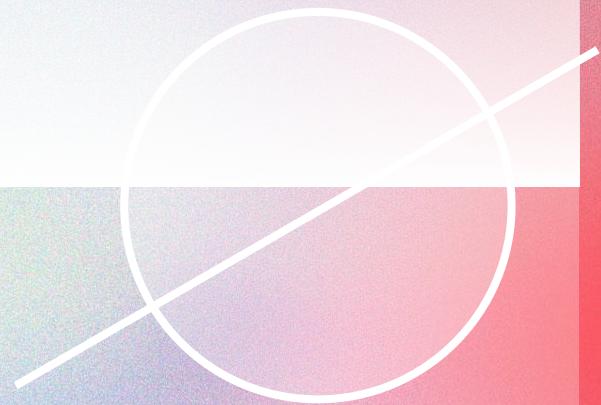


# CLUSTER ANALYSIS

cos'è e come funziona

Arenare Mattia 942766  
Drago Kevin 941701  
Pace Vittorio 941730



# INDICE



O1

## INTRODUZIONE

Origine, definizione e applicazioni della Cluster Analysis

O2

## CARATTERISTICHE

Tipologie di Clustering e metriche utilizzate

O3

## K-MEANS

Caratteristiche, pseudocodice ed esempio

O4

## ESEMPIO IN PYTHON

Applicazione di K-Means a un dataset medico

# O1

## INTRODUZIONE

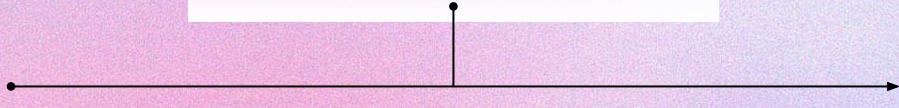
Origine, definizione e applicazioni della Cluster Analysis

# STORIA

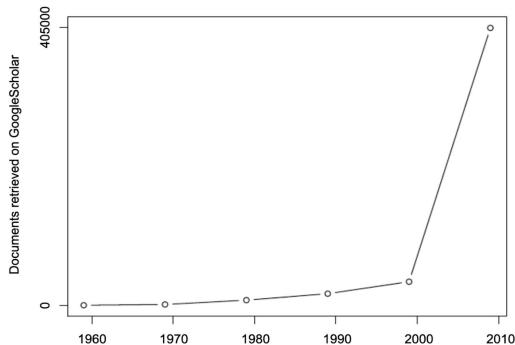
X

Prima apparizione della  
Cluster Analysis

**1954**



Search term 'Cluster Analysis'



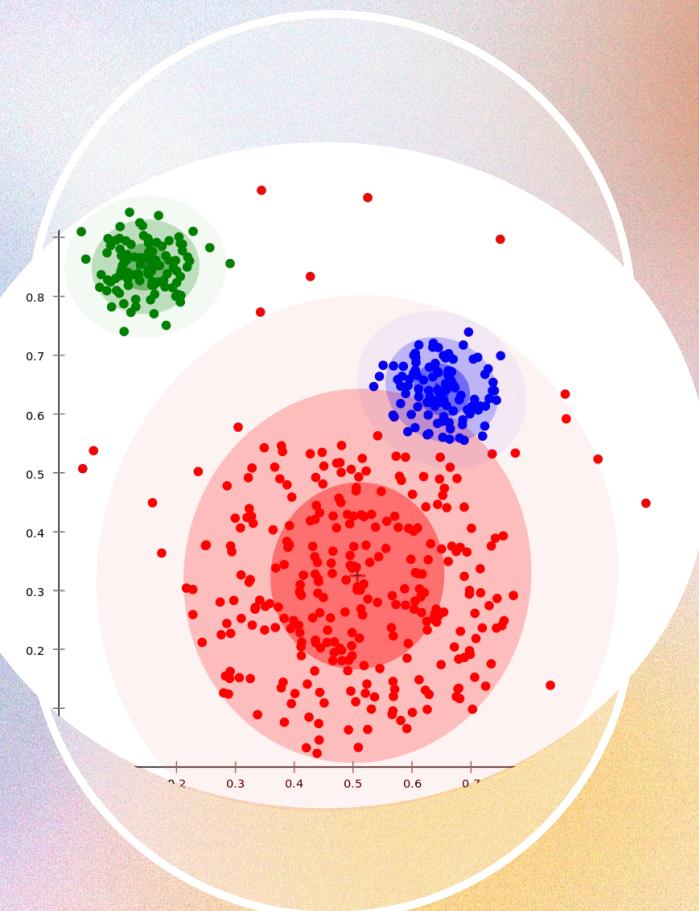
# **COS'È UN CLUSTER**

Raggruppamento di oggetti che hanno uno o più caratteristiche in comune.

Preso un insieme di oggetti aventi un certo numero di attributi, questi possono essere utilizzati per separare gli oggetti in un numero qualunque di cluster.

# CLUSTER ANALYSIS

Insieme di tecniche di analisi multivariata dei dati, che ha come obiettivo quello di selezionare e raggruppare le informazioni in base ad alcuni elementi omogenei tra i dati.



# SETTORI DI APPLICAZIONE



## MEDICINA

Diagnosi di quadri clinici,  
prevedere casi di morbilità...



## BIOLOGIA

Raggruppare la composizione  
degli amminoacidi...



## SCIENZE SOCIALI

Fotografare la società sotto il  
profilo demografico...



## MARKETING

Individuare insiemi di  
consumatori con peculiarità  
omogenee...

# POSSIBILI USI



## DATA REDUCTION

Riduzione di ampie moli di dati in gruppi più facilmente gestibili

## HYPOTHESIS GENERATION

Sviluppo di ipotesi sulla natura dei dati o verifica di ipotesi precedentemente formulate

## TAXONOMY DESCRIPTION

Identificazione di gruppi all'interno dei dati in esame

01

## DATA SIMPLIFICATION

Possibilità di analizzare gruppi di osservazioni simili anziché singoli dati

02

## RELATIONSHIP IDENTIFICATION

Identificazione di relazioni grazie alla struttura semplificata dei cluster

03

04

05

# COME FUNZIONA

L'obiettivo principale della cluster analysis è definire la struttura dei dati raggruppando le osservazioni simili.

