

Metodi Statistici per le decisioni

2024-2025

Vincenzo Nardelli



vincenzo.nardelli@unicatt.it

Alberi decisionali

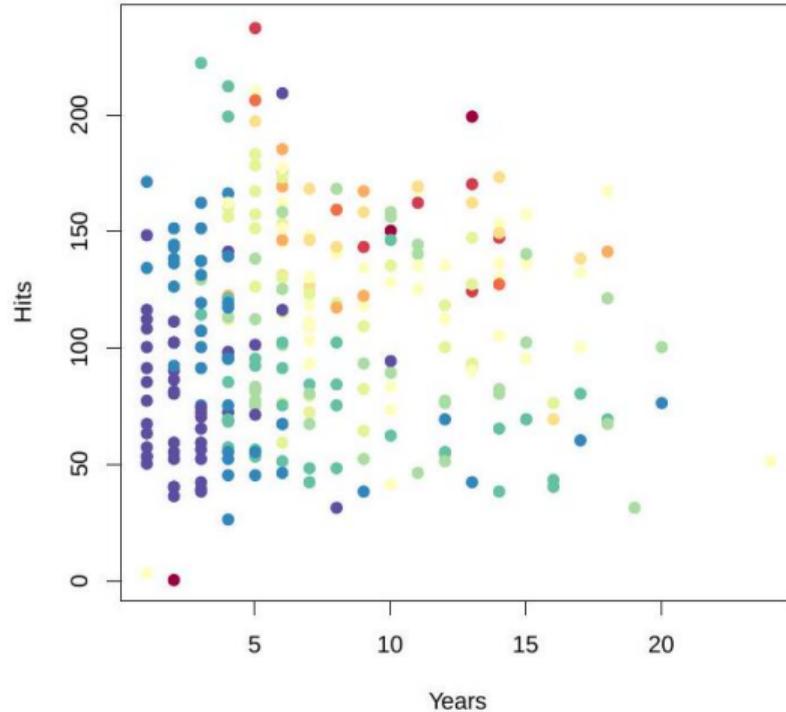
- ▶ I metodi basati sugli alberi decisionali prevedono di stratificare o segmentare lo spazio dei predittori in un certo numero di regioni semplici.
- ▶ Poiché l'insieme delle regole di divisione utilizzate per segmentare lo spazio dei predittori può essere riassunto in un albero, questi approcci sono noti come metodi basati sugli alberi decisionali.
- ▶ È possibile utilizzare questi metodi sia per regressione che classificazione.

Pro e Contro

- ▶ I metodi basati sugli alberi sono semplici e utili per l'interpretazione.
- ▶ Tuttavia, generalmente non sono competitivi con i nuovi approcci di apprendimento supervisionato in termini di accuratezza predittiva.
- ▶ Per questo motivo, discutiamo anche di bagging e foreste casuali (random forests). Questi metodi crescono molti alberi che vengono poi combinati per produrre una singola previsione di consenso.
- ▶ La combinazione di un gran numero di alberi può spesso migliorare notevolmente l'accuratezza predittiva, a scapito di una perdita di interpretabilità.

Dati sugli stipendi nel baseball: come li stratificheresti?

Gli stipendi sono codificati per colore dal basso blu, verde, giallo, rosso...



