
Alberi decisionali: spiegazione semplice + matematica di base

Gli **alberi decisionali** sono modelli di machine learning che prendono decisioni attraverso una serie di domande, creando "rami" che portano a una previsione (classificazione o regressione).

Come funziona un albero decisionale?

- A ogni nodo, l'albero sceglie **la domanda migliore** da fare sui dati (ad esempio: "le assenze sono più di 7?").
- Il percorso di ciascun esempio segue le risposte ("sì/no") fino ad arrivare a una **foglia**, che rappresenta la previsione finale.

Un po' di matematica: come decide le domande?

L'albero cerca di **dividere i dati** in modo che, dopo ogni domanda,
i gruppi siano il più "puri" possibile (cioè, con esempi della stessa classe).

Per misurare questa "purezza", si usano **indici matematici** come:

Indice di Gini (per classificazione):

$$Gini = 1 - \sum_{i=1}^{C} p_i^2$$

- Dove p_i è la proporzione di elementi della classe i nel nodo.
- **Esempio:**
  - Se in un nodo ci sono solo "promossi", Gini=0 (puro).
  - Se il nodo è mezzo "promossi" e mezzo "bocciati", Gini=0.5 (massimo miscuglio per 2 classi).

Entropia (altra misura di impurità):

$$Entropia = -\sum_{i=1}^{C} p_i \log_2 p_i$$

- Anche qui, valori bassi = nodo puro.

Come si costruisce l'albero?

1. **Per ogni domanda possibile** (es: assenze > X?), calcola quanto "migliora" la purezza dei gruppi creati.
2. **Scegli la domanda** che porta al miglioramento più grande.

3. **Ripeti** per ogni ramo finché i nodi sono “abbastanza puri” o si raggiunge un limite (profondità, numero minimo di casi, ecc.).

In sintesi:

- Gli alberi decisionali dividono i dati a tappe con domande “ottimali”.
- Usano indici come il **Gini** o l’**entropia** per scegliere le domande migliori.
- Sono facili da interpretare e molto utili per spiegare “come decide il modello”.

Mini esempio di calcolo dell’indice di **Gini** per un nodo di un albero decisionale, spiegato passo passo:

Immagina un nodo con 10 studenti:

- 7 sono **promossi**
- 3 sono **bocciati**

Calcolo della “purezza” (indice di Gini):

La formula è:

$$Gini = 1 - \sum_{i=1}^{C} p_i^2$$

dove:

- *C* = numero di classi (qui 2: promosso, bocciato)
- p_i = proporzione di elementi della classe i nel nodo

---

1. Calcolo delle proporzioni:

- $p_{promosso} = \frac{7}{10} = 0.7$
- $p_{bocciato} = \frac{3}{10} = 0.3$

---

2. Calcolo del Gini:

$$Gini = 1 - (0.7^2 + 0.3^2) = 1 - (0.49 + 0.09) = 1 - 0.58 = 0.42$$

Interpretazione

- **Gini = 0**: nodo perfettamente puro (solo una classe)
- **Gini = 0.5**: massimo miscuglio per due classi (es: 5 promossi, 5 bocciati)
- **Gini = 0.42**: il nodo è "abbastanza" puro, ma c'è ancora un po' di miscuglio

---

Altro esempio "estremo"

Se nel nodo ci fossero **tutti promossi** (10 su 10):

- $p_{promosso} = 1$
- $p_{bocciato} = 0$

$$Gini = 1 - (1^2 + 0^2) = 1 - 1 = 0$$
