

# Аналіз даних з використанням мови Python

# Оцінювання

$$R_D = 2 * R_{\text{мкр}} + 8 * R_{\text{л}}$$

$R_D$  - загальний рейтинг за дисципліну

$R_{\text{мкр}}$  - бали за модульні контрольні роботи (максимум 14)

$R_{\text{л}}$  - бали за лабораторні роботи (максимум 9)

Залік

# Структура курсу

- Основні поняття обробки даних
- Методи та алгоритми статистичної обробки даних
- Структури даних
- Методи візуалізації та групування даних
- Методи та алгоритми попередньої обробки даних
- Методи та алгоритми аналізу даних
- Основи машинного навчання.

# Література

1. Марченко О.О., Россада Т.В. Актуальні проблеми Data Mining: Навчальний посібник для студентів факультету комп'ютерних наук та кібернетики. – Київ. – 2017. – 150 с.
2. Ланде Д.В., Субач І.Ю., Бояринова Ю.Є. Основи теорії і практики інтелектуального аналізу даних у сфері кібербезпеки: навчальний посібник. – К.: ІСЗЗІ КПІ ім. Ігоря Сікорського», 2018. — 297 с.
3. Бахрушин В.Є. Методи аналізу даних : навчальний посібник для студентів / В.Є. Бахрушин. – Запоріжжя : КПУ, 2011. – 268 с.
4. Wes McKinney. Python for Data Analysis\_ Data Wrangling with Pandas, NumPy, and IPython. – O'Reilly Media, 2017. – 482 p.
5. Gayathri Rajagopalan. A Python Data Analyst's Toolkit. — apress, 2021. — 409 p.
6. Alex Campbell. Data Visualization Guide. — 2021. — 113 p.

# Основні поняття обробки даних

# Дані

Кожного дня в світі генеруються та зберігаються величезні об'єми даних.

По суті, все, що записується, зберігається, є даними.

Дані представляють собою факти, текст, графіки, картинки, звуки, аналогові або цифрові відео-сегменти.

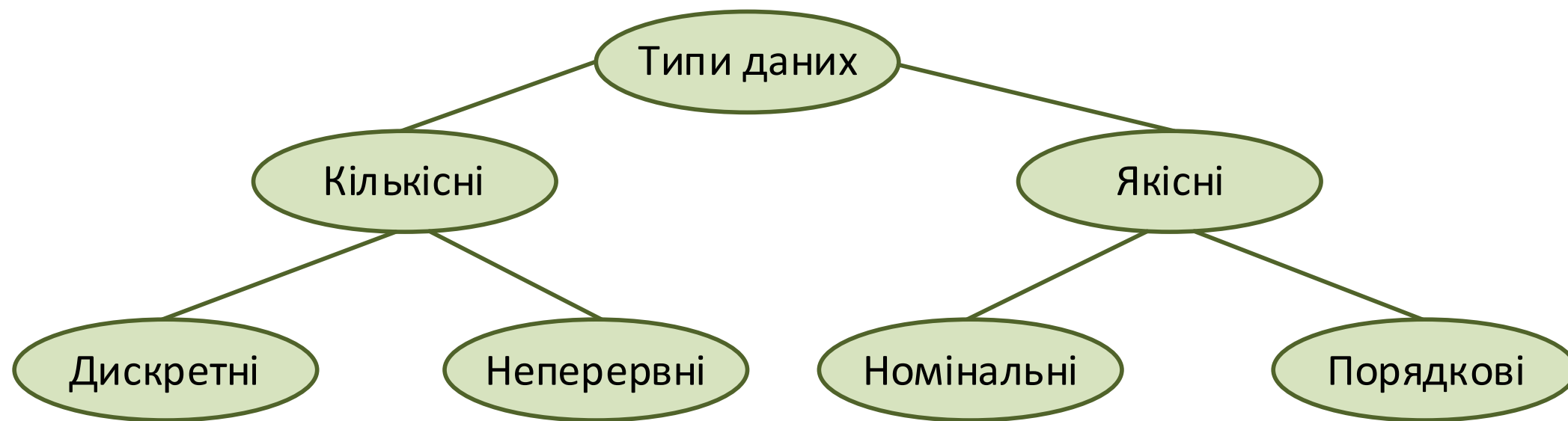
Іншими словами, дані — це необроблений матеріал, що використовується споживачами для формування інформації на основі даних.

# Дані

Дані бувають різних типів:

- Кількісні
  - Дискретні – мають обмежену кількість значень, наприклад, кількість студентів в групах
  - Неперервні – мають необмежену кількість значень, наприклад, вага
- Якісні (категоріальні)
  - Номінальні – не мають ранжування, наприклад, вид тварини
  - Порядкові – мають ранжування, наприклад, «малий», «середній», «великий»

# Дані





# Дані

Також дані можуть бути неструктуровані (наприклад, текст), напівструктуровані (наприклад, лог) і структуровані.

Структуровані дані часто представляють у вигляді таблиць, таблиці складаються з атрибутів (по вертикалі) та об'єктів (по горизонталі).