Albero decisionale per questi dati

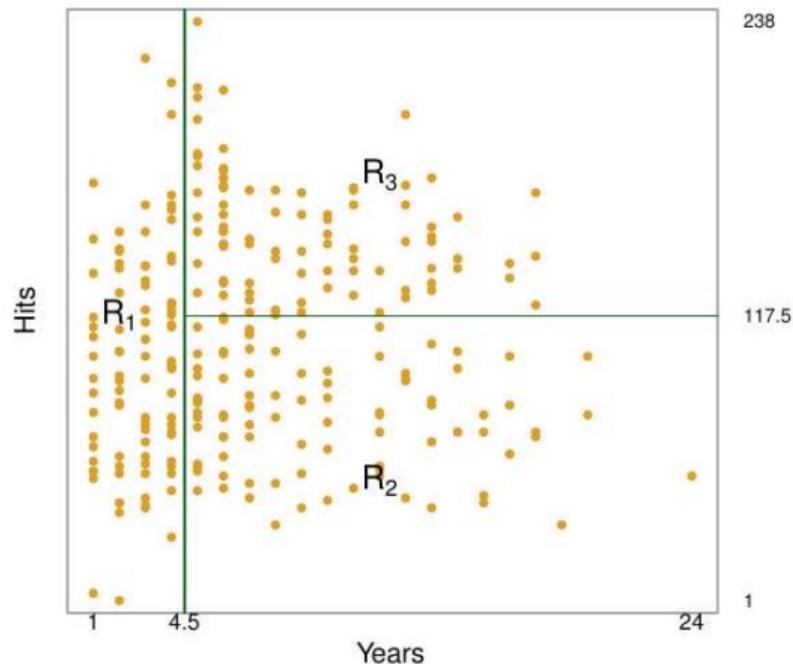


Dettagli della figura precedente

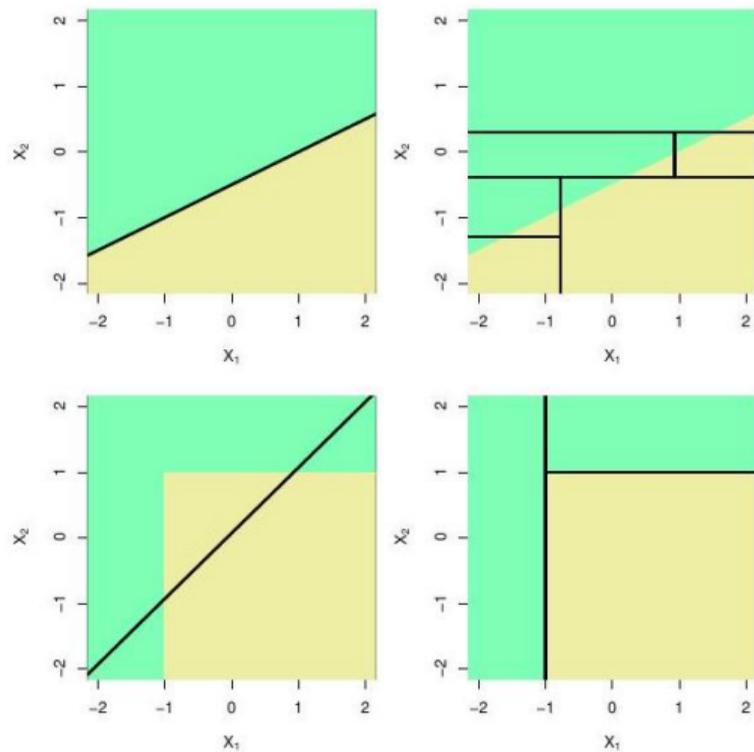
- ▶ Per i dati degli Hitters, un albero di regressione predice il logaritmo dello stipendio di un giocatore di baseball, basandosi sul numero di anni trascorsi nelle leghe maggiori e sul numero di battute valide (hits) effettuate nell'anno precedente.
- ▶ In un dato nodo interno, l'etichetta (del tipo $X_j < t_k$) indica che il ramo sinistro si riferisce alla condizione $X_j < t_k$, mentre il ramo destro corrisponde a $X_j \geq t_k$. Ad esempio, la divisione in cima all'albero produce due grandi rami: il ramo sinistro corrisponde a $\text{Years} < 4.5$, e il ramo destro corrisponde a $\text{Years} \geq 4.5$.
- ▶ L'albero ha due nodi interni e tre nodi terminali, o foglie. Il numero in ciascuna foglia rappresenta la media della risposta per le osservazioni che vi appartengono.

Risultati

- Complessivamente, l'albero stratifica o segmenta i giocatori in tre regioni dello spazio dei predittori: $R_1 = \{X \mid \text{Years} < 4.5\}$, $R_2 = \{X \mid \text{Years} \geq 4.5, \text{Hits} < 117.5\}$, e $R_3 = \{X \mid \text{Years} \geq 4.5, \text{Hits} \geq 117.5\}$.



