Методи кластеризації даних з категорійними ознаками

Виконала:
Мнухіна К. О.

НАУКОВИЙ КЕРІВНИК:
Д.Т.Н., ПРОФЕСОР КАФЕДРИ ММСА ДАНИЛОВ В.Я.

Актуальність задачі

- Визначення основних закономірностей в даних
- Покращення аналізу, завдяки побудові різних моделей для різних кластерів
- Зменшення обсягів вибірки, відбором «типового представника»
- Прогнозування
- Виявлення аномалій

Постановка задачі

- Задача: розбиття групи об’єктів на підгрупи, що містять схожі об’єкти та суттєво відрізняються між собою.
- Мета роботи: використання різних методів кластеризації для даних із категорійними ознаками, порівняння їх, виявлення сильних та слабких сторін.
- Об’єкт дослідження: сучасні алгоритми машинного навчання.
- Предмет дослідження: застосування методів машинного навчання для розв’язання задачі кластеризації категорійних даних.

Етапи розв'язання задачі

- Підготувати дані до аналізу і обробки
- Обрати метод кластеризації
- Обрати значення параметрів
- Побудувати моделі
- Обрати кращу модель кожного алгоритму
- Порівняти отримані результати

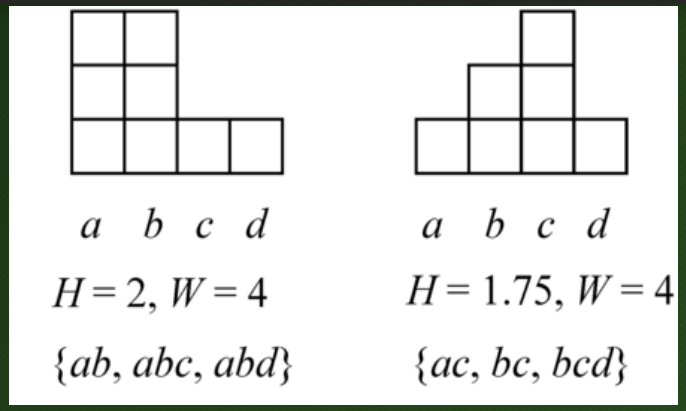
Перетворення категорійних ознак в числові

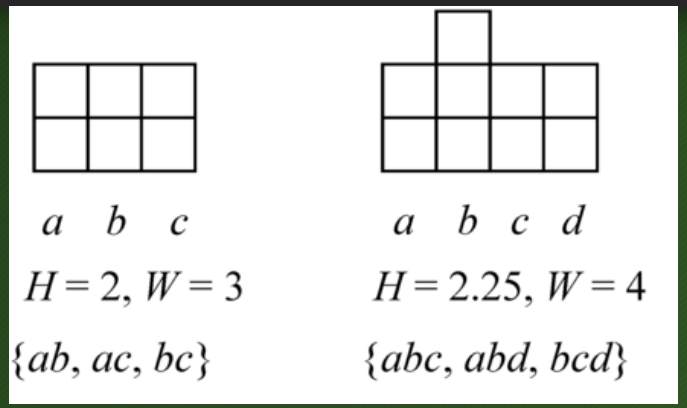
- 1. Label encoding: кожному значенню категоріальної ознаки ставиться у відповідність деяке унікальне число.

- 2. Hashing trick: кожній парі «ознака-значення» ставимо у відповідність значення обраної хеш-функції. Дозволяє працювати із заздалегідь невідомими значеннями деяких ознак.

- 3. One-hot encoding: нехай категоріальна ознака приймає відомі нам значення (всього k). Тоді замість цієї ознаки створюємо k ознак, з яких лише одна приймає значення 1, усі інші – 0.

Метод CLOPE





$H(C_i)$ – висота кластеру C_i,
$W(C_i)$ – ширина кластеру C_i

$$Profit(C) = \frac{\sum_{i=1}^{k} G(C_i) \times |C_i|}{\sum_{i=1}^{k} |C_i|}$$

, де C_i – і-й кластер, $G(C_i) = \frac{H(C_i)}{W(C_i)}$, k – кількість кластерів