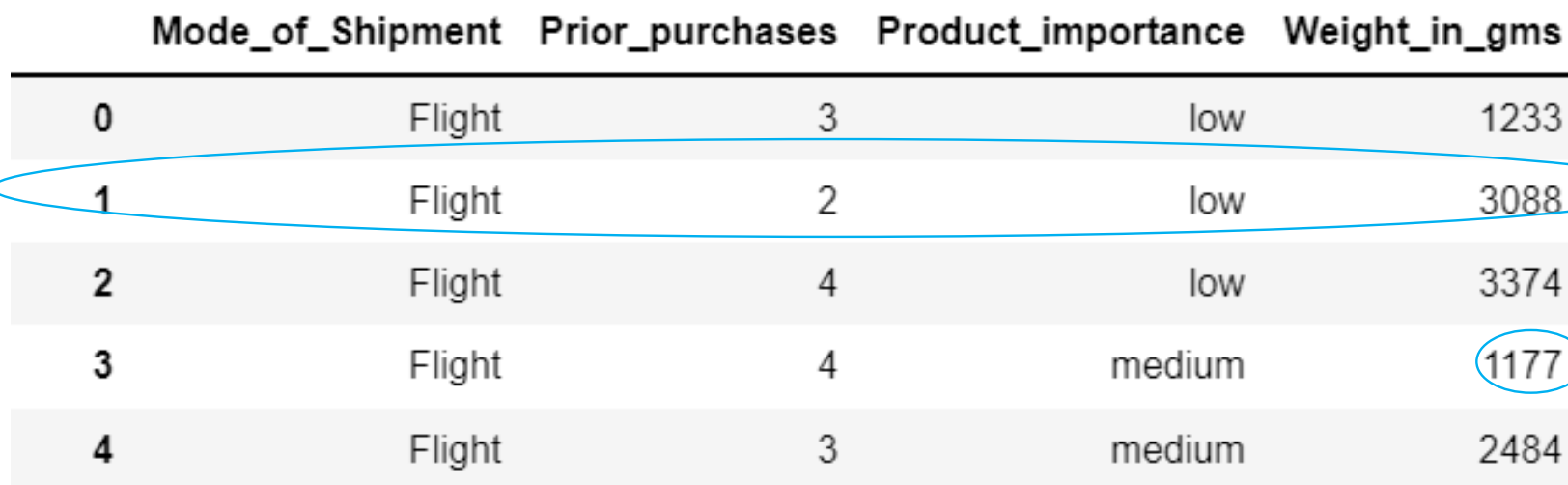
Атрибут (ознака, характеристика) – властивість, що характеризує об'єкт.

Атрибут – властивість або характеристика, загальна для всіх досліджуваних об'єктів, прояв якої може змінюватись від об'єкта до об'єкта.

# Дані

## Набір даних

Ознаки (атрибути)



	Mode_of_Shipment	Prior_purchases	Product_importance	Weight_in_gms
0	Flight	3	low	1233
1	Flight	2	low	3088
2	Flight	4	low	3374
3	Flight	4	medium	1177
4	Flight	3	medium	2484

Об'єкт

Значення ознаки

# Дані

Набір даних складається з даних, що відносяться до сукупності об'єктів, причому кожен об'єкт описаний в термінах набору атрибутів.

Зазвичай виділяють наступні типів атрибутів:

- Номінальні
- Порядкові
- Дискретні
- Неперервні

# Дані

	Номінальна	Дискретна	Порядкова	Неперервна
	Mode_of_Shipment	Prior_purchases	Product_importance	Weight_in_gms
0	Flight	3	low	1233
1	Flight	2	low	3088
2	Flight	4	low	3374
3	Flight	4	medium	1177
4	Flight	3	medium	2484

# Дані

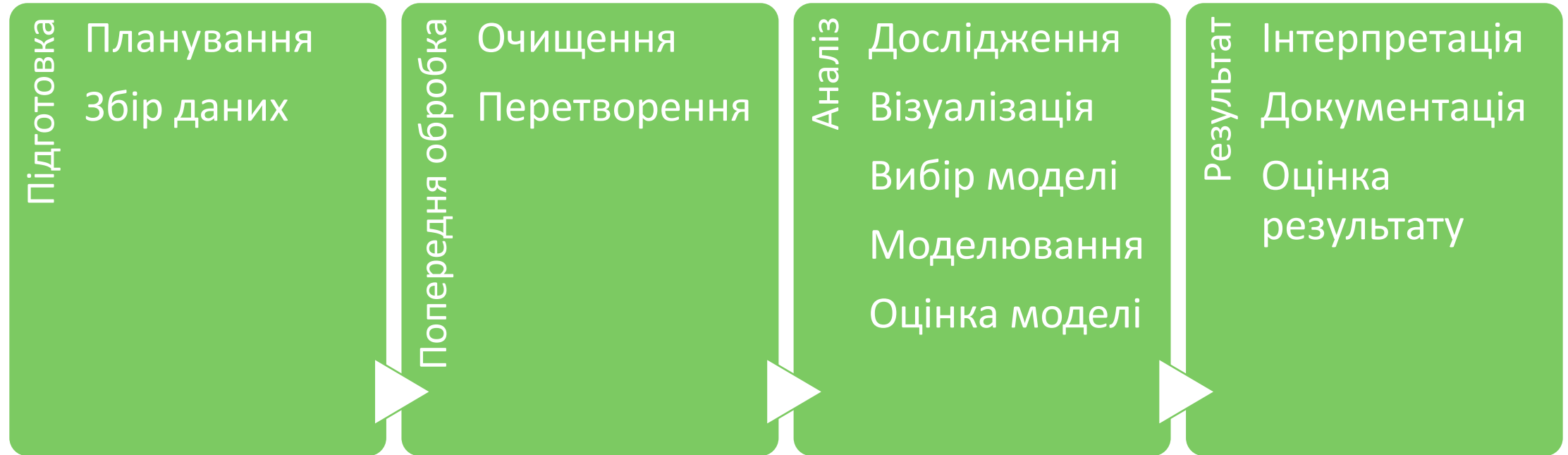
Тип атрибута впливає на методи аналізу та розуміння даних. Ці методи включають в себе як основну статистику, яку можна використовувати для опису розподілу значень атрибута, так і більш складні алгоритми, які застосовуються для виявлення закономірностей відносин між атрибутами.

