Progetto di Machine Learning – Febbraio 2021

ANALISI PREDITTIVA SULLA QUALITÀ DEL VINO

Team 01: Claudio Fadda[*], Marco Martines[*], Vittorio Menardo[*], Davide Mobilia[*]

[*]Università degli Studi di Milano Bicocca, CdLM Data Science

Sommario

L'Italia rappresenta il primo produttore mondiale di vino ed è il secondo esportatore più importante nell'Unione Europea. Ciò è dovuto in buona parte alla qualità del vino italiano, riconosciuta a livello locale e internazionale e in questo contesto diventa quindi sempre più importante riconoscere la qualità di un prodotto. La nostra ricerca ha lo scopo di sviluppare un modello predittivo di Machine Learning in grado di classificare la qualità del vino sulla base degli attributi che ne determinano le principali caratteristiche.

Indice

Introduzione

Il settore del vino in Italia è uno dei più importanti nell'economia agroalimentare del Paese, vedendo impiegate circa 2000 imprese industriali e detenendo una quota di circa 1/5 delle esportazioni mondiali, con una crescita continua di anno in anno[1].

Con il nostro progetto di Machine Learning ci si è posti l'obiettivo di riuscire ad individuare, tramite un modello di classificazione, i migliori vini rossi attraverso diversi attributi, e vedere quali di questi influenzano maggiormente la qualità del vino.

Sebbene i parametri che descrivono un vino non siano innumerevoli, essi determinano quelle caratteristiche che possono rendere un vino di scarsa qualità fino ad uno di ottima fattura con notevoli differenze di prezzo.

Diventa quindi importante poter identificare un prodotto di buona qualità anche sulla base di parametri oggettivi e processabili, ottenendo così analisi più veloci e attendibili.

L'articolo prevede inizialmente una breve analisi del dataset di riferimento, seguito dalla fase di preparazione dei dati. Successivamente vengono presentati i modelli di classificazione utilizzati, con le relative misure di performance. In conclusione vengono discussi e analizzati i risultati ottenuti e gli obiettivi raggiunti.

Descrizione del dataset

Il dataset "*winequality-red.csv*"[2], disponibile sulla piattaforma Kaggle[3], è composto da 12 attributi e contiene 1599 record. Tutti gli attributi sono di tipo quantitativo.
Gli attributi del dataset sono:

1 – fixed acidity (g/dm^2): acidità fissa, rappresentata dagli acidi non volatili;

2 – volatile acidity (g/dm^3): quantità di acido acetico (a livelli troppo alti può portare a uno sgradevole sapore di aceto);

3 – citric acid (g/dm^3): acido citrico (in piccole quantità può aggiungere freschezza e sapore ai vini);

4 – residual sugar (g/dm^3): zucchero rimanente dopo la fermentazione;

5 – chlorides (g/dm^3): quantità di sale;

6 – free sulfur dioxide (mg/dm^3): forma libera di SO_2;

7 – total sulfur dioxide (mg/dm^3): quantità totale di SO_2;

8 – density (g/cm^3): densità del vino (dipenda dal rapporto tra acqua, alcol e zuccheri);

9 – pH: acidità del vino, su una scala acido/basico;

10 – sulphates (g/dm^3): solfiti disciolti;

11 – alcohol (% in volume): il contenuto alcolico del vino;

12 – quality: basata su dati sensoriali, punteggio compreso tra 0 e 10;

Preprocessing

Dopo aver effettuato la lettura del dataset, si è prodotta un'analisi dell'attributo quality. L'analisi dell'attributo è stata effettuata tramite un istogramma (Fig.1). I valori dell' attributo quality (riportati sull'asse X) con le frequenze assolute maggiori (riportate sull'asse Y) sono 5 e 6. Nello specifico i valori minori di 6, che come abbiamo detto denotano una bassa qualità del vino, rappresentano il 47% della distribuzione, mostrando quindi come il dataset risulti composto per metà da vini di qualità, e per l'altra metà da vini di qualità non sufficiente. Infatti il solo valore 5 rappresenta in termini di frequenze relative il 42,5% della distribuzione, contrapposto al valore 6 che ne rappresenta il 39.9%.

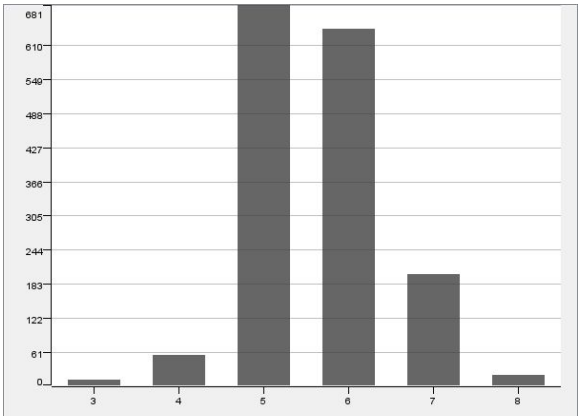

Figura 1: Distribuzione attributo quality

Una volta analizzato l'attributo della qualità e verificata la presenza di eventuali valori mancanti, l'attributo è stato binarizzato tramite il nodo R Snippet, in modo da creare un nuovo attributo binario good/low, attribuendo come good i valori con qualità =>6 e successivamente procedendo con l'analisi dei modelli.