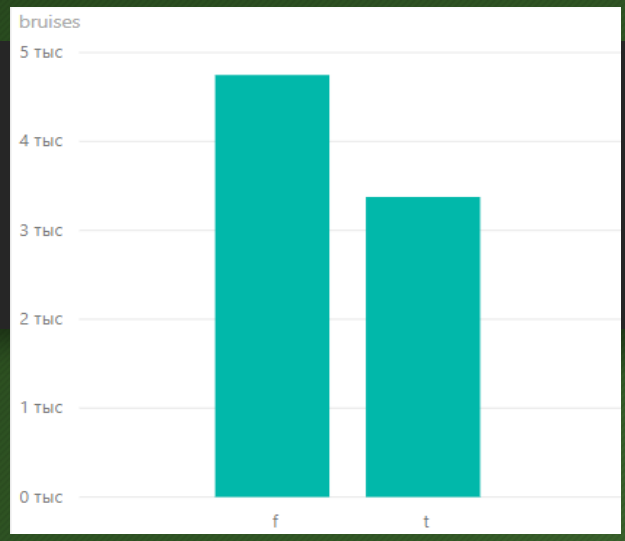
2/4 + 1.75/4 = 0.9375

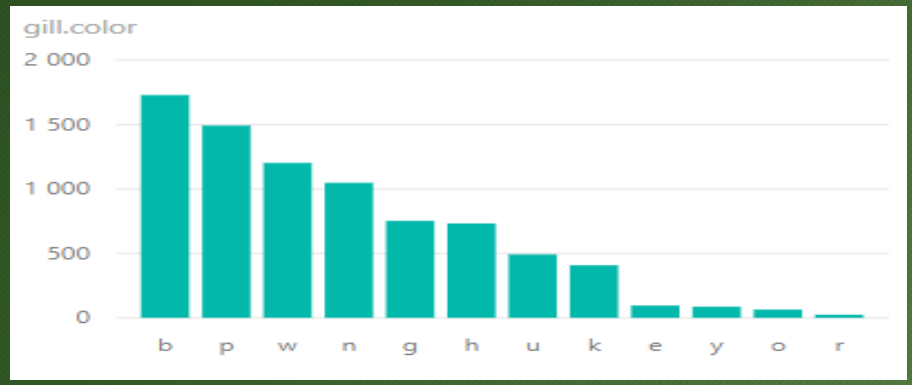
2/3 + 2.25/4 = 1.1229

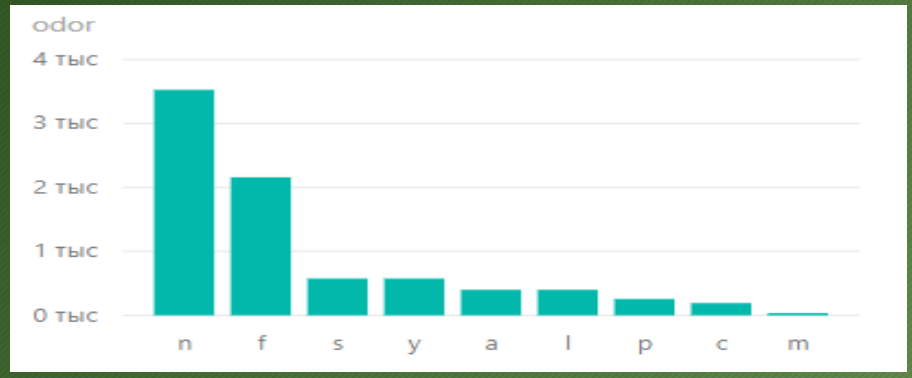
Візуалізація даних

Набір даних mushrooms.csv:
23 категорійні ознаки, 8124 об’єкти

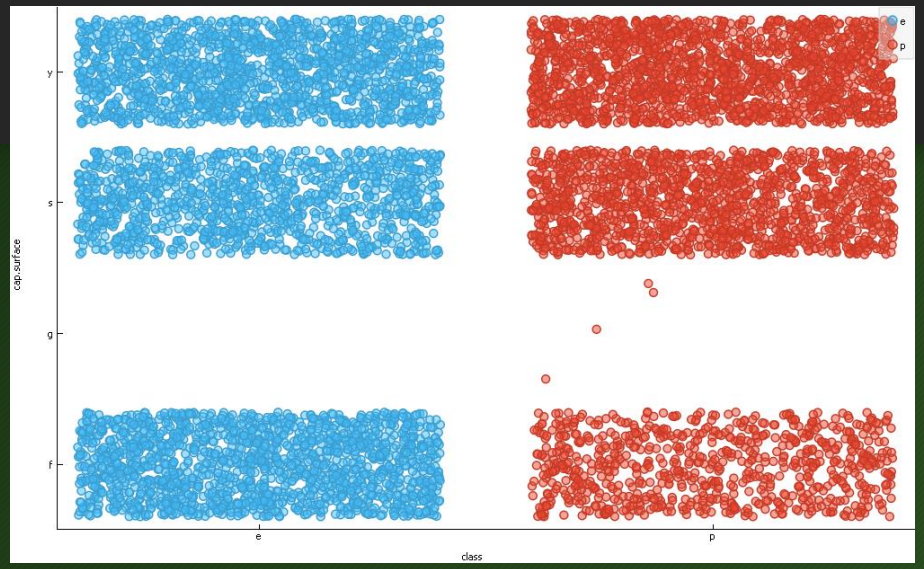
	class	p-sha	cap-color	bruises	odor	l-attachme	gill-spacing	gill-size	veil-type	veil-color	ring-number	population	habit
1	p	x	n	t	p	f	c	n	p	w	o	s	u
2	e	x	y	t	a	f	c	b	p	w	o	n	g
3	e	b	w	t	l	f	c	b	p	w	o	n	m
4	p	x	w	t	p	f	c	n	p	w	o	s	u
5	e	x	g	f	n	f	w	b	p	w	o	a	g
