QUALITA' DEL VINO: TECNICHE DI MACHINE LEARNING PER PREVEDERE IL LIVELLO IN BASE ALLE SUE CARATTERISTICHE FISICO-CHIMICHE

Laura Nembrini, Adele Zanfino, Antonio Lombardo, Gabriele Strano
Università degli Studi di Milano-Bicocca, CdLM Data Science

18 Gennaio 2021

Sommario

Definire la qualità del vino è un processo complicato. Per ottenere un buon prodotto, è necessario che ci sia un certo equilibrio tra le componenti chimiche. Ma quali sono le proprietà che consentono di classificare un vino 'di buona qualità'? Al fine di effettuare la giusta identificazione, sono stati implementati differenti modelli di classificazione con l'obiettivo di predire quando un vino è definibile 'di buona qualità', studiando gli aspetti fisico-chimici. In secondo luogo, si è andato ad analizzare quali sono le componenti fisico-chimiche che incidono maggiormente su un vino di 'buona qualità'. E' stato infine selezionato il modello migliore in termini di performance, che consente quindi di fornire una corretta distinzione tra vini di 'buona qualità' e 'medio-bassa qualità'.

1 Introduzione

L'industria del vino rosso mostra una recente crescita esponenziale. Al giorno d'oggi, gli operatori del settore utilizzano le certificazioni di qualità per promuovere i propri prodotti le quali richiedono tempo e valutazione fornita da esperti, il che lo rende molto costoso. Inoltre, il prezzo del vino rosso dipende da un concetto piuttosto astratto di apprezzamento da parte degli assaggiatori, l'opinione dei quali può avere un alto grado di variabilità. Un altro fattore che incide sulla certificazione del vino rosso e sulla valutazione della sua qualità sono i test fisico-chimici. Il mercato del vino rosso potrebbe essere semplificato se il fattore umano della degustazione fosse correlato alle proprietà chimiche del vino in modo da rendere i processi di certificazione e valutazione più controllati. A tal proposito, in questo studio verranno esplorate differenti tecniche di apprendimento automatico con l'obiettivo primario di stimare la qualità del prodotto, tenendo in considerazione le diverse componenti fisico-chimiche che lo compongono.
Nella prima parte del documento verrà introdotto il dataset utilizzato per svolgere le analisi e la fase di pre-processing iniziale. Nella seconda verranno invece esposti i modelli di apprendimento automatico analizzati, con le relative misure di performance. Verranno infine riportate le analisi svolte e i risultati ottenuti, fornendo una conclusione in merito agli obiettivi preposti.

2 Dataset

I dati utilizzati in questo studio sono relativi al vino *"Vinho Verde"*, un vino portoghese [2]. Sono state quindi riportate le caratteristiche fisico-chimiche più comuni, che rappresentano le variabili di input del modello. Il set di dati è composto da 1599 record e 12 variabili, tutte di tipo

quantitativo, comprensive del livello di qualità che assume un valore nell'intervallo [0,10].

- **fixed acidity**: acidi organici non volatili;
- **volatile acidity**: acido acetico presente nel vino;
- **citric acid**: acido organico presente nell'uva;
- **residual sugar**: zucchero non fermentato, rimasto quindi nel vino dopo la fermentazione;
- **chlorides**: quantità di sale nel vino;
- **free sulfur dioxide**: anidride solforosa libera;
- **total sulfur dioxide**: anidride solfosa totale;
- **density**: densità del vino, più il vino è dolce, maggiore è la densità;
- **pH**: livello di acidità;
- **sulphates**: anidride solfosa aggiunta al vino che agisce come antimicrobico e antiossidante;
- **alcohol**: quantità di alcol presente nel vino;
- **quality**: l'insieme di caratteri organolettici gradevoli, legati alla composizione chimica.

