Starry Night

Progetto di Machine Learning
A.A. 2023/2024

Moleri Andrea, 902011
Armani Filippo, 865939
Costantini Davide, 856114

Dominio di Riferimento ed Obiettivi

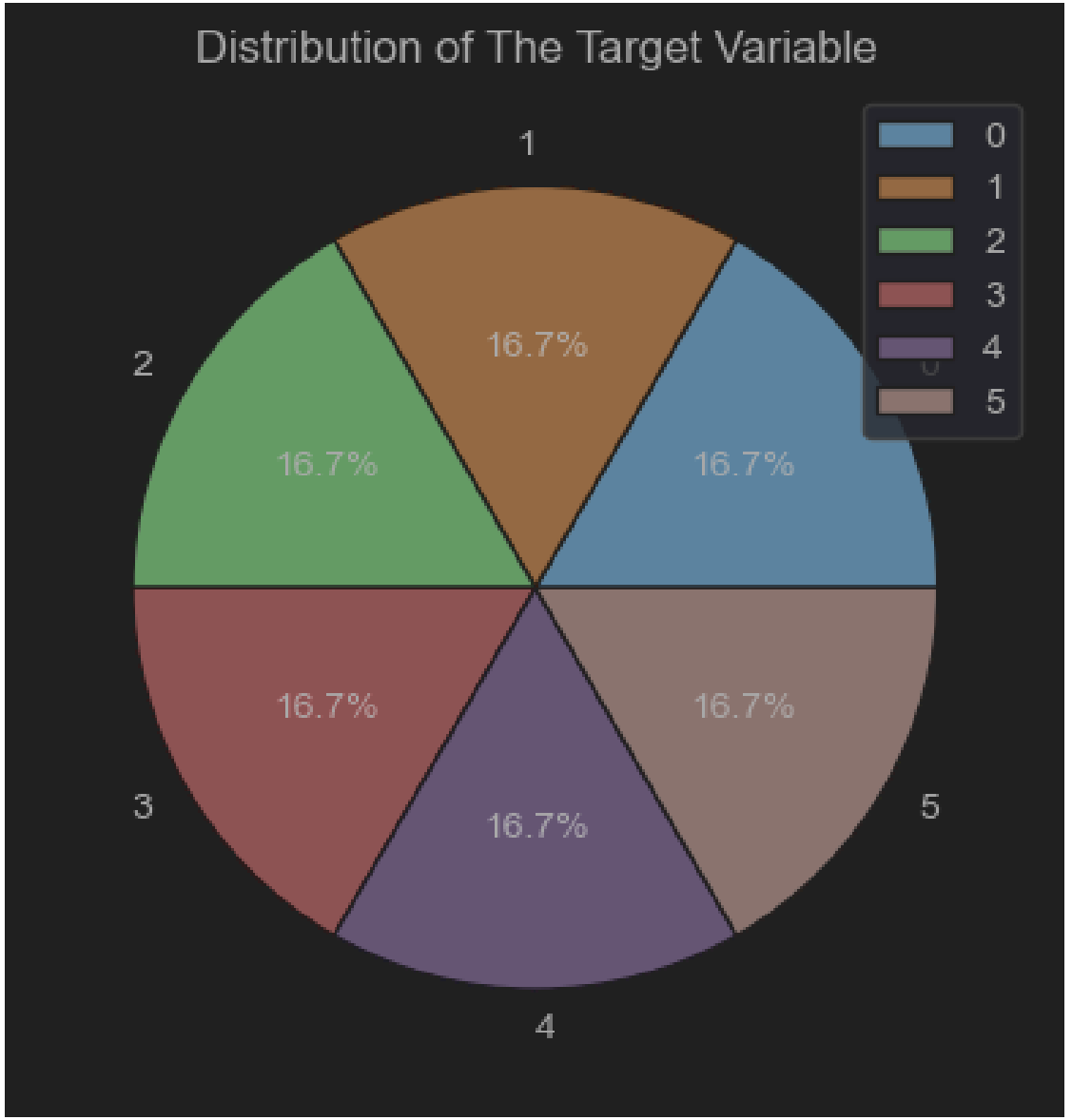
- Il dataset contiene varie informazioni su corpi celesti
  - Temperatura, luminosità, raggio, magnitudine, colore, classe spettrale e tipo
- Il nostro obiettivo è definire un modello per classificare il tipo delle stelle, seguendo la rappresentazione teorizzata dal diagramma di Hertzsprung-Russell

Descrizione del Dataset

- Il dataset, per ogni stella, contiene le seguenti informazioni:
  - Temperatura Assoluta
  - Luminosità Relativa (calcolata rispetto al sole)
  - Raggio Relativo (calcolato rispetto al sole)
  - Magnitudine Assoluta
  - Colore
  - Classe Spettrale
  - Tipo
- Il nostro target è il tipo della stella, "Star Type"
- Il set di test è stato creato prendendo il 30% dei valori del dataset, seguendo al 70/30 rule

