Unsupervised Learning

Mhd. Arsya Fikri | 191402066

Apa yang akan kita bahas?

What is Unsupervised Learning?

Principal Component Analysis

Factor Analysis

Clustering

.

.

01

Unsupervised Learning

What is it?

What is Unsupervised Learning?

Unsupervised ML	Supervised ML
Does not have accuracy metrics	Have accuracy metrics
Does not have labeled data	Can be evaluated with a test set that has labeled target data
No ‘absolute source of truth’	There is an ‘absolute source of truth’ to compare prediction

What is Unsupervised Learning?

Fraud Detection Case 1

You are a Data Scientist tasked to design a Machine Learning model that can identify **FRAUDULENT PAYMENTS**.

You are given:

- A dataset consisting of 1000 non-fraud payments and 100 fraud payments

Is this a supervised or unsupervised ML?

What is Unsupervised Learning?

Fraud Detection Case 2

You are a Data Scientist tasked to design a Machine Learning model that can identify **FRAUDULENT PAYMENTS**.

You are given:

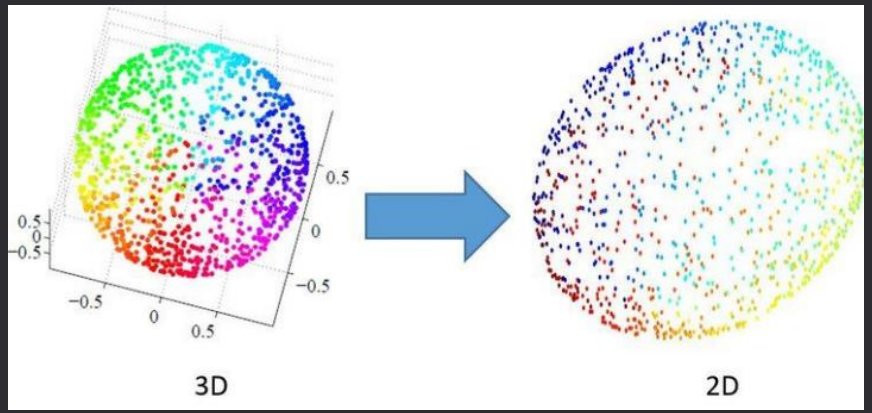
- A dataset consisting of 2000 payments, and are tasked to cluster these payments into 2 groups, fraud and non-fraud. You don't know which of these 2000 are actually fraudulent or not.

Is this a supervised or unsupervised ML?

Types of Unsupervised Learning

Type 1: Dimensionality Reduction

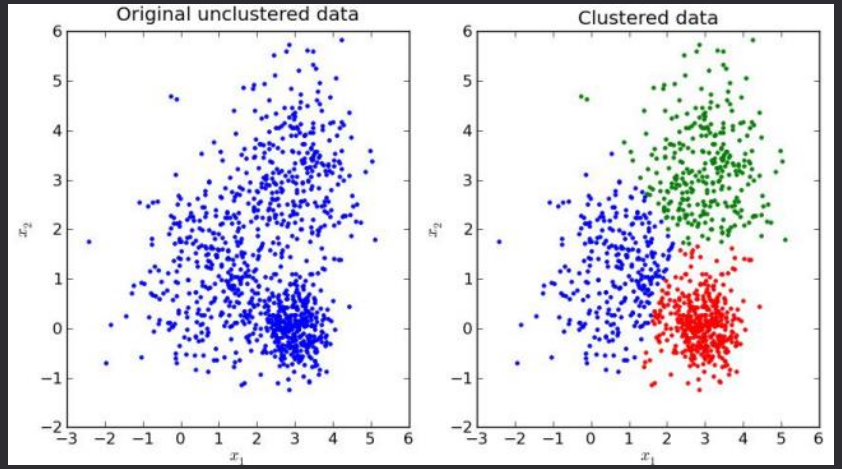
Given a dataset with a lot of columns, try to reduce the number of features but still retaining as much information as possible.



Types of Unsupervised Learning

Type 2: Clustering

Given a bunch of unlabeled data, create clusters that can group similar data together.