Per farlo dobbiamo rispondere a tre semplici domande:

## COME MISURARE LA SOMIGLIANZA?

Grado di corrispondenza tra gli oggetti

01

## COME FORMARE I CLUSTERS?

Identificare le due osservazioni più simili e combinarle

02

## QUANTI GRUPPI FORMARE?

Dipende, continuo a formarne finchè sono gruppi eterogenei

03

# ESEMPI



## SEGMENTAZIONE

Individuare la tipologia di clientela verso cui un brand può orientare una campagna marketing mirata



## CITOLOGIA

Classificare campioni ematici, per definire gruppi sanguigni e tipi di plasma



## DEMOGRAFIA

Suddividere i Paesi europei per causa di mortalità



## ECONOMIA

Indagare l'impatto economico del Covid-19

# CLUSTERING VS CLASSIFICATION

## CLUSTERING

Procedimento di tipo **non supervisionato**

Non esistono classi predeterminate, né esempi che le rappresentino

Estrapolare un certo numero di gruppi in cui è possibile separare gli oggetti di un insieme analizzando i valori dei loro attributi

## CLASSIFICATION

Procedimento di **apprendimento supervisionato**

Serie di classi note a priori

Capire a quale gruppo appartiene un oggetto osservando il valore dei suoi attributi

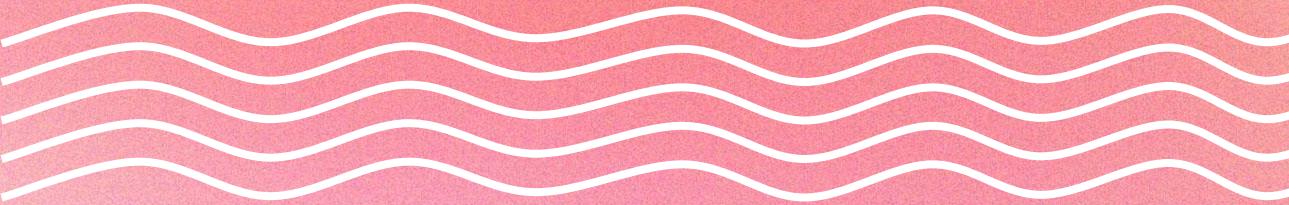
**Fase di addestramento**

Si parte da un insieme di oggetti dei quali si conosce già la categoria di appartenenza e si cerca di trovare un pattern comune

# O2

## CARATTERISTICHE

Tipologie di Clustering e metriche utilizzate



# TECNICHE DI CLUSTERING



## CLUSTERING GERARCHICO

Il risultato è una serie di partizioni innestate (un “dendrogramma”)

Mira ad evidenziare le relazioni tra i vari pattern del dataset

Non è necessario settare a priori il numero di cluster

Più informativo del partizionale, è improponibile per dataset grandi

## CLUSTERING PARTIZIONALE

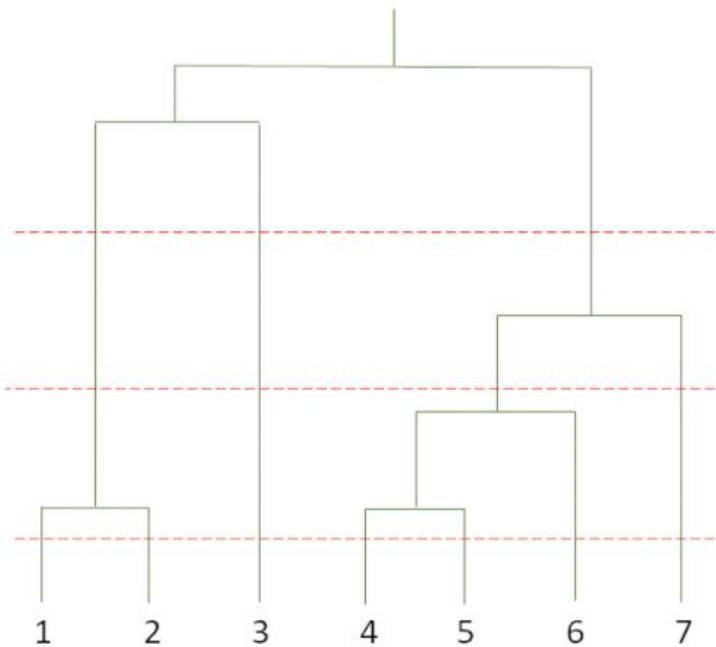
Il risultato è una singola partizione dei dati

Mira ad identificare i gruppi naturali presenti nel dataset

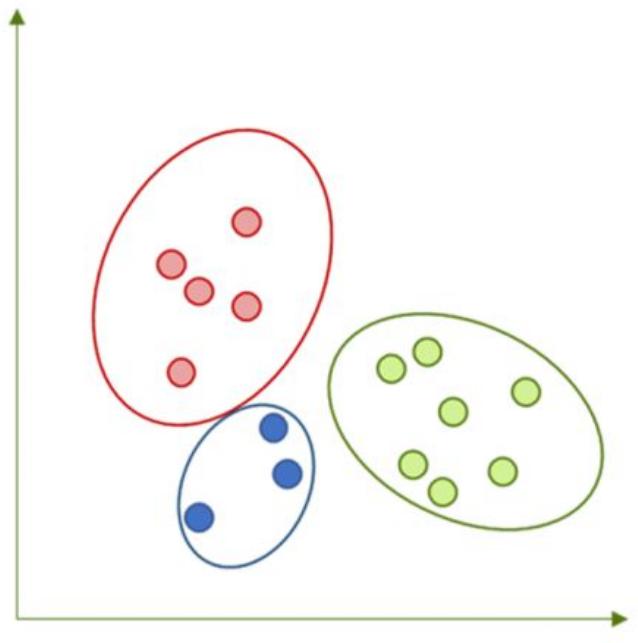
Tipicamente il numero di cluster deve essere dato a priori

Ottimo per dataset grandi

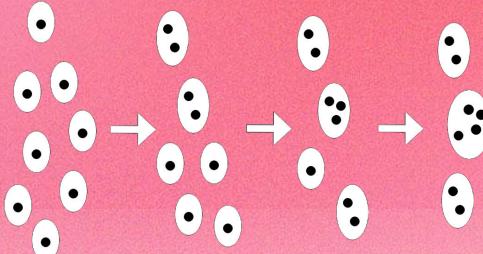
Hierarchical



Non-hierarchical

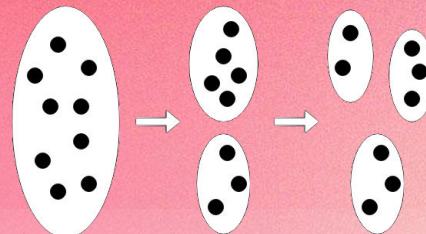


# CLUSTERING GERARCHICO



## ALGORITMI AGGLOMERATIVI (BOTTOM-UP)

Iniziano inserendo ogni oggetto dell'insieme in un proprio cluster per poi raggrupparli iterativamente fino al raggiungimento di una condizione specifica (es. numero di cluster desiderato)



