UNIVERSITÀ DEGLI STUDI FIRENZE

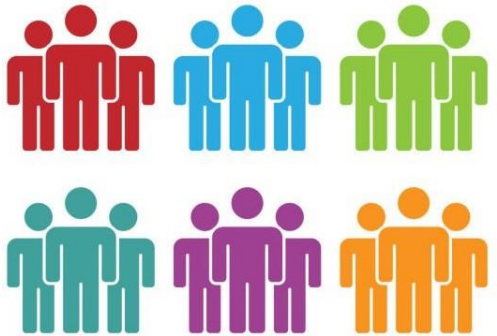
Data Mining – Customer Segmentation con Clustering e Classificazione

Michael Cavicchioli
Sofia Dami
Claudia Landi

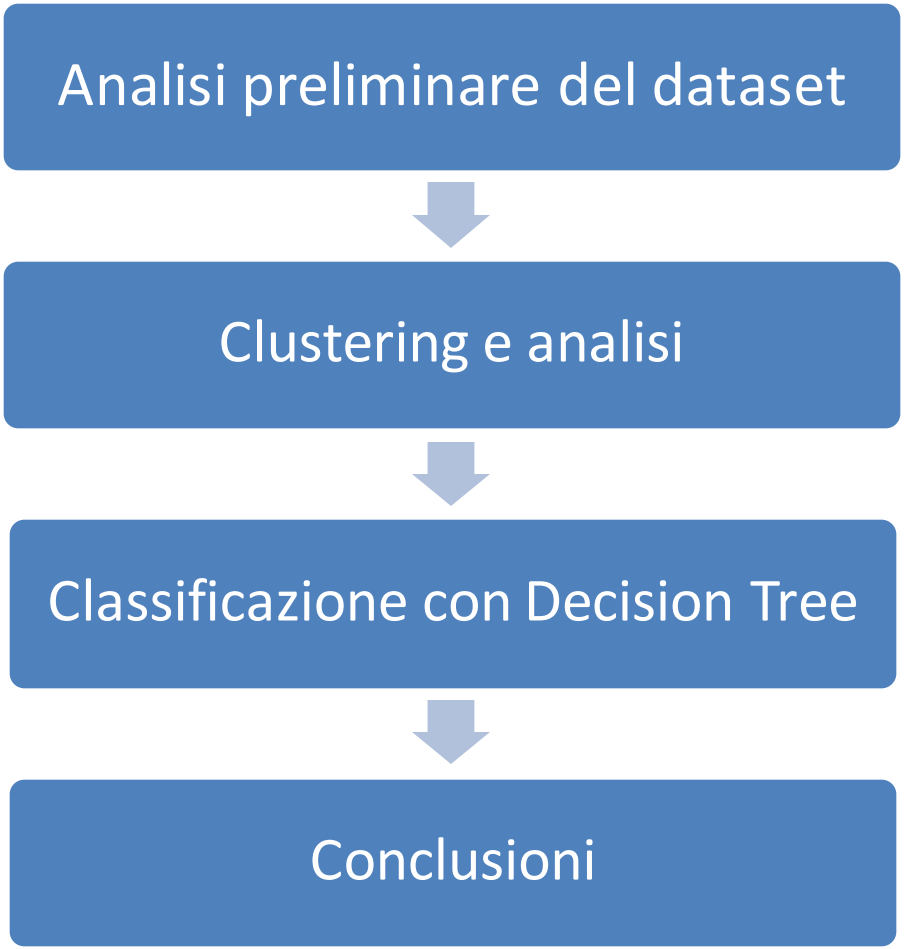
Firenze 20/06/2024

Customer Segmentation

- Strategia di marketing
- Suddivisione dei clienti in gruppi con caratteristiche comuni
- **Obiettivi:**
  - Campagne di marketing mirate
  - Aumentare le vendite
  - Fedeltà dei clienti