3 Esplorazione dei dati e Preprocessing

In modo da rendere il dataset maggiormente idoneo per le successive analisi, si è deciso di attuare una prima pulizia del dataset.
Dal risultato delle statistiche è emerso che il dataset non presenta valori mancanti per nessuno degli attributi. Successivamente, si sono studiate le correlazioni tra le varie componenti fisico-chimiche, in modo da avere una miglior comprensione della relazione tra le variabili.

Ponendo una maggiore attenzione alle correlazioni tra le variabili indipendenti e la variabile dipendente, ovvero la qualità, si può notare come l'alcol sia quella maggiormente correlata. Un vino con bassa gradazione alcolica, tendenzialmente non avrà un alto livello di qualità, seppure questa relazione non sia molto forte. Si può osservare inoltre la presenza di relazioni positive tra qualità e acido citrico e qualità e solfati. In secondo luogo, sono emerse relazioni inverse

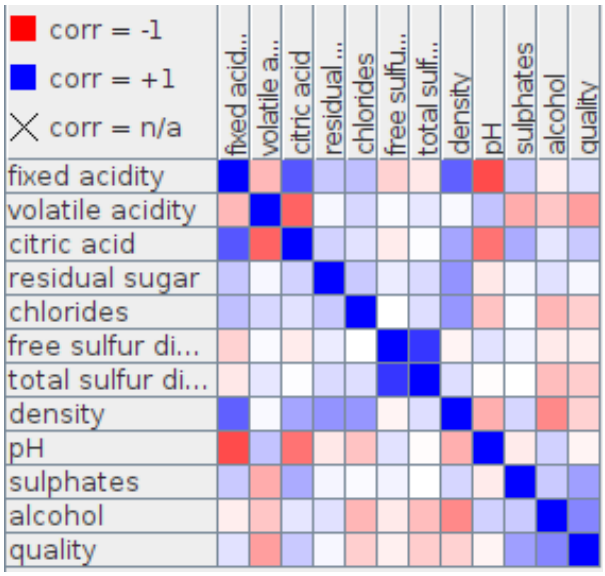

Figura 1: Correlazione variabili.

tra qualità e acidità volatile, densità e PH. Di fatti è ragionevole pensare che un vino di buona qualità abbia un basso livello di acidità e che siano preferibili vini meno dolci.
Per quanto riguarda la fase di pulizia iniziale del dataset, in accordo con il nostro scopo di ricerca, si è scelto di rendere binario l'attributo in modo tale da avere solamente due categorie. Per questo problema si deciso di suddividere l'attributo qualità in due range:

- il primo intervallo rappresentante i vini di qualità medio-bassa, con l'attributo "qualità" compreso tra 0 e 7 (escluso);
- il secondo rappresenta i vini di qualità ottima, con l'attributo "qualità" compreso tra 7 (incluso) e 10.

Dopo aver convertito la variabile di output in un output binario, si è verificata la presenza di un dataset sbilanciato.
Di fatti, dalla Figura 2, è possibile notare una netta sproporzione tra i vini di medio-bassa qualità, presenti in misura maggiore nel nostro set di dati, e i vini di alta qualità che rappresentano invece la classe rara. Per il nostro scopo, si è deciso di non applicare nessun tipo di ricampionamento, mantenendo quindi il dataset sbilanciato. Sono infatti molto più diffusi i vini di medio-bassa qualità rispetto a quelli di alta qualità. Successivamente sono stati eliminati gli outliers presenti nel dataset. Outlier è un termine utilizzato per definire, in un insieme di osservazioni, un valore anomalo, ovvero un va-

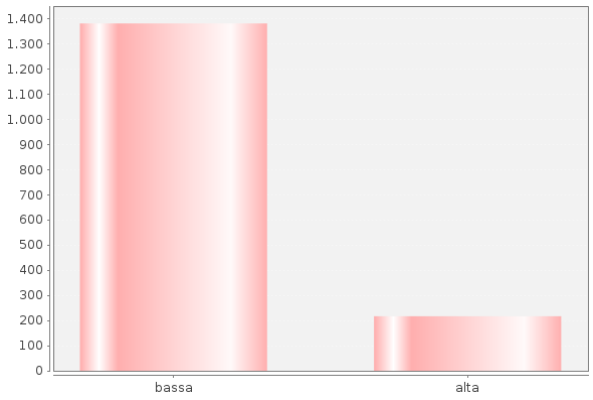

Figura 2: Dataset sbilanciato.

