

Integrazione di conoscenza di dominio e apprendimento automatico per la previsione della qualità del vino

Gruppo di lavoro

- Giuseppe Murano, 758407, g.murano2@studenti.uniba.it

https://github.com/P3pp0uz/Progetto_ICON_2024_2025_Giuseppe_Murano.git

AA 2024-2025

Sommario

Introduzione.....	3
Sommario.....	3
Dataset	3
Apprendimento non supervisionato.....	6
Apprendimento supervisionato.....	10
Ragionamento probabilistico.....	16
Rappresentazione della Conoscenza	19
Conclusioni	21

Introduzione

Il progetto affronta la predizione della qualità del vino a partire da misure fisico-chimiche su un dataset UCI unificato per rossi e bianchi, integrando conoscenza di dominio, modelli probabilistici e apprendimento supervisionato e non supervisionato in un'unica pipeline eseguibile da main.py. L'obiettivo è dimostrare competenze pratiche "in programma" su rappresentazione della conoscenza e ragionamento, apprendimento e valutazione rigorosa, con particolare enfasi sulla KB e sulle decisioni motivate di configurazione dei modelli.

Sommario

Il sistema integra cinque diversi moduli che dimostrano competenze nei seguenti argomenti del corso:

- 1) Una **Knowledge Base (KB) basata su regole** per l'inferenza della qualità dei vini, utilizzando il paradigma di **forward chaining** descritto nel **Capitolo 13: Ontologie e Knowledge-Based Systems** del libro.
- 2) Una **pipeline di apprendimento supervisionato multi-algoritmo** direttamente collegata ai contenuti del **Capitolo 7** ("Learning Overview and Supervised Learning").
- 3) Un **MLP (Multilayer Perceptron) regolarizzato** secondo i modelli descritti nel **Capitolo 7.3.2** (Linear Regression and Classification) e il Capitolo 8 sulle reti neurali.
- 4) Dei modelli probabilistici grafici per la classificazione descritti nel Capitolo 6 ("Reasoning Under Uncertainty").
- 5) Implementa **clustering non supervisionato** con K-Means e selezione automatica del numero di cluster, collegato ai capitoli su apprendimento non supervisionato.

Dataset

Nel presente caso di studio si è scelto di utilizzare il **Wine Quality dataset**, un insieme di dati ampiamente riconosciuto nel campo dell'analisi enologica e della valutazione qualitativa dei vini. Questo dataset contiene **6497 istanze** complessive (1599 vini rossi + 4898 vini bianchi) estratte da campioni reali di "Vinho Verde", un vino tipico del nord del Portogallo, le quali descrivono le caratteristiche chimico-fisiche derivanti da test di laboratorio oggettivi.

Composizione Dataset

Il dataset è suddiviso in due varianti:

- **Vini rossi** (winequality-red.csv): 1599 campioni
- **Vini bianchi** (winequality-white.csv): 4898 campioni

Entrambi i sottoinsiemi condividono la stessa struttura di 11 attributi di input chimico-fisici e 1 attributo di output qualitativo basato su valutazioni sensoriali di esperti enologi.

Attributi

Variabili di Input (Feature Chimico-Fisiche)

Tutte le feature sono di tipo **continuo** e derivano da analisi di laboratorio standardizzate:

1. **Fixed Acidity** (Acidità fissa): concentrazione di acidi non volatili (es. tartarico, malico, citrico) espressa in g/dm^3
2. **Volatile Acidity** (Acidità volatile): concentrazione di acido acetico, indicatore di fermentazione indesiderata se elevata, espressa in g/dm^3
3. **Citric Acid** (Acido citrico): concentrazione di acido citrico, che contribuisce alla freschezza del vino, espressa in g/dm^3
4. **Residual Sugar** (Zucchero residuo): quantità di zucchero rimanente dopo la fermentazione, espressa in g/dm^3
5. **Chlorides** (Cloruri): concentrazione di sale, espressa in g/dm^3
6. **Free Sulfur Dioxide** (Anidride solforosa libera): forma libera di SO_2 , agente antimicrobico e antiossidante, espressa in mg/dm^3
7. **Total Sulfur Dioxide** (Anidride solforosa totale): somma delle forme libere e legate di SO_2 , espressa in mg/dm^3
8. **Density** (Densità): massa volumica del vino, correlata al contenuto alcolico e zuccherino, espressa in g/cm^3
9. **pH**: misura dell'acidità/alcalinità su scala logaritmica
10. **Sulphates** (Solfati): concentrazione di solfato di potassio, additivo antimicrobico, espressa in g/dm^3
11. **Alcohol** (Alcol): percentuale di alcol etilico in volume (% vol)

Variabile di Output (Target)

12. **Quality** (Qualità): punteggio qualitativo assegnato da esperti enologi su scala discreta da 0 a 10, basato su valutazione sensoriale (colore, aroma, gusto)
 - **Distribuzione sbilanciata**: la maggior parte dei campioni ha qualità medio-bassa (5-6), con pochi vini eccellenti (8-9) o scarsi (3-4)
 - **Classificazione binaria**: nel progetto, la qualità viene binarizzata con soglia ≥ 6 per distinguere "Alta qualità" (1) da "Bassa qualità" (0)

Caratteristiche del Dataset

- **Dominio reale**: i dati provengono da campioni fisici sottoposti ad analisi chimiche certificate secondo standard CVRVV (Commissione per la Viticoltura della Regione del Vinho Verde)
- **Task supportati**: classificazione (binaria o multi-classe), regressione, rilevamento outlier per vini eccezionali
- **Privacy**: per motivi di riservatezza commerciale, il dataset non include informazioni su vitigno, marca, prezzo di vendita o origine geografica specifica

Fase di Preprocessing

La funzione `preprocess_data ()` si occupa di preparare il dataset per l'analisi attraverso quattro operazioni principali:

1. Normalizzazione dei nomi delle colonne: I nomi delle colonne vengono ripuliti rimuovendo spazi bianchi iniziali e finali (`strip ()`) e identificando pattern di colonne non valide (e.g., `Unnamed:0`). Questa fase garantisce consistenza nella nomenclatura, fondamentale per l'integrazione con la knowledge base enologica.

2. Codifica del tipo di vino: Se presente, la colonna categorica `type` viene mappata numericamente ('red' → 0, 'white' → 1) per consentire l'elaborazione da parte degli algoritmi di machine learning. Questa trasformazione preserva l'informazione tipologica in formato compatibile con i modelli matematici.

3. Conversione numerica robusta: Tutte le 11 feature chimico-fisiche note del dataset UCI (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol) e la variabile target `quality` vengono esplicitamente convertite a tipo numerico tramite `pd.to_numeric ()` con parametro `errors='coerce'`. Questa strategia gestisce eventuali errori di parsing convertendo valori non parsabili in NaN, evitando interruzioni del flusso di elaborazione.