6	e	x	y	t	a	f	c	b	p	w	o	n	g
7	e	b	w	t	a	f	c	b	p	w	o	n	m
8	e	b	w	t	l	f	c	b	p	w	o	s	m
9	p	x	w	t	p	f	c	n	p	w	o	v	g
10	e	b	y	t	a	f	c	b	p	w	o	s	m
11	e	x	y	t	l	f	c	b	p	w	o	n	g
12	e	x	y	t	a	f	c	b	p	w	o	s	m
13	e	b	y	t	a	f	c	b	p	w	o	s	g
14	p	x	w	t	p	f	c	n	p	w	o	v	u
15	e	x	n	f	n	f	w	b	p	w	o	a	g
16	e	s	g	f	n	f	c	n	p	w	o	y	u
17	e	f	w	f	n	f	w	b	p	w	o	a	g
18	p	x	n	t	p	f	c	n	p	w	o	s	g
19	p	x	w	t	p	f	c	n	p	w	o	s	u
20	p	x	n	t	p	f	c	n	p	w	o	s	u
21	e	b	y	t	a	f	c	b	p	w	o	s	m
22	p	x	n	t	p	f	c	n	p	w	o	v	g
23	e	b	y	t	l	f	c	b	p	w	o	s	m
24	e	b	w	t	a	f	c	b	p	w	o	n	m
25	e	b	w	t	l	f	c	b	p	w	o	s	m

