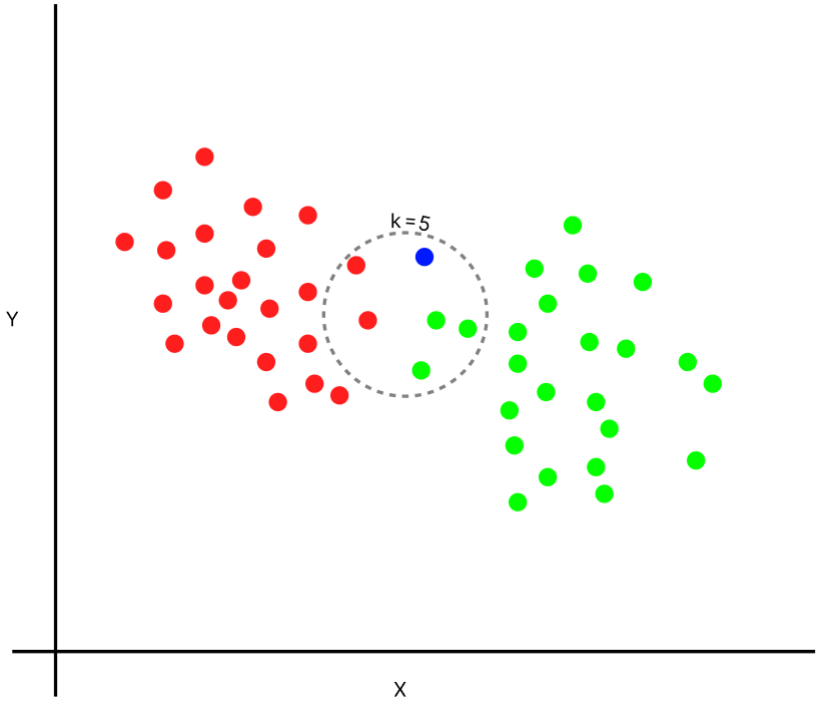
# Типові задачі та моделі машинного навчання

1. KNN
2. Регресійні моделі (лінійна регресія, SVR, TreeBased)
3. Класифікаційні моделі (Логістична регресія, SVC, TreeBased)
4. Ансамблі дерев (bagging, boosting)
5. Кластеризація та зменшення розмірності
6. Deep Learning

# KNN




**Приклад використання**

```
from sklearn.neighbors import NearestNeighbors
import numpy as np

X = np.array(
      [[-1, -1], [-2, -1], [-3, -2],
       [1, 1], [2, 1], [3, 2]])

nbrs = NearestNeighbors(n_neighbors=2,
      algorithm='ball_tree').fit(X)

distances, indices = nbrs.kneighbors(X)
```

**indices** *array([[0, 1], [1, 0], [2, 1], [3, 4], [4, 3], [5, 4]])*

**distances** *array([[0. , 1. ], [0. , 1. ],*
*[0. , 1.41421356], [0. , 1. ],*
*[0. , 1. ], [0. , 1.41421356]])*

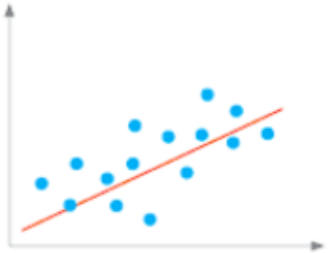
# Регресійні моделі (лінійна регресія), SVR, TreeBased)

**Регресія – це, по суті, спосіб передбачення майбутніх чисел на основі минулих чисел. Вона допомагає прогнозувати такі показники, як продажі, ціни, робити рекомендації, використовуючи те, що було раніше.**

## Лінійна регресія

**ВИКОРИСТАННЯ**: Найкраще підходить для простих передбачень із лінійними залежностями. Наприклад, прогнозування цін на будинки на основі їхнього розміру: більші будинки зазвичай коштують дорожче.

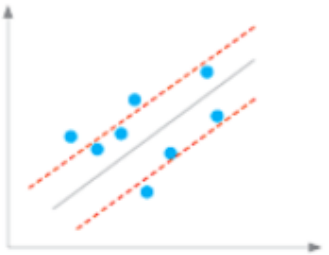
**КОНЦЕПЦІЯ**: Уявіть, що ви проводите пряму лінію через ряд точок на графіку так, щоб вона проходила якомога ближче до всіх точок.