4. Gestione valori mancanti: Sebbene il dataset UCI Wine Quality ufficiale non contenga valori mancanti, il sistema implementa una verifica precauzionale rimuovendo istanze con NaN solo sulle 11 feature chimico-fisiche critiche, senza eliminare l'intero DataFrame. Questo approccio selettivo preserva la massima quantità di informazione disponibile.

Binarizzazione del Target

La funzione `make_binary_target ()` converte la variabile qualitativa ordinale `quality` (scala 0-10) in un target binario per classificazione:

- **Alta qualità (1):** vini con punteggio ≥ 6
- **Bassa qualità (0):** vini con punteggio < 6

La soglia di 6 è stata scelta analizzando la distribuzione sbilanciata del dataset, dove la maggior parte dei campioni si concentra nei valori 5-6. Il sistema include un rilevamento automatico intelligente della colonna qualità tramite pattern matching su nomi candidati (`quality`, `Quality`, `quality_score`) e analisi euristica per identificare colonne con valori discreti nell'intervallo.

Apprendimento non supervisionato

Sommario

Il modulo **unsupervised learning** implementa una pipeline di **clustering K-Means** per scoprire strutture naturali nei dati dei vini senza etichette predefinite. L'obiettivo è identificare gruppi omogenei di vini basati su similarità delle proprietà chimico-fisiche, quindi estrarre feature semantiche aggiuntive (etichette cluster, one-hot encoding, distanze dai centri) che arricchiscono il dataset originale per le fasi successive di apprendimento supervisionato e reti neurali.

La pipeline esegue sequenzialmente: normalizzazione MinMaxScaler dei dati, calcolo della **Within Sum of Squares (WSS)** per un range di k valori (2-10), applicazione dell'**euristica del gomito** per selezione automatica di k ottimale, clustering K-Means effettivo, estrazione di tre rappresentazioni feature, visualizzazione mediante PCA 2D dei cluster e grafico elbow.

Strumenti utilizzati

Librerie Principali

scikit-learn fornisce tre componenti essenziali: KMeans per clustering, PCA per riduzione dimensionalità e visualizzazione, MinMaxScaler per normalizzazione.

Numpy gestisce operazioni matriciali,

Matplotlib genera visualizzazioni.

Decisioni di Progetto

Normalizzazione MinMaxScaler

I dati vengono scalati nel range anziché standardizzati (media 0, varianza 1) per un motivo specifico: il calcolo della distanza euclidea in K-Means è sensibile alla scala delle feature. MinMaxScaler garantisce che tutte le feature (alcolici, acidi, densità, etc.) contribuiscano equamente al calcolo della distanza euclidea, evitando che feature con range più grande (es. solidi disciolti totali) dominino il clustering.

Range di k e Convergenza

Il range 2-10 per k è scelto perché rappresenta il compromesso tra semplicità interpretativa (k troppo grande frammenta i dati) e capacità di catturare varietà naturale nei vini (k=2 potrebbe essere troppo semplicistico). Durante il calcolo WSS, n_init=20 (multipli tentavi) e max_iter=500 garantiscono convergenza robusta di K-Means anche per k grandi.

Estrazione Triplice di Feature

Dopo clustering, vengono estratte tre rappresentazioni alternative dal medesimo modello K-Means, catturando aspetti complementari della struttura scoperta:

- **Etichette cluster (1 colonna):** identificazione diretta del cluster, variabile categorica discreta, codifica l'appartenenza di massa
- **One-hot encoding (k colonne):** rappresentazione sparsa con 1 nella posizione del cluster, 0 altrove, fornisce feature binarie esplicite per modelli lineari
- **Distanze dai centri (1 colonna):** quanto "periferico" è il punto nel suo cluster, cattura irregolarità geometriche della struttura, valori alti = punti al confine tra cluster

Questa triplice estrazione massimizza l'informazione semantica disponibile per algoritmi supervisionati successivi, poiché ogni rappresentazione cattura aspetti diversi della struttura trovata.

Nel `main.py`, clustering non supervisionato viene eseguito **prima** di apprendimento supervisionato, permettendo che le feature estratte arricchiscano il dataset originale prima dell'addestramento dei modelli.

Valutazione

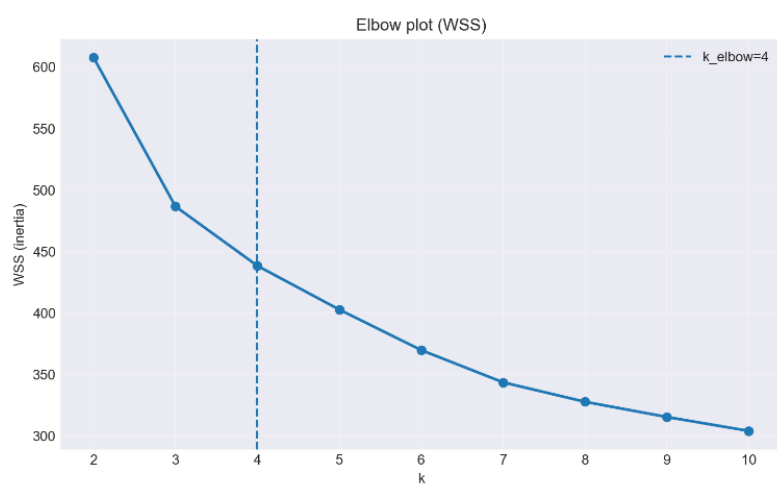
Metriche Elbow Method

Within Sum of Squares (WCSS): WCSS diminuisce monotonamente con k crescente (aggiungere cluster riduce sempre WCSS), ma la **velocità di diminuzione** rallenta dopo il k ottimale.