## ALGORITMI DIVISIVI (TOP-DOWN)

Iniziano inserendo tutti gli oggetti in un unico cluster per poi separarlo iterativamente in cluster più piccoli fino al raggiungimento di una condizione specifica (es. numero di cluster desiderato)

# ALTRÉ DISTINZIONI



## **ESCLUSIVO VS NON ESCLUSIVO**

Non esclusivo: i punti possono appartenere a più cluster

## **PARZIALE VS COMPLETO**

Parziale: alcuni punti potrebbero non appartenere a nessuno dei cluster

01

## **FUZZY VS NON FUZZY**

Fuzzy: un punto appartiene a tutti i cluster con un peso tra 0 e 1

03

## **ETEROGENEO VS OMOGENEO**

Eterogeneo: i cluster possono avere dimensioni, forme e densità molto diverse

02

04

# TIPI DI CLUSTER

X

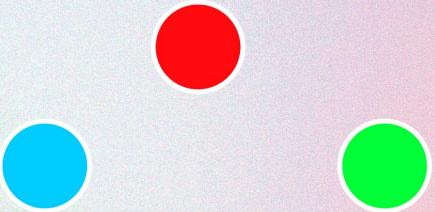
CENTER BASED

01



WELL  
SEPARATED

02



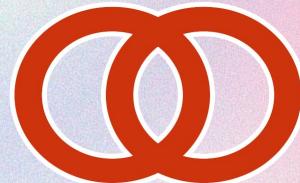
DENSITY BASED

03



CONCEPTUAL

04



# METRICA

Gli algoritmi di clustering si basano tutti su una **metrica**, puramente geometrica, che permetta di identificare quanto simili siano due oggetti fra di loro. Infatti gli oggetti in esame vengono visti come insiemi di valori reali che ne rappresentano le caratteristiche (colore degli occhi, altezza, peso, ecc.). Questi valori, a loro volta, possono essere raggruppati in modo da formare dei vettori che rappresentino punti in uno spazio euclideo.

## HAN

Altezza: 180 cm

Peso: 75 kg

## LEIA

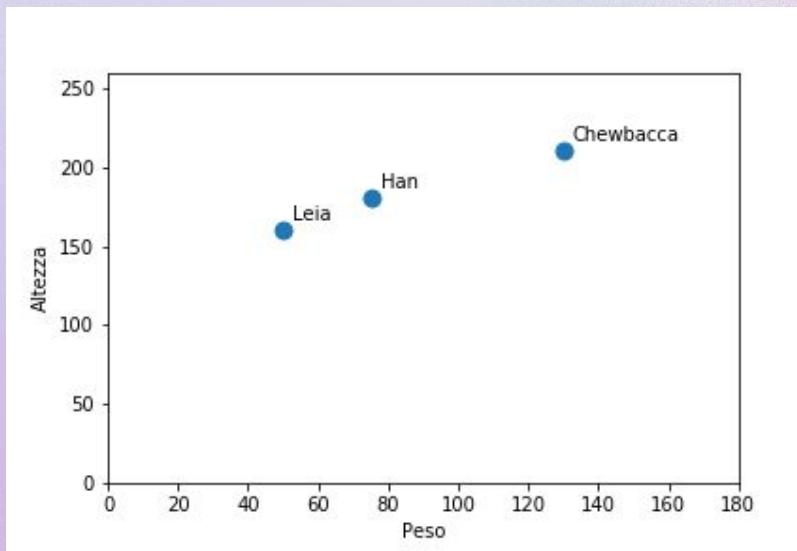
Altezza: 160 cm

Peso: 50 kg

## CHEWBACCA

Altezza: 210 cm

Peso: 130 kg

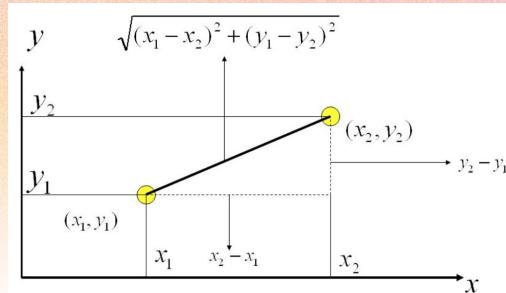


# PRINCIPALI METRICHE UTILIZZATE

O1

## DISTANZA EUCLIDEA

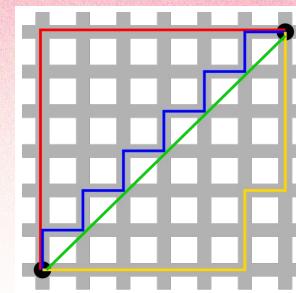
Misura della lunghezza del segmento avente per estremi i due punti