Analisi Esplorativa

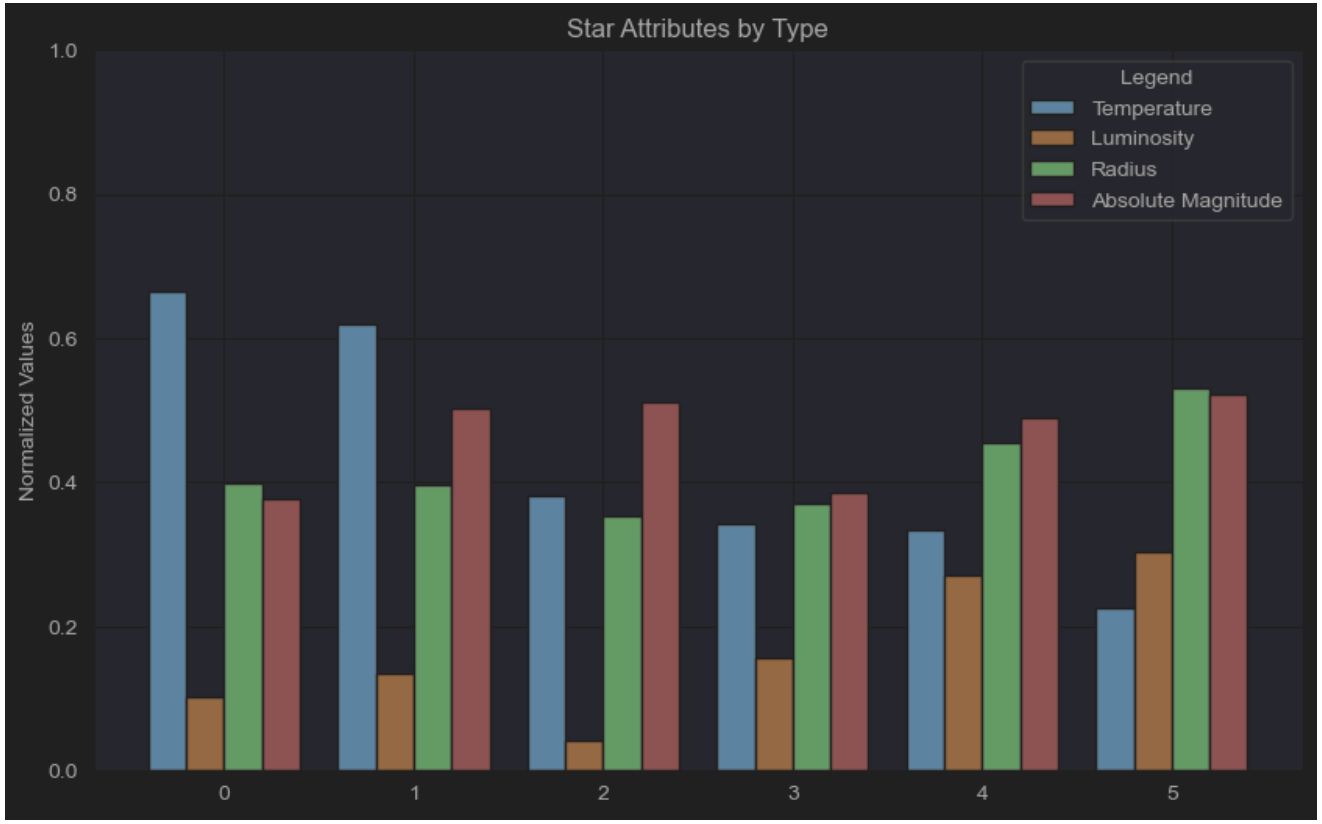
- Il dataset presenta 6 tipi di stelle
- I valori all'interno delle classi target si distribuiscono uniformemente, indicando un campionamento bilanciato nel dataset.



Analisi Esplorativa

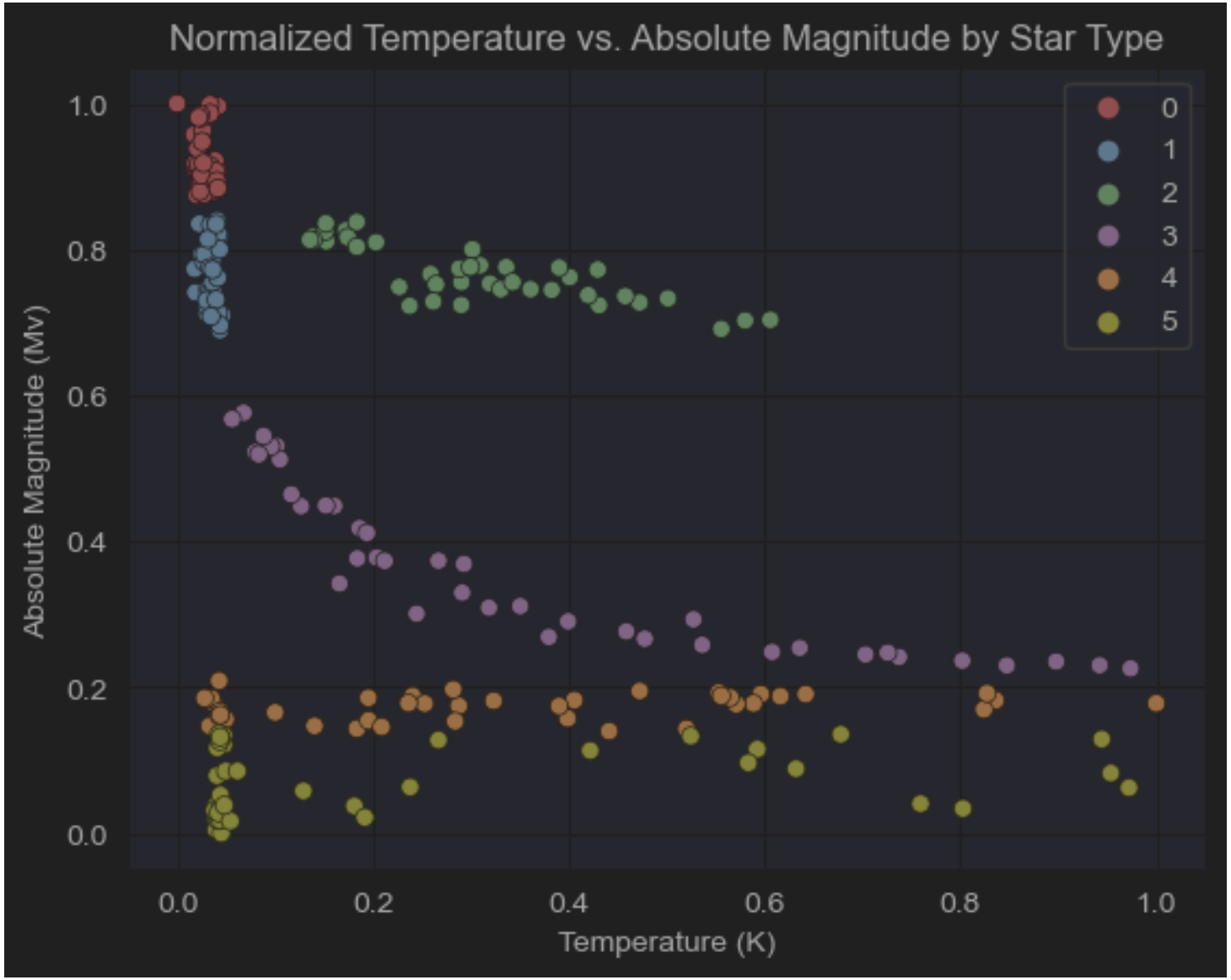
Dall’istogramma possiamo notare come alcuni valori, in particolare la Temperatura, mostrino differenze a seconda della classe della stella.

