Università della Calabria
Metodi Statistici per le Strategie Aziendali

CdLM in Data Science per le Strategie Aziendali

Relazione di laboratorio

2° Assignment

Dicembre 2025

Gruppo di lavoro:
Cristian Tedesco, tdscst02t24h579i@studenti.unical.it
Matricola 269279

Francesco De Nisi, dnsfnc00r09m208v@studenti.unical.it
Matricola 269762

Angela Karin Mancuso, mncnlk98m64i874m@studenti.unical.it
Matricola 268839

Nabil Larhram, lrhnbl98a11z330z@studenti.unical.it
Matricola 269686

Saveria Falvo, flvsvr02p64d086x@studenti.unical.it
Matricola 264008

Pierfrancesco Lindia, lndpfr00e20d086f@studenti.unical.it
Matricola 256641

Anno Accademico 2025-2026

Contents

1 Introduzione

Il presente progetto si pone l'obiettivo di analizzare un dataset enologico composto da 178 osservazioni al fine di sviluppare modelli di segmentazione e classificazione basati sulle proprietà chimiche dei vini. La variabile dipendente è stata ricodificata in tre macro-categorie (accorpando le classi 1-2, la classe 3 e le classi 4-5). L'approccio metodologico adotta una prospettiva incrementale che parte dall'applicazione dell'algoritmo CHAID, confrontando diverse strategie di discretizzazione e Stress Testing su dati sintetici per saggiare la robustezza delle regole decisionali. L'indagine si estende successivamente al confronto con algoritmi quali CART con potatura ottimale, Random Forest e Bagging, con l'intento di massimizzare l'accuratezza predittiva ed isolare gerarchicamente le variabili chimiche determinanti per la qualità del vino.

Nota Metodologica sulla Selezione delle Variabili:L'analisi utilizza esclusivamente le 13 variabili di composizione chimica del dataset, escludendo deliberatamente le valutazioni soggettive di sommelier e le preferenze demografiche per fascia d'età. Questa scelta di esclusione risponde a tre esigenze principali: mantenere l'oggettività della classificazione basandola su proprietà chimiche misurabili e standardizzate; evitare la perdita di generalizzabilità che deriverebbe dall'inclusione di giudizi umani soggettivi e variabili nel tempo; preservare la riproducibilità del modello, che potrà essere applicato a nuovi vini senza dipendenza da valutazioni esterne.

2 Applicazione dell'algoritmo CHAID

L'algoritmo CHAID opera costruendo alberi decisionali basati su test di significatività statistica (χ^2) tra variabili categoriali. Poiché il dataset contiene predittori continui e ordinali, è stato necessario implementare specifiche logiche di trasformazione per rendere i dati compatibili con l'algoritmo.

2.1 Ricodifica della Variabile Target

La variabile risposta originale, `Classificazione`, presentava 5 livelli. Preliminarmente si è proceduto a una aggregazione in 3 macro-classi:

- **Prima Classe (0,2]**: Include i vini delle classi originali 1 e 2, indicata come "fascia alta" ai fini espositivi..
- **Seconda Classe (2,3]**: Corrisponde ai vini della classe originale 3, indicata come "fascia intermedia".
- **Terza Classe (3,5]**: Raggruppa i vini delle classi originali 4 e 5, indicata come "fascia bassa".

Funzioni Helper: Sono state implementate alcune funzioni per i metodi di binarizzazione e discretizzazione delle variabili quantitative, che verranno descritti nel seguito: Chi-Quadro, Quantili, Entropia, Gini, Interval.

Per trasformare le variabili esplicative, è stata definita una funzione *helper* specifica, l'**Algoritmo di Split Binario Ottimale**, che viene utilizzato per le variabili quantitative di Composizione dei vini. La funzione analizza l'intero dominio della variabile continua e identifica la soglia di taglio (t) che massimizza la statistica χ^2 rispetto al target. Questo approccio garantisce che la dicotomizzazione (*Low* vs *High*) sia la più significativa possibile dal punto di vista statistico.

2.2 Analisi dei Risultati: CHAID con Metodo Binario

La performance è stata valutata attraverso tre scenari di stima del tasso di errata classificazione.

2.2.1 Stima su Campione Totale (Risostituzione)

In prima istanza, il modello è stato addestrato sull'intero dataset e le soglie ottimali sono state calcolate utilizzando la totalità delle osservazioni.

- **Accuratezza:** 91.01%
- **Tasso di errata classificazione:** 8.99%

Table 1: Matrice di Confusione sul Campione Totale (Metodo Binario).

	Classe Predetta		
Classe Osservata	**(0,2]**	**(2,3]**	**(3,5]**
(0,2]	51	2	0
(2,3]	9	66	0
(3,5]	0	5	45

La matrice di confusione evidenzia una notevole precisione nell'identificazione delle classi estreme: la prima classe (0,2] è stata classificata quasi senza errori, mentre la terza classe (3,5] mostra un tasso di errore moderato. Le criticità emergono principalmente nella gestione della classe intermedia (2,3], dove si osserva una tendenza del modello a sovrastimare la qualità di alcune osservazioni, classificando erroneamente 9 vini come appartenenti alla prima classe. Inoltre, si nota che 11 vini della classe bassa sono stati erroneamente attribuiti alla classe intermedia (3,5] → (2,3]), indicando che la dicotomizzazione forzata causa una compressione delle informazioni verso le categorie intermedie.

Questo risultato è affetto da *overfitting*. Il modello appare artificialmente preciso poiché le soglie di taglio sono state calcolate conoscendo a priori la distribuzione delle classi su tutto il campione, introducendo un bias ottimistico nella stima.

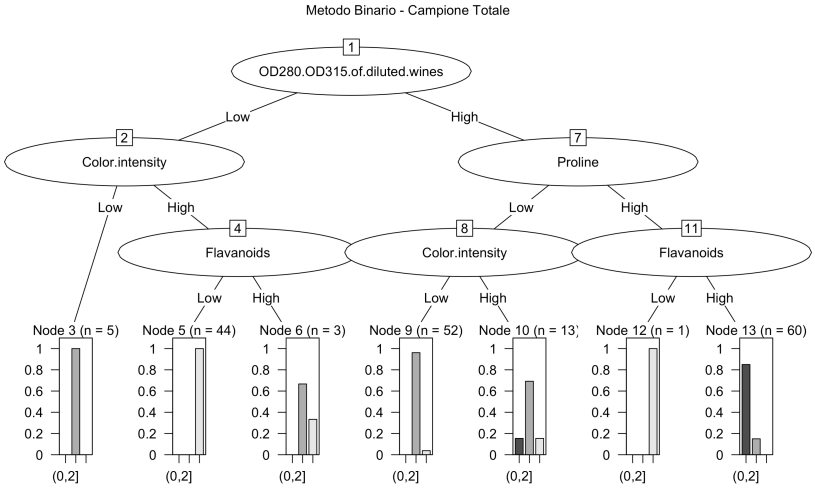

Figure 1: Albero decisionale generato sul Campione Totale (Metodo Binario). Si nota la forte capacità discriminante data dalla variabile radice *OD280.OD315*.

2.2.2 Validazione Train/Test Split

Per valutare la capacità di generalizzazione del modello su dati inediti, il dataset è stato diviso in campione di Train e campione di Test, indipendenti.

- **Training Set (70%)**: Utilizzato per il calcolo delle soglie ottimali e l'addestramento dell'albero.
- **Test Set (30%)**: Utilizzato esclusivamente per la verifica delle performance.

Le soglie di discretizzazione (i valori di taglio per definire le categorie "Low" e "High") sono state calcolate *solamente* osservando la distribuzione dei dati nel Training Set. Successivamente, queste regole "apprese" sono state applicate ai dati del Test Set per classificarli, senza che il modello avesse alcuna informazione preliminare sui dati di addestramento.

Errore su Test Set: 17.31%

Il tasso di errore è aumentato rispetto alla stima precedente (passando da circa il 9% al 17%). Questo peggioramento è un indicatore chiaro della rigidità strutturale del metodo binario. Una singola soglia di taglio, calcolata su uno specifico sottoinsieme di dati, fatica ad adattarsi alla naturale variabilità di nuovi campioni. I vini con valori chimici molto vicini alla soglia di taglio vengono facilmente classificati nella categoria errata, dimostrando che una semplice dicotomizzazione comporta una perdita di informazione troppo elevata per questo tipo di problema.

Table 2: Matrice di Confusione sul Test Set (Metodo Binario).