## Регресія опорних векторів

**ВИКОРИСТАННЯ**: Використовується для більш складних залежностей, але при цьому не чутлива до аномалій. Наприклад, прогнозування цін на акції, які можуть підійматися і опускатися в складних закономірностях.

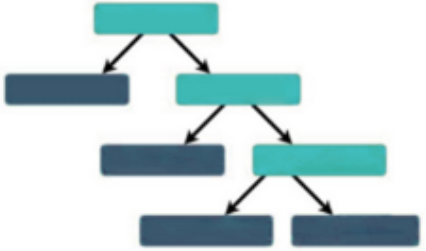
**КОНЦЕПЦІЯ**: Намагайтеся знайти таку криву, яка найкраще підходить через наібільшу кількість точок даних.



## Регресійне дерево

**ВИКОРИСТАННЯ**: Підходить для випадків, коли на певні чинники впливають на результат по різному в залежності від їх значень, наприклад, прогнозування витрат покупця на основі віку, доходу та історії покупок.

**КОНЦЕПЦІЯ**: Цей метод схожий на створення блок-схеми, що розділяє дані на різні гілки, щоб зробити прогноз на кожному кр



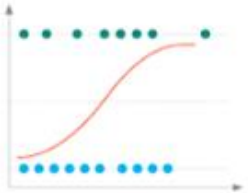
# Класифікаційні моделі (Логістична регресія, SVC, TreeBased)

**Методи класифікації допомагають передбачити, до якої категорії чи групи належить щось на основі минулих даних. Ці методи можуть допомогти визначити, чи є електронний лист спамом, або передбачити, чи має банк схвалити кредитну заявку.**

## Логістична регресія

**ВИКОРИСТАННЯ**: Найкраще підходить для прийняття простих рішень з відповіддю «так» або «ні». Наприклад, визначення того, купить клієнт товар чи ні.

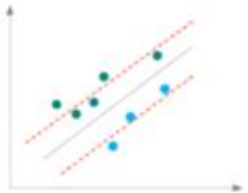
**КОНЦЕПЦІЯ**: Цей метод обчислює імовірність кожної категорії і вибирає ту, яка найбільш вірогідна. Уявіть, що ви малюєте лінію, яка розділяє «так» з одного боку і «ні» з іншого.



## Класифікація опорних векторів

**ВИКОРИСТАННЯ**: Використовується, коли рішення не з простих, а категорії складно розділити простою лінією. Наприклад, у випадку виокремлення різних типів клієнтів на основі їхніх купівельних звичок.

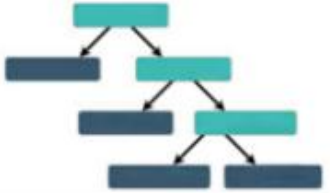
**КОНЦЕПЦІЯ**: Цей метод намагається знайти найкращу межу, що розділяє різні категорії, навіть у складних сценаріях. Уявіть, що ви малюєте криву, а не пряму лінію між групами.



## Класифікаційне дерево

**ВИКОРИСТАННЯ**: Метод корисний для прийняття рішень, які залежать від декількох чинників, наприклад, передбачення ймовірності відтоку клієнтів на основі численних атрибутів, як-от вік, історія транзакцій та використання послуг.

**КОНЦЕПЦІЯ**: Цей метод створює дерево рішень, у якому кожна гілка – це шлях прийняття рішення, що призводить до різних класифікацій на основі відповідей на кожному кроці.



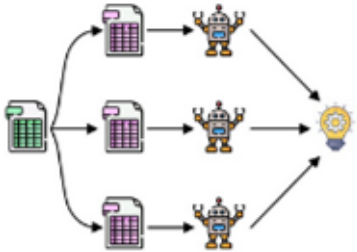
# Ансамблі дерев (bagging, boosting)

**Ансамблеві методи об'єднують кілька моделей для покращення прогнозів. Це наче запитувати групу експертів, а не одного з них, щоб отримати надійнішу відповідь.**

## Беггінг

**ВИКОРИСТАННЯ**: Метод допомагає зменшити кількість помилок завдяки використанню кількох моделей та усередненню їхніх прогнозів. Наприклад, метод випадкового лісу використовує багато дерев рішень, щоб краще прогнозувати дані, наприклад, про те, чи варто схвалювати кредит.