Fasi dello studio



Descrizione del dataset

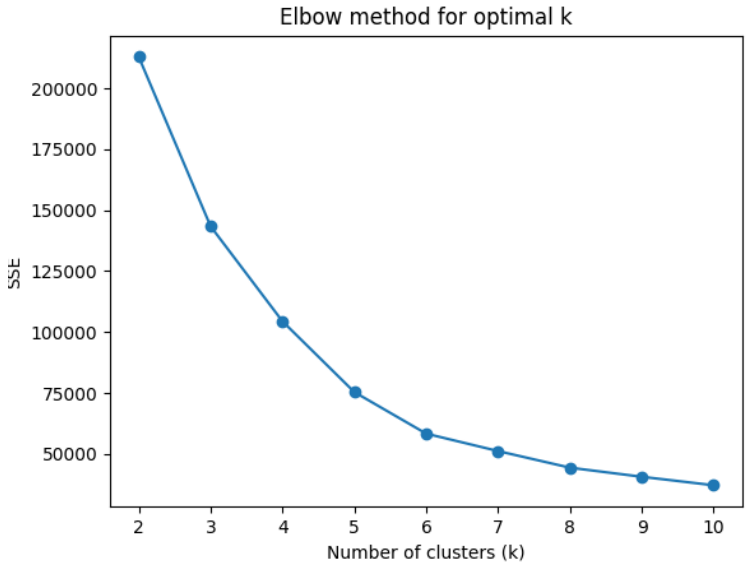
- 200 osservazioni, 5 features:
  - Customer ID
  - Genre: 88 maschi e 112 femmine
  - Age: da 18 a 70 anni
  - Annual Income: da 15k a 137k
  - Spending Score: punteggio da 1 a 100

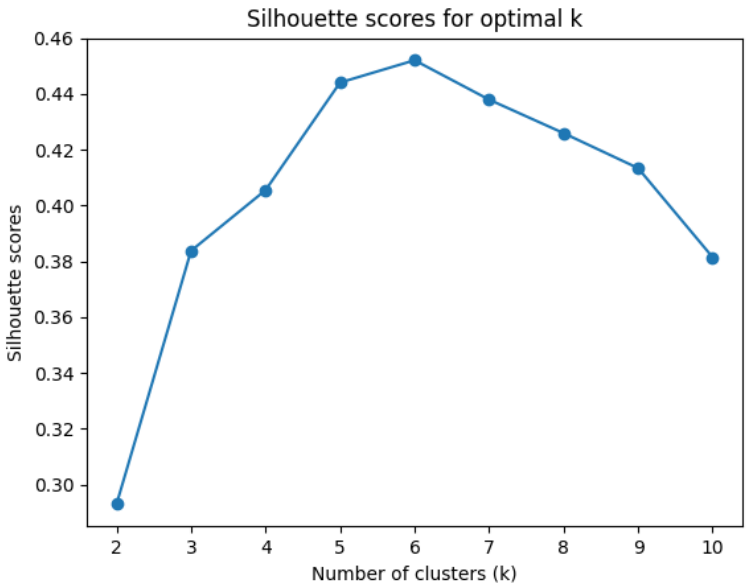
- **Preprocessing:**
  - Eliminazione Customer ID
  - Eliminazione eventuali valori nulli
  - Encoding di Genre