Alberi vs Modelli Lineari



Terminologia per gli alberi

- ▶ Seguendo l'analogia dell'albero, le regioni R_1 , R_2 e R_3 sono note come nodi terminali.
- ▶ Gli alberi decisionali sono solitamente rappresentati al contrario, nel senso che le foglie sono in basso nell'albero.
- ▶ I punti lungo l'albero dove lo spazio dei predittori viene diviso sono detti nodi interni.
- ▶ Nell'albero degli Hitters, i due nodi interni sono indicati dal testo $\text{Years} < 4.5$ e $\text{Hits} < 117.5$.

Interpretazione dei risultati

- ▶ Il numero di anni di esperienza (Years) è il fattore più importante nel determinare lo stipendio, e i giocatori con meno esperienza guadagnano stipendi più bassi rispetto a quelli più esperti.
- ▶ Dato che un giocatore è meno esperto, il numero di battute valide effettuate nell'anno precedente sembra avere poca influenza sullo stipendio.
- ▶ Ma tra i giocatori che sono stati nelle leghe maggiori per cinque o più anni, il numero di battute valide effettuate nell'anno precedente influisce sullo stipendio, e i giocatori che hanno realizzato più battute valide tendono a guadagnare di più.
- ▶ Sicuramente una semplificazione, ma rispetto a un modello di regressione è facile da visualizzare, interpretare e spiegare.

Dettagli del processo di costruzione dell'albero

1. Dividiamo lo spazio dei predittori, cioè l'insieme dei valori possibili per X_1, X_2, \dots, X_p , in J regioni distinte e non sovrapposte, R_1, R_2, \dots, R_J .
2. Per ogni osservazione che rientra nella regione R_j , facciamo la stessa previsione, che è semplicemente la media dei valori di risposta per le osservazioni di training in R_j .

Ulteriori dettagli sul processo di costruzione dell'albero

- ▶ In teoria, le regioni potrebbero avere qualsiasi forma. Tuttavia, scegliamo di dividere lo spazio dei predittori in rettangoli o box ad alta dimensione, per semplicità e facilità di interpretazione del modello predittivo risultante.
- ▶ L'obiettivo è trovare i box R_1, \dots, R_J che minimizzano l'RSS, dato da

$$\sum_{j=1}^J \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

dove \hat{y}_{R_j} è la media della risposta per le osservazioni di training all'interno della scatola j .

Ulteriori dettagli sul processo di costruzione dell'albero

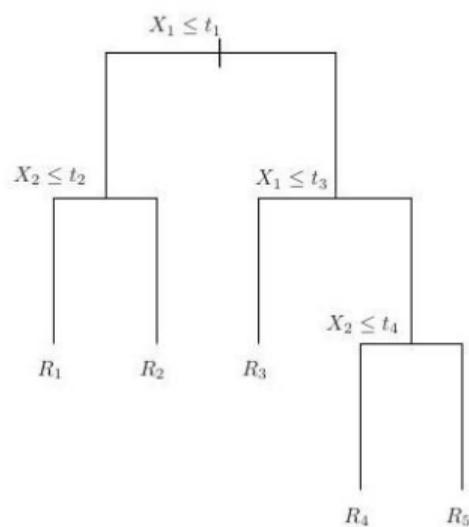
- ▶ Purtroppo, è computazionalmente impossibile considerare tutte le possibili partizioni dello spazio dei predittori in J box.
- ▶ Per questo motivo, adottiamo un approccio top-down noto come suddivisione binaria ricorsiva.
- ▶ L'approccio è top-down perché inizia dalla cima dell'albero e successivamente divide lo spazio dei predittori; ogni divisione è indicata da due nuovi rami più in basso nell'albero.
- ▶ Ad ogni passaggio del processo di costruzione dell'albero, viene effettuata la miglior divisione in quel particolare passaggio, invece di guardare avanti per scegliere una divisione che potrebbe portare a un albero migliore in un passaggio futuro (simile alla forward variable selection).