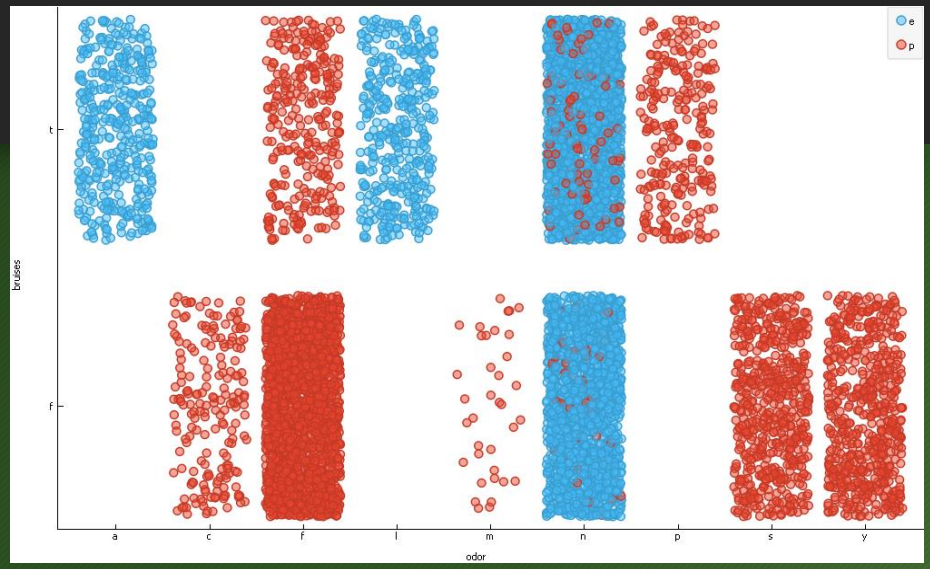


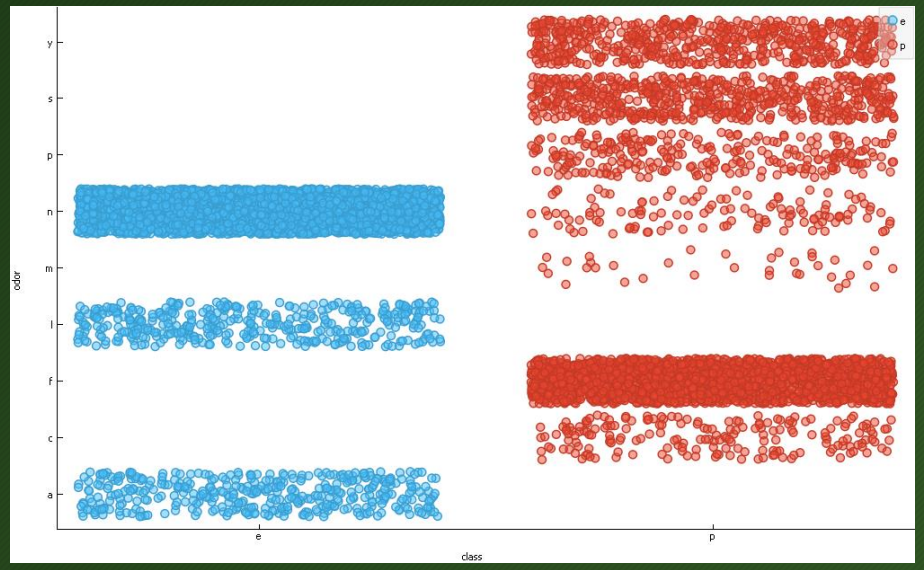


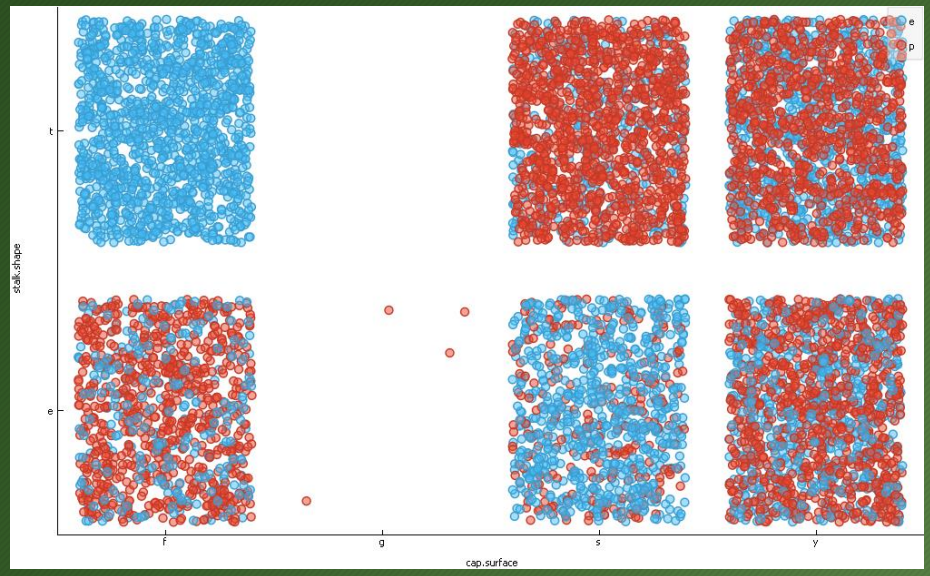


Візуалізація даних









Перетворення категорійних ознак в числові

cap-shape	cap-surface	cap-color	bruises
x	s	n	t
x	s	y	t
b	s	w	t
x	y	w	t
x	s	g	f
x	y	y	t
b	s	w	t

Початкові дані: 4 ознаки

Перетворені дані: 20 ознак

cap.shape=b	cap.shape=c	cap.shape=f	cap.shape=k	cap.shape=s	cap.shape=x	cap.surface=f	cap.surface=g	cap.surface=s	cap.surface=y	cap.color=b	cap.color=c	cap.color=e	cap.color=g	cap.color=n	cap.color=p	cap.color=r	cap.color=u	cap.color=w	cap.color=y
0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1
1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1
1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0