lore chiaramente distante dagli altri disponibili. Questi valori definiti anomali possono portare ad una distorsione della predizione e per questo motivo vengono esclusi. Come ultimo passaggio è stata applicata una normalizzazione delle variabili in modo da renderle tutte confrontabili. I dati verranno quindi trasformati in modo tale che la loro distribuzione abbia media 0 e deviazione standard pari a 1.

4 Modelli utilizzati e misure di performance

Diverse sono state le tecniche di classificazione implementate in questo studio, aventi come obiettivo comune quello di individuare la più adatta al nostro obiettivo.

1. **Modelli euristici**: sono algoritmi semplici ed intuitivi. I due modelli considerati sono stati Random Forest e K-Nearest-Neighbour(KNN). Il primo si avvale di un insieme di alberi decisionali la cui classificazione avviene sulla base della classe maggioritaria nel nodo finale. Il secondo è basato sul concetto di distanza tra osservazioni e permette quindi di classificare un'osservazione in base alle caratteristiche di quelle ad essa vicine;
2. **Modelli di regressione**: è stato preso in considerazione il modello di regressione logistica che utilizza una funzione logistica per modellare il risultato di una variabile dipendente categoriale, date una serie di variabili in input. Risulta essere un modello flessibile in quanto qualsiasi tipologia di variabili possono essere utilizzate come input del modello;
3. **Modelli di separazione**: questi modelli permettono di dividere lo spazio in regioni disgiunte, in modo tale da separare le osservazioni secondo la classe target. Support Vector Machine (SVM) è il modello utilizzato per questa categoria di classificatori;
4. **Modelli probabilistici**: i classificatori appartenenti a questa categoria sono in grado di prevedere una distribuzione di probabilità su un insieme di classi piuttosto che solo la classe più probabile a cui dovrebbe appartenere l'osservazione. Naive Bayes è uno dei metodi maggiormente utilizzati ed è basato sul teorema di Bayes. E' un metodo altamente scalabile e presenta come assunzione l'indipendenza tra gli attributi considerati.

4.1 Misure di performance

Nella nostra analisi sono stati utilizzati diversi criteri di valutazione; in particolare, abbiamo scelto di calcolare Accuracy, Recall, Precision, F1-measure e la curva ROC.

La ***Accuracy*** è definita come la percentuale di osservazioni sia positive che negative previste correttamente e misura quindi la capacità del modello di dare classificazioni affidabili sui nuovi record da prevedere. Essa è calcolata dalla formula:

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \qquad (1)$$

dove:

- TP e TN indicano il numero di record correttamente classificati come appartenenti rispettivamente alla classe positiva e a quella negativa;
- FP e FN indicano, invece, il numero di record erroneamente predetti.

In generale, un valore di Accuracy maggiore rappresenta un modello di classificazione migliore. Tuttavia il dataset utilizzato nella nostra analisi, come precedentemente accennato, risulta sbilanciato. Abbiamo riscontrato che un dataset di questo tipo può interferire sui risultati prodotti da questa misura di performance.

Di conseguenza, abbiamo ritenuto necessario tenere in considerazione altre misure di va-

lutazione. L'indicatore ***Recall*** misura la frazione di record positivi correttamente predetti dal modello; un valore alto di questa misura indica la presenza di pochi record appartenenti alla classe positiva classificati erroneamente. In particolare essa è definita come:

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

L'indicatore ***Precision*** determina invece la frazione di record che si rivela essere effettivamente positiva all'interno del gruppo che il classificatore definisce come classe positiva.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

Vi è un legame molto stretto tra questi due indici, per cui ci si potrebbe trovare nella situazione in cui si abbia un valore elevato della Recall e un basso livello di Precision. Considerando ciò, si utilizza una misura capace di riassumerle.