Dettagli - Continuazione

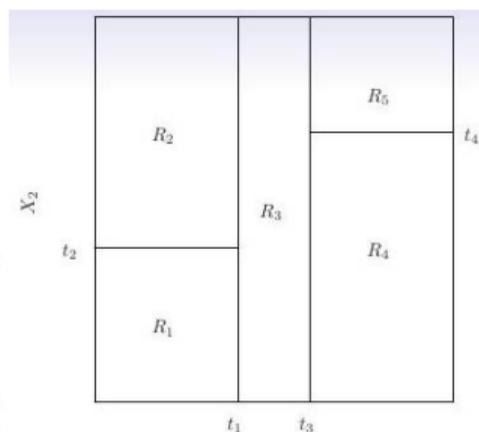
- ▶ Selezioniamo il predittore X_j e il punto di taglio s tali che dividendo lo spazio dei predittori nelle regioni $\{X \mid X_j < s\}$ e $\{X \mid X_j \geq s\}$ si ottenga la maggiore riduzione possibile dell'RSS.
- ▶ Successivamente, ripetiamo il processo, cercando il miglior predittore e il miglior punto di taglio per dividere ulteriormente i dati in modo da minimizzare l'RSS all'interno delle regioni risultanti.
- ▶ Questa volta, invece di dividere l'intero spazio dei predittori, dividiamo una delle due regioni precedentemente identificate. Ora abbiamo tre regioni.
- ▶ Ancora, cerchiamo di dividere una di queste tre regioni, per minimizzare l'RSS. Il processo continua fino a raggiungere un criterio di arresto; ad esempio, possiamo continuare finché nessuna regione contiene più di cinque osservazioni.

Previsioni

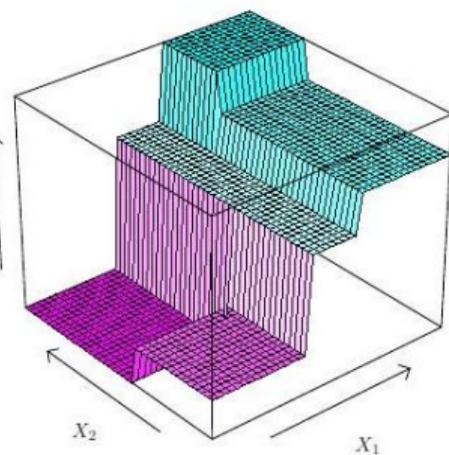
Prevediamo la risposta per una determinata osservazione di test usando la media delle osservazioni di training nella regione a cui appartiene tale osservazione di test.



Il risultato della suddivisione binaria ricorsiva.



L'albero corrispondente alla partizione.



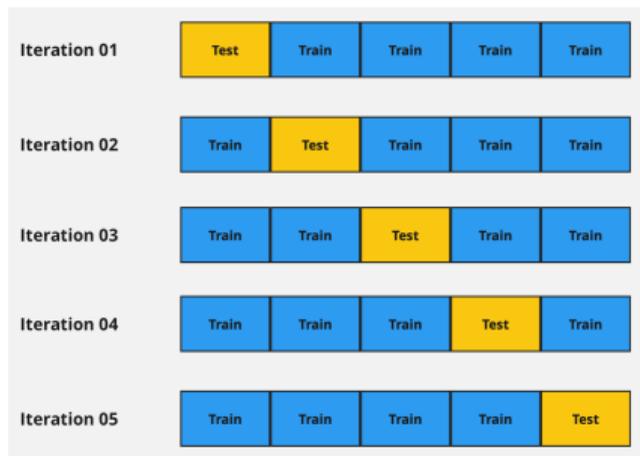
Un grafico prospettico della superficie predittiva.

Pruning di un albero

- ▶ Il pruning semplifica un albero complesso eliminando rami non necessari, migliorando interpretabilità e riducendo il rischio di sovradattamento.
- ▶ Si parte da un albero grande e complesso e si rimuovono progressivamente i rami meno significativi, mantenendo solo quelli utili per la predizione.
- ▶ L'obiettivo è trovare un equilibrio tra complessità e accuratezza, evitando di includere dettagli inutili che potrebbero peggiorare le performance sui nuovi dati.