K-Means: determinare il numero di cluster

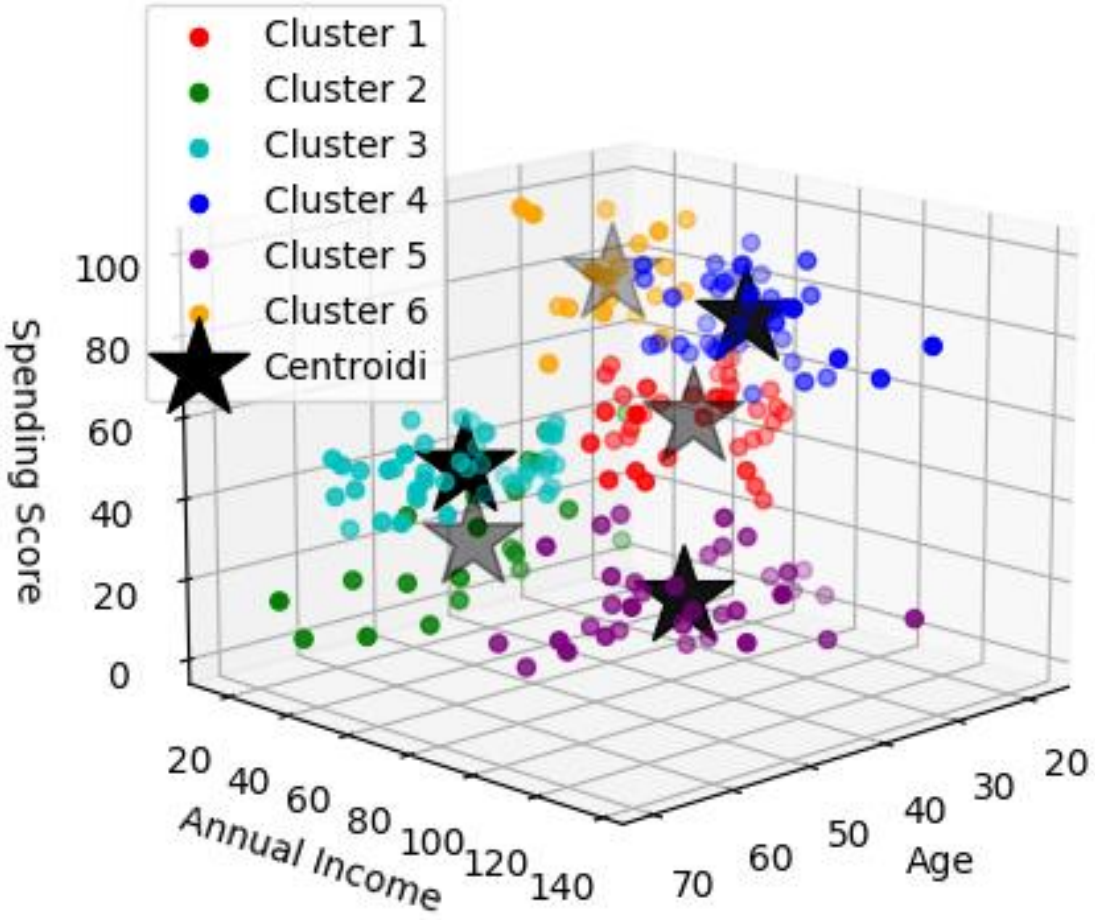
- SSE con elbow method
- Silhouette score: ricerca del massimo

→ Numero ottimale di cluster: 6.