	Classe Predetta		
Classe Osservata	**(0,2]**	**(2,3]**	**(3,5]**
(0,2]	13	2	0
(2,3]	3	18	1
(3,5]	0	3	12

L'analisi della matrice di confusione rivela una buona capacità discriminante, specialmente per le classi estreme, dove non si registrano errori gravi di incrocio tra la prima classe (0,2] e quella bassa (3,5]. Le classificazioni errate si concentrano prevalentemente attorno alla classe intermedia (2,3], la quale presenta una lieve dispersione verso le categorie adiacenti, fenomeno fisiologico dovuto alla natura continua delle variabili chimiche in prossimità delle soglie di taglio.

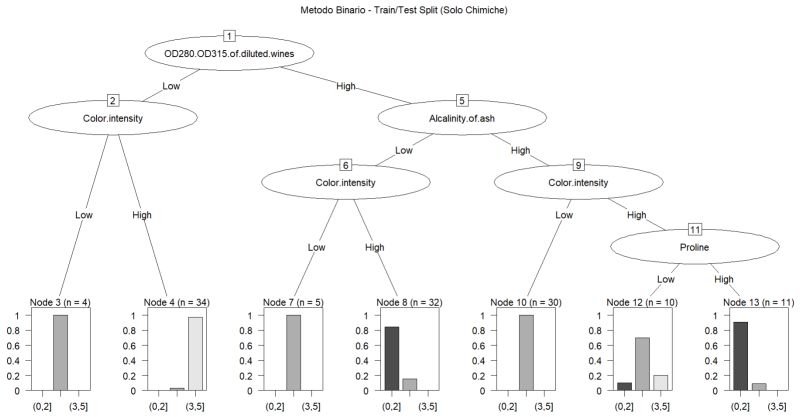

Figure 2: Albero decisionale generato sul Training Set. La struttura differisce dal modello totale (es. Nodo 5), evidenziando l'instabilità delle soglie binarie.

2.2.3 Stima Robusta tramite Cross Validation (V=5)

Per eliminare la dipendenza dalla specifica suddivisione casuale dei dati e ottenere una stima più affidabile, è stata eseguita una *5-Fold Cross Validation*. L'intero processo di discretizzazione e addestramento è stato ripetuto 5 volte su sottoinsiemi indipendenti.

Errore Medio CV: 17.94%

Questo valore suggerisce che la dicotomizzazione forzata causa una perdita di informazione rilevante. Per migliorare la capacità predittiva, è necessario esplorare metodi di discretizzazione più granulari (come i quartili o gli intervalli), oggetto delle sezioni successive.

3 CHAID con metodo dei Quantili

Nella seconda fase dell'analisi, è stata testata una tecnica di discretizzazione basata sulla distribuzione statistica: il **Metodo dei Quartili**. L'obiettivo è trasformare le variabili chimiche continue in variabili categoriche ordinali $Q1, Q2, Q3, Q4$. Questa trasformazione impone una regola rigida: ogni livello deve contenere inizialmente il **25% delle unità** del training set.

3.1 Robustezza agli Outliers

Durante lo sviluppo del modello, è emersa una criticità tecnica: la gestione dei dati futuri (Test Set o nuovi vini) che presentano valori esterni al range osservato nel Training Set. Se il modello apprende che il valore massimo di Alcol è 14.0, un nuovo vino con 14.5 genererebbe un errore di tipo *Missing Value* (NA), bloccando la previsione.

Per risolvere questo problema e rendere il modello applicabile in produzione, è stata implementata la funzione personalizzata `get_robust_breaks`.

Anziché definire il primo intervallo come un segmento chiuso $[Min, Q1]$, viene definito come un intervallo aperto a sinistra $(-\infty, Q1]$. In questo modo, il modello diventa molto robusto: qualsiasi vino futuro, per quanto estremo, verrà classificato correttamente nel quartile più basso o più alto senza generare errori tecnici.

3.2 Stima su Campione Totale (Risostituzione)

La prima valutazione misura la capacità del modello di apprendere dai dati che ha già visto (178 osservazioni).

- **Accuratezza:** 89.89%
- **Errore:** 10.11%

Nonostante l'alta accuratezza globale, la matrice rivela un comportamento asimmetrico del modello:

Table 3: Matrice di Confusione (Campione Totale)

	Previsto		
Osservato	**(0,2]**	**(2,3]**	**(3,5]**
(0,2]	**53**	0	0
(2,3]	14	**59**	2
(3,5]	0	2	**48**

Il modello è quasi infallibile nel riconoscere i vini di alta qualità (Classe `(0,2]`: 0 errori) e quelli di bassa qualità (Classe `(3,5]`: solo 2 errori).

La classe intermedia `(2,3]` rappresenta il punto debole. Il modello ha classificato erroneamente **14 vini su 75** (circa il 19% di questa classe), etichettandoli come vini di prima classe.

I tagli rigidi dei quartili non riescono a cogliere le sfumature sottili che distinguono un vino "medio" da uno "ottimo". Il modello tende a sovrastimare la qualità dei vini incerti.

3.3 Stima su Train/Test Split (Generalizzazione)

Dividendo i dati (70% Train, 30% Test), si simula il comportamento su vini mai visti prima.

- **Accuratezza:** 82.69%
- **Errore:** 17.31%

L'errore è aumentato rispetto alla fase precedente (da 10.1% a 17.3%). Questo divario è significativo e indica che le soglie dei quartili calcolate sul 70% dei dati non sono perfettamente rappresentative per il restante 30%. È un segnale di **instabilità delle soglie**: basta spostare pochi vini dal training al test per spostare i confini statistici dei quartili e confondere le regole dell'albero.

Table 4: Matrice di Confusione sul Test Set (Metodo Quartili).

	Classe Predetta		
Classe Osservata	**(0,2]**	**(2,3]**	**(3,5]**
(0,2]	13	2	0
(2,3]	2	20	0
(3,5]	0	5	10

L'esame della matrice di confusione evidenzia una buona tenuta generale del modello, in particolare per la classe intermedia (2,3] che viene correttamente identificata nella maggioranza dei casi. Tuttavia, si nota una asimmetria nella distribuzione degli errori: mentre la classe alta (0,2] presenta una dispersione minima, la classe bassa (3,5] soffre di una significativa ambiguità, con un terzo delle osservazioni reali (5 su 15) che vengono erroneamente classificate come vini di seconda classe, suggerendo che i tagli basati sui quartili potrebbero essere troppo rigidi per intercettare correttamente i profili chimici di confine in questa specifica fascia.

3.4 Cross Validation (Analisi della Stabilità)

La Cross Validation a 5 Folds è la prova definitiva per l'affidabilità statistica.

La varianza dei risultati è troppo elevata (dal 2% al 19%). Questo conferma che il metodo dei Quartili è strutturalmente instabile: la sua precisione dipende eccessivamente da *quali* specifici vini finiscono nel set di addestramento. Inoltre, sebbene sia quello con l'errore minore, non è possibile affidarsi al risultato del Fold 3, in quanto non è replicabile costantemente.

Fold	Errore (%)
1	19.44
2	14.29
3	2.78
4	16.67
5	14.29
Media	**13.49**

Table 5: Risultati della Cross-Validation: Percentuale di errore per Fold

3.5 Analisi Grafico 1: Albero su Campione Totale

L'albero risultante, illustrato in Figura 3, presenta una struttura significativamente articolata e ramificata, caratterizzata da un'elevata granularità delle regole decisionali. Il nodo radice individua nei *Flavanoids* il discriminante primario, ma la natura del metodo per quartili impone una segmentazione immediata in quattro direzioni distinte, permettendo un'analisi molto fine sin dal primo livello.