02

Unsupervised Learning:
Dimentionality Reduction

Principal Component Analysis (PCA)

PCA

The basic type of Dimensionality Reduction.

Suppose you have a dataset that has 50 columns. You want to reduce it to only 2-3 columns. If you delete 40+ columns, you lose 80% of information.

How to still retain the information, but reduce the dataset size?

Principal Component Analysis.

PCA

What for?

1. **Visualization**. If you have data with 5-10 columns, but you want to plot them in a 2D graph (x and y coordinate), you need to reduce them into having 2 principal components.

2. **Reduction in size**. If you have a huge dataset, you might want to reduce the size to speed up model training process.

PCA

Limitation

Cannot be used on categorical/one-hot encoded data. The aim of PCA is to preserve variance within numerical features so it still retains the information. Categorical/one-hot data simply does not have this.

Further read:
https://towardsdatascience.com/pca-is-not-feature-selection-3344fb764ae6

03

Unsupervised Learning:
Dimentionality Reduction

Factor Analysis

Factor Analysis

Factor Analysis adalah sebuah teknik untuk melakukan Dimensionality Reduction dalam Unsupervised Machine Learning.

Jikalau dalam PCA kita mencari principal component, di Factor Analysis kita mencari factors, yaitu variabel laten yang mampu menjelaskan hubungan antara variabel bebas.

Factor Analysis

Analogi:

IQ siswa diukur dan dicatat dalam sebuah tabel. Kemudian, nilai Bahasa Inggris, German, nilai Matematika, dan nilai Fisika juga dicatat dalam tabel tersebut.

Maka, menggunakan FA, kita bisa meng'ekstrak' 2 factor utama yang memengaruhi IQ, yaitu:

- Factor 1: Kemamuan Berbahasa
- Factor 2: Kemampuan Sains

FA akan membuat 2 'variable baru'.

Factor Analysis

Di dalam contoh yang kompleks, mungkin penemuan ‘factor’ tidak sesederhana contoh di analogi kita. Oleh sebab itu, algoritma ini dapat membantu kita menganalisa apakah terdapat ‘factor-factor’ yang ‘tersembunyi’ di dalam data yang multivariate.

PCA vs FA

Perbedaan PCA vs FA

Principal Component dibentuk dari kombinasi linear masing-masing variabel bebas.

Factor adalah yang membentuk masing-masing variabel bebas.

PCA vs FA

Perbedaan PCA vs FA

Kapan pakai Factor Analysis?
Ketika kita berasumsi bahwa ada ‘factor-factor’ yang melatarbelakangi banyaknya variable bebas.

Kapan kita pakai PCA?
Ketika tujuan utama kita adalah mereduksi dimensi dengan mempertahankan sebanyak mungkin ‘informasi’ (variance) dalam data.