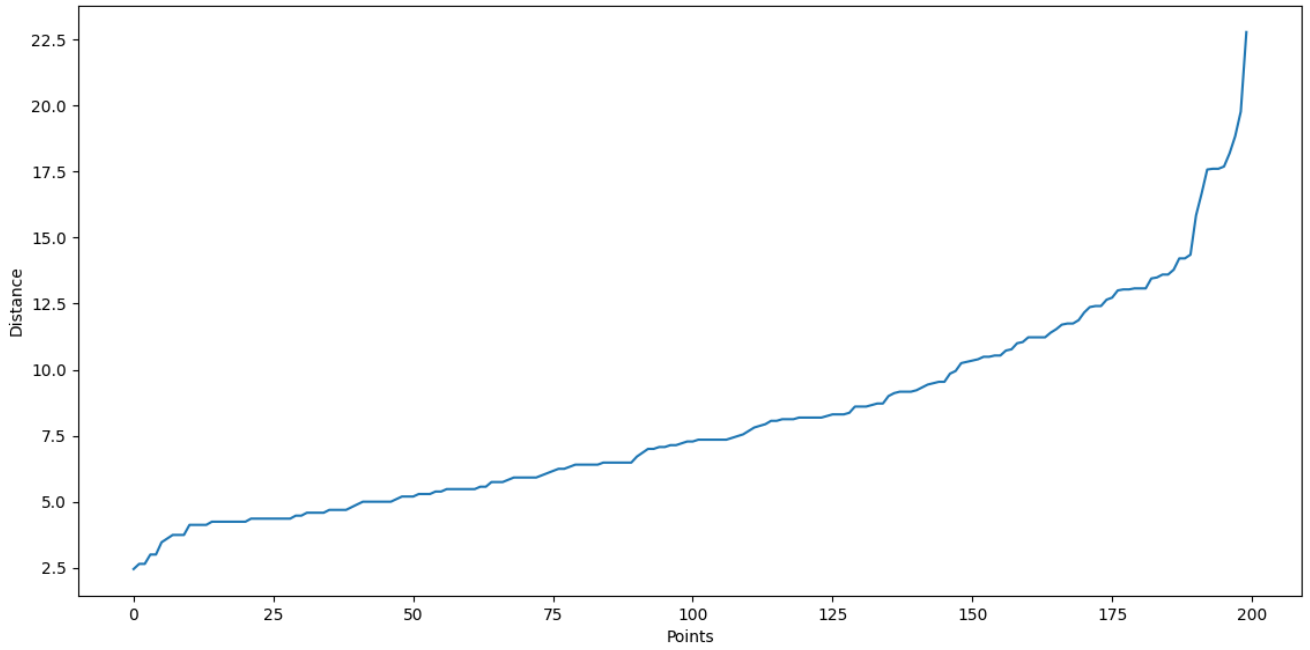


K-Means: risultati

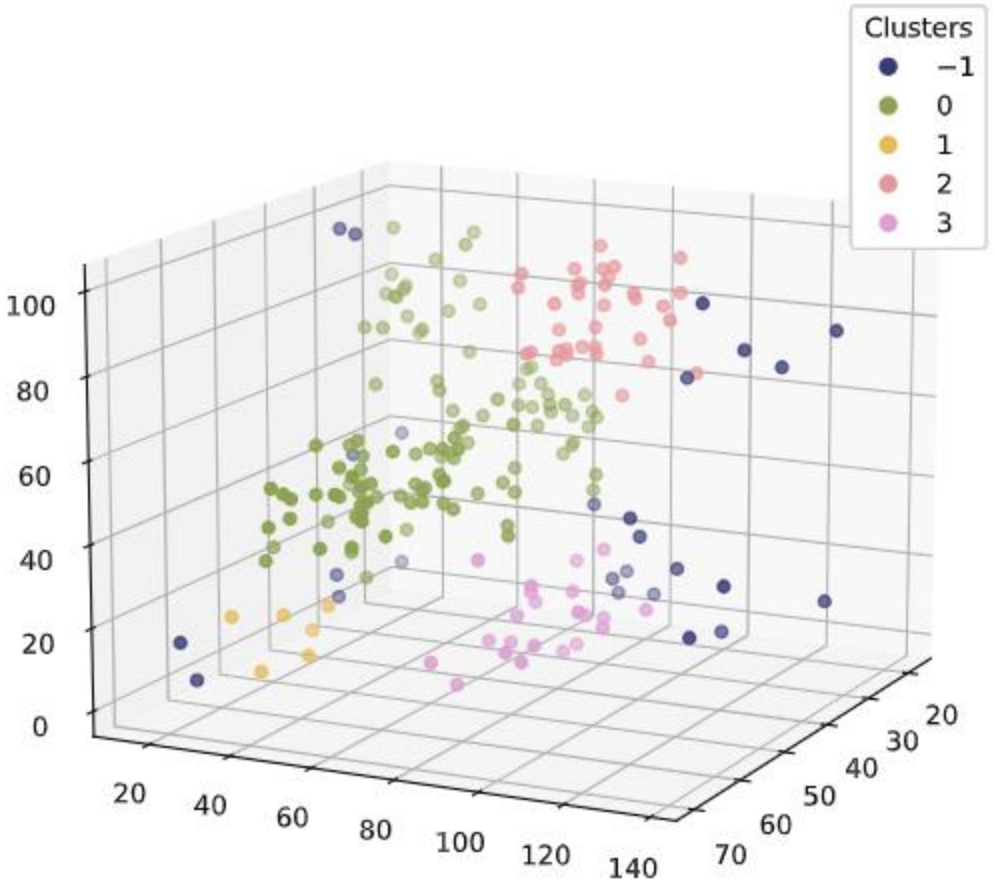


DBscan: ricerca di MinPts e Eps

- Ricerca del miglior **MinPts** con più esecuzioni e selezione i base al miglior silhouette score → 3
- KNN per trovare il miglior **Eps** → 14