La ***F-measure*** rappresenta la media armonica tra Recall e Precision. In particolare:

$$F_1 = \frac{2 \cdot r \cdot p}{r + p} \tag{4}$$

Nel modo in cui è definita, un valore alto di questa misura implica alti valori sia nell'indicatore Recall che nella Precision.

5 Analisi e risultati

In questa sezione verranno presentate le analisi effettuate e i risultati ottenuti in merito ad esse. Le analisi sono state svolte utilizzando il software KNIME [3].

5.1 Prima domanda di ricerca

Per quanto riguarda la prima domanda di ricerca, l'obiettivo principale del nostro studio consiste nel prevedere la qualità del vino in base ai suoi attributi fisico-chimici. L'attributo scelto come variabile di risposta è l'attributo quality binned, che rappresenta l'attributo quality dopo essere stato binarizzato nella fase di preprocessing , mentre tutte le altre variabili presenti nel dataset vengono utilizzate come variabili esplicative. Per questa prima domanda di ricerca , sono stati confrontati 5 diversi modelli di classificazione: Random Forest, KNN, Logistic Regression, SVM e Naive Bayes. In accordo con i nostri obiettivi, siamo andati a verificare quali di essi risultava essere il più performante in termini di predizione della classe 'alta qualità'. Qui di seguito vengono mostrati i vari risultati ottenuti attraverso i tre approcci impiegati.

5.1.1 Classificazione con metodo Holdout

Nella prima fase dell'analisi il dataset è stato suddiviso in un training set formato dal 75% dei record e in un test set con il restante 25%. I classificatori creati utilizzando i cinque modelli precedentemente descritti sono sono stati addestrati con il training set e validati attraverso il test set. Si è deciso di fare un confronto tra due partizionamenti utilizzando un seed [1] differente. I risultati ottenuti sono riportati nella seguente Tabella 1.

Classificatore	Recall	Precision	F1	Accuracy
RandomForest	0.595	0.733	0.657	0.923
Logistic	0.432	0.533	0.478	0.883
KNN	0.351	0.542	0.426	0.883
NaiveBayes	0.649	0.471	0.545	0.866
SVM	0.351	0.812	0.491	0.91
RandomForest	0.541	0.541	0.541	0.886
Logistic	0.297	0.44	0.355	0.866
KNN	0.351	0.52	0.419	0.88
NaiveBayes	0.649	0.429	0.516	0.849
SVM	0.324	0.706	0.444	0.9

Tabella 1: Misure di performance ricavate con il metodo Holdout (le prime con seed pari a 234, le seconde con seed pari 444).

Dai risultati si può notare come, per ogni classificatore, l'accuracy è molto alta. I migliori in termini di accuracy risultano essere Random Forest per i modelli euristici e SVM per i modelli di separazione. In particolare se osserviamo il modello SVM si può notare come il valore della Precision è pari a 0.812 mentre la Recall è molto bassa, pari a 0.324, quindi ci sono molti vini di buona qualità che vengono classificati erroneamente come vini di medio-bassa qualità. L'F-measure, che combina le due misure, è di fatti molto bassa per questo classificatore. Se andiamo invece a considerare Random Forest, notiamo come anche in questo caso la

[1] Espediente utile a rendere riproducibili i risultati del classificatore

Precision è abbastanza alta, pari a 0.733 mentre la Recall in questo caso è maggiore. Facendo un confronto tra i valori ottenuti con le due differenti partizioni del dataset, è possibile notare come i valori associati alle varie metriche di performance cambiano notevolmente al variare del seed specificato. Questo è dato dal fatto che il metodo holdout presenta un'elevata varianza e può dipendere fortemente da quali dati vengono utilizzati per addestrare il modello e quali per testarlo. Questo metodo quindi non lo consideriamo in quanto è molto variabile per un dataset di piccole dimensioni come quello considerato.