Методи кластеризації

- Ієрархічні методи

- K-means
- PAM (Partitioning Around Medoid)
- CLARA (Clustering LARge Applications)

- DBSCAN (Density Based Spatial Clustering of Applications with Noise)

- CLOPE

Порівняння

Виконується кластеризація набору даних різними методами. Для їх порівняння розглядаємо наступні характеристики:

- Час виконання алгоритму
- Кількість параметрів
- Визначена кількість кластерів
- Якість кластеризації

Порівняння

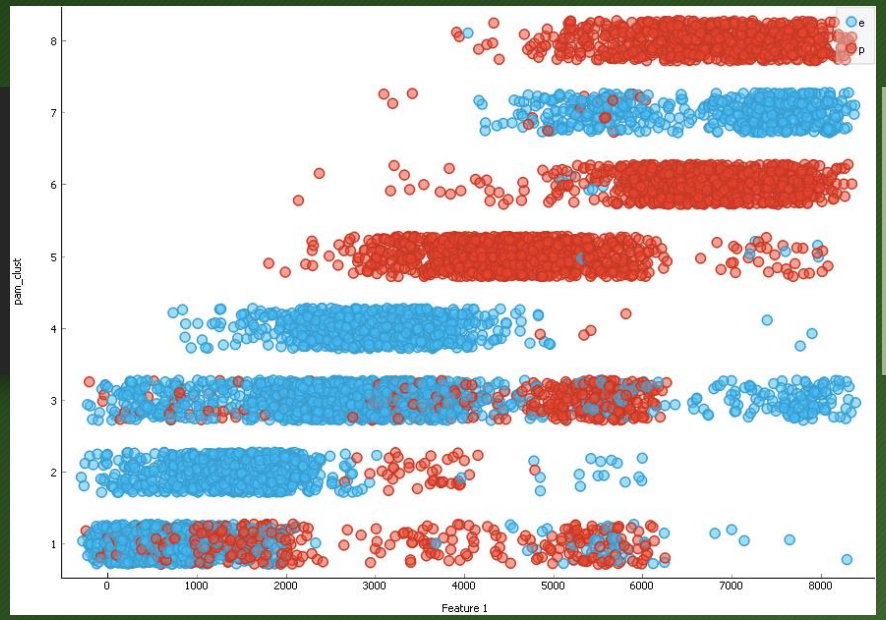
	KMeans	PAM	CLARA	DBSCAN	CLOPE
Час роботи	0.22 с	1.89 хв	0.19 с	13.18 с	0. 24 с
К-сть кластерів	6	8	10	10	11
К-сть параметрів	1	1	1	2	1
Чисті/брудні кластери	0 / 6	5 / 3 (з 5 чистих 4 з домішками)	7 / 3 (з 7 чистих 1 з домішками)	8 / 2 (але з 8 чистих багато «малочисельних кластерів)	8 / 3

Візуалізація результатів

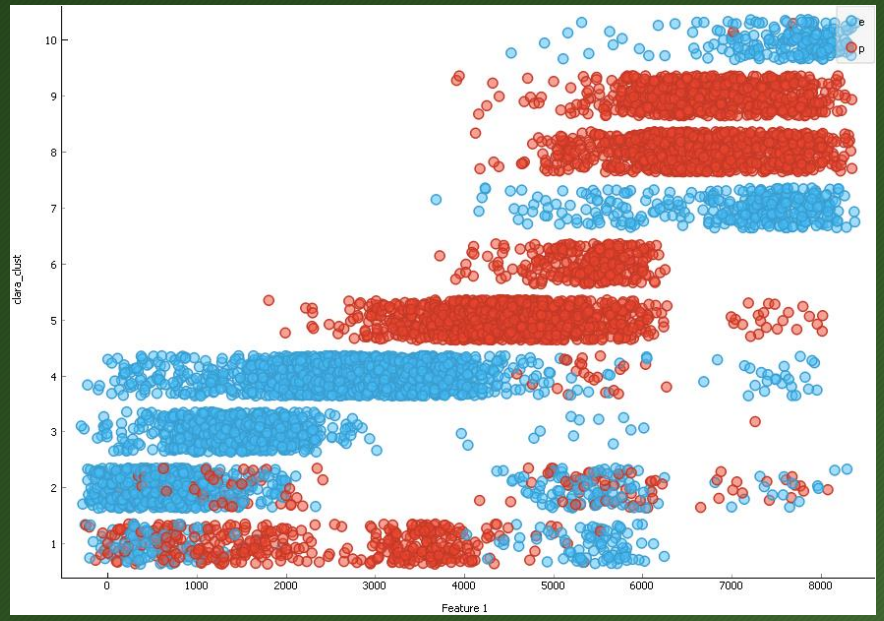
CLOPE

```
cluster # 0          cluster # 6
1272 eatable         0 eatable
272 poisoned         36 poisoned

cluster # 1          cluster # 7
2024 eatable         144 eatable
228 poisoned         0 poisoned

cluster # 2          cluster # 8
192 eatable          288 eatable
0 poisoned           0 poisoned

cluster # 3          cluster # 9
0 eatable            0 eatable
1296 poisoned        36 poisoned

cluster # 4          cluster # 10
0 eatable            48 eatable
288 poisoned         0 poisoned

cluster # 5          created 11 clusters
240 eatable          4208 eatable
1760 poisoned        3916 poisoned
```

