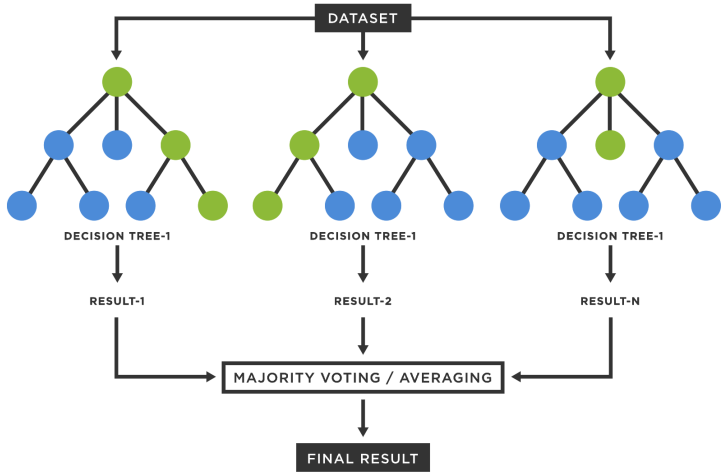
Coding e Big Data

2024-2025

Vincenzo Nardelli

vincenzo.nardelli@unicatt.it

Cos'è una Random Forest



Cos'è una Random Forest

- La Random Forest è un metodo che utilizza tanti piccoli alberi di decisione per fare previsioni più accurate.
- Ogni albero dà la sua previsione, e la Random Forest combina tutte le risposte per scegliere la migliore.
- Gli alberi vengono costruiti usando:
  - Dati scelti a caso dal dataset.
  - Alcune variabili scelte a caso per ogni divisione.
- Questo approccio rende la Random Forest stabile e precisa.

Differenze rispetto a un Albero Singolo

- **Albero Singolo**:
  - Facile da capire, ma può adattarsi troppo ai dati, dando risultati meno affidabili (overfitting).
  - Sensibile ai dati rumorosi o piccoli.
- **Random Forest**:
  - Combina tanti alberi per ridurre il rischio di errori.
  - Funziona meglio su dati complessi e variabili.

Esempio: Albero di Decisione

- Modello costruito con 10 nodi terminali.
- Residual mean deviance = 1.017.
- Misclassification error rate = 24%.

Confusion Matrix

	No	Yes
No	191	51
Yes	45	113

- Accuratezza = $\frac{191+113}{400}$ = 76%.

Esempio: Random Forest

- Costruita con 500 alberi e 2 variabili selezionate a ogni split.
- **Errore OOB (Out-of-Bag)**:
  - L'errore OOB stima la performance del modello utilizzando campioni non selezionati (out-of-bag) durante il bootstrapping.
  - Funziona come una valida approssimazione per l'errore su un set di test indipendente.
  - Errore OOB stimato = 32.75%.

Confusion Matrix (OOB)

	No	Yes
No	176	60
Yes	71	93

- Accuratezza stimata OOB = $\frac{176+93}{400}$ = 67.25%.

Esempio: Random Forest

- **Accuratezza finale sul training set**:
  - Previsioni sul training set mostrano un'accuratezza del 99.5%.
  - Confusion Matrix:

Confusion Matrix (Training Set)

	No	Yes
No	236	2
Yes	0	162

- Accuratezza finale = $\frac{236+162}{400}$ = 99.5%.

Importanza delle Variabili

- La Random Forest fornisce metriche per valutare l'importanza delle variabili:
  - **Mean Decrease Accuracy (MDA)**: Riduzione dell'accuratezza media eliminando la variabile.
  - **Mean Decrease Gini (MDG)**: Riduzione media della devianza degli split per ogni variabile.

Variabile	No	Yes	MDA	MDG
Price	19.28	22.46	27.37	76.01
Advertising	16.13	18.04	26.55	45.03
Income	4.13	3.54	5.34	57.87
US	7.02	3.15	9.02	5.27

Vantaggi della Random Forest

- Alta accuratezza anche con dataset complessi.
- Robusta rispetto a valori mancanti e outlier.
- Riduce il rischio di overfitting grazie alla media delle previsioni.
- Fornisce una stima dell'importanza delle variabili.

Svantaggi della Random Forest

- Meno interpretabile rispetto a un singolo albero.
- Richiede più risorse computazionali.
- Può perdere precisione su dati altamente sbilanciati senza tecniche di bilanciamento adeguate.

Conclusioni

- La Random Forest è un potente strumento per problemi di classificazione e regressione.
- Offre una migliore accuratezza rispetto agli alberi singoli.
- Fornisce insight sull'importanza delle variabili, utile per l'analisi esplorativa e il feature engineering.