Dall’istogramma possiamo notare che certi attributi, in particolare le Temperature, sono ben distinti in base alle tipologie di stella ai quali appartengono



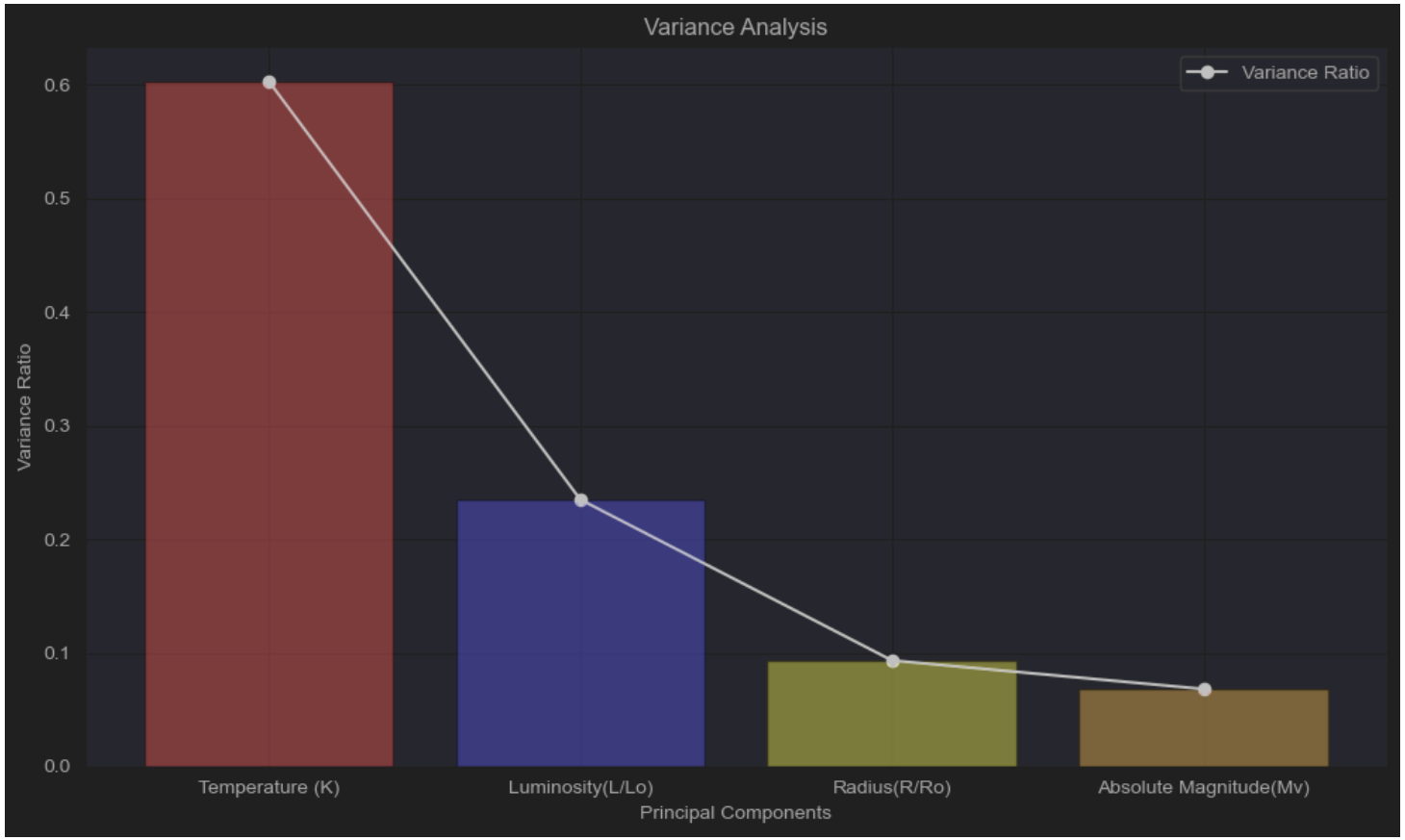
Analisi Esplorativa

Disponendo gli attributi in uno scatter plot, notiamo che la scelta degli attributi Magnitudine Assoluta e Temperatura crea la migliore separazione lineare per Star Type