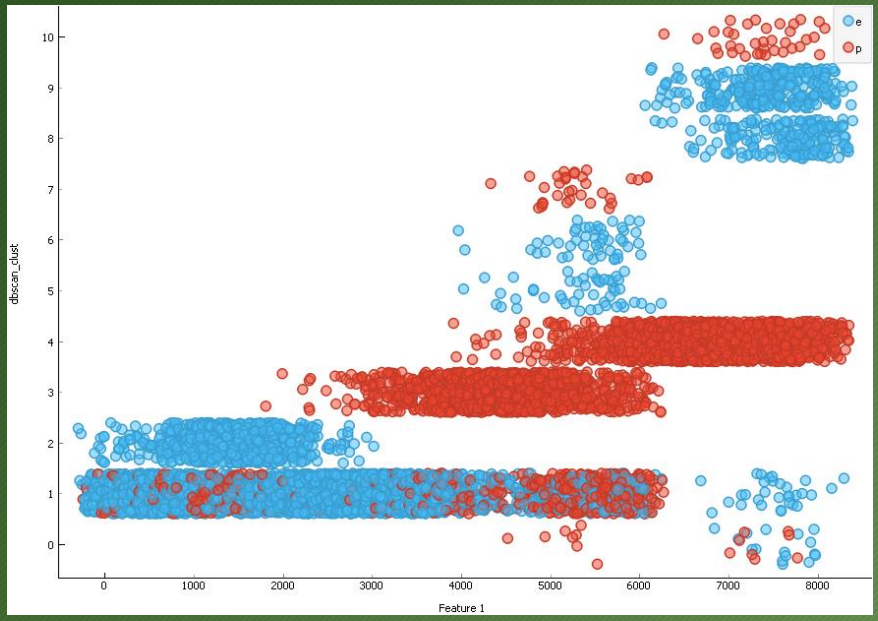
PAM



CLARA



DBSCAN



Порівняльна таблиця

	Переваги	**Недоліки**
CLOPE	Робота із категорійними ознаками, швидка обробка великих БД, малі витрати оперативної пам’яті.	Неочевидний вплив коефіцієнта відштовхування на кількість кластерів
CLARA	Швидкість обробки даних	Необхідність задання к-сті кластерів
DBSCAN	Швидкість обробки даних	Чутливість до викидів
PAM	Простота алгоритму, низька чутливість до викидів	Великі витрати часу, необхідність задавати кількість кластерів
KMeans	Простота і зрозумілість алгоритму	Чутливість до викидів, необхідність задання к-сті кластерів, залежність від початкового стану

Кращий алгоритм – CLOPE

- Працює із категорійними даними
- Регулюється одним параметром -- коефіцієнтом відштовхування
- Автоматично підбирає кількість кластерів
- Висока швидкість роботи
- Малі витрати пам’яті
- Інтуїтивно зрозумілий алгоритм

Висновки

- Проаналізовано ефективність методів кластеризації числових даних для перетворених даних із категорійними ознаками;
- Порівняно різні методи обробки даних, різні метрики та алгоритми кластеризації;
- Реалізовано метод кластеризації CLOPE, що працює безпосередньо із категорійними ознаками, досліджено його ефективність.

Дякую за увагу!