DBscan: risultati



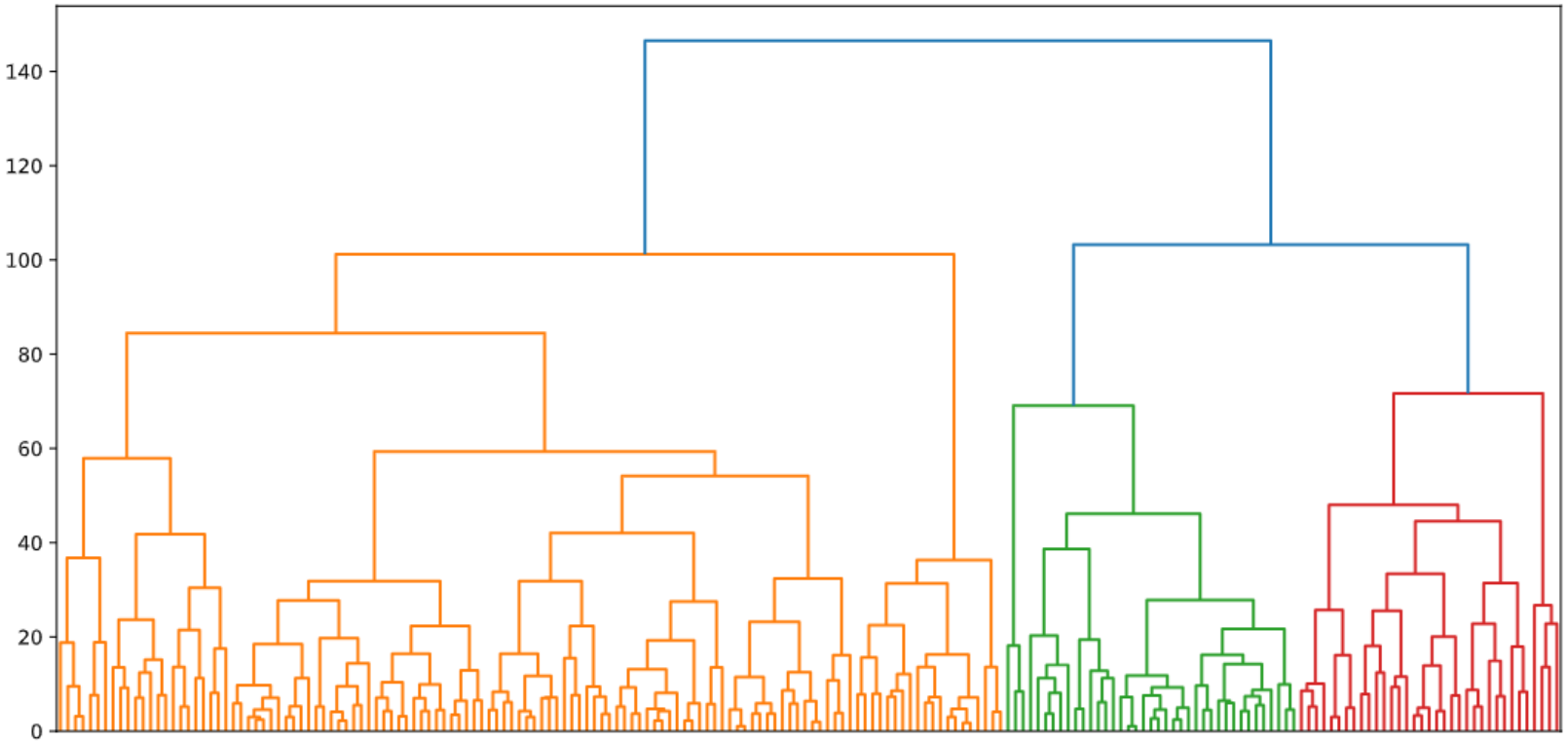
Clustering gerarchico agglomerativo

- MIN o Single Link
- MAX o Complete Linkage
- Group Average

- Taglio per ottenere 6 cluster e confrontare con K-Means
- È stato poi ricercata la migliore strategia in base al miglior **silhouette score**