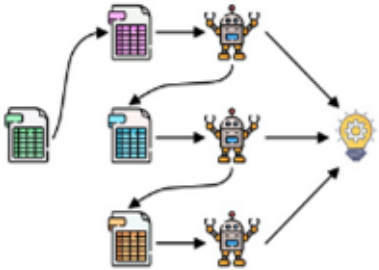
**КОНЦЕПЦІЯ**: Уявіть, що кілька експертів паралельно приймають незалежні рішення, а тоді усереднюють свої висновки, щоб прийняти остаточне рішення.



## Бустінг

**ВИКОРИСТАННЯ**: Цей метод зосереджений на послідовному навчанні на помилках попередніх моделей для підвищення точності. Він корисний в таких випадках, як прогнозування відтоку клієнтів, де точність грає вирішальну роль.

**КОНЦЕПЦІЯ**: Кожен новий експерт вчиться на помилках попереднього, послідовно вдосконалюючи процес прийняття рішень крок за кроком.



# Кластеризація та зменшення розмірності

## Кластеризація

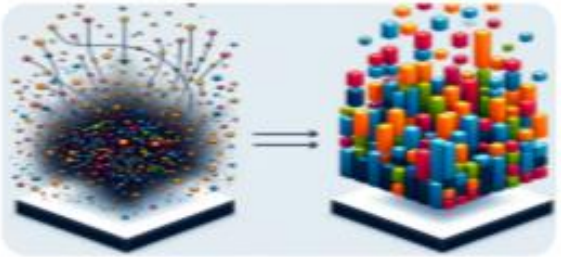
**ВИКОРИСТАННЯ**: Ідеально підходить для сегментації ринку, коли компанії класифікують клієнтів на основі спільних характеристик, як-от купівельні звички, без попередньої інформації про групи.

**КОНЦЕПЦІЯ**: Метод групує схожі дані. Мета – мінімізувати відмінності всередині кожної групи та максимізувати відмінності між різними групами.

## Зниження розмірності

**ВИКОРИСТАННЯ**: Цей метод корисний для аналітики клієнтів, де зменшення кількості змінних полегшує аналіз та візуалізацію складних наборів даних. Наприклад, спрощення даних опитувань клієнтів для кращого розуміння ключових чинників, що впливають на задоволеність.

**КОНЦЕПЦІЯ**: Метод спрощує великі набори даних, зменшуючи кількість змінних, що розглядаються, але зберігає основну інформацію, яка грає найбільшу роль у варіації даних.



# Deep Learning

## Глибоке навчання

**Типи моделей** – Глибокі (складні) Нейронні мережі

**Проектування ознак** – Моделі вивчають та знаходять необхідні ознаки самостійно із "сирих" даних

**Формат даних** – Зазвичай дані у форматі, яким користується людина (зображення, текст), іноді складні табличні (панельні) дані.

**Вимоги до даних та потужностей** – Вимагають більше обчислювальних потужностей та багато даних для навчання

**Точність** – Зазвичай більш точні для складних задач

## Машинне навчання

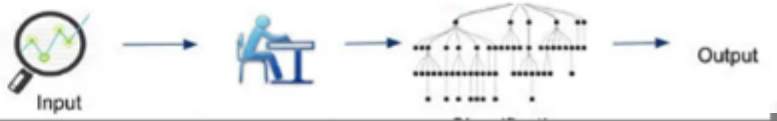
**Типи моделей** – Різні статистичні моделі, прості нейронні мережі

**Проектування ознак** – Важливий крок в побудові моделі, виконує експерт. Дуже впливає на точність.

**Формат даних** – Зазвичай табличні дані, іноді прості дані у природньому форматі (що несуть експерту непрості страждання)

**Вимоги до даних та потужностей** – Зазвичай додсатньо середнього по потужності комп'ютера, можуть дати хороші результати навіть коли даних мало

**Точність** – Можуть бути дуже точними, особливо для простих задач



Питання?