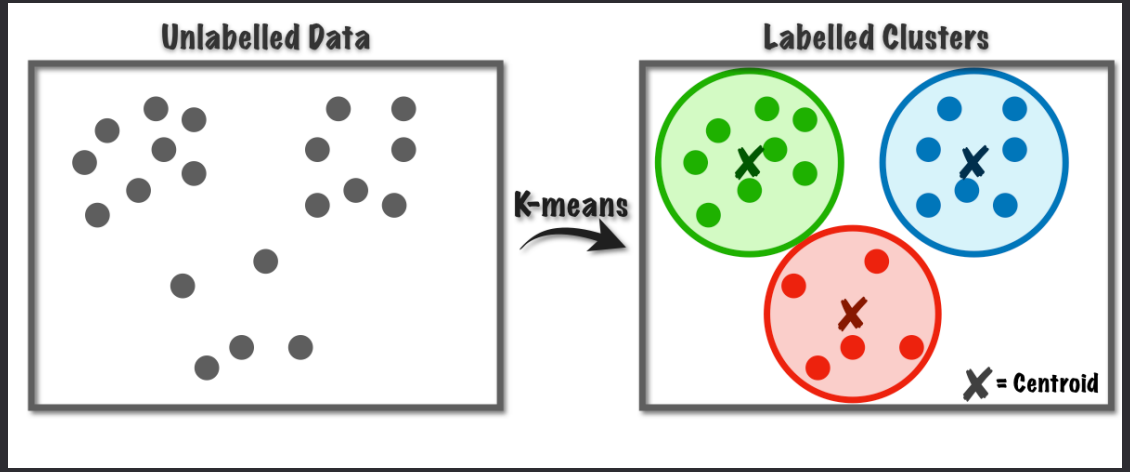
04

Unsupervised Learning: Clustering

K-Means Clustering

K-Means Clustering

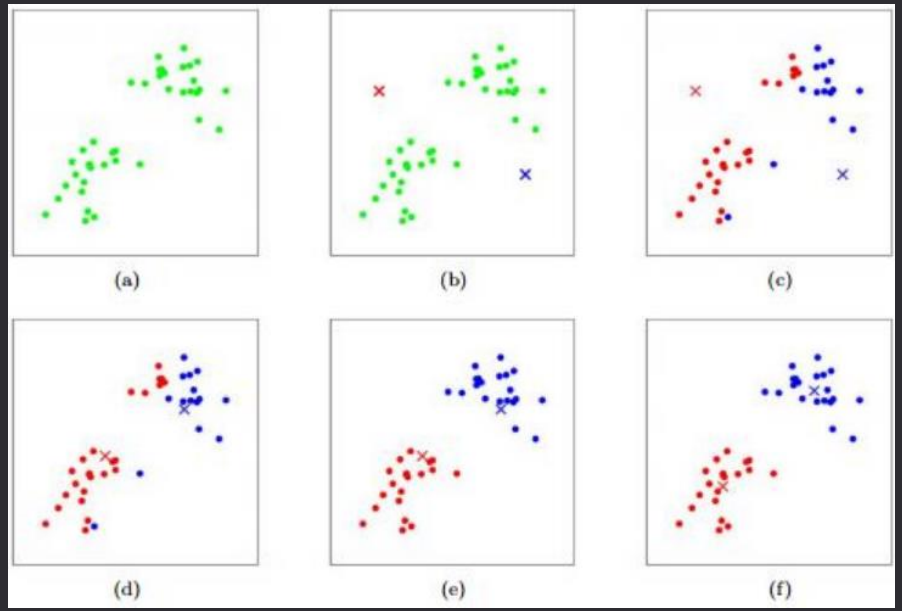
Clustering is a process of dividing entire data into groups (known as clusters) based on similarity and observed patterns.



K-Means Clustering

Step-By-Step K-Means Clustering:

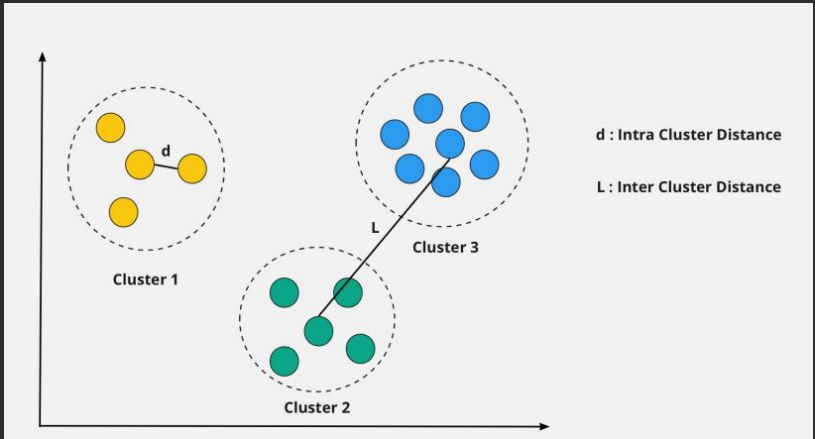
1. Memilih ‘k’ posisi ‘acak’ untuk dijadikan ‘cluster center’
2. Data points dipisahkan berdasarkan jarak mereka ke masing-masing ‘cluster center’
3. Lokasi cluster center diubah dengan mencari ‘titik tengah’ dari titik-titik yang telah dikelompokkan
4. Karena ‘cluster center’ berubah lokasi, maka pengelompokkan pun akan berubah
5. Lakukan Step 2-3-4 sampai tidak ada titik yang ‘berubah kelompok’ lagi.