O2

## DISTANZA MANHATTAN

Somma del valore assoluto delle differenze delle coordinate dei punti

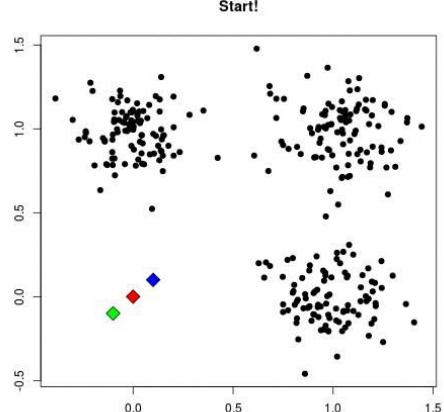


# 03

## K-MEANS

Caratteristiche, pseudocodice ed esempio

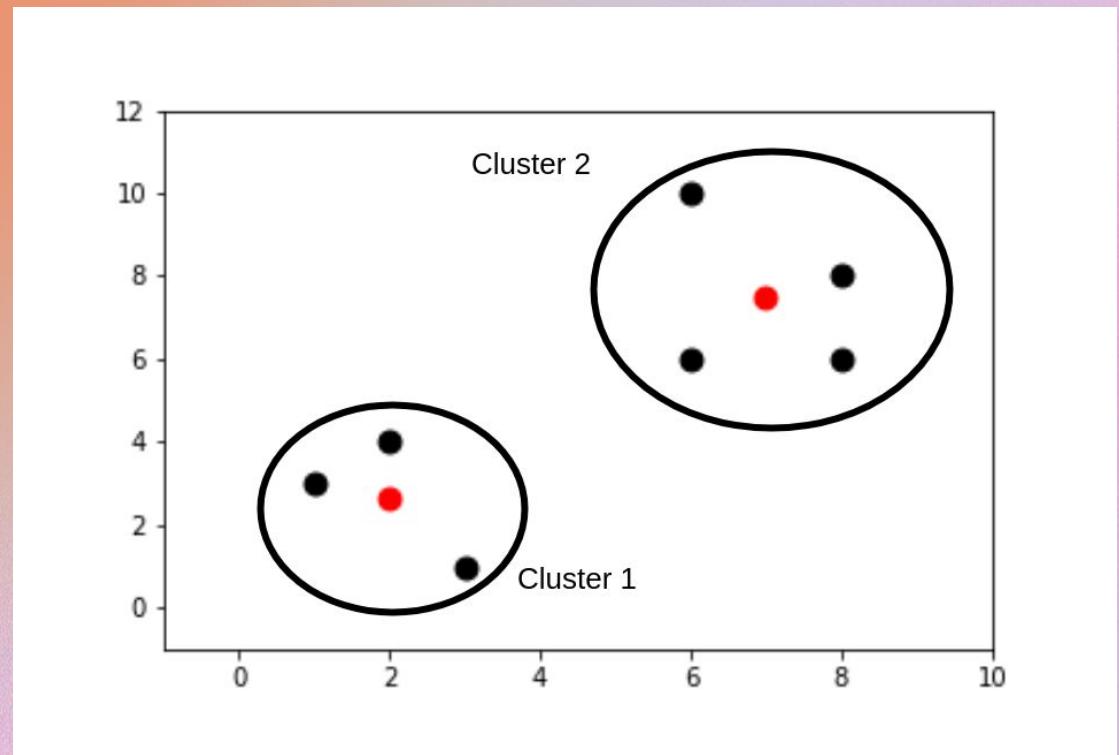




# K-MEANS

Tecnica di clustering partizionale che raffina la suddivisione degli oggetti a ogni ciclo e che si basa sul concetto di centroide.

Un **centroide** è un punto nello spazio che rappresenta, sostanzialmente, un cluster e che corrisponde al punto medio dei punti del cluster stesso.



# PSEUDOCODICE

Input:

- N oggetti caratterizzati da vettori di dimensione d
- Un intero C che rappresenta il numero di cluster desiderati
- C vettori di dimensione d che rappresentano i centroidi rappresentativi dei cluster iniziali

Repeat:

**Step 1 (Assignment):**

Assegna ogni oggetto al cluster il cui centroide ha la distanza euclidea più bassa (in sostanza il più vicino)

**Step 2 (Update):**

Per ogni cluster calcolare il punto medio delle osservazioni a esso associate; questo punto sarà il centroide di questo cluster per l'iterazione successiva

# SCELTA DEI CENTROIDI



## RANDOM

K punti casuali selezionati dal dataset e utilizzati come centroidi iniziali

Non assicura che i centroidi selezionati siano ben posizionati in tutto lo spazio dati

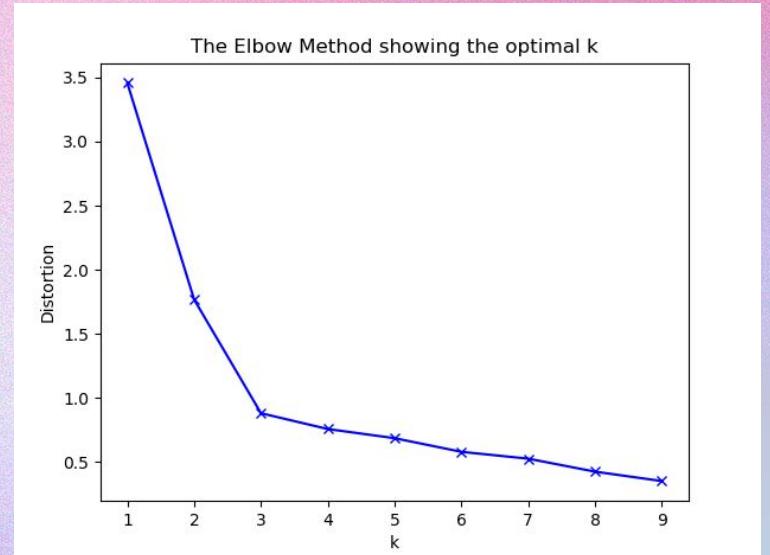
## K-MEANS++

Viene assegnato il primo centroide alla posizione di un punto selezionato **casualmente** e quindi scegliendo i centroidi successivi dai punti dati rimanenti in base a una probabilità **proporzionale alla distanza al quadrato** dal baricentro esistente più vicino di un dato punto

# TECNICA DEL GOMITO



Viene iterato il K-means per diversi valori di K ed ogni volta si calcola la somma delle distanze al quadrato tra ogni centroide ed i punti del proprio cluster.



# ESEMPIO

Dividiamo questi 8 punti in 3 cluster

A1 (2,10)

A2 (2,5)

A3 (8,4)

A4 (5,8)