5.1.2 Classificazione con metodo Iterated Holdout

Una delle tecniche utilizzate per rendere più efficiente il metodo Holdout consiste nel ripetere iterativamente la procedura. Nel nostro caso sono state effettuate 10 iterazioni. Abbiamo quindi implementato questo metodo con i classificatori utilizzati precedentemente nell'Holdout, ottenendo i risultati riportati nella seguente Tabella 2.

Classificatore	Recall	Precision	F1	Accuracy
RandomForest	0.546	0.671	0.602	0.911
Logistic	0.373	0.651	0.474	0.898
KNN	0.33	0.555	0.414	0.884
NaiveBayes	0.635	0.429	0.512	0.85
SVM	0.338	0.74	0.464	0.903

Tabella 2: Misure di performance ricavate con il metodo Iterated Holdout.

L'accuracy risulta anche in questo caso particolarmente elevata a causa della natura sbilanciata del dataset. Per questo motivo diamo maggior rilievo alle altre metriche di valutazioni, in particolare alla Precision. Questo perché viene posta una maggiore attenzione nell'evitare di classificare un vino come di "buona qualità" quando questa definizione non rispecchia l'effettivo livello del prodotto, in quanto l'investimento su di esso si rivelerebbe senz'altro fallimentare. I valori della Recall anche in questo caso sono bassi, tranne che per Random Forest che risulta pari a 0.546 e Naive Bayes pari a 0.635. Andando ad analizzare i valori della Precision, notiamo come SVM e Random Forest sono i due classificatori migliori, rispettivamente con risultati pari a 0.671 e 0.74. Osservando però la F-measure, il classificatore Random Forest presenta un valore più elevato rispetto ad SVM in quanto effettua meno errori nel classificare un vino di alta qualità come vino di bassa qualità. Prediligendo però come metrica la Precision, consideriamo come classificatori migliori in questo caso sia SVM che Random Forest.

5.1.3 Classificazione con metodo Cross Validation

Come ultima valutazione è stato utilizzato il metodo Cross Validation. Come valore di partizionamento è stato utilizzato $k = 10$, ciò significa che ogni record è presente nel test set una sola volta mentre nel 9 volte nel dataset di train. Nella Tabella 3 sono riportate le misure di performance ottenute.

Classificatore	Recall	Precision	F1	Accuracy
RandomForest	0.612	0.634	0.623	0.909
Logistic	0.374	0.64	0.472	0.897
KNN	0.381	0.566	0.455	0.888
NaiveBayes	0.667	0.458	0.543	0.862
SVM	0.293	0.694	0.411	0.897

Tabella 3: Misure di performance ricavate con il metodo Cross Validation.

Osservando i risultati, ci troviamo davanti ad una diminuzione, seppur di poco, del valore dell'Accuracy, rispetto ai modelli precedenti. Anche il valore di Precision è diminuito in modo lieve rispetto al metodo precedente. Infatti se prima potevamo osservare un valore di Precision pari a 0.671 per Random Forest, ora presenta un valore pari a 0.634 ma in compenso è aumentata la Recall. Considerando invece SVM, in questo caso la Precision ottenuta è pari a 0.694 mentre la Recall è ulteriormente diminuita. In termini di Precision, con questo metodo risulta performante anche la regressione logistica.

Considerando tutte le analisi effettuate e valori di performance per le classificazioni possiamo osservare che il metodo Iterated Holdout permette di ottenere valori migliori rispetto agli altri metodi utilizzati. In particolare Random Forest e SVM risultano essere i classificatori più performanti.

5.1.4 Intervalli di confidenza

Date le misure di performance ottenute abbiamo effettuato un confronto con l'obiettivo di comparare i diversi classificatori attraverso un'analisi più approfondita. Abbiamo quindi calcolato gli intervalli di confidenza, ad un livello del 95%, per il metodo Iterated Holdout e K-folds Cross Validation, ottenendo i seguenti risultati.