K-Means Clustering

Prinsip K-Means Clustering:

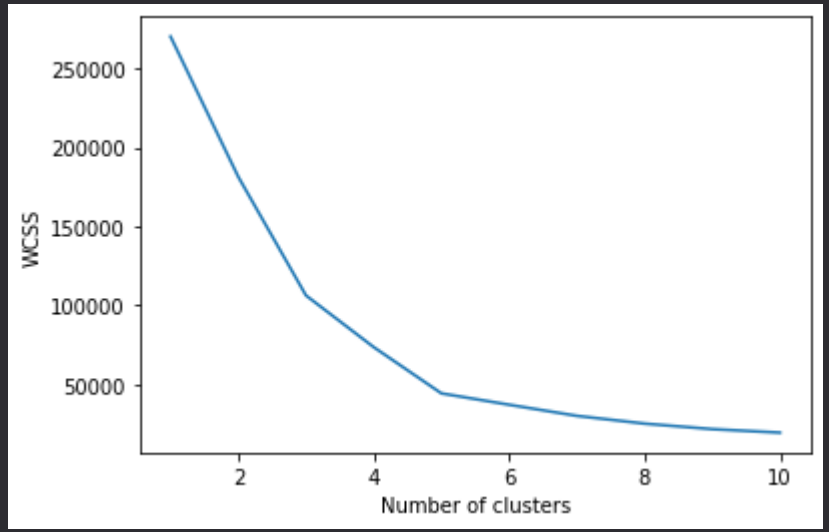
- Meminimalkan ‘intra-cluster distance’
- Memaksimalkan ‘inter-cluster distance’



K-Means Clustering

Elbow Method

The elbow method is a graphical representation of finding the optimal 'K' in a K-means clustering. It works by finding WCSS (Within-Cluster Sum of Square) i.e. the sum of the square distance between points in a cluster and the cluster centroid.

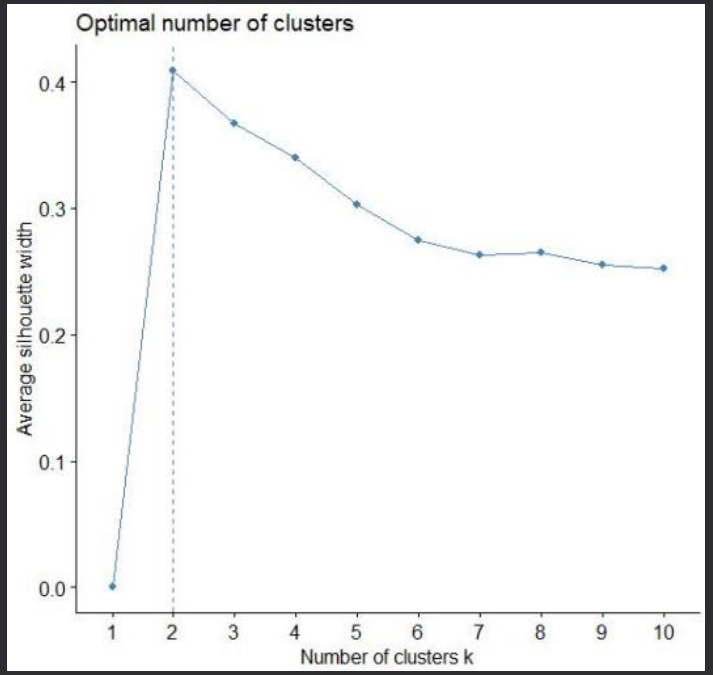


K-Means Clustering

Silhouette Method

Memiliki nilai kecil jika intercluster distance kecil.

Memiliki nilai tinggi jika intercluster distance tinggi.



K-Means Clustering

Challenges

- Terkadang ‘k’ hasil Elbow Method berbeda dengan ‘k’ optimal hasil Silhouette Method
- Terkadang, stakeholder yang ingin menetapkan ‘banyaknya cluster yang harus dibentuk’
- Penentuan ‘k’ optimal memang proses yang tidak mudah dan memerlukan banyak pertimbangan technical maupun ‘business side’

Thanks!

Do you have any questions?

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**