Principal Component Analysis

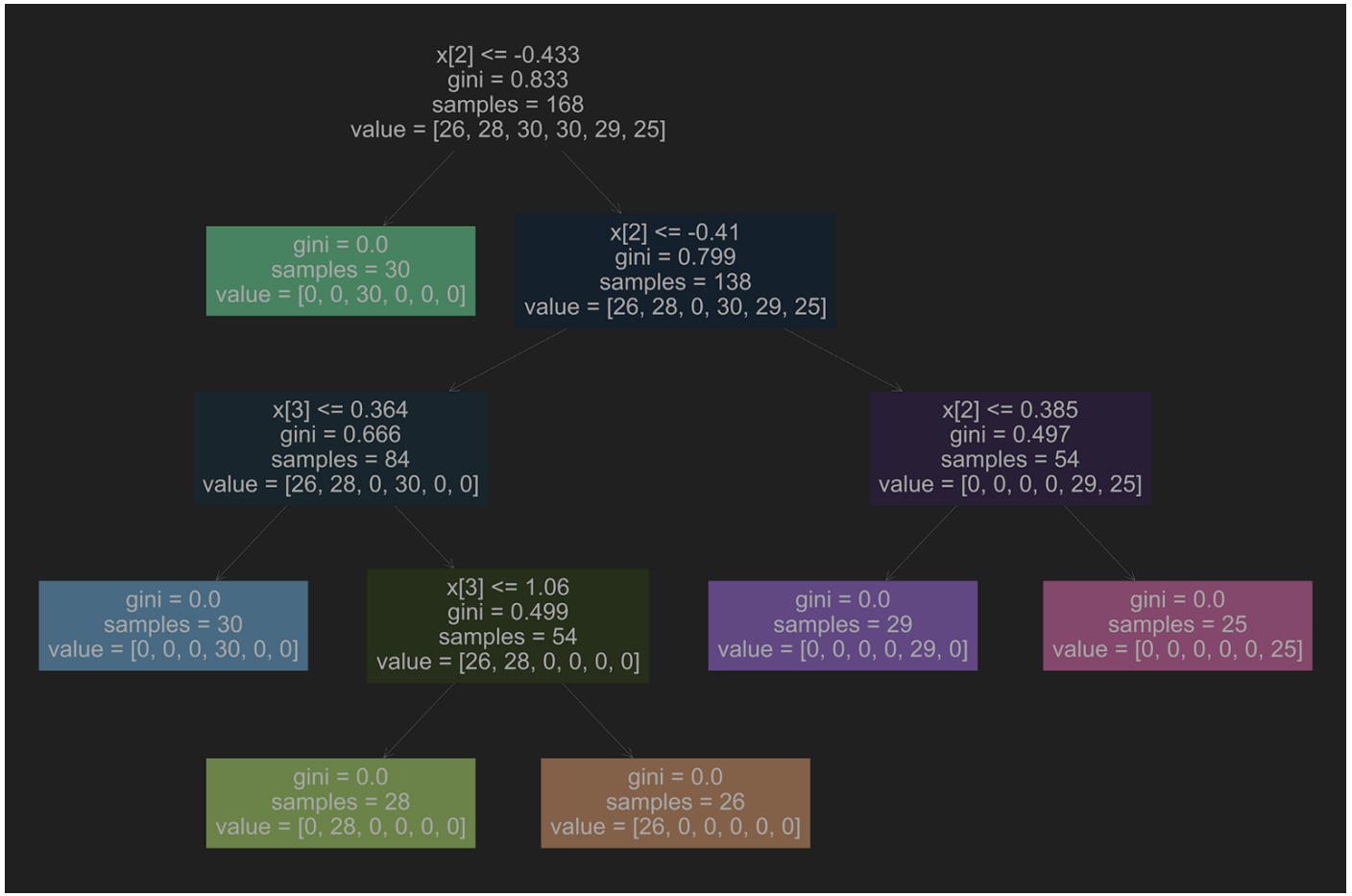
- Le varianze della luminosità, del raggio e della magnitudine assoluta indicano dei dati piuttosto omogenei
- La varianza della temperatura indica che sono relativamente omogenee, ma sono comunque presenti differenze significative