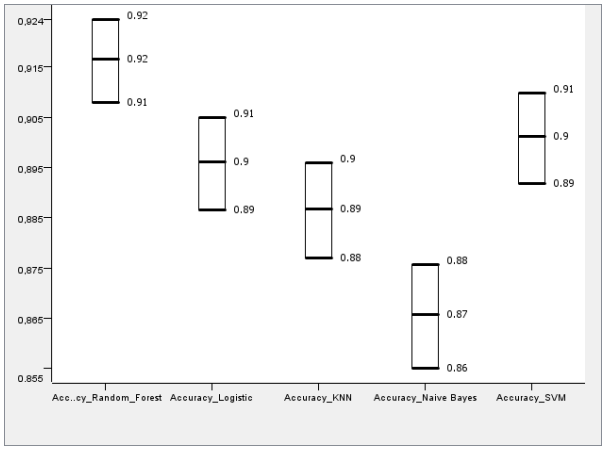

Figura 3: Intervalli di confidenza dell'Accuracy con Iterated Holdout.

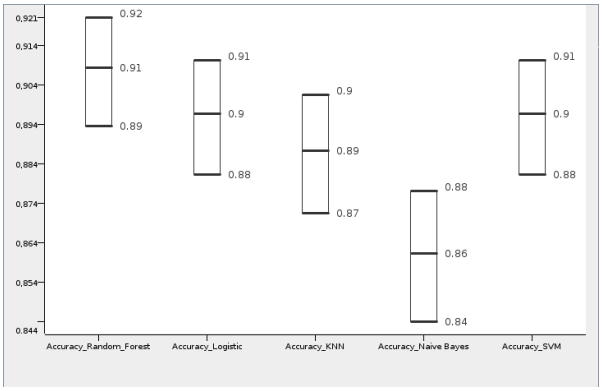

Figura 4: Intervalli di confidenza dell'Accuracy con k-folds.

Esaminando le figure 3 e 4 possiamo notare come alcuni intervalli di confidenza sembrano coincidere. Per non dover condizionare la valutazione di un modello ad un unico valore della soglia, che è, in aggiunta, problematico definire quando i dati presentano un forte sbilanciamento fra le classi, una valida alternativa è offerta dalla curva ROC, uno strumento che permette di studiare analiticamente il risultato di un modello di classifcazione indipendentemente dalla soglia [4]. Sono state quindi confrontate, per i due modelli migliori, Random Forest e SVM, le corrispondenti curve ROC, la cui caratteristica principale è quella di non tener conto della distribuzione della variabile risposta. Sono stati ritenuti buoni entrambi i modelli utilizzati in quanto le aree sottese alle curve sono maggiori dell'area sottesa alla retta rappresentante il modello "Zero Rule" cioè il modello casuale che non aggiunge nessuna informazione. La curva,

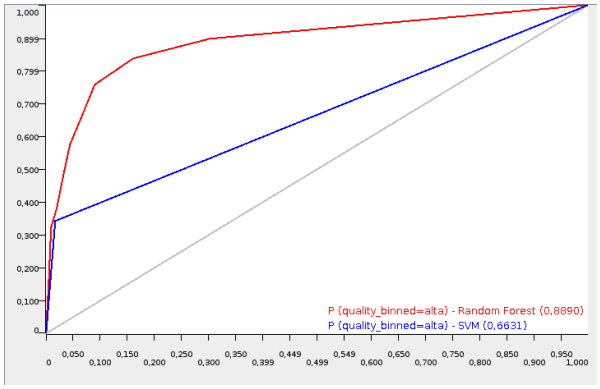

Figura 5: Curva ROC.

di cui si riporta un esempio nella Figura 5, è ottenuta ponendo in ascissa il tasso di falsi positivi e in ordinata quello di veri positivi, calcolati al variare della soglia. Dal grafico è già possibile notare come il classificatore Random Forest sia migliore rispetto a SVM in quanto la sua curva è sempre al di sopra. Andando a valutare l'AUC, ovvero l'area sottesa alla curva, risulta infatti maggiore per il classificatore Random Forest valutato tramite la procedura Iterated Holdout.