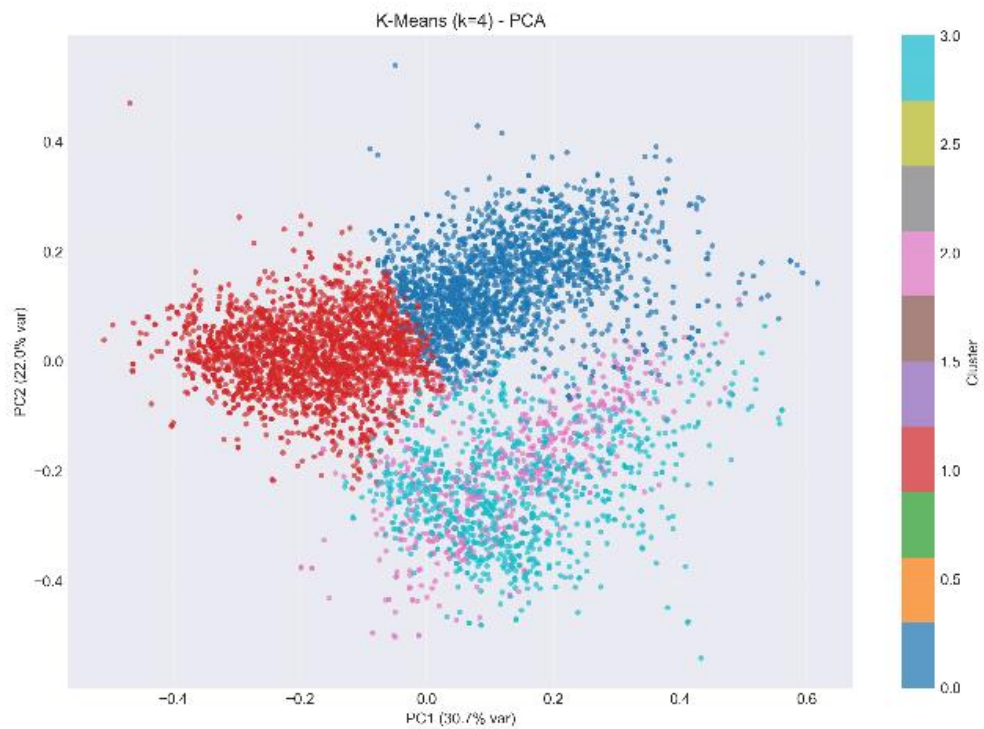
Punto gomito: identificato da `_elbow_k()` come il valore di k con massima distanza perpendicolare dal segmento che connette $(2, WSS_2)$ a $(10, WSS_{10})$. Rappresenta il trade-off ottimale tra semplicità (pochi cluster) e fit (WCSS basso).

Visualizzazioni Generate

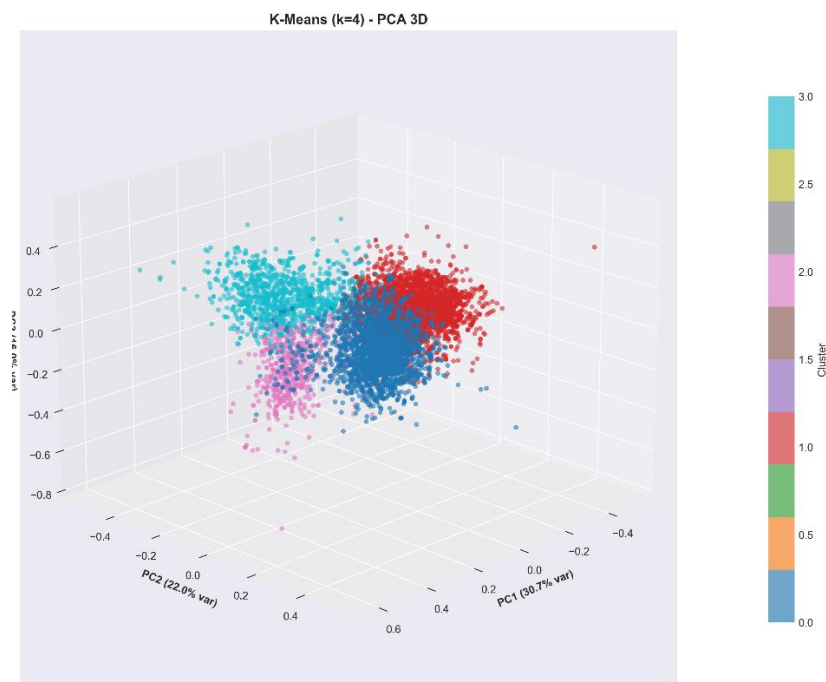
Elbow Plot: Grafico lineare WCSS vs k con linea verticale al k_{elbow} scelto. Permette verifica visiva della scelta automatica e identificazione manuale di gomiti alternativi se la curvatura è ambigua.



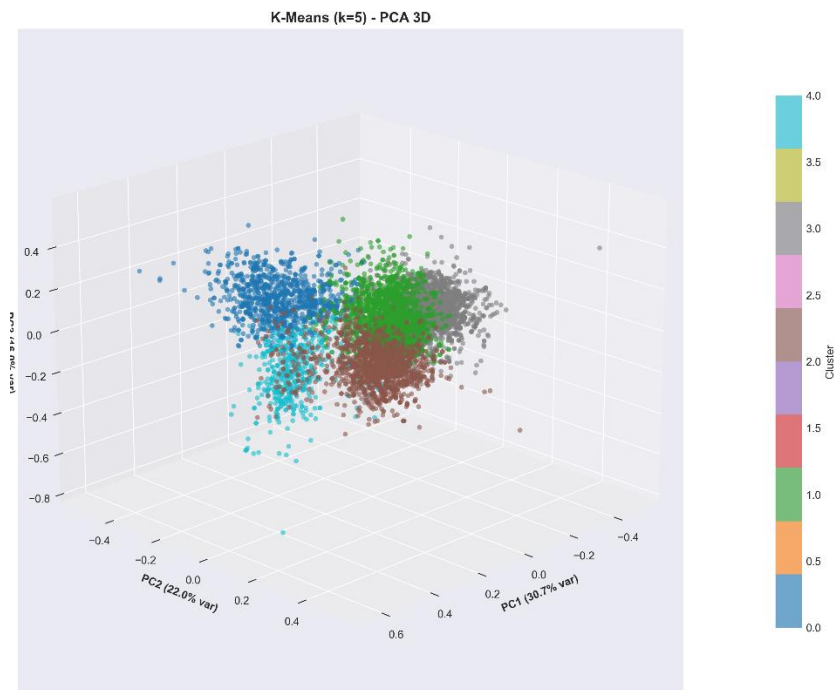
Clustering Visualization 2D via PCA: Scatter plot che proietta i dati sui primi due principal components (PC_1 e PC_2), con colori differenziati per cluster assegnato. Rivela la **struttura geometrica effettiva** del clustering: cluster ben separati appaiono come regioni spaziali distinte, cluster sovrapposti appaiono intrecciati.



dal pca 2d non si riesce a capire se il cluster rosa e blu sono un unico cluster o due cluster divisi quindi analizziamo il pca 3d



Dalla visualizzazione 3d si può notare la divisione tra i cluster



Dal pca con $k = 5$ notiamo che si hanno comunque 4 divisioni nette quindi si continua con 4 cluster

Apprendimento supervisionato

Sommario

La pipeline completa di classificazione combina due paradigmi complementari di machine learning per la predizione della qualità dei vini (alta/bassa).

Apprendimento Supervisionato implementa una comparazione di 5 algoritmi classici (KNN, Decision Tree, Random Forest ottimizzato, Gradient Boosting, SVM) con regolarizzazione anti-overfitting, eseguendo feature selection, normalizzazione, ottimizzazione iperparametri con RandomizedSearchCV, e cross-validation stratificata 5-fold.