Cross-validation

Obiettivo: Scegliere il modello che funziona meglio su dati non visti.



- ▶ La validazione incrociata è una tecnica per valutare le performance di un modello suddividendo i dati in più parti (folds).
- ▶ Ogni fold viene utilizzato a turno come set di test, mentre i rimanenti fungono da set di training.
- ▶ Permette di stimare l'accuratezza del modello su nuovi dati, evitando il rischio di sovradattamento.
- ▶ Nel contesto del pruning degli alberi, aiuta a identificare il sottoalbero con il miglior compromesso tra semplicità e capacità predittiva.

Selezione del miglior sottoalbero

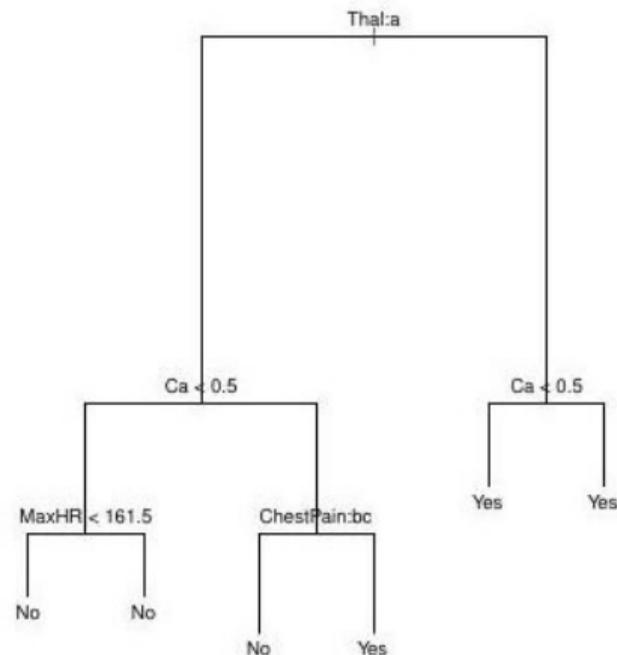
- ▶ La scelta del miglior albero avviene attraverso la validazione incrociata, confrontando le performance di diversi sottoalberi.
- ▶ Ogni sottoalbero viene valutato per identificare quello che offre la migliore combinazione tra semplicità e capacità predittiva.
- ▶ Una volta identificato l'albero ottimale, lo si applica ai dati completi per migliorare l'accuratezza finale.

Alberi di Classificazione

- ▶ Molto simili a un albero di regressione, con la differenza che vengono utilizzati per prevedere una risposta qualitativa anziché quantitativa.
- ▶ Per un albero di classificazione, prevediamo che ogni osservazione appartenga alla classe più comune tra le osservazioni di training nella regione a cui appartiene.
- ▶ Come nel contesto della regressione, utilizziamo la suddivisione binaria ricorsiva per far crescere un albero di classificazione.
- ▶ Nel contesto della classificazione, l'RSS non può essere utilizzato come criterio per effettuare le suddivisioni binarie:
 - ▶ **Indice Gini:** Misura l'impurità di una regione in base alla probabilità che una osservazione sia classificata in modo errato.
 - ▶ **Devianza:** Una metrica basata sull'entropia, che misura la quantità di disordine o incertezza in una regione. La devianza è calcolata come una funzione logaritmica delle probabilità delle classi e penalizza le suddivisioni meno utili.

Esempio: dati cardiaci

- ▶ Questi dati contengono un esito binario **HD** per 303 pazienti che si sono presentati con dolore toracico.
- ▶ Un valore di esito pari a **Yes** indica la presenza di malattia cardiaca basata su un test angiografico, mentre **No** significa assenza di malattia cardiaca.
- ▶ Ci sono 13 predittori, inclusi **Age**, **Sex**, **Chol** (una misura del colesterolo) e altre misure della funzione cardiaca e polmonare.