→ Complete Linkage

Complete Linkage



Analisi dei cluster

- Analizzati i cluster di **K-Means** e **Complete Linkage**
- Cluster non identici ma con caratteristiche simili

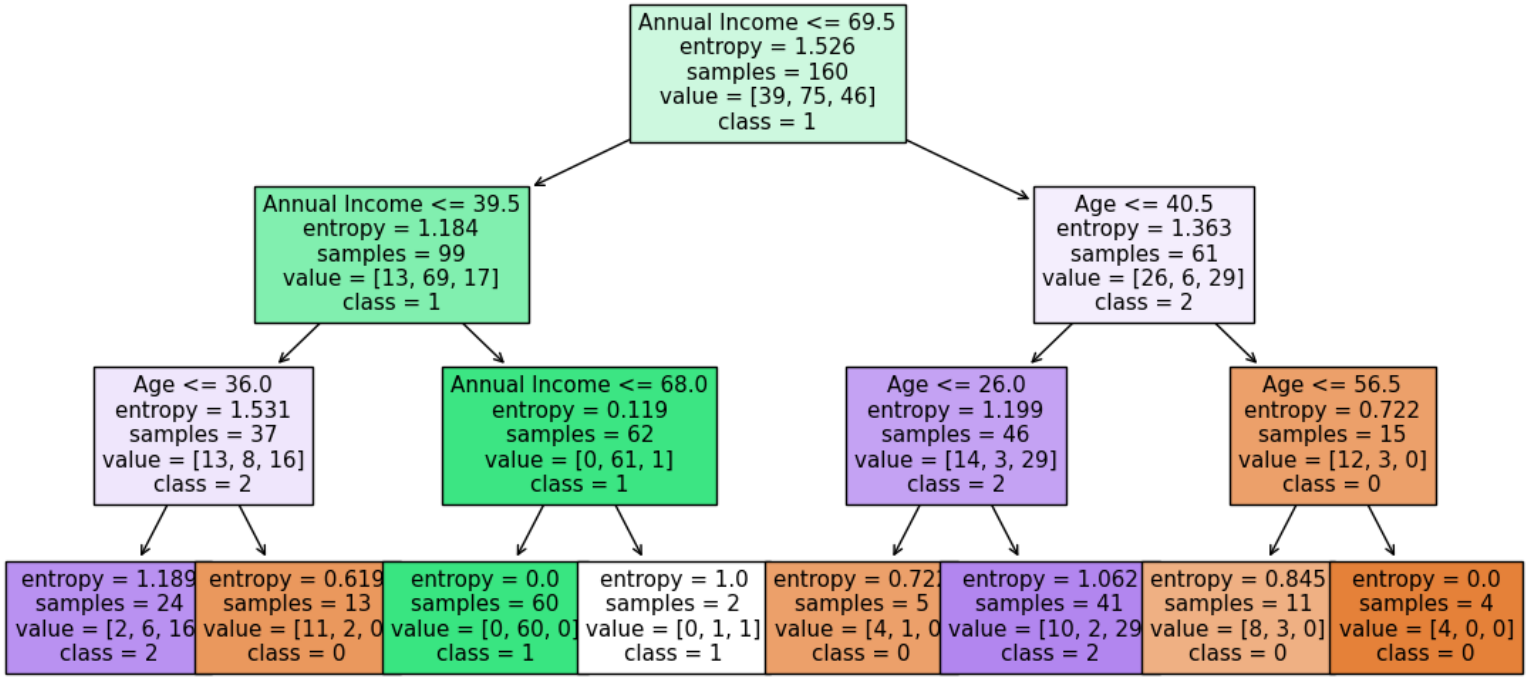
- Osservazioni:
  - Il sesso non è discriminante per il clustering
  - Non sempre lo spending score è in linea con l'annual income
  - Persone giovani con basso reddito e spending score alto
  - Viceversa, persone con alto reddito ma basso spending score