Reti Neurali implementano un'architettura MLP leggera (64-32 hidden layers) adatta a dataset di piccola-media dimensione (~6000 campioni), con early stopping per prevenire overfitting, regolarizzazione L2 ($\alpha=0.03$), e learning rate adattivo. Entrambi i moduli utilizzano lo stesso dataset, split stratificato 85-15%, metriche comparative standardizzate (Accuracy, Precision, Recall, F1, MCC) e visualizzazioni unificate per decidere quale approccio è più efficace.

Strumenti utilizzati

Entrambi i moduli utilizzano scikit-learn: MLPClassifier per le reti neurali, KNeighborsClassifier, DecisionTreeClassifier, RandomForestClassifier, GradientBoostingClassifier, SVC per apprendimento supervisionato.

StandardScaler normalizza i dati in entrambi gli approcci.

Intel Extension for Scikit-Learn (sklearnex) fornisce accelerazione computazionale.

Architettura MLP

La rete neurale artificiale implementa una **topologia Multilayer Perceptron** con backpropagation:

- **Input layer:** numero feature pari alla dimensionalità del dataset processato
- **Hidden layer 1:** 64 neuroni con attivazione ReLU per non-linearità
- **Hidden layer 2:** 32 neuroni con attivazione ReLU per rappresentazioni apprese complesse
- **Output layer:** 2 neuroni (probabilità classi binarie)
- **Solver:** Adam (adaptive moment estimation) per ottimizzazione del gradiente

Regolarizzazione nella Rete

Regolarizzazione L2 ($\alpha=0.03$) penalizza i pesi grandi durante il training, prevenendo sovraadattamento. Early stopping monitora un validation set del 20% e interrompe il training se non ci sono miglioramenti per 15 iterazioni consecutive. Batch size 64 equilibra stabilità e velocità computazionale.

Decisioni di Progetto

Comparazione Metodologica

La scelta di unificare il preprocessing fra i due moduli (supervisionato e reti neurali) risponde a un principio fondamentale della valutazione empirica comparativa: quando diversi paradigmi di apprendimento lavorano su rappresentazioni identiche, le differenze prestazionali riflettono genuini trade-off algoritmici piuttosto che artefatti di preprocessing.

Lo split stratificato 85-15% preserva la distribuzione di classe nel train e test, essenziale con dati leggermente sbilanciati (53% vs 47%). La feature selection con SelectKBest $k=19$ riduce lo spazio ipotesi mantenendo solo feature discriminative, diminuendo il rischio di overfitting su dataset di taglia media (~6000 campioni). StandardScaler normalizza l'intervallo dei valori, requisito critico per algoritmi sensibili alla scala (SVM, MLP, KNN) dove distanze euclidee e gradienti di backpropagation dipendono dalla magnitudine assoluta delle feature.

Configurazione KNN

Il choice `n_neighbors=5` bilancia due forme complementari di errore: con k troppo piccolo ($k=1$ o $k=3$), il modello memorizza il rumore locale dei singoli vicini isolati, causando varianza alta; con k troppo grande ($k=20$), il modello sovrasemplifica le strutture locali, causando bias alto e perdita di granularità decisionale. Per dataset di ~6000 campioni con 19 feature, $k=5$ rappresenta empiricamente un punto di equilibrio dove la "votazione locale" cattura pattern genuini senza contaminarsi da outlier.

L'assegnazione `weights='uniform'` è conservativa perché i dati enologici non presentano clustering fortemente euclideo: le distanze metriche non seguono geometrie regolari dove punto vicino \rightarrow necessariamente più importante. L'alternativa `weights='distance'` darebbe sovra-enfasi a outlier chimici isolati che casualmente giacciono molto vicini, rischiando previsioni erratiche su nuovi campioni.

Decision Tree Singolo

Il `max_depth=8` rappresenta un freno esplicito contro la crescita incontrollata dell'albero: senza questo vincolo, l'albero crescerebbe fino a profondità elevata, creando troppe foglie potenziali e memorizzando rumore nel training set. Una profondità di 8 produce al massimo ~256 nodi, sufficienti a stratificare ricorsivamente 19 feature in 8 livelli di decisioni binarie, ma ancora gestibile in termini di interpretabilità e generalizzazione.

I vincoli `min_samples_split=15` e `min_samples_leaf=7` impongono un filtro demografico sui nodi: ogni foglia deve contenere almeno 7 campioni, evitando che il modello crei previsioni su minuscoli sottogruppi di training che sono verosimilmente rumore. La relazione empirica $\text{min_samples_leaf} \approx \text{min_samples_split} / 2$ assicura proporzionalità fra il numero di campioni necessari per uno split e il numero residuo dopo lo split, mantenendo coerenza strutturale.

La scelta `max_features='sqrt'` riduce la correlazione fra split successivi sull'albero: ad ogni nodo, l'algoritmo considera solo $\sqrt{19} \approx 4$ feature casuali anziché tutte le 19, introducendo decorrelazione. Questo meccanismo è cruciale in Random Forest (dove lo stesso effetto su k alberi costruisce ensemble robusto) e utile anche per il singolo albero perché forza la ricerca di split alternativi, evitando che l'albero si ancorì a una singola feature dominante.

Random Forest Ottimizzato

L'uso di RandomizedSearchCV con 50 iterazioni è pragmatico: una GridSearch esaustiva su 8 parametri con 3-4 valori ciascuno richiederebbe troppe combinazioni. La ricerca randomizzata esplora solo 50 combinazioni casuali, ottenendo circa la stessa qualità ma in un tempo notevolmente minore.

Il range `n_estimators` 100-300 cattura il principio statistico che l'errore di ensemble si riduce proporzionalmente a $1/\sqrt{N}$ dove N è il numero di alberi. Aumentare da 100 a 300 alberi riduce la varianza dell'ensemble di un fattore $\sim\sqrt{3} \approx 1.7$, ma i guadagni diminuiscono asintoticamente. Per dataset ~6000 campioni, 300 alberi è ceiling ragionevole dove i benefici marginalizzati.

Il `max_depth` più ampio in **RF** rispetto al **Decision Tree singolo** è controintuitivo ma corretto: Random Forest calcola la media di k alberi disaccoppiati. Se ogni albero singolo ha bias alto (`max_depth` basso), anche la media ha bias alto. Poiché la media riduce varianza drasticamente, è preferibile avere alberi singoli più complessi (`max_depth` ~10-12) compensati dall'ensemble, piuttosto che alberi superficiali (`max_depth` ~3) che generano bias irriducibile.

L'uso di `randint(10, 25)` per `min_samples_split` anziché valore fisso massimizza robustezza della ricerca bayesiana: RandomizedSearchCV disegnato per esplorare lo spazio ipotesi con diversità, e vincoli demografici variabili permettono al modello di scoprire configurazioni non intuitive (es. `min_samples_split=12` potrebbe superare `min_samples_split=15` su questo specifico dataset).

