
La matematica dietro Random Forest (in modo semplice)

- Una **Random Forest** è un insieme di tanti **alberi decisionali** quindi la matematica è la stessa degli alberi decisionali singoli.
- Ogni **albero** si costruisce dividendo i dati in rami usando "domande" del tipo:
  "Le assenze sono più di 6?" oppure *"Le ore di studio sono almeno 5?"*
- **Come decide dove dividere?**
  L'albero sceglie le domande che **separano meglio i dati**.
  Per misurare quanto una divisione è "buona", usa delle formule di **purezza**, come l'**indice di Gini** o l'**entropia**.
- Ad esempio, **l'indice di Gini** misura quanto sono "misti" i dati in un nodo:

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

dove p_i è la percentuale di ogni classe nel nodo. Più il valore è basso, più il nodo è "puro".

- Ogni albero della foresta viene costruito su **dati scelti a caso** e usando **solo alcune variabili** scelte casualmente a ogni divisione ("random").
- Quando si deve prevedere una risposta, **ogni albero vota**:
  la classe che ottiene più voti (maggioranza) è la risposta finale della foresta.

In sintesi:

La Random Forest usa la matematica per dividere i dati nel modo più "ordinato" possibile,
poi combina tante decisioni diverse per ottenere una previsione più solida e affidabile.

Vuoi anche un esempio numerico sul calcolo dell'indice di Gini o entropia per una divisione?