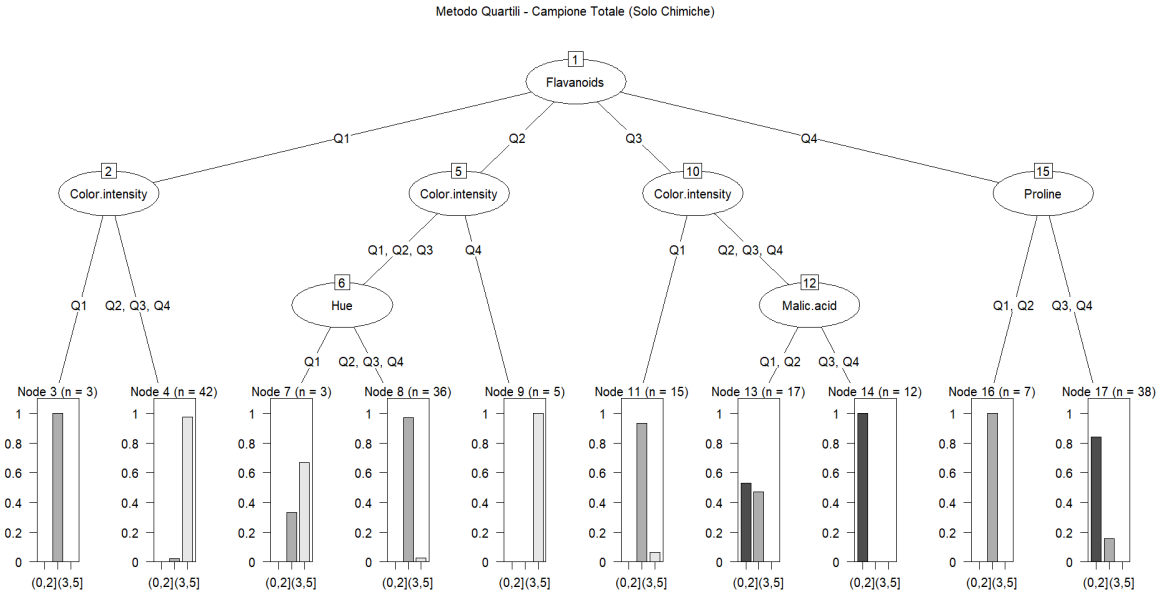

Figure 3: Albero decisionale CHAID generato sull'intero campione utilizzando la discretizzazione per Quartili. Si nota l'elevata profondità e larghezza della struttura.

Osservando i rami generati, si nota come il primo quartile (bassi flavonoidi) intercetti i vini di qualità inferiore, operando un ulteriore filtraggio tramite l'*Intensità del Colore*. Spostandosi verso i quartili intermedi e superiori ($Q2, Q3, Q4$), la struttura rivela interazioni complesse che coinvolgono variabili come il *Magnesium* e nuovamente l'*Intensità del Colore*. In particolare, il *Magnesium* assume un ruolo centrale nei rami a più alto contenuto fenolico, suggerendo che questa variabile contribuisca in modo determinante alla caratterizzazione delle sfumature di qualità superiore. L'albero si sviluppa fino a cinque livelli di profondità, generando un numero elevato di nodi terminali. Sebbene questa configurazione offra una descrizione estremamente dettagliata delle sottocategorie di vini, l'eccessiva frammentazione dei dati in nodi con poche osservazioni suggerisce che l'approccio per quartili, pur essendo esplorativamente potente, possa risultare meno immediato da interpretare operativamente rispetto a strategie di discretizzazione più sintetiche.

3.6 Analisi Grafico 2: Albero Train/Test Split

L'albero in figura 4 rappresenta la capacità media del modello di generalizzare (Errore $\approx 17\%$).

Analisi della Logica Decisionale:

1. **La Variabile dominante (Flavanoids):** L'albero sceglie i Flavonoidi come discriminante primario (Nodo 1).
   - *Ramo Sinistro ($Q1$ - Bassi Flavonoidi):* Identifica immediatamente i vini di bassa qualità.
   - *Ramo Destro ($Q3, Q4$ - Alti Flavonoidi):* Al contrario, vengono identificati vini di alta qualità.
2. **Il Percorso dell'Eccellenza:** Per essere classificato come vino Top `(0,2]`, un vino deve soddisfare tre criteri: Flavonoidi Alti → Magnesio Medio-Alto → Prolina Alta (Nodo 13). La Prolina agisce come raffinatore finale per i vini di qualità.
3. **Segnali di Incertezza:** Si osservino i nodi centrali (es. Nodo 3 e Nodo 12). Le barre non sono pure, ma miste (grigio chiaro e scuro). Questo conferma visivamente perché l'errore è alto: la divisione in quartili è troppo grossolana per separare nettamente i casi borderline.

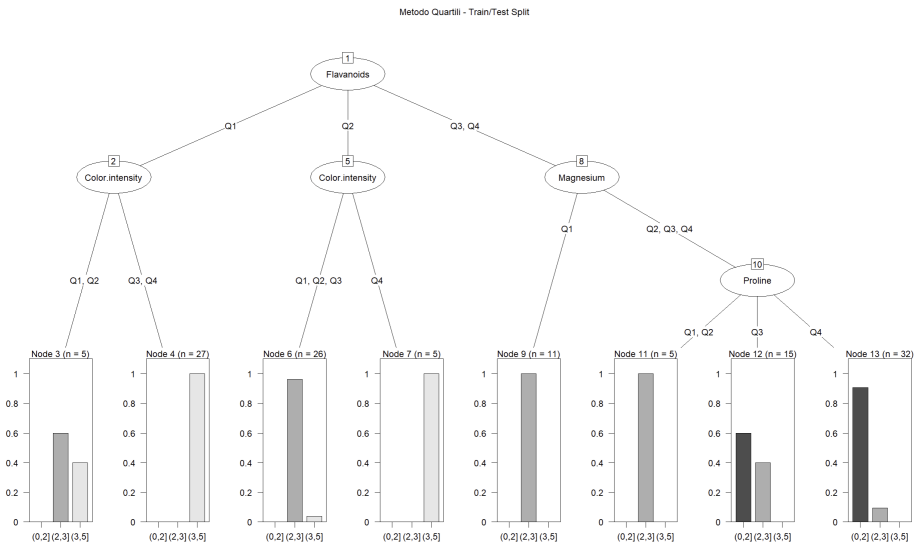

Figure 4: Albero Decisionale Train/Test Split (Errore 17.31%)

3.7 Analisi Grafico 3: Albero CV Fold 3

Questo è l'albero del caso "eccezionale" della Cross Validation (Errore 2.78%).

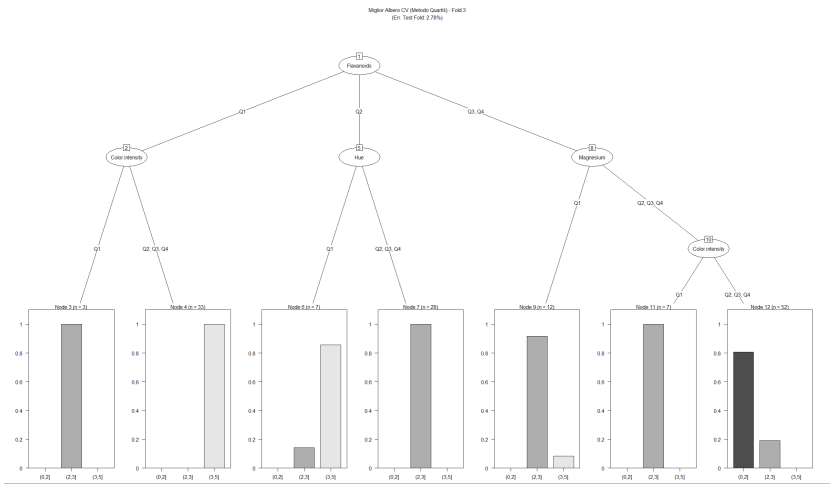

Figure 5: Miglior Albero della CV - Fold 3 (Errore 2.78%)

Confronto Strutturale:

1. **Purezza Assoluta:** A differenza dell'albero precedente, i nodi terminali sono purissimi, come si nota guardando il **Nodo 12** o il **Nodo 7**: una sola barra scura domina completamente. In questo specifico campione, i tagli matematici dei quartili sono caduti esattamente negli spazi vuoti tra le classi di vino.
2. **Cambio di Variabili (Sintomo di Instabilità):** Mentre la radice rimane *Flavanoids* (conferma della sua importanza chimica fondamentale), la struttura interna cambia. Qui compare la variabile **Hue** (Tonalità) al nodo 5, che non era determinante nell'albero precedente.

Il fatto che le variabili secondarie cambino da un test all'altro dimostra che il modello non ha trovato regole universali, ma si sta adattando troppo allo specifico set di dati usato in quel momento (overfitting).

3.8 Analisi Critica e Limiti del Modello

L'analisi condotta sull'applicazione del metodo di discretizzazione per frequenza (**Quartili**) ha restituito un quadro duale, caratterizzato da un'eccellente tenuta tecnica ma da evidenti limiti in termini di modellazione statistica. Dal punto di vista implementativo, il modello si è dimostrato **estremamente robusto**.

Tuttavia, i risultati della Cross Validation hanno evidenziato che la discretizzazione per quartili non è la strategia ottimale per questo dataset. La causa principale risiede nella **rigidità strutturale** del metodo: imporre che ogni classe contenga forzatamente il 25% delle osservazioni ha portato a spezzare gruppi naturali di dati. Questo è evidenziato dalla confusione nella classe media, che il modello fatica ad isolare poichè i quartili creano confini artificiali dove, in realtà, le proprietà chimiche variano in modo continuo. Inoltre, l'elevata oscillazione dell'errore tra i vari fold dimostra l'ipersensibilità del modello al training set in base a piccole variazioni.