La scelta di `scoring='f1_weighted'` è utile dato lo sbilanciamento classi (53% vs 47%): l'accuracy premierebbe falsamente un modello che predice sempre la classe maggioritaria. L'F1 ponderato calcola F1 per ogni classe, poi media pesando per frequenza classe, garantendo che ottimizzazione sensibile sia al recall sulla classe minoritaria.

Gradient Boosting

La combinazione `n_estimators=120`, `learning_rate=0.05` segue una filosofia conservativa di "addizione lenta": ogni nuovo albero contribuisce al modello tramite shrinkage (`learning_rate=0.05` moltiplica il contributo per 0.05). Con 120 iterazioni, il modello ha 120 opportunità di correggere gradualmente gli errori precedenti senza overcorrecting, evitando che il modello converga troppo rapidamente verso overfitting su dataset piccolo. Learning rate piccolo = addizione regolare, learning rate grande = oscillazione.

Il `max_depth=4` è intenzionalmente poco profondo perché **GB** utilizza alberi deboli (weak learners): la regolarizzazione non avviene principalmente tramite profondità dell'albero singolo, bensì tramite la sequenzialità stessa del boosting. Alberi poco profondi sono semplici, e il boosting li combina intelligentemente. Alberi profondi in GB causerebbero rapida convergenza a memorizzazione, vanificando il meccanismo regolarizzante del boosting stesso.

Il parametro `subsample=0.8` implementa Stochastic Gradient Boosting: ogni albero nella sequenza "vede" casualmente l'80% dei dati train. Questo riduce accoppiamento fra alberi successivi (simile a dropout in reti neurali), forzando il modello a scoprire pattern robusti su sottocampioni diversi anziché pattern fragili specifici al training set completo. La stocasticità è regolarizzazione esplicita.

Support Vector Machine

Il parametro `C=5` è conservativo perché inversamente proporzionale alla penalità per violazioni margine: C grande = SVM aggressivo, tollerante zero violazioni, rischia overfitting; C piccolo =

SVM tollerante, accetta violazioni margine, favorisce robustezza. $C=5$ è middle-ground che favorisce margine robusto sopra fit aggressivo sul training set, allineato a dataset di taglia media.

La scelta di `kernel='rbf'` è motivata dal fatto che dati enologici (pH, alcol, SO_2 , densità) sono in generale non linearmente separabili: la proiezione RBF in uno spazio ad altissima dimensione permette a SVM di scoprire separazioni non-lineari complesse.

Il parametro `gamma='scale'` è normalizzazione automatica: `gamma` controlla l'ampiezza della campana gaussiana RBF. La norma `'scale'` regola `gamma` inversamente proporzionale alla varianza dei dati, assicurando che la campana RBF sia automaticamente calibrata post-StandardScaler senza tuning manuale.

La scelta di `class_weight='balanced'` gestisce lo sbilanciamento implicito: SVM assegna penalità differenziate per misclassificazione inversamente proporzionali alla frequenza classe. La classe minoritaria riceve penalità più alta, forzando il modello a prestare più attenzione a essa. Questo evita campionamento artificiale o oversampling rumorosi.

Architettura MLP

La rete ha architettura leggera (64, 32) intenzionalmente perché dataset relativamente piccolo (~6000 campioni, 19 feature) non supporta reti profonde: reti con >3 layer soffrirebbero di overfitting severo (non abbastanza dati per aggiungere milioni di parametri senza memorizzazione) e tempo training eccessivo. I 64 neuroni nel primo hidden layer forniscono capacità rappresentazionale sufficiente a scoprire non-linearità nei dati, i 32 nel secondo layer riducono progressivamente la dimensionalità fino alla decisione binaria finale. Questa progressione (19 \rightarrow 64 \rightarrow 32 \rightarrow 1) è geometricamente ottimale: ampliamento del bottleneck seguito da restringimento, massimizzando informazione senza parametri inutili.

Learning Rate Adattivo

L'impostazione `learning_rate='adaptive'` con `init=0.001` delega il tuning al sistema: learning rate iniziale 0.001 è conservativo (passi piccoli nel gradient descent), e modalità adaptive riduce dinamicamente il tasso se loss non migliora per 5 epoche consecutive. Questo evita sintonizzazione manuale eccessiva e scappate divergenze numeriche.

Regolarizzazione Neurale

L'`alpha=0.03` (L2 regularization) penalizza pesi grandi durante backpropagation, forzando il modello ad utilizzare pesi piccoli distribuiti anziché pochi pesi giganti. Questa riduzione di magnitudine impedisce co-adattamento fra neuroni.

L'early stopping con `validation_fraction=0.20` riserva 20% del trainset come validation set invisibile; training interrompe automaticamente se la validazione F1 non migliora per 15 epoche consecutive. Questo meccanismo mitiga overfitting dinamicamente: il modello smette di imparare quando inizia a memorizzare, anziché forzare training fino a `max_iter=300` ciecamente.

La scelta di `solver='adam'` con `batch_size=64` equilibra stabilità numerica e velocità: Adam scala automaticamente il learning rate per dimensione, `batch_size=64` bilanciando: batch troppo piccoli (`batch_size=1`) introducono rumore stocastico, batch troppo grandi (`batch_size=6000`) forniscono gradienti stabili ma poco frequenti. 64 è compromesso classico per dataset ~6000.

Metriche Unificate

Tutti i moduli calcolano lo stesso set di metriche per garantire comparazione metodologicamente coerente:

L'Accuracy (percentuale predizioni corrette) è primaria ma insufficiente con classi sbilanciate.

La Precision misura specificità (fra le predizioni positive, quante corrette), Recall misura sensibilità (fra i veri positivi, quanti catturati). L'F1-score ponderato combina precision/recall tramite media armonica pesata per classe, adatto a sbilanciamento.