Decision Tree

- **Obiettivo**: modello per previsione dello spending score

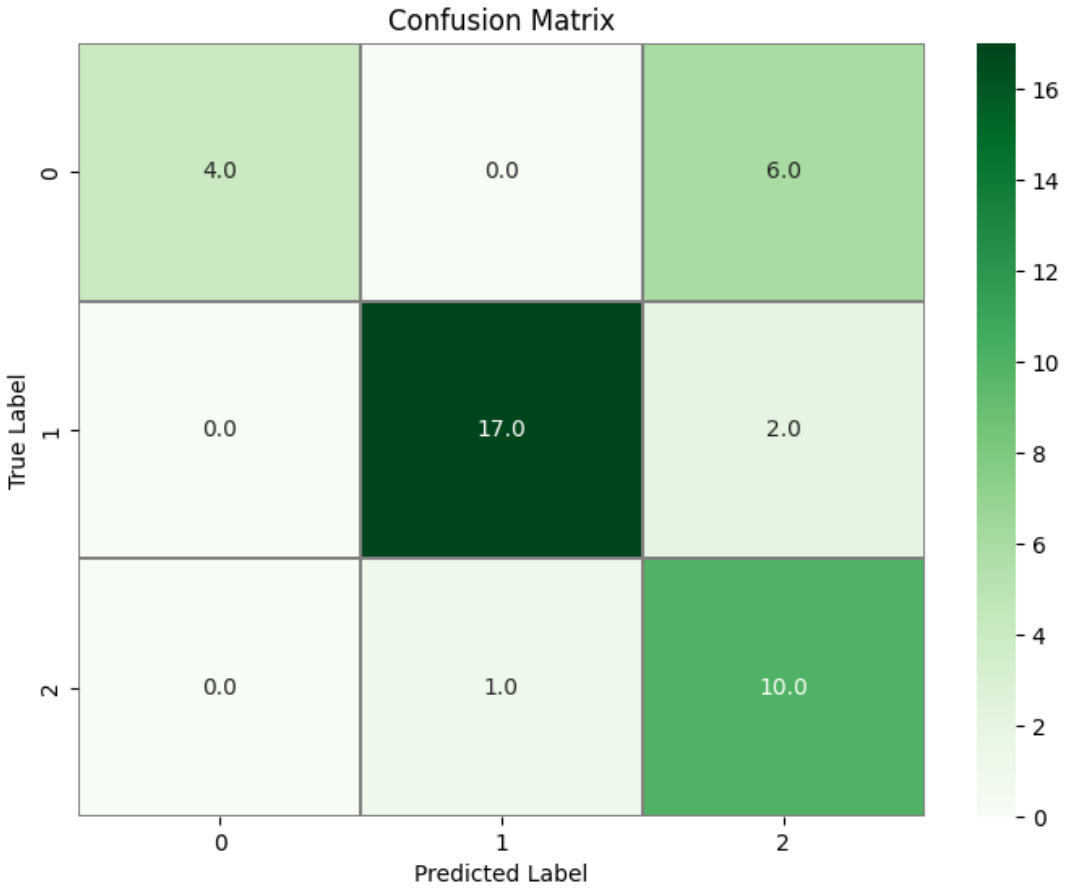
Decision Tree: miglior modello

Profondità massima 3, criterio dell'entropia e 3 classi di spending score.



Decision Tree: matrice di confusione

Accuracy: 0.79



Conclusioni

Clustering

- K-Means e Complete Linkage restituiscono risultati simili.
- Il sesso non influisce sui cluster.
- Non sempre il reddito si riflette sullo spending score.

Classificazione

- Miglior modello: 3 classi, profondità massima 3, criterio entropy
- Difficoltà nel classificare la classe 0.
- Possibili miglioramenti aumentando le dimensioni del dataset con nuove osservazioni o utilizzando altri modelli.

Grazie per l’attenzione.