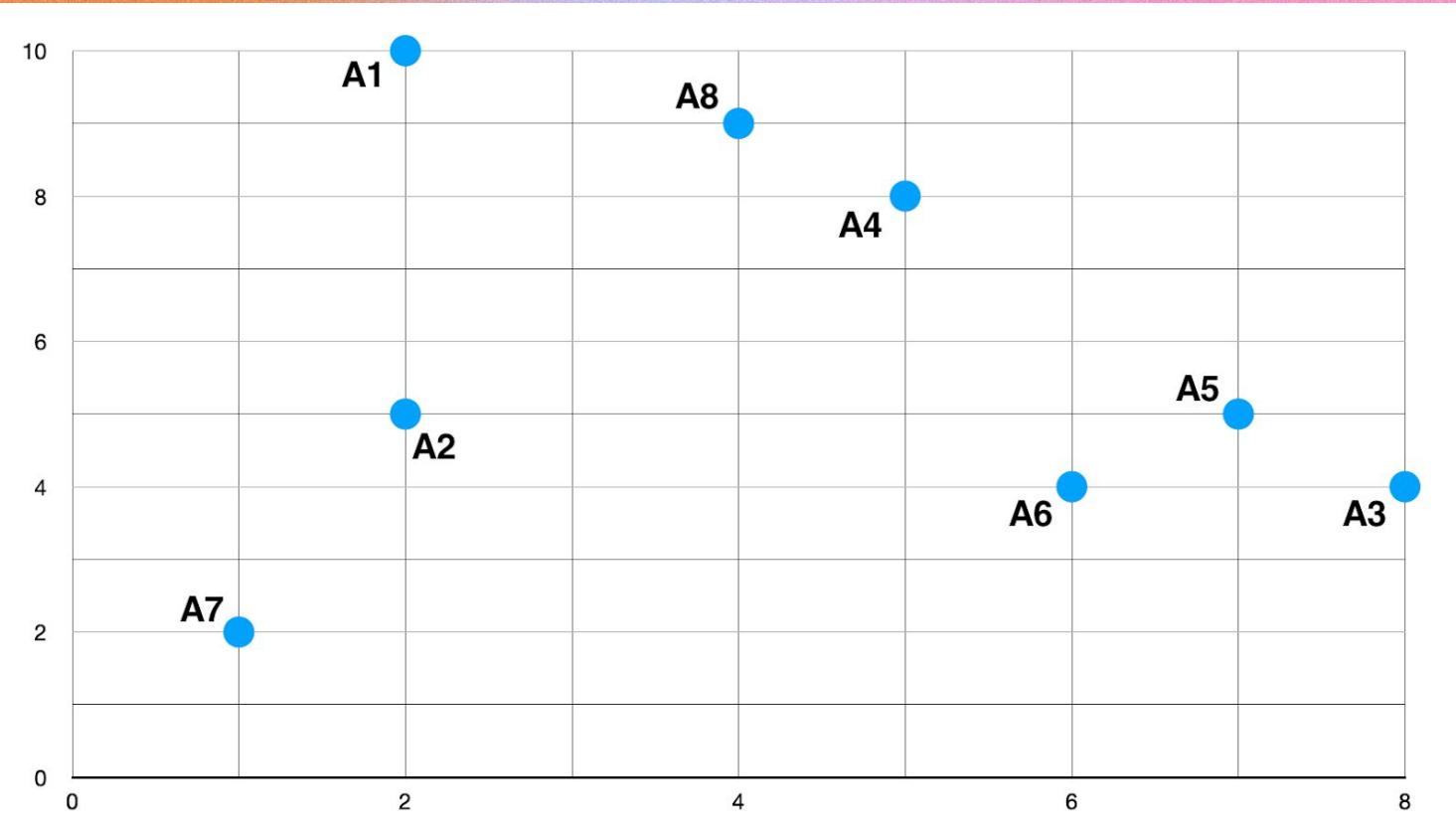
A5 (7,5)

A6 (6,4)

A7 (1,2)

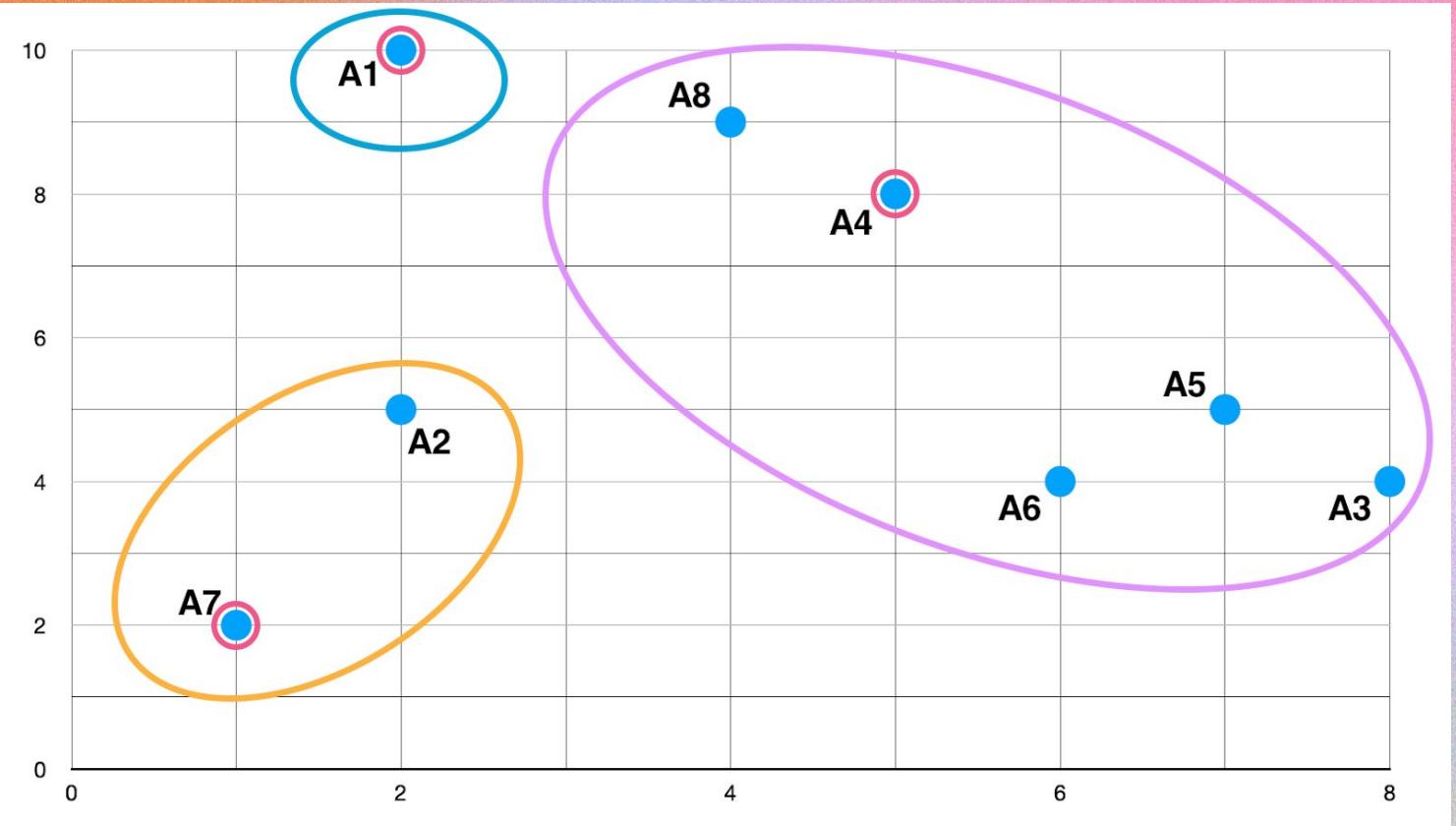
A8 (4,9)