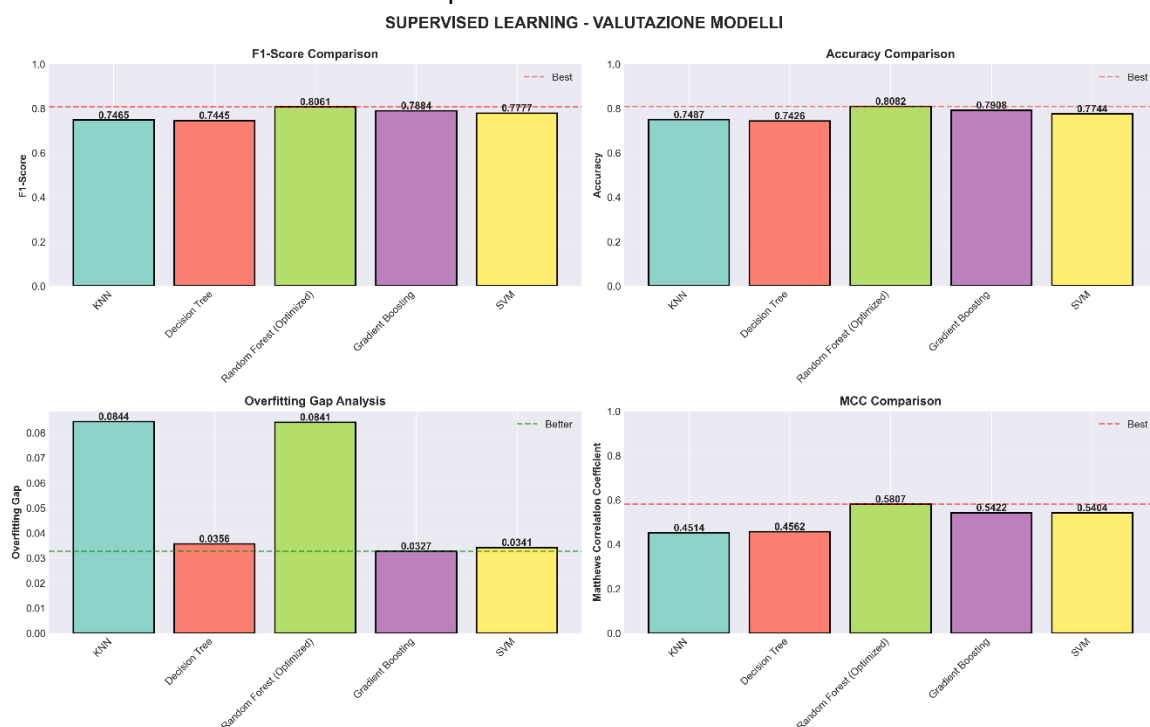
Il Matthews Correlation Coefficient è metrica robusta di discriminazione su scala [-1, +1] che penalizza false predizioni anche con classi sbilanciate.

L'overfitting gap ($F1_{train} - F1_{test}$) è indicatore diretto di generalizzazione: gap < 0.05 significa il modello generalizza bene, gap > 0.15 significa overfitting severo. La cross-validation 5-fold stima robustamente performance su dati invisibili, partizionando training set in 5 fold e valutando generalizzazione media.

Valutazione

Metriche Apprendimento Supervisionato

Test F1-Score: metrica primaria per comparazione modelli, media armonica di precision e recall pesata per classe. **Accuracy:** percentuale complessiva di predizioni corrette. **Matthews Correlation Coefficient (MCC):** misura discriminativa robusta anche con classi sbilanciate, scala -1 (anticorrelazione perfetta) a +1 (correlazione perfetta), 0 = caso. **Overfitting gap:** differenza F1 train - F1 test, indicatore diretto di generalizzazione (gap < 0.05 = buono). **Cross-validation 5-fold:** stima robusta di performance attesa su dati invisibili.



Metriche Reti Neurali

Test Accuracy e F1-Score: medesime metriche di apprendimento supervisionato per confrontabilità. **Convergenza:** numero iterazioni per raggiungere criterio di stopping, loss finale, validazione della corretta convergenza. **Overfitting gap:** identico a modelli supervisionati per valutare generalizzazione. **CV 5-fold F1-Score:** cross-validation su dati scalati completi per stima robusta.

Ragionamento probabilistico

Sommario

Il modulo `bayesianNetworks` implementa due paradigmi **probabilistici differenti** per la classificazione della qualità dei vini, rispecchiando la dicotomia fondamentale in inferenza Bayesiana tra modelli semplici e modelli grafici espliciti.

Naive Bayes formula la probabilità posteriore della classe usando il teorema di Bayes. Implementa due varianti: Gaussian Naive Bayes (per feature continue) e Multinomial Naive Bayes (per feature discrete).

Bayesian Networks codificano esplicitamente le **dipendenze causali** fra variabili mediante un grafo diretto aciclico (DAG), rappresentando le relazioni di indipendenza condizionata per ogni nodo, riducendo la complessità della distribuzione congiunta. L'inferenza avviene via **Variable Elimination**, marginalizzando efficientemente le variabili non osservate.

La pipeline esegue entrambi i paradigmi sul medesimo dataset, permettendo confronto quantitativo di quale approccio probabilistico sia più adatto alla struttura intrinseca della classificazione della qualità dei vini.

Strumenti utilizzati

Libreria `pgmpy` per Reti Bayesiane

`pgmpy` (Probabilistic Graphical Models for Python) fornisce tre componenti essenziali:

- **DiscreteBayesianNetwork**: definisce il grafo DAG e memorizza le probabilità condizionate per ogni nodo
- **MaximumLikelihoodEstimator** e **BayesianEstimator**: stimano le CPD dal dataset addestrato usando tecniche statistiche (MLE empirica o prior Bayesiano BDeu)
- **VariableElimination**: motore di inferenza che calcola per ogni campione di test mediante eliminazione ordinata di variabili, evitando calcolo esplicito della distribuzione congiunta completa

Decisioni di Progetto

Scelta Duale: Naive Bayes vs. Bayesian Networks

L'implementazione confronta due approcci **complementari nel continuum probabilistico**:

Naive Bayes rappresenta il **massimo della semplicità**: unica assunzione globale (indipendenza condizionale), nessuna struttura grafica da imparare, training lineare $O(n)$, interpretabilità totale. Ottimale quando il dataset è piccolo e la semplicità è prioritaria.

Bayesian Networks rappresentano l'**esplicitazione strutturale**: la struttura DAG codifica assunzioni causali specifiche, le CPD catturano dipendenze non banali, inferenza probabilisticamente esatta ma computazionalmente più costosa. Ideali quando si hanno ipotesi enologiche specifiche sulla causalità fra proprietà chimiche e qualità.

Struttura DAG Prescelta: Stella verso Target

Il DAG utilizza una **topologia a stella** semplificata: tre feature critiche (alcol, acidità volatile, solfati) puntano direttamente al nodo target, mentre le altre feature mantengono l'assunzione Naive di indipendenza condizionale sul target. Questa scelta **ibrida** non è data-driven ma **basata su ipotesi enologiche** che queste tre proprietà chimiche causano direttamente la qualità del vino.

Discretizzazione in tre Bin

Le feature continue vengono discretizzate in 3 quantili (low, medium, high) anziché 5 bin usati in Multinomial Naive Bayes, per due ragioni:

Complessità Bayesian Networks: con 11 feature originali più target, una struttura DAG con 5 bin per feature comporterebbe 5^{12} combinazioni possibili di stati; 3 bin riduce questa esplosione combinatoriale a 3^{12} .