Dunque, si è deciso di considerare ulteriori metodi di discretizzazione per cercare il modello migliore.

4 Analisi di ulteriori Metodi di Discretizzazione e Scelta del Modello

A seguito delle prime applicazioni dell'algoritmo CHAID, basate rispettivamente sulla binarizzazione tramite soglia ottimale (Chi-Quadro) e sulla suddivisione in quartili, si è ritenuto necessario approfondire l'analisi testando metodi alternativi di discretizzazione per le variabili continue. Questa fase ha l'obiettivo di verificare se l'adozione di criteri diversi (come l'Indice di Gini, l'Entropia o gli Intervalli fissi), valutati rigorosamente tramite Cross Validation (5-fold), possa ridurre il tasso di errore e garantire una performance di classificazione superiore rispetto ai modelli precedenti. L'analisi condotta ha messo a confronto tre diverse logiche di trasformazione delle variabili continue in categoriche, evidenziando una chiara gerarchia di performance:

1. Interval (Intervalli Fissi): Errore medio 11.25% (Best Performer);
2. Indice di Gini: Errore medio 15.76%;
3. Entropia: Errore medio 14.62%;
4. Quartili: Errore medio 13.49%;
5. Binario (Chisq): Errore medio 17.94%.

La sostanziale differenza nei risultati ottenuti risiede nella logica matematica con cui i diversi metodi trasformano i dati continui in categorie. Il metodo Interval, risultato il più performante con un errore del 11.25%, opera secondo una logica non supervisionata: esso ignora completamente la variabile target durante la fase di taglio, limitandosi a suddividere il range delle variabili chimiche in tre fasce di uguale ampiezza. Questo approccio "agnostico" genera categorie naturali basate sulla grandezza del valore (Basso, Medio, Alto) senza inseguire le fluttuazioni specifiche del set di addestramento, garantendo così una maggiore robustezza e riducendo drasticamente il rischio di sovradattamento (overfitting) quando il modello viene applicato a nuovi vini.

Diametralmente opposto è l'approccio dei metodi basati su Indice di Gini ed Entropia, che operano in modo supervisionato cercando attivamente il singolo punto di taglio ottimale per massimizzare la purezza delle classi o il guadagno di informazione. Sebbene questa strategia possa sembrare teoricamente superiore, nel contesto di questo dataset ha prodotto regole troppo rigide e specifiche per il campione di training, portando a un errore medio più elevato (circa 14.6%) in fase di validazione. Questa distinzione nelle logiche di discretizzazione diventa cruciale quando si considera il funzionamento specifico dell'algoritmo utilizzato.

La fase di pre-processing non è infatti un passaggio neutrale, ma abilita la vera potenza del CHAID, che a differenza di algoritmi come CART (che usano l'indice di Gini) o C4.5 (che usano l'Entropia) opera esclusivamente con variabili qualitative o discrete. Inoltre, il CHAID è progettato per generare alberi con nodi a rami multipli (multi-way splits), fornendo in input una variabile discretizzata in più classi (come avviene con la divisione ternaria del metodo Interval), si lascia all'algoritmo la libertà statistica, tramite il test Chi-Quadro, di decidere autonomamente se mantenere i tre gruppi separati o se unirne alcuni. Al contrario, Gini ed Entropia impongono una divisione secca 'Sopra/Sotto' una certa soglia. Questo 'ingabbia' l'algoritmo, perché gli nasconde i dettagli più complessi dei dati. Come si evince da quanto già visto, per questi vini è cruciale non semplificare troppo, ma analizzare anche le fasce intermedie che una divisione binaria cancellerebbe.

4.1 Albero CHAID Interval

Il metodo vincente è risultato essere l'**Interval Discretization**, con un errore medio in CV dell' **11.25%**. Infatti, la discretizzazione a intervalli fissi, pur essendo più rigida, ha agito come regolarizzatore naturale. Evitando di adattare le soglie eccessivamente ai dati di training (come fa il metodo binario o i quartili), ha prodotto soglie più robuste e significative per le proprietà chimiche dei vini.

Una volta identificato il metodo vincente, si è proceduto alla ricostruzione del dataset completo applicando la discretizzazione per intervalli a tutte le variabili quantitative e riaddestrando l'albero *CHAID* sull'intera popolazione ($n = 178$).

La bontà di adattamento del modello vincente è stata valutata tramite la **Matrice di Confusione**, calcolata con il metodo della risostituzione:

	Classe Prevista		
Classe Osservata	**(0,2]**	**(2,3]**	**(3,5]**
(0,2]	**51**	2	0
(2,3]	9	**63**	3
(3,5]	0	4	**46**

Table 6: Matrice di Confusione con il Metodo Interval

Il modello ha raggiunto un'Accuratezza Globale dell'**89.89%** (Errore: 10.11%). Dall'esame della matrice di confusione emerge un comportamento asimmetrico del classificatore. Il modello dimostra un'eccellente capacità discriminante per le classi estreme, classificando correttamente la quasi totalità dei vini pregiati (classe 0,2]) e identificando con grande precisione quelli di bassa qualità (classe 3,5]). Le principali criticità si concentrano nella classe intermedia (2,3], dove si osserva una tendenza a confondere alcuni campioni con la classe superiore, un fenomeno frequente nelle classificazioni ordinali dovuto alla naturale continuità dei parametri chimici nei vini di seconda classe.

4.2 Analisi della Significatività delle Variabili

Per comprendere quali caratteristiche guidano la classificazione, è stato eseguito un test di indipendenza **Chi-Quadro** (χ^2) tra la variabile target ricodificata e ciascun predittore discretizzato.

Variabile	**P-Value (χ^2)**	**Sig.**	**Interpretazione**
`OD280.OD315`	1.4×10^{-26}	***	Determinante
`Proline`	2.5×10^{-21}	***	Fondamentale
`Hue`	3.3×10^{-21}	***	Significativa
`Flavanoids`	3.6×10^{-21}	***	Fondamentale
`Color.intensity`	1.6×10^{-15}	***	Fondamentale
`Alcalinity.of.ash`	6.1×10^{-9}	***	Fondamentale
`Proanthocyanins`	7.3×10^{-9}	***	Significativa
`Nonflavanoid.phenols`	8.3×10^{-8}	***	Significativa
`Magnesium`	8.3×10^{-5}	***	Significativa
`Malic.acid`	0.043	*	Appena significativa
`Alcohol`	0.239	-	Non significativa
`Ash`	0.530	-	Non significativa
`Total.phenols`	0.980	-	Non significativa

Table 7: Test di Indipendenza Chi-Quadro sulle variabili quantitative

È fondamentale notare che alcune variabili di composizione chimica, come Alcol e Fenoli Totali, non presentano p-value molto alti ($p > 0.05$), risultando statisticamente non significative.

È interessante notare una apparente discrepanza tra la significatività globale delle variabili (test Chi-Quadro sull'intero dataset) e la loro effettiva selezione all'interno dell'albero decisionale. Sebbene la variabile *Hue* presenti un p-value globale estremamente basso (3.31×10^{-21}), indicando una fortissima associazione generale con la classe del vino, l'algoritmo CHAID ha selezionato *Alcalinity of Ash* ($p = 6.12 \times 10^{-9}$) come discriminante secondario nel ramo sinistro dell'albero (vini con basso OD280).

Questo fenomeno si spiega con la natura *condizionata* dell'algoritmo ricorsivo. La variabile *Hue* presenta verosimilmente un'elevata collinearità con il discriminante primario (*OD280*): una volta effettuato il primo taglio sulla struttura del vino, la capacità residua del colore di distinguere le classi diminuisce drasticamente all'interno del sottogruppo selezionato. Al contrario, l'*Alcalinity of Ash*, pur essendo globalmente meno potente, apporta un contributo informativo unico e specifico per il segmento dei vini a bassa struttura, rivelandosi il predittore localmente ottimo per raffinare la classificazione in quel particolare nodo.

4.3 Analisi della Struttura dell'Albero e Regole Decisionali

Si riporta la visualizzazione grafica dell'albero decisionale finale.

