржанська О. С. Кластерний аналіз як інструмент прикладних досліджень. Зб. наук. праць за матеріалами Всеукр. наук.-практ. конф., 18-19 травня 2017 р. Вінниця : ФОП Рогальська І. О., 2017. С. 228-231.
Vakoliuk H. A., O. S. Turzhanska (2017). Klasternyi analiz iak instrument prykladnykh doslidzhen [Cluster analysis as an applied research tool] Zb. nauk. prats za materialamy Vseukr. nauk.-prakt. konf., 18-19 travnia 2017 r. Vinnytsia : FOP Rohalska I. O. – Collection of scientific papers based on the materials of the All-Ukrainian Scientific and Practical Conference. 228-231. [in Ukrainian].
3. Кунанець Н. Е., Камінський Р. М. Кластерний аналіз як методологічний інструментарій дослідження бібліотек. *Вісник Національного університету «Львівська політехніка». 2014. № 783 : Інформаційні системи та мережі.* С. 435–443.
Kunanets N. E., Kaminskyi R. M. (2014). Klasternyi analiz iak metodolohichnyi instrumentarii doslidzhennia bibliotek [Cluster analysis as a methodological tool for library research] / N. E. Kunanets, Visnyk Natsionalnoho universytetu «Lvivska politekhnika» : Informatsiini systemy ta merezhi. – Bulletin of Lviv Polytechnic National University: Information systems and networks. 783. 435–443. [in Ukrainian].
4. Лахно, В., Волошин, С., Мамченко, С., Кулініч, О., & Касаткін, Д. (2024). Кластерний аналіз для дослідження цифрових слідів студентів закладів освіти. *Електронне фахове наукове видання «Кібербезпека: освіта, наука, техніка», 3(23), 31–41.* https://doi.org/10.28925/2663-4023.2024.23.3141
Lakhno, V., Voloshyn, S., Mamchenko, S., Kulinich, O., & Kasatkin, D. (2024). Klasternyy analiz dlya doslidzhennya tsyfrovykh slidiv studentiv zakladiv osvity [Cluster analysis for researching digital footprints of students in educational institutions] Elektronne fakhove naukove vydannia «Kiberbezpeka: osvita, nauka, tekhnika» – Electronic Professional Scientific Edition «Cybersecurity: Education, Science, Technique». 3(23), 31–41. https://doi.org/10.28925/2663-4023.2024.23.3141 [in Ukrainian].
5. Ahmed, M.; Seraj, R.; Islam, S.M.S. The *k-means* Algorithm: A Comprehensive Survey and Performance Evaluation. *Electronics* 2020, *9*, 1295. https://doi.org/10.3390/electronics9081295
Ahmed, M.; Seraj, R.; Islam, S.M.S. (2020). The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. Electronics. 9. 1295. https://doi.org/10.3390/electronics9081295
6. Климчук В.О. Кластерний аналіз: використання у психологічних дослідженнях. *Практична психологія та соціальна робота. 2006. №4.* С. 30-36.
Klymchuk V.O. (2006). Klasternyi analiz: vykorystannia u psykholohichnykh doslidzhenniakh [Cluster analysis: use in psychological research] Praktychna psykholohiia ta sotsialna robota – Practical psychology and social work. 4. 30-36. [in Ukrainian].
7. Климчук В.О. Математичні методи у психології. Навчальний посібник для студентів психологічних спеціальностей. К.: Освіта України. 2009. 288 с.
Klymchuk V.O. (2009). Matematychni metody u psykholohii. Navchalnyi posibnyk dlia studentiv psykholohichnykh spetsialnostei [Mathematical methods in psychology. A study guide for students of psychology specialties]. K. Osvita Ukrainy. 288. [in Ukrainian].
8. Пістунов І.М., Антонюк О.П., Турчанінова І.Ю. Кластерний аналіз в економіці: навч. посібник. Дніпропетровськ: Національний гірничий університет, 2008. 84 с.
Pistunov I.M., Antoniuk O.P., Turchaninova I.Yu. (2008). Klasternyi analiz v ekonomitsi: Navch. posibnyk [Cluster analysis in economics: a study guide] – Dnipropetrovsk: Natsionalnyi hirnychyi universytet. 84. [in Ukrainian].

Horoshko Yurii

https://orcid.org/0000-0001-9290-7563
Researcher ID GQB-3684-2022
Scopus-Author ID 57952935600

Doctor of Pedagogical Sciences, Professor,
Head of Department of Computer Science and Engineering
Taras Shevchenko National University «Chernihiv Colehium»
(Chernihiv, Ukraine) E-mail: horoshko_y@ukr.net

Tsybko Hanna

https://orcid.org/0000-0002-1861-3003
Researcher ID AAC-6021-2021
Scopus-Author ID 57952055100

Candidate of Pedagogical Sciences, Associate Professor,
Associate Professor of Department of Computer Science and Engineering
Taras Shevchenko National University «Chernihiv Colehium»
(Chernihiv, Ukraine) E-mail: a.tsb@ukr.net

Kostiuchenko Andrii

https://orcid.org/0000-0002-6178-6444
Researcher ID GPX-1175-2022

Candidate of Pedagogical Sciences,
Senior Lecturer of Department of Computer Science and Engineering
Taras Shevchenko National University «Chernihiv Colehium»
(Chernihiv, Ukraine) E-mail: kost_andrey@ukr.net

USING CLUSTER ANALYSIS IN THE INTELLIGENT DATA PROCESSING

The article considers the basic concepts of such a section of big data science (Data Science) as cluster analysis. The theoretical foundations and practical aspects of the application of cluster analysis in various fields are highlighted. A selection of freely distributed software tools for cluster analysis, appropriate for use in the educational process and practical activities, is made. Elements of methodology of teaching the basics of cluster analysis to future computer science teachers, computer science and social sciences specialists are proposed.

***Article's purpose** is to analyze and select didactically appropriate freely distributed tools for execution cluster data analysis, and to develop certain components of methodology of teaching this topic to future computer science teachers, computer and social science specialists.*

***Methodology.** Study and analysis of scientific papers, educational and methodological publications, comparative analysis of software, generalization of the experience of specialists in the field of education, computer and social sciences, modeling and synthesis of components of the teaching methodology, a systematic approach to teaching computer science.*

***Scientific novelty.** Appropriate freely distributed tools for cluster analysis have been selected and certain components of the methodology for teaching future specialists have been developed.*

***Conclusions.** The paper considers the basic concepts of cluster analysis. The essence and purpose of cluster analysis are highlighted, a review of sources on this topic is made, freely distributed tools are selected and methodological approaches to teaching the basics of cluster analysis to future computer science teachers and specialists in computer and social sciences are given. The indicated methodological approaches to teaching modern methods and tools of cluster analysis are aimed at forming in future specialists special competencies necessary for intelligent data analysis. Such formation can be carried out when studying the discipline «Fundamentals of Artificial Intelligence and Intelligent Data Analysis», that actualizes the topic of the specified course.*

***Keywords:** intelligent data analysis, cluster analysis, KNIME, Python.*

Стаття надійшла до редакції 02.12.2025

Рецензент: доктор педагогічних наук, професор ***Торубара О.М.***