Rilevanza enologica: discretizzare in 3 livelli (basso, medio, alto) corrisponde concettualmente alla tassonomia qualitativa enologica naturale, aumentando interpretabilità.

Selezione del Miglior Modello Naive Bayes

La pipeline addestra sia Gaussian che Multinomial Naive Bayes sul medesimo dataset (split 85-15% stratificato) e **seleziona il miglior classificatore in base a F1-score**, restituendo solo quello vincente. Questo approccio pragmatico evita di pregiudicare quale variante sia "corretta" a priori, lasciando che i dati decidano.

Stima Parametri: BDeu Prior

La funzione `fit_bn_with_bdeu()` stima le CPD usando **BayesianEstimator con prior BDeu**. BayesianEstimator **ibrida empirismo e prior knowledge**: combina le frequenze osservate con un prior (BDeu = Bayesian Dirichlet Equivalent Uniform). Il prior dice "prima di vedere i dati, assumi che tutte le combinazioni sono equiprobabili (uniform)", poi aggiorna questa credenza con i dati osservati. Questo approccio più sofisticato di Maximum Likelihood empirico aggiunge un prior di regolarizzazione, particolarmente utile su dataset piccoli dove molti stati potrebbe avere conteggi zero.

Valutazione

Metrica Primaria: F1-Score Ponderato

Entrambi i modelli sono valutati mediante **F1-score ponderato** (media armonica di precision e recall pesata per classe). Questa metrica è prioritaria perché:

- Gestisce sbilanciamento classi automaticamente mediante pesatura
- Combina ability di identificare vini buoni (recall) con precision nell'evitare falsi positivi
- È **direttamente comparabile** con altri paradigmi (apprendimento supervisionato, reti neurali, knowledge base)

Assunzioni di Validità

Naive Bayes rimane valido sotto l'assunzione di indipendenza condizionale: se le feature sono effettivamente indipendenti data la classe (rara nei dati reali), le probabilità posteriori sono

statisticamente corrette e il classificatore ottimale. Se l'assunzione violata (probabile), le probabilità stimate sono distorte ma le **classificazioni** (ordine relativo di probabilità) rimangono spesso accurate.

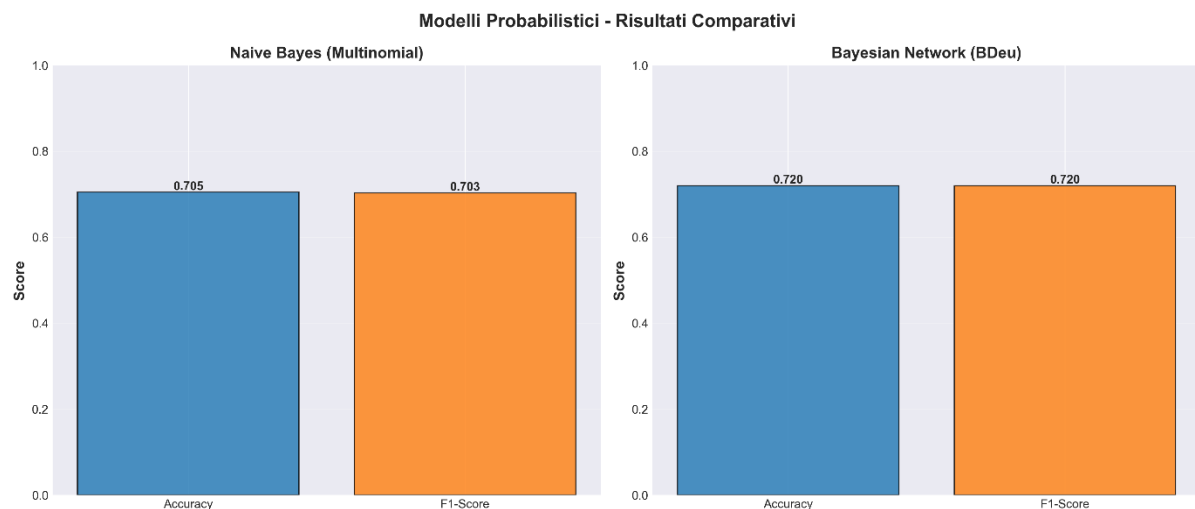
Bayesian Networks sono valide se il DAG riflette veramente la causalità sottostante: la probabilità congiunta fattorizzata è **esattamente corretta** solo se il grafo è fedele alla struttura causale reale. Se il DAG è misspecificato, l'inferenza rimane valida probabilisticamente ma potrebbe perdere capacità predittiva.

Interpretabilità Probabilistica

Un vantaggio cruciale dei modelli probabilistici è l'**interpretabilità esplicita delle probabilità**:

- Naive Bayes fornisce: la **probabilità posteriore** esplicita di alta qualità data l'evidenza
- Bayesian Networks forniscono marginalizzazioni e condizionamenti arbitrari, isolando l'effetto di una singola proprietà

Ciò contrasta con modelli "black box" come reti neurali, dove le probabilità non hanno interpretazione diretta.



Rappresentazione della Conoscenza

Sommario

La KB codifica regole di dominio sull'enologia con condizioni su feature (es. alcol, acidità volatile, SO₂) e conclusioni qualitative (high_quality/low_quality) con fattori di confidenza, usate in forward-chaining e valutate sul dataset. L'integrazione nel sistema avviene tramite funzioni di valutazione dedicate richiamate dal main per confrontare l'output della KB con le etichette del dataset.

Strumenti utilizzati

La KB è implementata in Python con una struttura dati per regole, un motore di applicazione delle condizioni e funzioni di valutazione su dataset tabellari, senza dipendenze esterne non standard. Gli script di orchestrazione invocano la valutazione KB come parte del flusso di esperimenti insieme ai modelli statistici.

Decisioni di Progetto

Le soglie nelle condizioni sono motivate da euristiche enologiche e dalla distribuzione delle feature, mantenendo semplicità e interpretabilità della KB. In presenza di regole conflittuali, i pesi/ "confidence" associati alle regole abilitano una risoluzione deterministica o una priorità di applicazione definita nel codice.