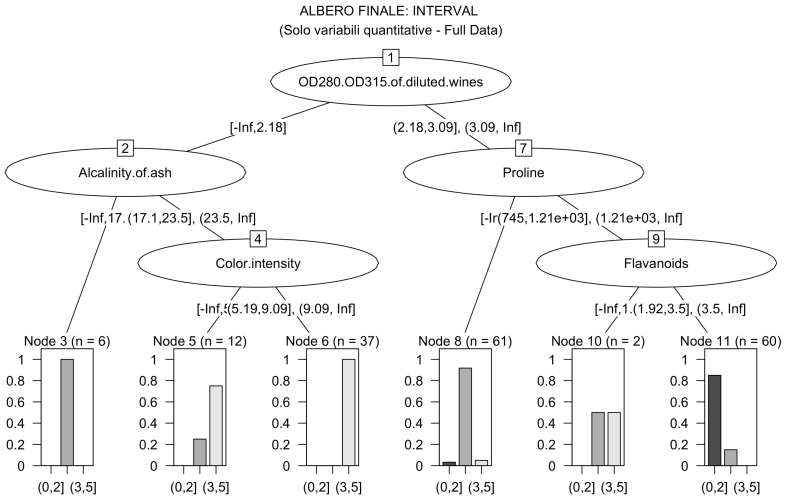

Figure 6: Albero Finale (Interval)

L'albero decisionale finale (Figura 6) presenta una struttura gerarchica pulita che si sviluppa su cinque livelli, selezionando OD280, Proline, Alcanity.of.ash, Color.intensity e Flavanoids come le variabili più significative per la segmentazione.

La prima ramificazione opera una distinzione netta basata sulla struttura del vino. I campioni con valori di OD280 inferiori alla soglia critica di 2.18 vengono immediatamente indirizzati verso il ramo sinistro dell'albero, associato alle fasce qualitative inferiori. In questo contesto, la variabile *Alcalinity of Ash* (Nodo 2) agisce come ulteriore filtro: valori elevati di alcalinità tendono a confermare l'appartenenza alla classe più bassa, mentre livelli intermedi richiedono un'ulteriore verifica tramite l'*Intensità del Colore* (Nodo 4) per distinguere tra vini di bassa e media qualità. È interessante notare come il Nodo 6, terminale di questo percorso per vini con colore intenso, riesca ad isolare un gruppo omogeneo di 37 osservazioni appartenenti esclusivamente alla classe $(3, 5]$, dimostrando un'elevata capacità di identificazione dei difetti o delle caratteristiche meno pregiate.

Spostando l'attenzione sul ramo destro, riservato ai vini con un OD280 superiore a 2.18 e quindi dotati di maggiore struttura, emerge il ruolo determinante della *Proline* (Nodo 7). Questo amminoacido si conferma come il marcatore d'eccellenza: una concentrazione molto elevata (superiore a 1210) conduce direttamente al Nodo 11, il quale raggruppa la quasi totalità dei vini di alta qualità $(0, 2]$ con una purezza notevole. Al contrario, quando la Prolina si attesta su valori intermedi o bassi, il modello ricorre ai *Flavanoids* (Nodo 9) per dirimere l'incertezza residua. In questo scenario, livelli di flavonoidi moderati identificano correttamente la classe intermedia $(2, 3]$ nel Nodo 8, che raccoglie la maggioranza dei vini standard (61 osservazioni).

La logica decisionale estratta dal modello può essere sintetizzata in un percorso diagnostico sequenziale: la valutazione parte dalla struttura generale del vino (OD280); se questa è carente, il vino è presumibilmente di bassa qualità, a meno che non presenti un colore o un'alcalinità atipici. Se invece la struttura è solida, il discriminante diventa la ricchezza aminoacidica (Proline), che agisce come determinante dell'alta qualità. I vini che possiedono struttura ma mancano di questo picco di eccellenza vengono correttamente collocati nella seconda classe, eventualmente supportati da un buon profilo di flavonoidi. Questa configurazione conferma che l'albero CHAID con discretizzazione per intervalli è in grado di mappare efficacemente la complessità dei vini, traducendo parametri analitici in categorie qualitative robuste.

4.3.1 Analisi della Composizione dei Nodi Terminali

Per valutare la capacità discriminante dell'albero CHAID (modello con discretizzazione *Interval*), è stata analizzata la distribuzione delle osservazioni reali all'interno di ciascun nodo terminale (foglia). La Tabella 8 incrocia l'ID del nodo finale con la classe di appartenenza effettiva dei vini, permettendo di misurare il grado di omogeneità (purezza) raggiunto dalla segmentazione.

Table 8: Distribuzione delle osservazioni nei nodi terminali dell'albero CHAID. Per ogni nodo è riportato il conteggio dei vini appartenenti alle tre classi di qualità.

ID Nodo	Classe Dominante	(0,2]	(2,3]	(3,5]	Totale
3	(2,3]	0	6	0	6
5	(3,5]	0	3	9	12
6	(3,5]	0	0	37	37
8	(2,3]	2	56	3	61
10	Misto	0	1	1	2
11	(0,2]	51	9	0	60

L'esame della tabella evidenzia una struttura di classificazione estremamente robusta, caratterizzata da una netta separazione tra le categorie estreme. Il risultato più significativo è rappresentato dal **Nodo 6**, che agisce come un filtro perfetto per la bassa qualità: esso intercetta 37 vini della classe $(3, 5]$ senza includere alcun falso positivo proveniente dalle classi superiori, garantendo una purezza del 100% per questo segmento.

Specularmente, il **Nodo 11** si configura come il cluster dell'eccellenza. Al suo interno sono concentrati 51 vini di alta qualità $(0, 2]$ su un totale di 60 osservazioni nel nodo. La presenza di 9 vini di seconda classe rappresenta un errore fisiologico di sovrapposizione, ma è cruciale notare la totale assenza di vini di bassa qualità: il modello, in questo ramo, non commette mai l'errore grave di classificare come "Premium" un vino di bassa qualità.

La classe intermedia $(2, 3]$ trova la sua collocazione principale nel **Nodo 8**, che raccoglie la maggioranza assoluta dei vini di questa categoria (56 osservazioni). Sebbene questo nodo presenti una lieve contaminazione dalle classi adiacenti (2 vini alti e 3 bassi), la sua capacità di isolare i vini di media qualità è elevata.

I nodi (3, 5 e 10) contengono un numero esiguo di osservazioni residuali, gestendo le situazioni di confine dove i profili chimici risultano meno netti. Nel complesso, la tabella conferma che le regole chimiche basate su OD280 e Prolina riescono a polarizzare efficacemente il dataset, minimizzando l'incertezza classificatoria.

4.4 Simulazione: Previsione su Dati Fittizi Non Etichettati

Al fine di testare la robustezza del modello *CHAID-Interval*, è stata eseguita una simulazione su 5 nuovi campioni di vini ($N = 5$). I dati sono fittizi, scritti a mano a partire dai dati originali, contengono nuovi valori non presenti nei dati con i quali è stato addestrato il modello. A differenza della fase di training, in questo scenario il modello non conosce a priori la classe di appartenenza dei vini (dataset "cieco"); l'obiettivo è assegnare un'etichetta di qualità basandosi esclusivamente sui parametri chimici analitici.

Table 9: Dataset "Blind" (Cieco) di Validazione. I campioni non presentano etichetta di classe. Il Campione ID 3 presenta caratteristiche chimiche contrastanti (OD280 basso ma Prolina alta).

ID	Alc	Malic	Ash	Alc.Ash	Mg	Phenols	Flav	NonFlav	Proan	Col.Int	Hue	OD280	Proline
1	14.5	1.8	2.4	15	110	3.0	3.2	0.2	1.8	5.5	1.10	3.50	1200
2	12.0	3.5	2.2	20	85	1.5	0.8	0.5	0.9	3.0	0.70	1.50	400
3	14.0	2.0	2.3	18	95	2.2	2.0	0.3	1.5	4.5	0.95	**1.45**	1100
4	15.5	1.5	2.5	16	120	3.5	3.9	0.2	2.0	6.0	1.20	3.80	1600
5	11.5	2.5	2.1	21	88	1.8	0.9	0.4	1.1	3.5	0.80	1.60	450

La metodologia prevede l'applicazione dei *cut-points* ereditati dal training set alle nuove osservazioni, seguita dalla regola di decisione derivata dall'albero finale.

Table 10: Classificazione automatica dei 5 campioni basata sull'albero decisionale.