Vantaggi e Svantaggi degli Alberi

Vantaggi

- ▶ Gli alberi sono facili da spiegare, anche più della regressione lineare.
- ▶ Riflettono in modo intuitivo il processo decisionale umano.
- ▶ Sono facilmente visualizzabili e interpretabili, soprattutto se piccoli.
- ▶ Gestiscono predittori qualitativi senza necessità di variabili dummy.

Svantaggi

- ▶ Gli alberi hanno generalmente una minore accuratezza predittiva rispetto ad altri approcci.
- ▶ La loro performance può essere migliorata aggregando più alberi (es. ensemble methods).

Introduzione al Bagging

- ▶ **Bagging** (Bootstrap Aggregation) è una tecnica per migliorare l'accuratezza e la stabilità di modelli predittivi.
- ▶ Combina le previsioni di molti modelli, ognuno costruito su una versione diversa dello stesso dataset.
- ▶ Riduce la varianza del modello, rendendolo più robusto ai cambiamenti nei dati.
- ▶ Particolarmente utile per modelli ad alta varianza come gli alberi decisionali.

Come Funziona il Bagging

1. Si crea un grande numero (B) di sottoinsiemi casuali del dataset originale (**campioni bootstrappati**).
2. Per ogni campione, si costruisce un modello indipendente (ad esempio, un albero decisionale).
3. Si calcolano le previsioni per ciascun modello su nuovi dati.
4. Le previsioni finali sono ottenute facendo la media (per regressione) o votazione maggioritaria (per classificazione) tra tutti i modelli.

Vantaggi del Bagging

- ▶ **Riduzione della varianza:** Combinando molti modelli, si ottengono previsioni più stabili.
- ▶ **Robustezza agli outlier:** Gli effetti degli outlier vengono diluiti tra i diversi modelli.
- ▶ **Adattabilità:** Funziona con una vasta gamma di algoritmi, non solo alberi decisionali.
- ▶ **Efficace con dati complessi:** Ideale per gestire dataset rumorosi o con molte variabili.

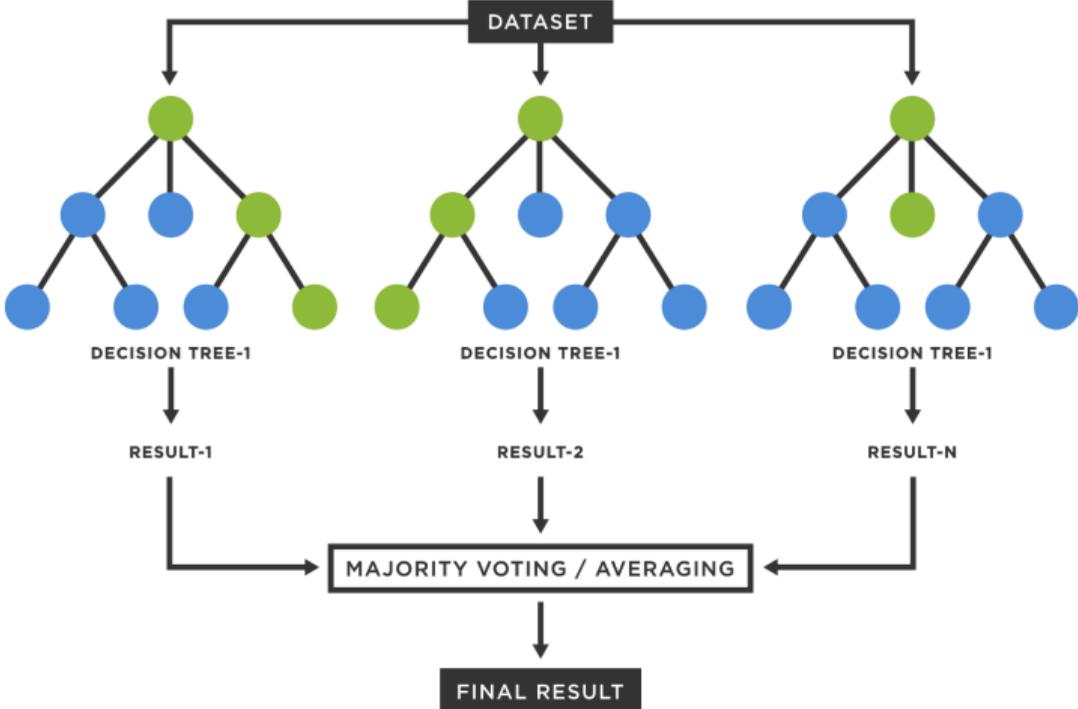
Errore Out-of-Bag (OOB)