Primo Approccio: Albero di Decisione

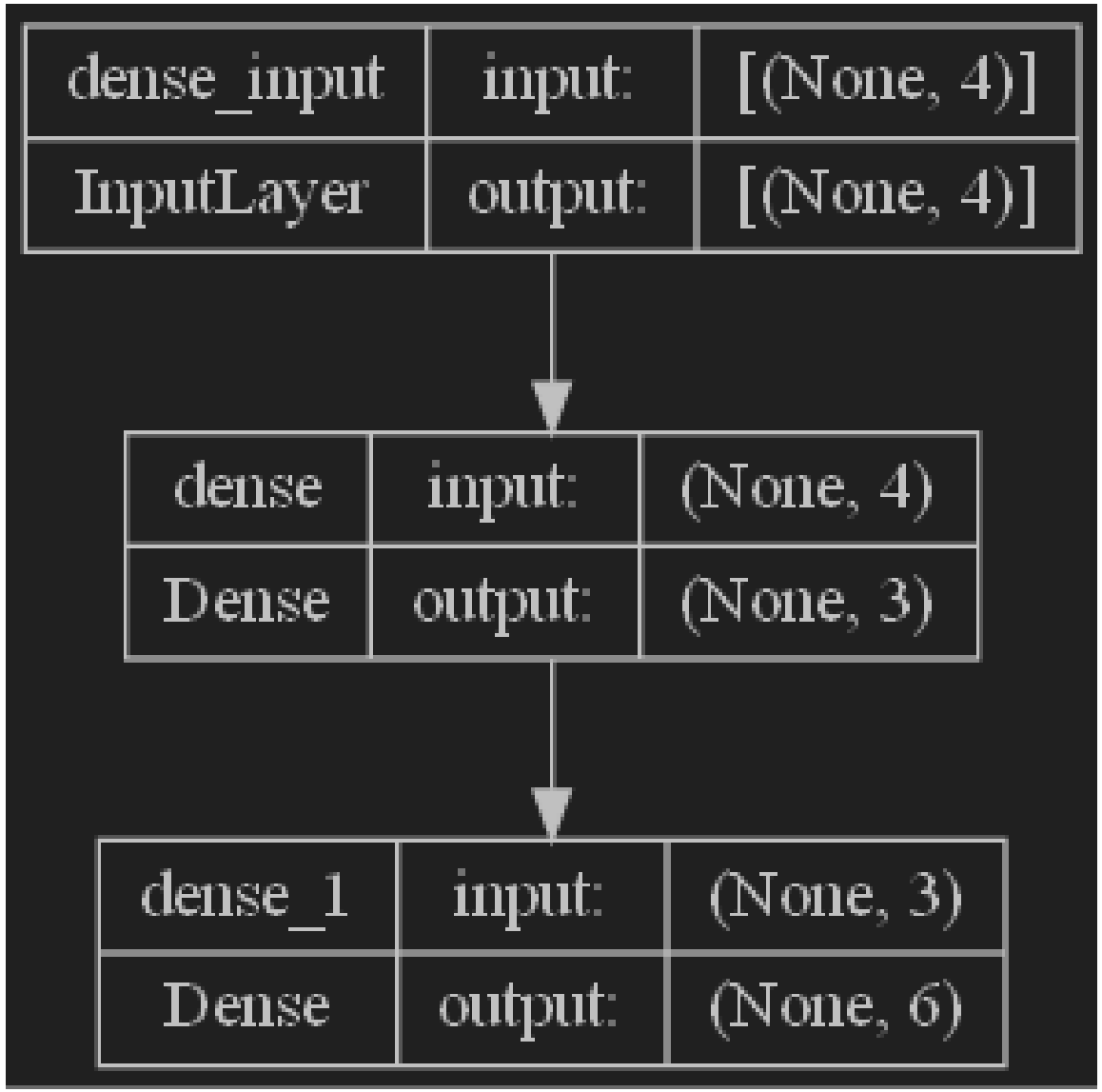
- Motivazioni
  - Eccellono nella manipolazione di dati categorici
  - Efficienza computazionale e scalabilità
  - Prevenzione dell’overfitting
- Il modello risultante ha una precisione del 100%

Albero di Decisione



Secondo Approccio: Reti Neurali

- Motivazioni:
  - Capacità di riconoscere pattern nei dati
  - Adattabili a diversi tipi di dati
  - Scalabile all'aumentare dei dati a disposizione
  - La maggior parte delle classi sono separate nello spazio, dunque facciamo uso degli iperpiani separatori
- La rete risultante è formata da tre strati
- Dopo 500 epoche, il modello ha una precisione del 95.83%