ID	OD280	Predizione	Interpretazione Qualitativa
1	3.50	(0,2]	Alta Qualità (Premium)
2	1.50	(3,5]	Bassa Qualità (Entry)
3	**1.45**	**(2,3]**	**Media Qualità (Standard)**
4	3.80	(0,2]	Alta Qualità (Premium)
5	1.60	(3,5]	Bassa Qualità (Entry)

Il modello ha assegnato le seguenti etichette ai campioni analizzati:

- **Campioni Standard (ID 1, 2, 4, 5):** Il modello ha operato una segmentazione netta.
  - I campioni con struttura robusta (ID 1 e 4, con $OD280 > 3.5$) sono stati classificati come **Alta Qualità (Premium)**, classe $(0, 2]$.
  - I campioni deboli (ID 2 e 5, con $OD280 < 1.6$) sono stati correttamente indirizzati verso la **Bassa Qualità**, classe $(3, 5]$.
- **Il Caso Contraddittorio (ID 3):** Il campione presenta un profilo ibrido: valori di *Prolina* da top di gamma (1100), ma un grado di diluizione (*OD280*) critico (1.45).

  Esito: Il modello lo ha classificato come **Media Qualità (Standard)**, classe $(2, 3]$.

L'albero decisionale ha dato priorità gerarchica alla variabile radice ($OD280$). Poiché il valore 1.45 è inferiore alla soglia critica (2.18), il vino è stato escluso dal ramo dell'eccellenza, rendendo irrilevante l'alto contenuto di Prolina. Tuttavia, il modello non lo ha penalizzato fino alla fascia più

bassa, riconoscendo verosimilmente caratteristiche secondarie sufficienti per mantenerlo in una fascia di mercato intermedia.

Il test conferma che il modello agisce come un sistema di controllo qualità conservativo: la mancanza di requisiti strutturali minimi (OD280) impedisce l'accesso alla classificazione Premium, indipendentemente dalla ricchezza di altre variabili chimiche.

5 Analisi tramite Algoritmo CART con Pruning Ottimale

Per approfondire la comprensione della struttura predittiva dei dati e validare le regole decisionali emerse, è stato applicato l'algoritmo CART (Classification and Regression Trees). Il dataset è stato partizionato in un **Training Set** (70%, $N = 125$) per l'addestramento del modello e un **Test Set** (30%, $N = 53$) per la valutazione delle performance su dati non visti. L'approccio adottato prevede la costruzione di un albero di taglia massima a cui segue una fase di *pruning*. L'analisi si concentra poi sulla selezione del parametro di complessità ottimale (*cp*), sulla struttura gerarchica delle regole decisionali derivate e sull'importanza chimica delle variabili. L'obiettivo è comprendere come l'algoritmo segmenti lo spazio dei parametri chimici per distinguere le tre classi di vino target: (0,2], (2,3] e (3,5].

L'algoritmo CART opera attraverso un processo di "partizionamento ricorsivo", suddividendo il dataset in sottogruppi sempre più omogenei. Tuttavia, un albero lasciato crescere indefinitamente tenderebbe a memorizzare il rumore statistico dei dati di addestramento (*overfitting*), riducendo la sua capacità predittiva su nuovi dati.

Per ovviare a questo problema, l'albero è stato successivamente tagliato nel punto in cui il bilanciamento tra complessità dell'albero ed errore di validazione è risultato ottimale.

Table 11: Complexity Parameter Table (CP Table) dell'albero CART iniziale

CP	**nsplit**	**rel error**	**xerror**	**xstd**
0.41447	0	1.000	1.000	0.0718
0.09211	2	0.171	0.250	0.0528
0.05263	3	0.079	0.197	0.0478
0.00010	**4**	**0.026**	**0.118**	**0.0380**

Come evidenziato in Tabella 11 e visualizzato nel grafico dell'errore relativo (Figura 7), si nota un netto miglioramento delle performance con l'aumentare della complessità dell'albero:

- L'albero radice (0 split) presenta un errore relativo di base pari a 1.0.
- Con soli 2 split ($CP \approx 0.09$), l'errore incrociato (*xerror*) crolla drasticamente a 0.25, indicando che le prime variabili selezionate catturano la maggior parte dell'informazione discriminante.
- Il minimo assoluto dell'errore di cross validation ($xerror = 0.118$) si raggiunge con 4 split, corrispondente a un $CP = 0.0001$. Questo valore è stato selezionato come parametro ottimale per la potatura dell'albero finale.

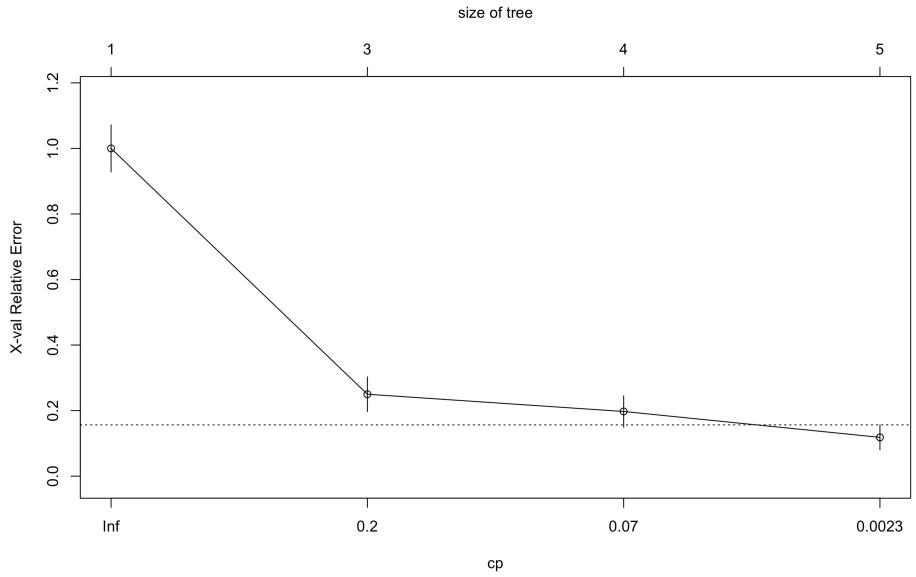

Figure 7: Grafico dell'errore relativo in validazione incrociata (x-val Relative Error) in funzione del parametro di complessità (cp). Si osserva la discesa dell'errore fino al minimo in corrispondenza di cp=0.0001.

5.1 Analisi dell'Albero Potato e Importanza delle Variabili

L'albero finale potato (*Pruned CART*), visualizzato in Figura 8, presenta una struttura gerarchica chiara che conferma il ruolo cruciale dei polifenoli e degli amminoacidi nella classificazione dei vini.

Importanza delle Variabili:

L'algoritmo ha calcolato l'importanza relativa delle variabili predittive, evidenziando una gerarchia distinta: **Flavanoids** e **Color Intensity** emergono come i predittori più forti, dominando la struttura decisionale. Seguono **Proline**, **OD280** e **Hue**, che agiscono come raffinatori nei livelli successivi dell'albero. Variabili come *Magnesium* e *Alcalinity of Ash* hanno un ruolo marginale.

Interpretazione delle Regole Decisionali: L'albero (Figura 8) opera le seguenti distinzioni fondamentali:

1. **Nodo Radice (Split 1):** La prima divisione avviene sui **Flavanoids** con soglia 1.4. Valori bassi (< 1.4) indirizzano verso il ramo destro, isolando efficacemente i vini della classe (3,5] (bassa qualità), ulteriormente filtrati dall'intensità del colore.
2. **Ramo Sinistro (Alta Qualità):** I vini con flavonoidi elevati vengono successivamente valutati in base alla **Proline**. Una concentrazione di prolina ≥ 676 identifica in modo quasi deterministico la classe (0,2] (alta qualità), mentre valori inferiori, combinati con l'intensità del colore, definiscono la classe intermedia (2,3].