Utilizziamo la distanza di Manhattan



# ITERAZIONE 1

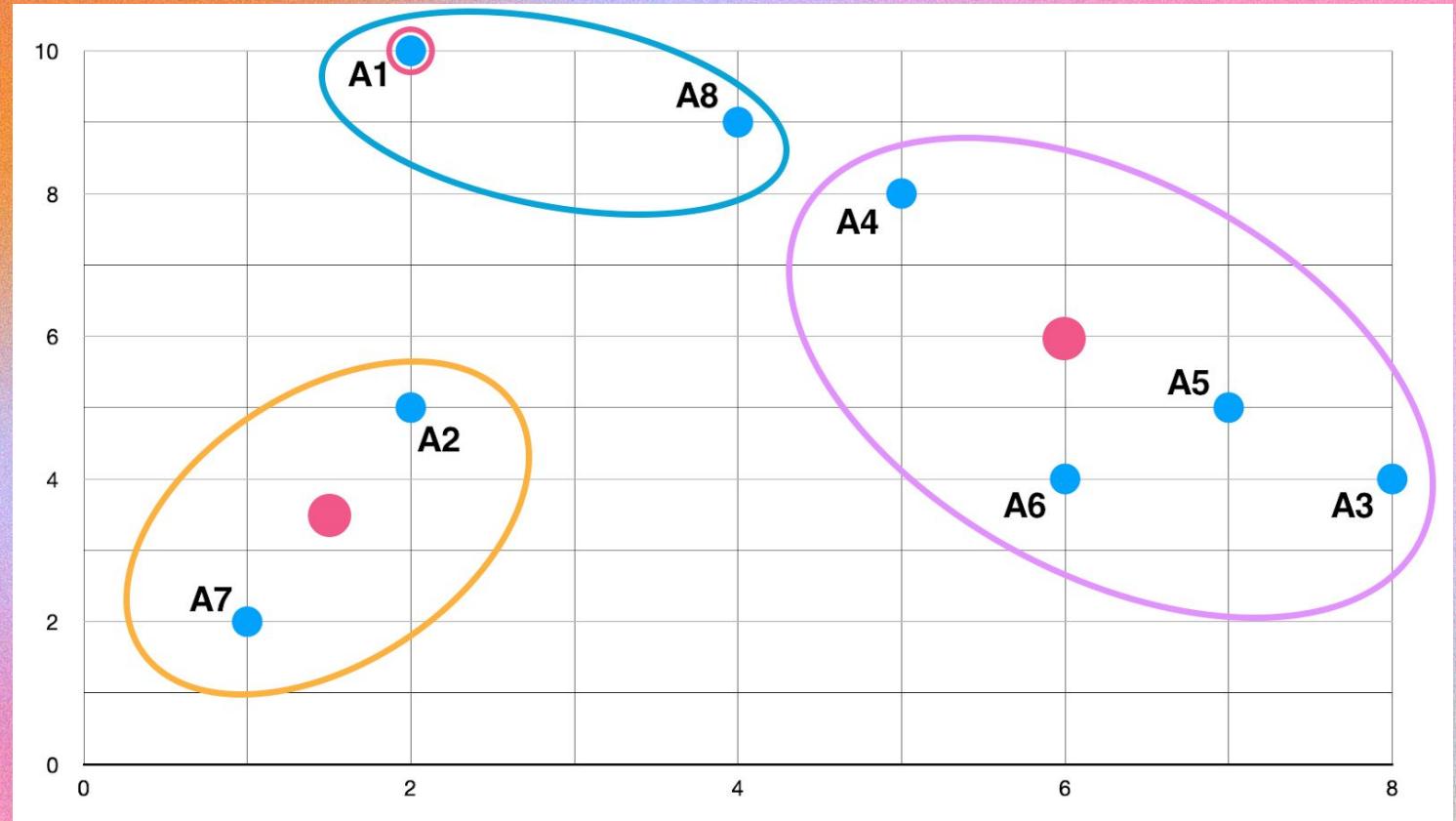
		DISTANZA CENTROIDE 1 (2, 10)	DISTANZA CENTROIDE 2 (5, 8)	DISTANZA CENTROIDE 3 (1, 2)	cluster
A1	(2, 10)	0	5	9	1
A2	(2, 5)	5	6	4	3
A3	(8, 4)	12	7	9	2
A4	(5, 8)	5	0	10	2
A5	(7, 5)	10	5	9	2
A6	(6, 4)	10	5	7	2
A7	(1, 2)	9	10	0	3
A8	(4, 9)	3	2	10	2