- ▶ Durante il processo di bagging, ogni albero utilizza solo una parte del dataset.
- ▶ Il resto delle osservazioni (**Out-of-Bag**) viene utilizzato per testare l'accuratezza del modello.
- ▶ L'errore OOB è calcolato combinando le previsioni fatte sugli OOB di tutti gli alberi.
- ▶ Permette di stimare la performance del modello senza necessità di un set di test separato.

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.
- ▶ Le random forests forniscono un miglioramento rispetto agli alberi bagged tramite una piccola modifica che decorrela gli alberi. Questo riduce la varianza quando mediamo gli alberi.
- ▶ Questo approccio rende la Random Forest stabile e precisa.

Cos'è una Random Forest



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.

Importanza delle Variabili: Random Forest per Classificazione

- ▶ Nei modelli di classificazione, la Random Forest offre tre metriche principali per misurare l'importanza delle variabili:
 - ▶ **Class-Specific Importance:** Fornisce un'analisi dettagliata dell'importanza delle variabili rispetto a ciascuna classe target.
 - ▶ **Mean Decrease Accuracy (MDA):** Quantifica la diminuzione dell'accuratezza del modello quando i valori di una variabile sono permutati casualmente.
 - ▶ **Mean Decrease Gini (MDG):** Misura la riduzione complessiva della devianza degli split associati a ciascuna variabile, rappresentando la loro capacità discriminante.
- ▶ Queste metriche consentono di identificare le variabili chiave che influenzano le previsioni del modello, ottimizzando le prestazioni.

Importanza delle Variabili: Random Forest per Regressione

- ▶ Due metriche principali per valutare l'importanza delle variabili:
 - ▶ **% Increase in Mean Squared Error (%IncMSE)**: Indica quanto aumenta l'errore quadratico medio delle previsioni quando i valori di una variabile sono permutati casualmente.
 - ▶ **Increase in Node Purity (IncNodePurity)**: Quantifica la riduzione complessiva della devianza (o impurità) nei nodi grazie a ciascuna variabile.
- ▶ Queste metriche aiutano a comprendere il contributo relativo di ogni variabile nel migliorare la qualità delle previsioni.

Random Forest: Vantaggi e Svantaggi

Vantaggi

- ▶ 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

- ▶ 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.

LAB Credit Analysis: Obiettivi

Questo laboratorio riprende l'analisi dei dati bancari introdotta nel Capitolo 1, con l'obiettivo di approfondire ulteriormente:

- ▶ La relazione tra variabili socioeconomiche e il saldo medio delle carte di credito (**Balance**).
- ▶ Il confronto tra due approcci di modellazione: **regressione lineare** e **regression tree**.

L'obiettivo finale è identificare il modello più efficace per fornire raccomandazioni strategiche alla banca.

LAB Credit Analysis: Procedura

- ▶ Ripeti l'analisi introdotta nel Capitolo 1, utilizzando la regressione lineare per predire il saldo medio delle carte di credito (`Ba1ance`).
- ▶ Estendi l'analisi costruendo un **regression tree** con lo stesso dataset e confronta i due modelli.
- ▶ Per ciascun modello, valuta:
 - ▶ L'accuratezza predittiva tramite il **Mean Squared Error (MSE)**.
 - ▶ La capacità di interpretare le relazioni tra variabili.
- ▶ Discuti quale modello è più adatto per l'obiettivo della banca.