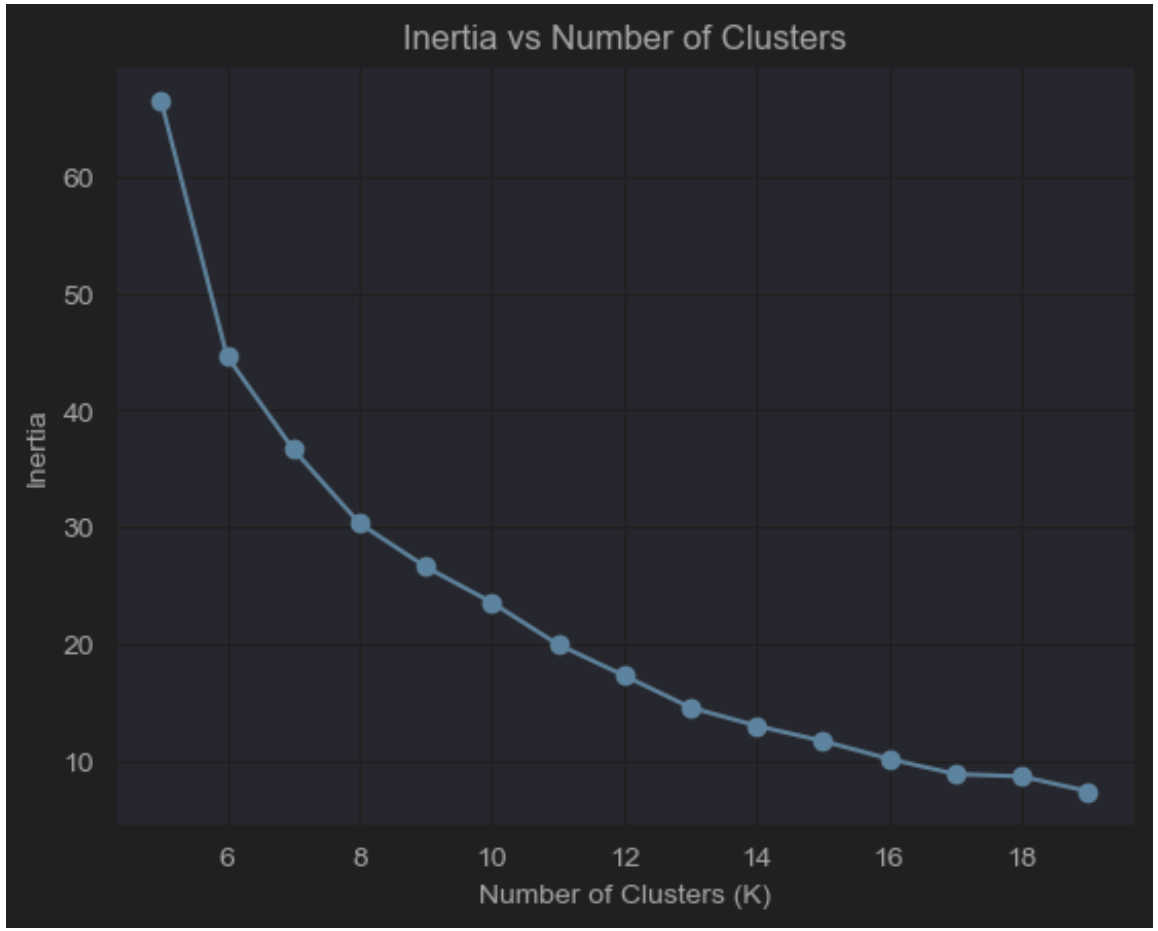
Terzo Approccio: Clustering

- Motivazioni:
  - Eccelle nel separare le feature del dataset
  - Vantaggioso con dataset in continua evoluzione
  - La maggior parte delle classi sono separate nello spazio, quindi è possibile operare tramite raggruppamento
- Feature utilizzate: magnitudine assoluta, raggio e temperatura
- Il modello risultante ha una precisione del 94.44%
- Matrice di confusione:

```
[[14  0  0  0  0  0]
 [ 0 12  0  0  0  0]
 [ 0  0 10  0  0  0]
 [ 0  0  0  8  2  0]
 [ 0  0  0  2  9  0]
 [ 0  0  0  0  0 15]]
```

Clustering: Elbow Method

- Applichiamo al grafico il metodo del gomito
- Non potendo identificare il numero di cluster, applichiamo un altro metodo