5.1.5 Matrice di costo

Il secondo metodo utilizzato per trattare un dataset sbilanciato è stato l'analisi dei costi effettuato tramite l'utilizzo di una matrice di costo applicata al classificatore. Essa assegna dei costi di errata classificazione diversi per le unità delle due classi, forzando anche in questo modo il processo a concentrarsi sui casi più rari. Con essa è possibile stabilire in base ai falsi positivi e ai falsi negativi il costo che dovrebbe sostenere un'ipotetica azienda per la classificazione di un certo vino. Per far ciò, è stato adoperato il nodo *Weka CostSensitiveClassifier* basato su un approccio *Cost Sensitive Learning*. Si è fatto uso sia del partizionamento del dataset tramite

10-folds Cross Validation che tramite la procedura Iterated Holdout. Per scegliere i costi relativi alle classificazioni errate abbiamo variato, utilizzando un approccio brute force, per ogni modello di classificazione adoperato, i costi dei true positive e dei false positive in un ciclo. I costi relativi ad una corretta classificazione sono stati posti uguali a 0; infatti una corretta classificazione di un'unità non comporta alcun costo [1]. I parametri di costo che ottimizzano la misura di performance selezionata vengono in seguito adoperati nella matrice del rispettivo classificatore.

Classificatore	Recall	Precision	F1	Accuracy
RandomForest(IteratedHoldout)	0.238	0.815	0.368	0.899
RandomForest(CrossValidation)	0.252	0.822	0.385	0.901

Tabella 4: Misure di performance con matrice dei costi.

I risultati ottenuto tramite la matrice di costo non sono risultati soddisfacenti. Infatti, ancora una volta, risulta essere più performante il modello Random Forest generato attraverso la procedura Iterated Holdout. Una possibile spiegazione di questo fatto è che non è l'algoritmo a fornire la matrice dei costi ma viene fornita da noi in input e per questo motivo potremmo non aver identificato la matrice di costo ottimale.

5.2 Seconda domanda di ricerca

Il secondo obiettivo preposto è quello di individuare le componenti fisico-chimiche che incidono maggiormente su un vino di 'buona qualità'. Si è deciso di procedere quindi con la Feature Selection, un'operazione che permette di stabilire quali variabili esplicative possono essere ritenute irrilevanti, non contenendo alcuna informazione significativa per la variabile risposta.
Dei possibili diversi approcci si è deciso di utilizzare i metodi Filter e Wrapper e confrontare poi i valori anche con la Regressione Lineare Multipla.

5.2.1 Filter

Il metodo Filter permette di stimare la bontà degli attributi usando euristiche basate sulle caratteristiche generali dei dati, che possono essere studiati secondo un modello multivariato od univariato. L'approccio univariato considera ogni caratteristica in modo isolato mentre l'approccio multivariato tiene in considerazione potenziali interdipendenze. Per entrambi gli approcci, è stato utilizzato il nodo *Weka AttributeSelectedClassifier* utilizzando come tecnica di filtro l'*Information Gain* [2] e *CfsSubsetEval* [3] rispettivamente per il caso univariato e multivatiato. Entrambi i metodi sono stati applicati al classificatore Random Forest. Le variabili più importanti risultano essere le seguenti:

- univariato: *alcohol, citric acid, sulphates, volatile acidity, density, chlorides, total sulfur dioxide*;
- multivariato: *volatile acidity, citric acid, chlorides, density, sulphates, alcohol.*

5.2.2 Wrapper

L'approccio Wrapper applica un algoritmo di apprendimento, usato successivamente per la classificazione, per stimare il valore dei parametri. Se da un lato i metodi di Wrapper sono più efficienti, dall'altro soffrono della dipendenza del classificatore da cui invece non sono vincolati i metodi di filtro, che quindi risultano più flessibili e con un costo computazionale inferiore. Anche per questo metodo è stato utilizzato il nodo *Weka AttributeSelectedClassifier* e come tecnica di filtro *WrapperSubsetEval.*
Le variabili più importanti selezionate da questo modello risultano essere: *volatile acidity, density, pH, sulphates, alcohol.*