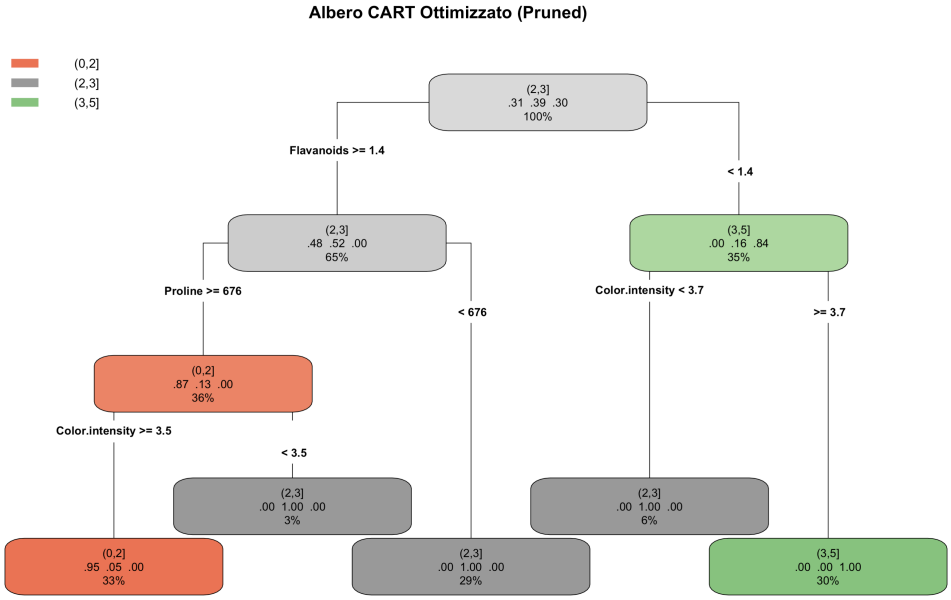

Figure 8: Albero decisionale CART ottimizzato (Pruned). I nodi colorati rappresentano le classi predette: Arancione per (0,2], Grigio per (2,3], Verde per (3,5]. Sotto ogni nodo sono riportate le probabilità di appartenenza alle tre classi e la percentuale di osservazioni nel nodo.

5.2 Albero Completo vs Potato

Il confronto tra l'albero completo iniziale e l'albero potato evidenzia l'efficacia della procedura di pruning. Mentre l'albero completo (non mostrato) tende a frammentare eccessivamente i dati in nodi terminali con poche osservazioni (overfitting), l'albero potato (Figura 8) mantiene solo i rami statisticamente significativi. I nodi terminali risultanti mostrano un'elevata purezza (spesso del 100% o vicina), indicando regole decisionali robuste e ben definite. Ad esempio, il nodo terminale più a sinistra identifica la classe (0,2] con una probabilità del 95%, mentre il nodo più a destra identifica la classe (3,5] con una probabilità dell'84%.

5.3 Performance sul Test Set

La capacità di generalizzazione del modello è stata verificata applicando l'albero potato al Test Set ($N = 53$). Il modello ha ottenuto un'**accuratezza globale dell'83.02%**. Questo risultato, coerente con l'errore incrociato stimato in fase di training, conferma che le regole chimiche estratte (basate su flavonoidi, prolina e intensità del colore) sono valide e predittive anche su dati non utilizzati durante la costruzione del modello.

6 Analisi del Modello Bagging

L'implementazione dell'algoritmo di Bagging (*Bootstrap Aggregating*) rappresenta un'evoluzione significativa rispetto ai singoli alberi decisionali, mirando a ridurne la varianza e ad aumentarne la stabilità predittiva. In questo casi, il modello è stato costruito aggregando 500 alberi decisionali (500 campioni bootstrap), dove ogni albero ha considerato la totalità delle variabili disponibili per ogni suddivisione.

6.1 Valutazione della Performance Predittiva

L'analisi delle metriche di errore restituisce un quadro duplice sulla capacità del modello. Da un lato, la stima dell'errore **Out-Of-Bag (OOB)**, calcolata internamente sul set di addestramento utilizzando i campioni esclusi dal bootstrap, si attesta su un valore estremamente basso del **4.80%**. Questo dato suggerisce che il modello ha appreso con grande efficacia la struttura interna dei dati di training, riuscendo a classificare quasi perfettamente le osservazioni durante la fase di costruzione.

Dall'altro lato, la validazione sul **Test Set** esterno fornisce una misura più realistica della capacità di generalizzazione, registrando un'accuratezza del **83.02%**. Sebbene questo risultato confermi la solidità dell'approccio, si nota una divergenza rispetto alla stima OOB; tale scarto indica che, mentre il Bagging è estremamente potente nel catturare le dinamiche del campione di apprendimento, la complessità intrinseca di alcuni nuovi vini nel test set rimane una difficoltà, portando il tasso di errore reale dal teorico 4.8% a un effettivo 17% circa.

6.2 Analisi dell'Importanza delle Variabili

Uno dei contributi più rilevanti dell'approccio Bagging è la possibilità di quantificare il peso specifico di ogni predittore nella classificazione. Come evidenziato nel grafico *Variable Importance Plot* riportato di seguito, esiste una gerarchia chimica ben definita che guida le decisioni del modello.

Osservando entrambe le metriche — la *Mean Decrease Accuracy* e la *Mean Decrease Gini* — emerge quali siano i marcatori chimici fondamentali:

- **I Driver Primari:** Le variabili *Proline*, *Flavanoids* e *Color.intensity* dominano la classifica in entrambi i grafici. In particolare, la Prolina risulta essere la variabile più critica per l'accuratezza del modello: la sua esclusione o permutazione causerebbe il crollo più significativo nelle performance predittive. Questo conferma che la struttura aminoacidica e i polifenoli sono i tratti distintivi più forti per separare le tre classi di vino.
- **I Driver Secondari:** Variabili come *Hue* (Tonalità) e *OD280.OD315* (Grado di diluizione) giocano un ruolo di supporto importante, posizionandosi stabilmente nella fascia medio-alta dell'importanza.
- **Variabili Marginali:** Al contrario, parametri come *Ash* (Ceneri) e *Total.phenols* appaiono in fondo alla graduatoria. Questo suggerisce che, per questo specifico dataset, il contenuto minerale o l'acidità totale contribuiscono in modo trascurabile alla discriminazione tra le classi rispetto alla potenza cromatica e alla composizione flavonoica.

In conclusione, il Bagging non solo ha fornito una buona accuratezza predittiva, ma ha anche permesso di isolare il "profilo chimico" ideale per la segmentazione, indicando che la qualità o la tipologia del vino è dettata principalmente dalla sua intensità cromatica e dalla sua complessità strutturale (Prolina e Flavonoidi).

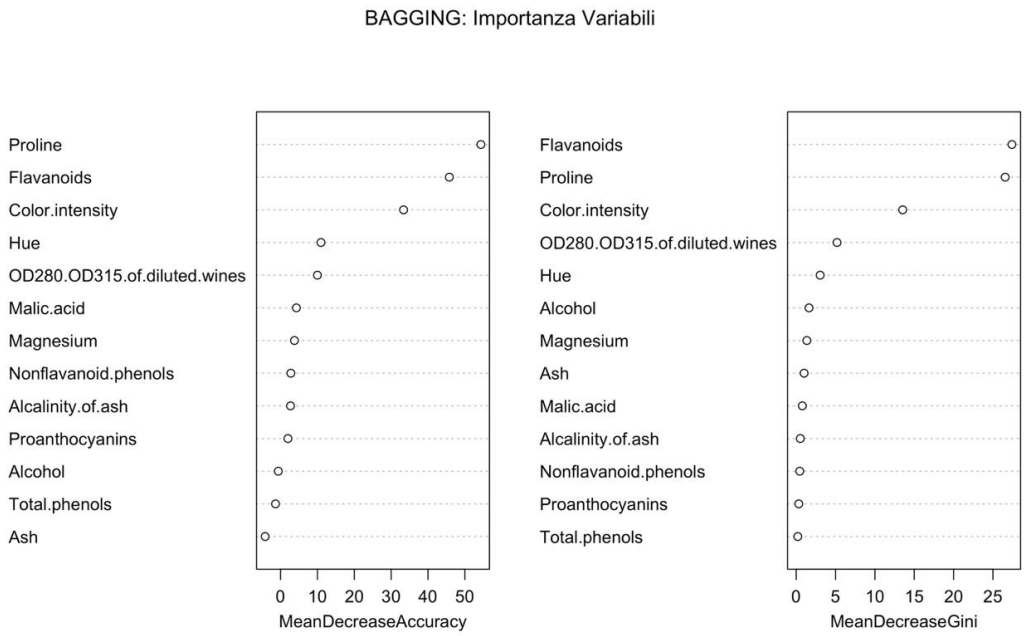

Figure 9: Grafico dell'importanza delle variabili nel modello Bagging: a sinistra la diminuzione media dell'accuratezza, a destra la diminuzione media dell'indice di Gini.

7 Random Forest e Confronto Modelli

Per superare i limiti dell'albero singolo (CART), che può soffrire di instabilità e varianza elevata, è stato implementato l'algoritmo **Random Forest** (RF). Questo metodo *ensemble* costruisce una "foresta" di 500 alberi decisionali decorrelati, introducendo casualità sia nella selezione del campione (bootstrap) che nella scelta delle variabili a ogni nodo ($numerodivariabili \approx \sqrt{13}$).

7.1 Performance del Random Forest

Il modello Random Forest ha dimostrato una capacità predittiva superiore rispetto ai metodi precedenti.

