

Classe:

Nome e Cognome:

---

## L’algoritmo k–NN

Si ha a disposizione il seguente **dataset di training** con riportati i dati relativi alla playlist di un utente.

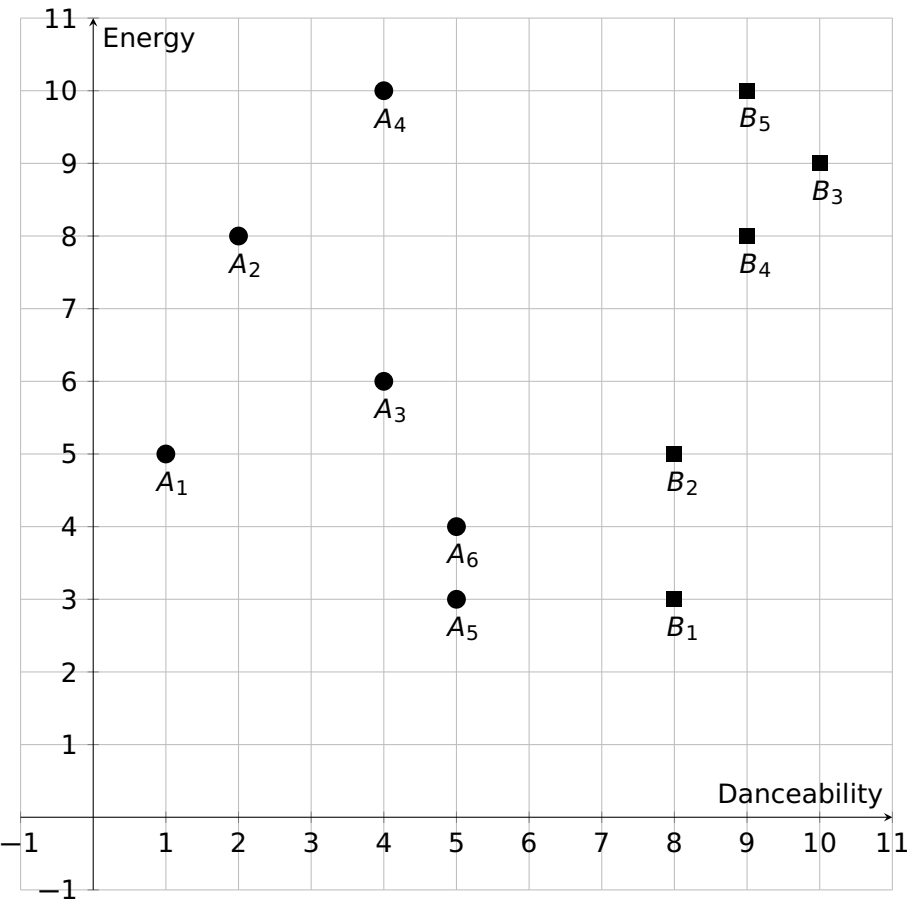
Danceability	Energy	Label
1	5	Like
2	8	Like
4	6	Like
4	10	Like
5	3	Like
5	4	Like
8	3	Don’t like
8	5	Don’t like
10	9	Don’t like
9	8	Don’t like
9	10	Don’t like

La prima *feature* (colonna) rappresenta la danceability di una canzone, la seconda l’energy. Le prime due colonne costituiscono la mia matrice  $X_{train}$ . All’utente della tua piattaforma solo alcune canzoni piacciono. L’informazione è riportata nella terza colonna, *label*. Questa colonna è il vettore  $y_{train}$  che voglio provare a predire in base ai dati della matrice  $X_{train}$ .

Si ha poi a disposizione un secondo dataset su cui testare il modello, **dataset di test**.

Superficie (·100m <sup>2</sup> )	Prezzo (·10 <sup>5</sup> €)	Label
5	8	Don’t like
7	4	Like
10	4	Don’t like
10	8	Don’t like
3	7	Like

1. Consideriamo nel **dataset di training** le due *feature* come le coordinate  $(x,y)$  di un punto nel piano cartesiano come riportato nel seguente grafico.



Classifica i seguenti punti  $P_1(5; 8)$ ,  $P_2(7, 4)$ ,  $P_3(10, 4)$ ,  $P_4(10, 8)$ ,  $P_5(3, 7)$  considerando  $k = 3$ .

- (a) Come si calcola la distanza tra due punti non allineati?  
.....
- (b) E nel caso di punti sono allineati?  
.....

(c) Riporta le distanze nella seguente tabella

	A <sub>1</sub>	A <sub>2</sub>	A <sub>3</sub>	A <sub>4</sub>	A <sub>5</sub>	A <sub>6</sub>	B <sub>1</sub>	B <sub>2</sub>	B <sub>3</sub>	B <sub>4</sub>	B <sub>5</sub>
P <sub>1</sub>											
P <sub>2</sub>											
P <sub>3</sub>											
P <sub>4</sub>											
P <sub>5</sub>											

(d) Ordinare le distanze in ordine crescente, calcolare la frequenza per ogni label e classificare il punto.

	Distanze ordinate			$f_{\text{Like}}$	$f_{\text{Don't like}}$	label
$P_1$						
$P_2$						
$P_3$						
$P_4$						
$P_5$						

2. Sai che i cinque punti che hai classificato precedentemente in realtà sono rispettivamente Don't like, Like, Don't like, Don't like, Like.

(a) Completare la seguente matrice di confusione

		Valori predetti	
Valori reali			

(b) Come si calcola l'accuratezza?

.....

(c) In questo caso l'accuratezza risulta .....

3. Ponendo  $k=5$

(a) il punto  $P_2$  come viene classificato? Riportare i conti svolti.

(b) Come cambia la matrice di confusione? E l'accuratezza?

		Valori predetti	
Valori reali			

In questo caso l'accuratezza risulta .....