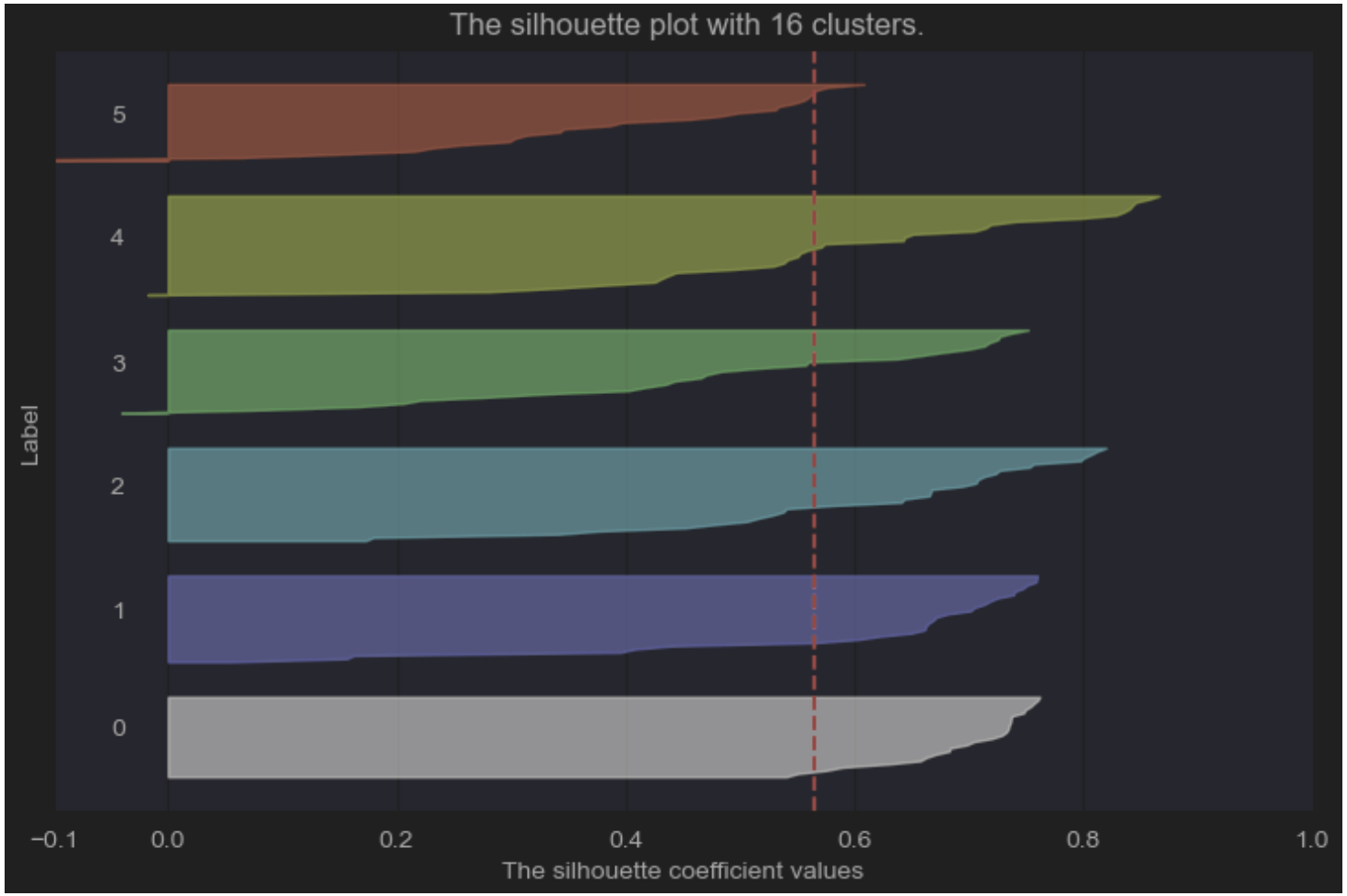


Clustering: Analisi della silhouette

Criteri:

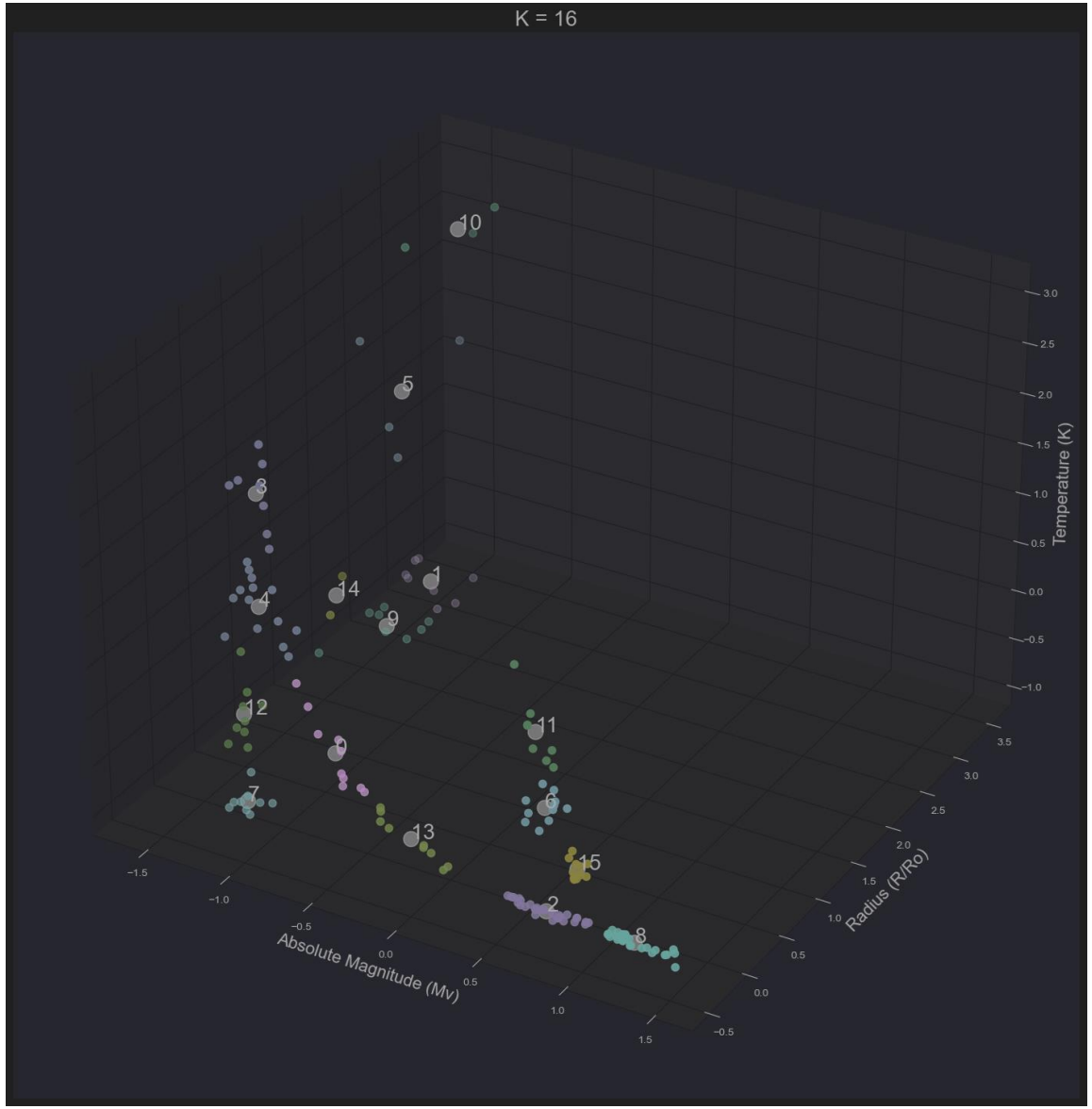
- Suddivisione del training set in label uniformi
- Per ogni label il valore massimo deve essere maggiore della linea rossa
- Nel caso di grafici simili viene scelto quello con la silhouette media più grande