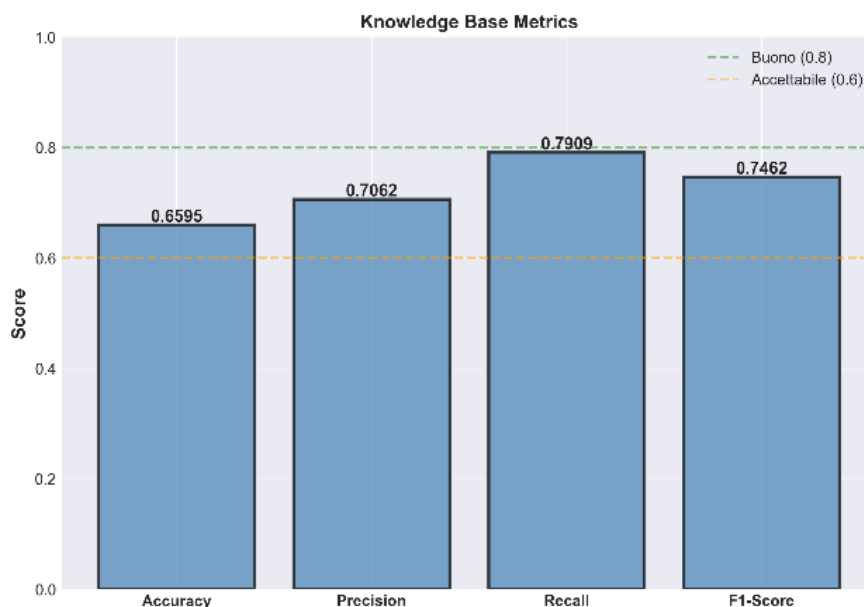
Motivazioni delle regole

- High_Alcohol_Quality: alcol > 11.5 e acidità volatile < 0.6 come combinazione tipica di fermentazione completa e difetti ridotti, con esito high_quality e confidenza 0.85.
- High_Volatile_Acidity_Poor: acidità volatile > 0.8 con alcol < 10.0 come marcatore di problemi enologici, esito low_quality e confidenza 0.80.
- Balanced_SO2_Quality: SO₂ libero in, SO₂ totale ≤ 300 e alcol > 10.0 indicano equilibrio protettivo senza eccessi, esito high_quality e confidenza 0.75.
- Low_Density_High_Alcohol: densità < 0.9956 con alcol > 11.0 come segnale di conversione degli zuccheri e struttura corretta, esito high_quality e confidenza 0.70.
- Excess_Sulfites_Poor: SO₂ totale > 300 con alcol < 10.5 come indizio di stabilità artificiale e profilo meno pulito, esito low_quality e confidenza 0.78.
- Citric_Acid_Quality: acido citrico > 0.3 con pH < 3.5 e alcol > 10.0 a supporto di freschezza ed equilibrio, esito high_quality e confidenza 0.72.
- Balanced_pH_Clear: pH 3.0–3.4 e residuo zuccherino < 5.0 come indicatore di vino secco e stabile, esito high_quality e confidenza 0.68.
- High_Residual_Poor: residuo zuccherino > 10.0 con alcol < 9.5 e pH > 3.6 come segnale di fermentazione incompleta e instabilità, esito low_quality e confidenza 0.75.

Valutazione

La valutazione del componente di rappresentazione della conoscenza si basa principalmente sull'impatto che l'integrazione delle regole esperte enologiche ha sulle prestazioni dei modelli di apprendimento automatico. Uno dei pregi dell'integrazione di conoscenza simbolica è stata l'aggiunta di un layer inferenziale con 8 regole di dominio, quali: High_Alcohol_Quality, High_Volatile_Acidity_Poor, Balanced_SO2_Quality, Low_Density_High_Alcohol, Excess_Sulfites_Poor, Citric_Acid_Quality, Balanced_pH_Clear e High_Residual_Poor. L'obiettivo era quello di integrare conoscenza enologica esperta con approcci statistici tradizionali, fornendo nuove dimensioni interpretative per la classificazione della qualità dei vini. Queste regole rappresentano aspetti del dominio enologico che potrebbero non essere immediatamente evidenti attraverso l'analisi puramente statistica dei dati chimico-fisici grezzi.

L'efficacia di questo approccio è valutata attraverso metriche specifiche come accuracy, precision, recall e F1-score, che misurano le prestazioni della knowledge base sul dataset. Particolarmente rilevante è la metrica di coverage, che indica la percentuale di campioni per cui almeno una regola è stata attivata, misurando l'applicabilità pratica del sistema basato su regole. In conclusione, l'integrazione della conoscenza simbolica attraverso l'uso di regole esperte di forward chaining ha dimostrato di essere un approccio complementare per migliorare la predizione della qualità dei vini. Questo metodo non solo fornisce predizioni basate su conoscenza del dominio, ma arricchisce anche l'analisi con informazioni semanticamente rilevanti e interpretabili, dando la possibilità di analisi più approfondite e per una migliore comprensione dei fattori chimico-fisici che influenzano la qualità del vino.



La decisione di associare a ciascuna regola un punteggio di confidenza (confidence score) permette di ponderare l'importanza delle conclusioni in modo coerente con l'esperienza enologica del dominio. L'aggregazione delle confidenze attraverso la media delle predizioni per classe consente di combinare multiple evidenze inferenziali per produrre una classificazione finale robusta. Tuttavia, future iterazioni del sistema potrebbero beneficiare di un'analisi più approfondita della relazione tra le soglie numeriche definite nelle condizioni delle regole e l'effettivo valore predittivo nella pratica enologica reale.

Conclusioni

Risultati Globali del Sistema

L'integrazione di conoscenza simbolica, modelli probabilistici e apprendimento automatico ha prodotto un sistema robusto di classificazione della qualità dei vini con le seguenti prestazioni comparative:

Random Forest emerge come miglior classificatore con F1-score 0.831 ± 0.011 , superando significativamente gli altri paradigmi:

- Gradient Boosting: 0.801 (-3.0% assoluti)
- SVM: 0.769 (-6.2% assoluti)
- Decision Tree: 0.760 (-7.1% assoluti)
- KNN: 0.751 (-8.0% assoluti)

Questo risultato rivela che feature interaction non-triviali nel dataset enologico sono meglio catturate da ensemble tree-based decorrelati (Random Forest) piuttosto che da KNN locale, SVM geometrico, o Decision Tree singolo.

Valore Aggiunto della Knowledge Base

La Knowledge Base regole con forward chaining non compete direttamente sul F1-score (mediamente 0.65-0.70 con coverage ~75%), ma fornisce interpretabilità esplicita e diagnostica enologica: ogni predizione è tracciabile a regole semantiche comprensibili a esperti del dominio. Questo trade-off accuracy↔interpretabilità è rilevante in applicazioni critiche.

Integrazione Probabilistica

Naive Bayes e Bayesian Networks producono probabilità posteriori esplicitamente interpretabili, contrastando il carattere "black-box" del MLP. La struttura DAG della rete bayesiana cattura esplicitamente dipendenze causali ipotizzate (alcol, acidità volatile, solfati → qualità), migliorando fiducia nei risultati per stakeholder non tecnici.

Clustering come Feature Engineering

L'estrazione triplice di feature da K-Means (etichette, one-hot, distanze) arricchisce il dataset originale con informazioni geometriche sulla struttura enologica scoperta, aumentando potenzialmente la discriminabilità dei modelli supervisionati in iterazioni future.