5.2.3 Regressione Lineare Multipla

Ultima valutazione è stata effettuata attravero il modello di Regressione Lineare Multipla. Come variabile dipendente è stata utilizzata 'quality binned' mentre i restanti attributi sono stati utilizzati come variabili indipendenti. In particolare è stato utilizzato il test ANOVA, applicato sui valori normalizzati, il quale permette di andare a verificare se c'è una relazione significativa tra la variabile dipendente e l'insieme delle variabili esplicative. Come ipotesi nulla ha che la relazione non esiste, ossia è nulla. Dal test

[2] Il sottoinsieme individuato è costituito da quegli attributi che hanno IG maggiore di 0.

[3] Sono preferiti sottoinsiemi di funzionalità altamente correlate con la classe pur avendo una bassa intercorrelazione.

effettuato risulta che le 6 variabili che influenzano maggiormente sulla qualità di un vino sono: *volatile acidity, density, suphates, alcohol, total sulfur dioxide, chlorides.*

Combiando i due metodi di Feature Selection utilizzati e confrontandoli con i risultati della Regressione Lineare Multipla, è emerso che le varibili che in tutte e tre le metodologie risultano importanti per la qualità di un vino sono: *alcohol, sulphates, density, chlorides* e *volatile acidity.* Tenendo in considerazione queste variabili, è stato quindi testato il metodo di classificazione Random Forest, risultato precedentemente come il più performante, sia con la procedura Iterated Holdout che con Cross Validation ottenendo però risultati meno performanti rispetto al modello completo, cioè considerando tutte le variabili a nostra disposizione.

Classificatore	Recall	Precision	F1	Accuracy
RandomForest(IteratedHoldout)	0.551	0.644	0.594	0.907
RandomForest(CrossValidation)	0.592	0.659	0.624	0.912

Tabella 5: Misure di performance con Feature Selection.

6 Conclusioni e sviluppi futuri

Riprendendo ciò che era stato affermato nell'introduzione, la previsione della qualità del vino rimane un processo complicato, che non può essere limitato a poche caratteristiche sebbene il seguente studio ne abbia trovate alcune che sembrano incidere più di altre. D'altra parte non possiamo non considerare l'impatto di alcuni risultati, in particolar modo di quelli ottenuti tramite il classificatore Random Forest e SVM. In relazione al primo obiettivo, infatti, si è voluta identificare la tecnica di classificazione di Machine Learning migliore per la catalogazione dei vini con lo scopo di prevederne la qualità in relazione alle componenti chimico-fisiche che lo caratterizzano. Dal primo obiettivo è emerso che in particolare il classificatore Random Forest aveva performance migliori in confronto agli altri. Essendo in presenza di un dataset sbilanciato, si è effettuata un'analisi dei costi a seguito della quale non si sono riscontrati significativi miglioramenti. Successivamente, in accordo con il secondo scopo, è stata applicata una Feature Selection con l'utilizzo di diversi approcci quali Filter e Wrapper che, confrontati poi con i risultati generati dalla Regressione Lineare Multipla, hanno evidenziato l'importanza di 5 variabili alle quali poi è stato testato il modello Random Forest senza generare, anche in questo caso, performance migliori.
La presente ricerca potrebbe avere numerose implicazioni future principalmente legate alla produzione e al commercio di vino a livello mondiale e potrebbe essere condotta in modo più efficiente ed efficace considerando non solo la varietà di vino contenuta in questo dataset, ma estendendo quindi la ricerca ad ulteriori produzioni di questo alcolico e utilizzando approcci meno comuni come l'adozione modelli più complessi.

Riferimenti bibliografici

[1] Charles Elkan. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd, 2001.

[2] https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009.

[3] https://www.knime.com/ .

[4] F Provost and T Fawcett. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions in: Proc of the 3rd international conference on knowledge discovery and data mining. pages 43–48, 1997.