Utilizziamo 16 cluster



Clustering: Analisi dei Dati

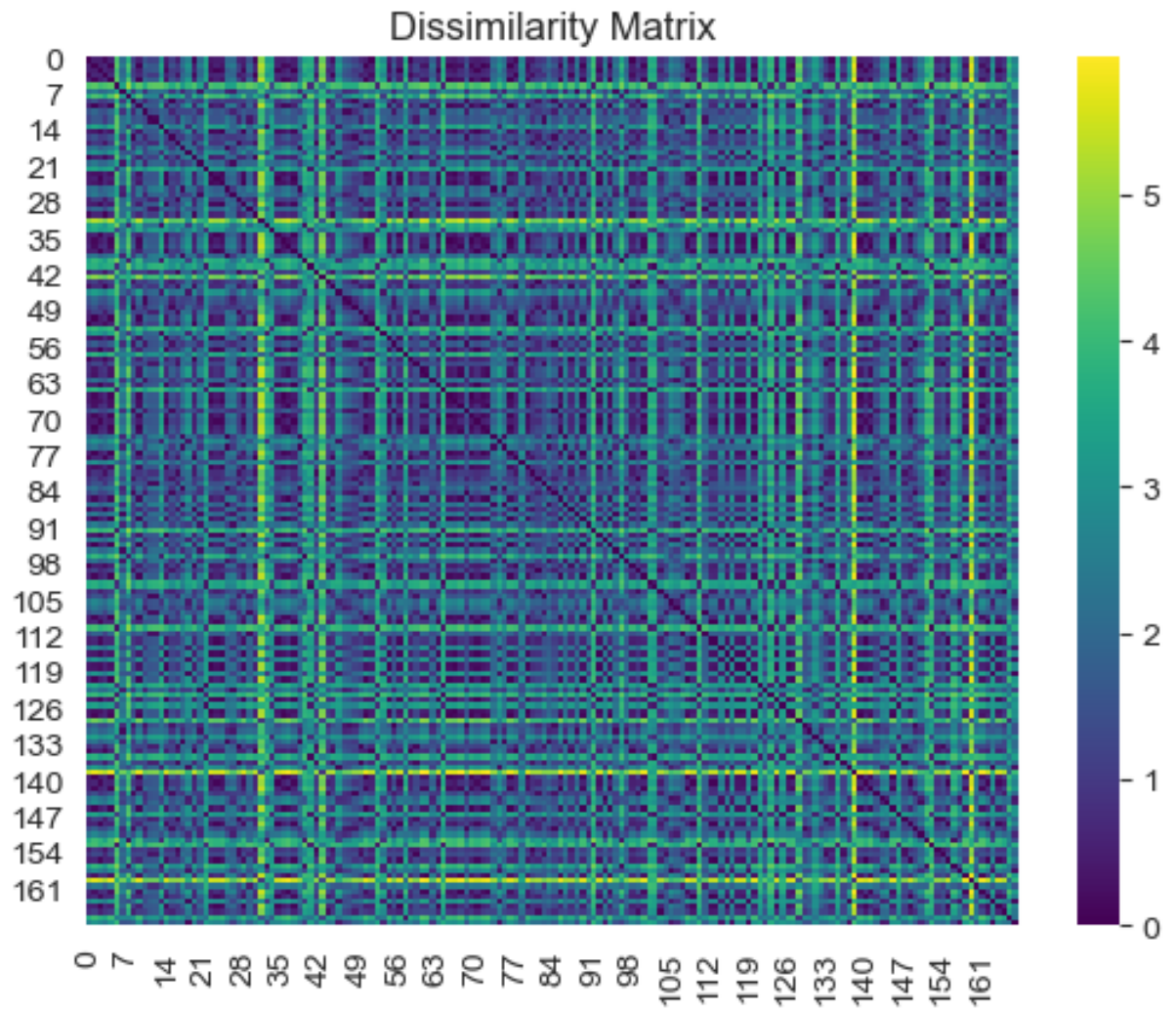
Disponendo i punti in uno spazio tridimensionale abbiamo una rappresentazione della suddivisione in cluster.



Clustering: Matrice di Dissimilarità

La matrice di dissimilarità ci permette di rappresentare la dissimilarità all'interno del dataset, evidenziando pattern comuni.

Una diagonale di colore scuro indica che i cluster sono omogenei.



Analisi dei Risultati Ottenuti

- Parametri utilizzati
  - Accuratezza
  - Richiamo
  - F1-score
  - Tempo di training
- L’albero di decisione è il modello più preciso
  - La rete neurale ed il clustering sono quasi equivalenti
- La rete neurale è il modello più lento nella fase di training
  - L’albero decisionale è il modello più veloce
- Gli errori di classificazione sono dati dal fatto che alcuni attributi presentano per classi diverse valori simili

Analisi dei Risultati Ottenuti

	precision	recall	f1-score	support
0	1.00	1.00	1.00	14
1	1.00	1.00	1.00	12
2	1.00	1.00	1.00	10
3	1.00	1.00	1.00	10
4	1.00	1.00	1.00	11
5	1.00	1.00	1.00	15
accuracy			1.00	72
macro avg	1.00	1.00	1.00	72
weighted avg	1.00	1.00	1.00	72

Albero decisionale
Tempo: 0.01s

	precision	recall	f1-score	support
0	1.00	1.00	1.00	14
1	1.00	1.00	1.00	12
2	1.00	1.00	1.00	10
3	0.90	1.00	0.95	9
4	1.00	0.92	0.96	12
5	1.00	1.00	1.00	15
accuracy			0.99	72
macro avg	0.98	0.99	0.98	72
weighted avg	0.99	0.99	0.99	72

Reti Neurale
Tempo: 41.68s

	precision	recall	f1-score	support
0	1.00	1.00	1.00	14
1	1.00	1.00	1.00	12
2	1.00	1.00	1.00	10
3	0.90	0.82	0.86	11
4	0.82	0.90	0.86	10
5	1.00	1.00	1.00	15
accuracy			0.96	72
macro avg	0.95	0.95	0.95	72
weighted avg	0.96	0.96	0.96	72

K-Means
Tempo: 0.22s