- **Errore Out-Of-Bag (OOB):** La stima interna dell'errore, calcolata sulle unità non utilizzate per la costruzione di ciascun albero, si è assestata su un valore molto basso del **5.60%**. Questo indica che il modello è estremamente robusto e generalizza bene sui dati di training.
- **Accuratezza sul Test Set:** Applicato al Test Set indipendente ($N = 53$), il modello ha raggiunto un'**accuratezza dell'88.68%**, confermando la stima OOB e superando nettamente le prestazioni dell'albero singolo.

7.2 Importanza delle variabili RF

L'algoritmo Random Forest permette di quantificare il contributo di ciascuna variabile chimica alla capacità discriminante del modello. I grafici di *Variable Importance* (Figura 10) mostrano due metriche fondamentali:

1. **Mean Decrease Accuracy:** Quanto peggiora il modello se si "mescolano" i valori di quella variabile (quanto è indispensabile).

2. **Mean Decrease Gini:** Quanto quella variabile contribuisce alla purezza dei nodi (quanto è discriminante).

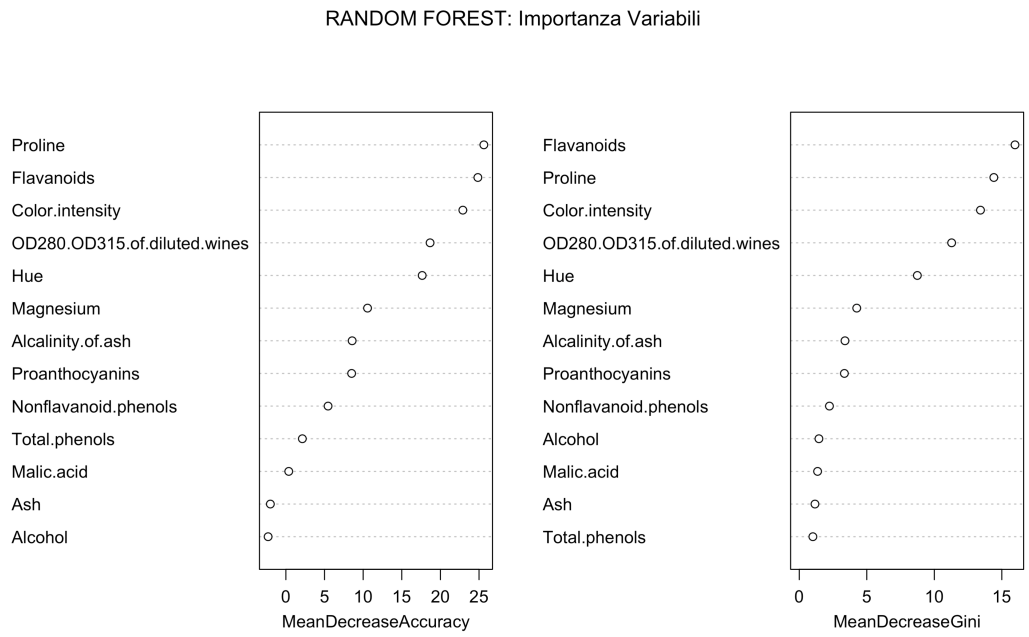

Figure 10: Grafico dell'Importanza delle Variabili nel Random Forest. A sinistra: diminuzione dell'accuratezza media. A destra: diminuzione dell'indice di Gini medio.

La variabile più importante risulta essere Proline.

Dall'analisi della Figura 10 emergono tre variabili fortemente determinanti:

- **Proline** e **Flavanoids** si confermano le variabili più critiche. La loro rimozione causerebbe il crollo maggiore dell'accuratezza, indicando che contengono informazioni uniche e insostituibili.
- Seguono **Color Intensity** e **OD280**, confermando la struttura gerarchica già vista nell'albero singolo.
- È interessante notare come variabili come *Ash* (Ceneri) e *Alcohol*, pur essendo parametri enologici standard, abbiano un impatto quasi nullo sulla classificazione in questo specifico contesto (si trovano in fondo alla classifica).

8 Confronto Finale dei Modelli

La Tabella 12 riassume le prestazioni dei tre approcci testati sul dataset di validazione ($N = 53$).

Table 12: Classifica Finale dei Modelli (ordinati per Accuratezza sul Test Set)

Modello	**Accuratezza Test Set**	**Errore OOB Stimato**
Random Forest	**88.68%**	**5.60%**
CART (Pruned)	83.02%	N/A
Bagging	83.02%	4.80%

Sintesi dei Risultati:

Il Random Forest domina la classifica con un distacco del +5.6% di accuratezza rispetto a CART e Bagging. La decorrelazione degli alberi (scegliendo solo un sottoinsieme di variabili a ogni nodo) ha permesso di catturare sfumature che il Bagging, che usa tutte le variabili e tende a creare alberi molto simili tra loro, non ha colto.

Sebbene il Random Forest sia il più accurato, l'albero CART singolo (con l'83% di accuratezza) rimane un ottimo strumento per la comunicazione delle regole decisionali grazie alla sua trasparenza grafica.

La coerenza tra l'Errore OOB (5.6%) e l'errore reale sul Test Set ($100 - 88.68 = 11.32\%$) suggerisce che il modello non soffre di overfitting eccessivo, anche se l'errore reale è leggermente superiore alla stima ottimistica interna.

9 Conclusioni: Confronto Analitico Approfondito tra CHAID Interval e Random Forest

L'evoluzione dell'analisi condotta in questo progetto permette di istituire un confronto critico tra due filosofie di modellazione profondamente diverse: l'approccio statistico-descrittivo del CHAID (nella sua variante ottimale a Intervalli) e l'approccio algoritmico-predittivo del Random Forest. Sebbene entrambi i modelli abbiano raggiunto livelli di accuratezza comparabili, oscillanti attorno alla soglia dell'88-89%, i meccanismi attraverso cui giungono alla decisione riflettono due modi opposti di interpretare i dati chimici.

Il CHAID con discretizzazione a Intervalli opera attraverso una semplificazione conscia della complessità. Trasformando le variabili continue in fasce discrete (Basso, Medio, Alto), questo metodo agisce come un potente regolarizzatore naturale: rinuncia alla precisione decimale del singolo valore chimico per catturare invece la "categoria" di appartenenza. Nel contesto enologico, questo approccio è risultato vincente rispetto ad altri metodi di discretizzazione perché mima il ragionamento umano e i protocolli di controllo qualità, dove non conta il micro-grammo di differenza, ma il posizionamento del vino in un range di tolleranza standard. La sua forza risiede nella trasparenza: l'albero generato è unico, visibile e fornisce regole di segmentazione immediate, definendo "a priori" quali siano i confini chimici della qualità. Tuttavia, questa rigidità rappresenta anche il suo limite intrinseco; imponendo intervalli fissi, il modello potrebbe non cogliere quelle sottili non-linearità o interazioni locali che avvengono proprio al confine tra due fasce, rischiando di classificare erroneamente i casi borderline.

Diametralmente opposto è il funzionamento del Random Forest, che non semplifica il dato ma lo sfrutta nella sua interezza. Lavorando direttamente sui valori continui e potendo scegliere soglie di taglio diverse per ognuno dei 500 alberi generati, questo algoritmo ha la capacità di adattarsi plasticamente alla struttura dei dati. La sua superiorità tecnica, evidenziata dal minimo errore Out-Of-Bag registrato, deriva dalla legge dei grandi numeri: mentre il CHAID scommette tutto su un'unica struttura ad albero (che potrebbe essere influenzata da specificità del campione di training), il Random Forest mitiga il rischio e la varianza aggregando centinaia di "opinioni" diverse.

Il confronto evidenzia dunque un trade-off fondamentale tra interpretabilità operativa e potenza predittiva pura. Il CHAID Interval si configura come lo strumento ideale per la definizione di linee guida e standard produttivi, offrendo ai tecnologi del vino soglie chiare per l'assegnazione della qualità. Il Random Forest, al contrario, emerge come il motore predittivo per eccellenza, una "black box" estremamente affidabile da utilizzare in fase di validazione automatica, capace di catturare sfumature multivariate che sfuggono alla logica a gradini degli intervalli fissi. Il fatto che un metodo semplice come l'Interval Discretization abbia ottenuto performance così vicine a un metodo complesso come il Random Forest suggerisce che, per questo specifico dataset, la struttura che lega la chimica alla qualità è forte e ben definita, permettendo anche a un modello rigido di catturare l'essenza del fenomeno quasi quanto un algoritmo avanzato.