# ITERAZIONE 2

X

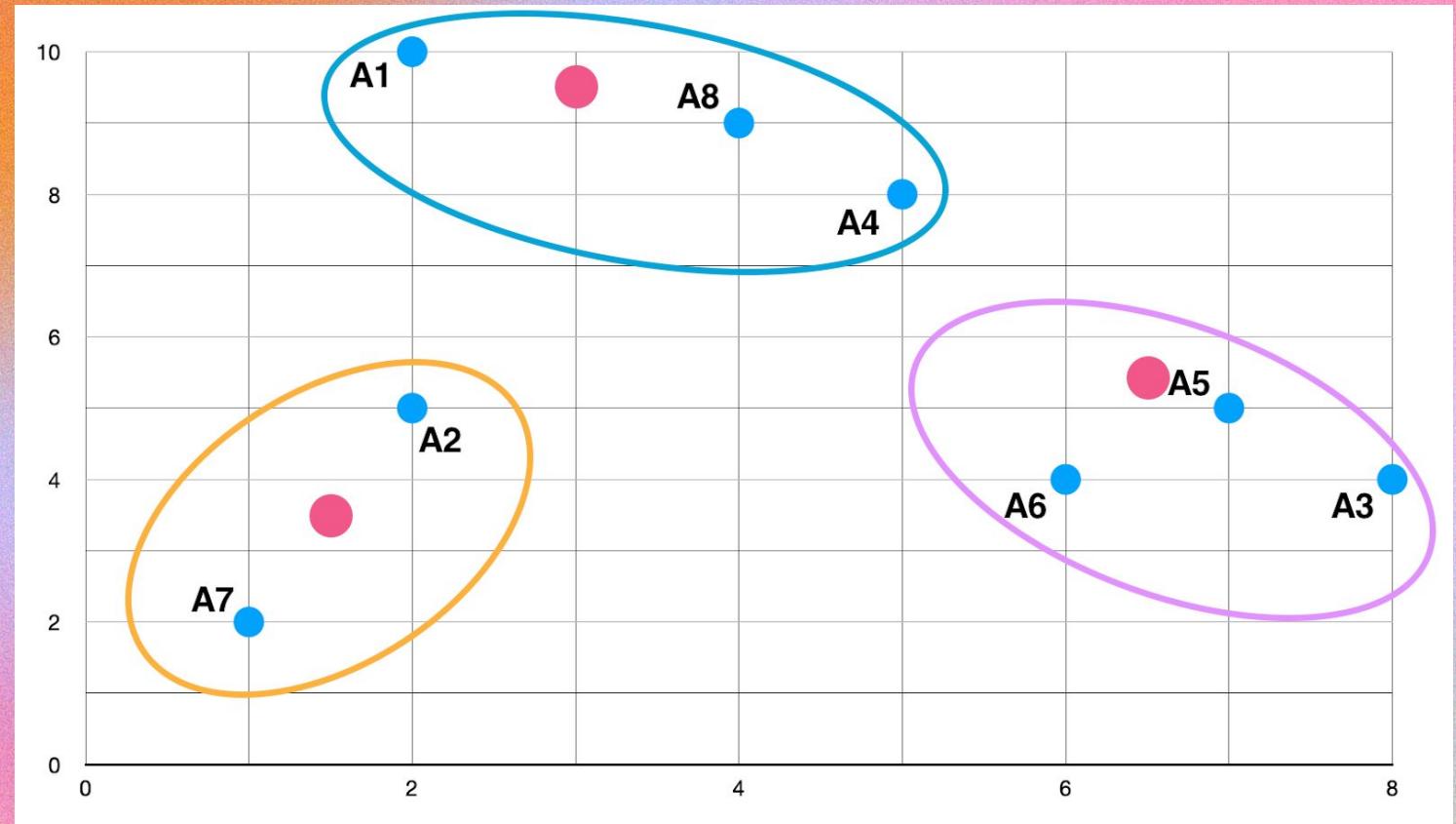
		DISTANZA CENTROIDE 1 (2, 10)	DISTANZA CENTROIDE 2 (6, 6)	DISTANZA CENTROIDE 3 (1.5, 3.5)	CLUSTER
A1	(2, 10)	0	8	7	1
A2	(2, 5)	5	5	2	3
A3	(8, 4)	12	4	7	2
A4	(5, 8)	5	3	8	2
A5	(7, 5)	10	2	7	2
A6	(6, 4)	10	2	5	2
A7	(1, 2)	9	9	2	3
A8	(4, 9)	3	5	8	1



# ITERAZIONE 3

X

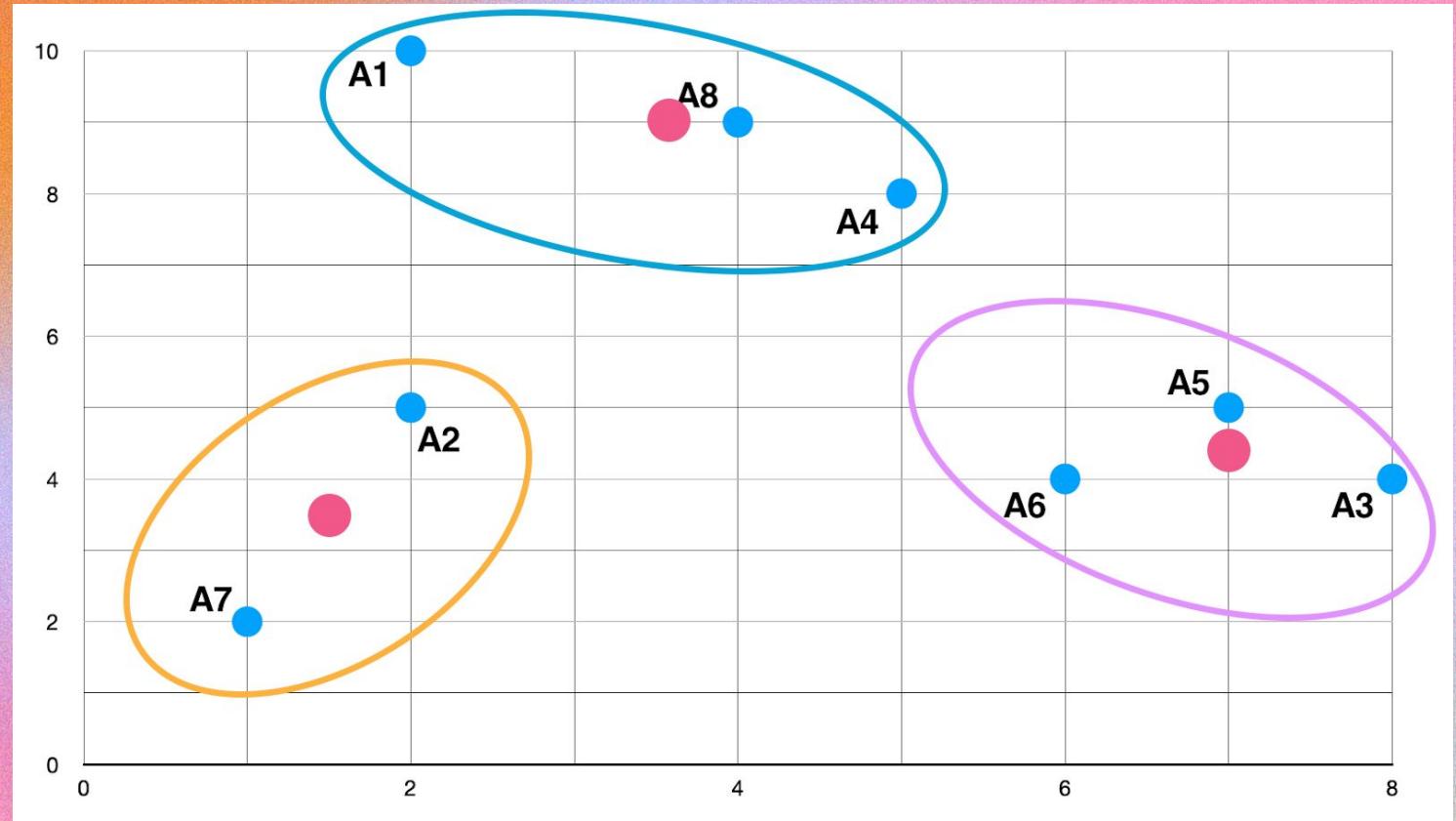
		DISTANZA CENTROIDE 1 (3, 9.5)	DISTANZA CENTROIDE 2 (6.5, 5.25)	DISTANZA CENTROIDE 3 (1.5, 3.5)	CLUSTER
A1	(2, 10)	1.5	9.25	7	1
A2	(2, 5)	5.5	4.74	2	3
A3	(8, 4)	10.5	2.75	7	2
A4	(5, 8)	3.5	4.25	8	1
A5	(7, 5)	8.5	0.75	7	2
A6	(6, 4)	8.5	1.75	5	2
A7	(1, 2)	9.5	8.75	2	3
A8	(4, 9)	1.5	6.25	8	1