# Аналіз даних

Аналіз даних – це діяльність з перевірки, попередньої обробки, дослідження, опису та візуалізації даного набору даних. Основною метою процесу аналізу даних є виявлення необхідної інформації для прийняття рішення. Аналіз даних пропонує набір підходів, інструментів і методів, усі з яких можна застосувати до різних галузей, таких як бізнес, соціальні науки та фундаментальні науки.

Аналіз даних – це процес отримання інформація з даних шляхом створення моделей і застосування математичного апарату для пошуку закономірностей.

# Етапи процесу аналізу даних



# Підготовка

## Планування:

- Формулювання задачі
- Вибір даних та їх джерел
- Корисність результатів
- Критерії оцінки результатів
- Способи представлення результатів

## Збір даних



# Попередня обробка даних

## Очищення:

- Помилки при введенні, неможливі значення
- Відсутні значення
- Відхилення
- Різні одиниці вимірювання даних

## Перетворення:

- Агрегування даних
- Екстраполяція
- Вибір похідних атрибутів
- Створення додаткових змінних
- Скорочення змінних

# Аналіз даних

## Статистичний аналіз даних:

- Дескриптивна статистика
- Перевірка гіпотез
- Кореляційний аналіз
- Регресійний аналіз

## Машинне навчання:

- Класифікація
- Кластеризація
- Зниження розмірності
- Глибоке навчання

## Візуалізація

# Аналіз даних

Класифікація:

— віднесення об'єктів (спостережень, подій) до одного з класів;

Регресія (в тому числі задачі прогнозування):

— встановлення залежності неперервних вихідних від вхідних змінних;

Кластеризація:

— групування об'єктів (спостережень, подій) на основі даних (властивостей), що описують сутність цих об'єктів, у кластери. Об'єкти всередині кластера повинні бути «близькими» один на одного в сенсі відстані і відрізнятися від об'єктів, що увійшли в інші кластери. Чим більше ближчі об'єкти всередині кластера і чим більше відмінностей між кластерами, тим точніша кластеризація.

# Аналіз даних

Не існує універсальних методів аналізу або алгоритмів, що придатні для обробки будь-яких об'ємів інформації. Методи аналізу даних істотно відрізняються один від одного по продуктивності, якості результатів, зручності застосування і вимогам до даних.

# Аналіз даних

Серед основних властивостей і характеристик методів аналізу можна виділити наступні:

- Точність
- Масштабованість
- Інтерпретованість
- Можливість перевірки
- Трудомісткість
- Гнучкість
- Швидкість

# Бібліотеки Python для роботи з даними



# NumPy

**Numerical Python** є основною бібліотекою для наукових обчислень.

- Об'єкт багатовимірного масиву `ndarray`
- Функції для роботи з елементами одного масиву
- Функції для математичних операцій з декількома масивами
- Засоби читання та запису наборів даних, представлених у вигляді масивів
- Операції лінійної алгебри

# SciPy

**SciPy** - набір пакетів, призначених для вирішення різних стандартних обчислювальних задач.

## **scipy.stats**

- Стандартні розподіли
- Статистичні критерії
- Дескриптивна статистика



# matplotlib

Бібліотека **matplotlib** - інструмент для створення графіків і інших способів візуалізації двовимірних даних.

- Гістограми
- Стовпчасті діаграми
- Секторні діаграми
- Лінійні діаграми
- Діаграми розсіювання

**Seaborn** базується на matplotlib і спрощує створення графіків.

# pandas

Бібліотека **pandas** надає структури даних і функції, покликані зробити роботу зі структурованими даними простою та швидкою.

- Дескриптивна статистика
- Обробка відсутніх даних
- Комбінування, злиття наборів даних
- Перетворення даних
- Агрегування даних та групові операції
- Робота з часовими рядами

# scikit-learn

Бібліотека **scikit-learn** має множину інструментів для статистичного моделювання та машинного навчання.

- Класифікація
- Кластеризація
- Регресія
- Зменшення розмірності