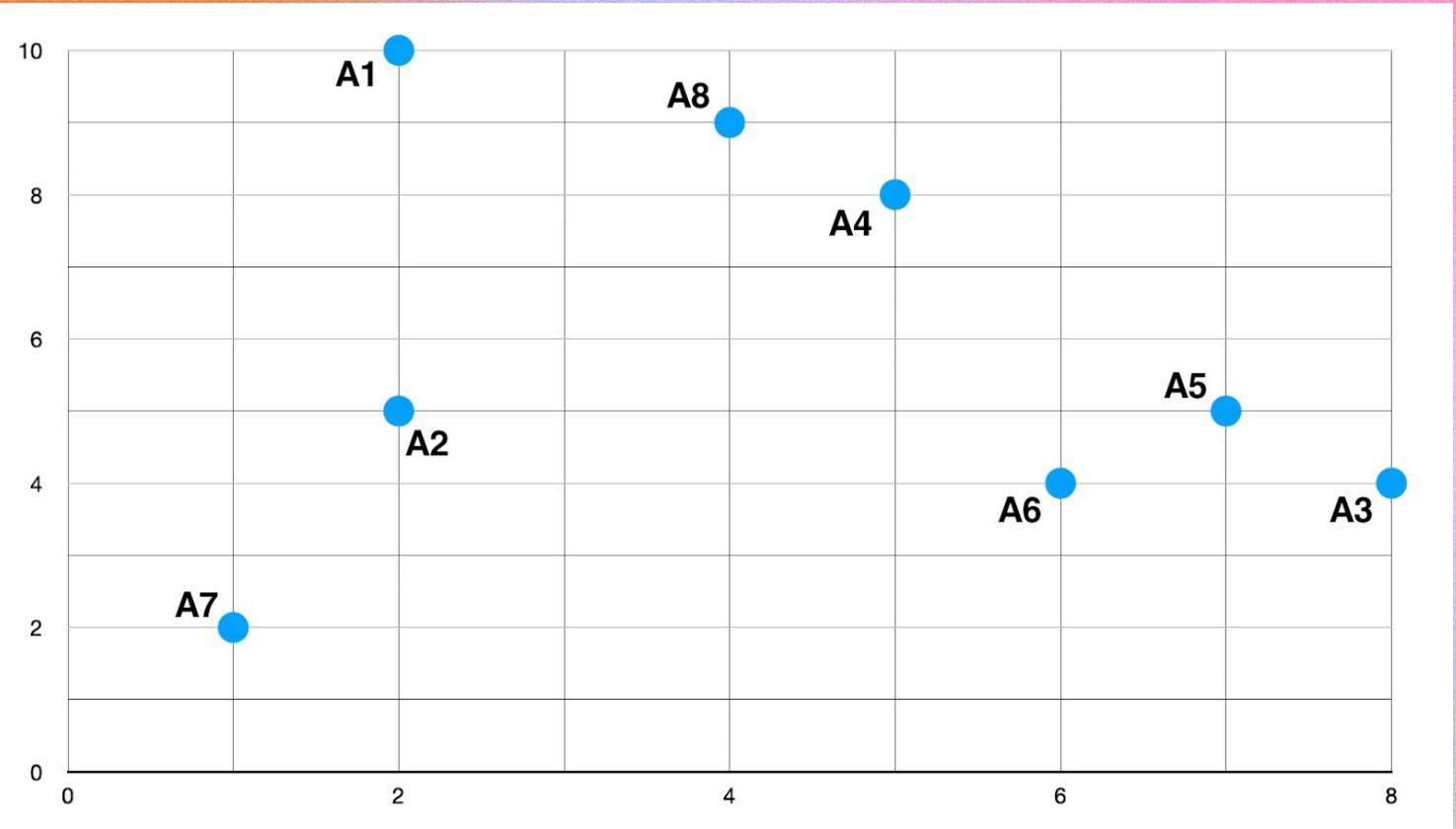


# ITERAZIONE 4

X

		DISTANZA CENTROIDE 1 (3.67, 9)	DISTANZA CENTROIDE 2 (7, 4.3)	DISTANZA CENTROIDE 3 (1.5, 3.5)	cluster
A1	(2, 10)	2.67	10.7	7	1
A2	(2, 5)	5.67	5.7	2	3
A3	(8, 4)	9.33	1.3	7	2
A4	(5, 8)	2.33	5.7	8	1
A5	(7, 5)	7.33	0.7	7	2
A6	(6, 4)	7.33	1.3	5	2
A7	(1, 2)	9.67	8.3	2	3
A8	(4, 9)	0.33	7.7	8	1





# O4

## ESEMPIO IN PYTHON

Applicazione di K-Means a un dataset medico

# **NOTEBOOK COLAB**

# GRAZIE

**CREDITS:** This presentation template was created by [Slidesgo](#), including icons by [Flaticon](#) and infographics & images by [Freepik](#)