Feature Selection

Con le feature selection vengono selezionati gli attributi che più influiscono sulla variabile target (Quality) in modo da ridurre gli attributi di input: una volta individuati gli attributi desiderati si procederà all'eliminazione degli altri attributi considerati irrilevanti, portando ad un aumento della prestazioni del modello. Attraverso diversi Attribute Selected sono stati individuati gli attributi più influenti sulla variabile di output, valutandoli quindi uno ad uno e producendo come risultato un ranking della loro rilevanza rispetto alla variabile desiderata. I cinque attributi in ordine di importanza, sono: *alcohol, total sulfur dioxide, density, volatile acidity* e *sulphates*. Gli attributi giudicati irrilevanti dalle features selection, vengono rimossi portando un minor rischio di overfitting ed un miglioramento dei tempi di training del modello stesso.

```
Ranked attributes:
 0.0885      11 alcohol
 0.0789      7 total sulfur dioxide
 0.0541      8 density
 0.0529      2 volatile acidity
 0.0497      10 sulphates
 0.0316      5 chlorides
 0.0315      3 citric acid
 0.0302      1 fixed acidity
 0           4 residual sugar
 0           9 pH
 0           6 freesulfurdioxide
```

Figura 2: Risposta output Attribuite Selected GainRatio

Modelli Utilizzati

Per fornire delle risposte sulle qualità dei vini sono stati utilizzate diverse tecniche di classificazione, con l'obiettivo di valutare la più adatta. Sono stati proposti i seguenti modelli:

- Modelli probabilistici

Sono i metodi basati su ipotesi probabilistiche come il classificatore *Naive Bayes* che usa il teorema di Bayes, il *Bayes Net* che incorpora il concetto di rete Bayesiana e l'*NBTree* che si basa sempre sulla formula di Bayes e sul concetto di albero.

- Modelli euristici

Nonostante non garantiscono di giungere a risultati ottimali, permettono di ottenere soluzioni approssimate e ragionevoli. Questi metodi fanno riferimento in particolar modo agli alberi decisionali che vengono sviluppati a partire da un sottoinsieme di dati iniziali (training set) per il quale è nota la classe output. La classificazione avviene sulla base della classe maggioritaria all'interno del nodo finale. Tra questi modelli si è concentrata l'attenzione sul classificatore *Random Forest*, sull'albero di regressione *J48*, e sul *Decision Tree Learner*.

- Modelli di regressione

Sono modelli basati sulla regressione logistica, il loro vantaggio è rappresentato dalla possibilità di considerare qualsiasi tipo di input, risultando così estremamente flessibili. Questi classificatori tendono a misurare l'effetto di ciascun input sull'output calcolando la probabilità a posteriori dell'attributo di classe dati i valori degli attributi esplicativi. Per questa categoria è stato scelto il modello *Logistic* e il *Simple Logistic*.

- Modelli di separazione

Fanno parte di questa categoria i modelli che vanno a partizionare lo spazio degli attributi. Sono stati scelti: *SPegasos*, SVM (Support Vector Machine) con kernel Puk e infine il *Multilayer Perceptron*.

Valutazione dei risultati

Nella nostra analisi sono stati utilizzati più criteri in grado di valutare la performance. Nello specifico, visto che il nostro dataset risulta bilanciato, ci siamo concentrati sulla valutazione dell'Accuracy, andando a fare un confronto tra i vari classificatori. Inoltre sono state analizzate la ROC curve, la Cumulative Gains e il Lift Charts dei vari modelli di classificazione.

In questa prima fase è stato impiegato il metodo dell' Iterated Holdout, versione più performante dell'holdout. Con la sua implementazione si riesce, infatti, a mitigare notevolmente il bias che caratterizza questo tipo di partizionamento, troppo dipendente dal subset considerato con il rischio di sovra/sottostimare la reale accuratezza.

Grazie all'Iterated Holdout la partizione dei dati viene divisa, K volte, in un set di training (67% dei record) e in un set di test (33% dei record). L'accuratezza viene quindi calcolata come media delle accuratezze stimate per ognuna di queste istanze dell'Holdout.

Da specificare il fatto che è stato deciso di utilizzare un iterated holdout per le dimensioni ridotte del dataset. Una cross validation, infatti, avrebbe ridotto troppo le dimensioni dei training set, portando a instabilità nella valutazione dell'accuracy.

Accuracy

L'**Accuracy** indica la percentuale di osservazioni positive e negative previste correttamente:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Dove TP e TN indicano il numero di istanze classificate correttamente come appartenenti rispettivamente alla classe positiva e negativa; FP e FN indicano il numero di istanze positive e negative classificate erroneamente. In generale, i modelli con Accuracy più alta sono valutati come migliori.

Per comparare i diversi classificatori in termini di accuratezza, è stata svolta un'analisi dell'Accuracy calcolando gli intervalli di confidenza attraverso il metodo Wilson con un livello di confidenza pari al 95%. Il limite inferiore e il limite superiore dell'intervallo sono stati implementati in Knime attraverso dei nodi Math Formula utilizzando la formula dell'intervallo di confidenza di Wilson, come riportato in basso.

$$\left(\frac{acc + \frac{Z^2_{1-\frac{\alpha}{2}}}{2 \cdot N} - Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{acc}{N} - \frac{acc^2}{N} + \frac{Z^2_{1-\frac{\alpha}{2}}}{4 \cdot N^2}}}{\left(1 + \frac{Z^2_{1-\frac{\alpha}{2}}}{N}\right)}, \frac{acc + \frac{Z^2_{1-\frac{\alpha}{2}}}{2 \cdot N} + Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{acc}{N} - \frac{acc^2}{N} + \frac{Z^2_{1-\frac{\alpha}{2}}}{4 \cdot N^2}}}{\left(1 + \frac{Z^2_{1-\frac{\alpha}{2}}}{N}\right)} \right)$$

Di seguito riportiamo i valori di accuracy dei valori modelli di classificazione (Figura 3).

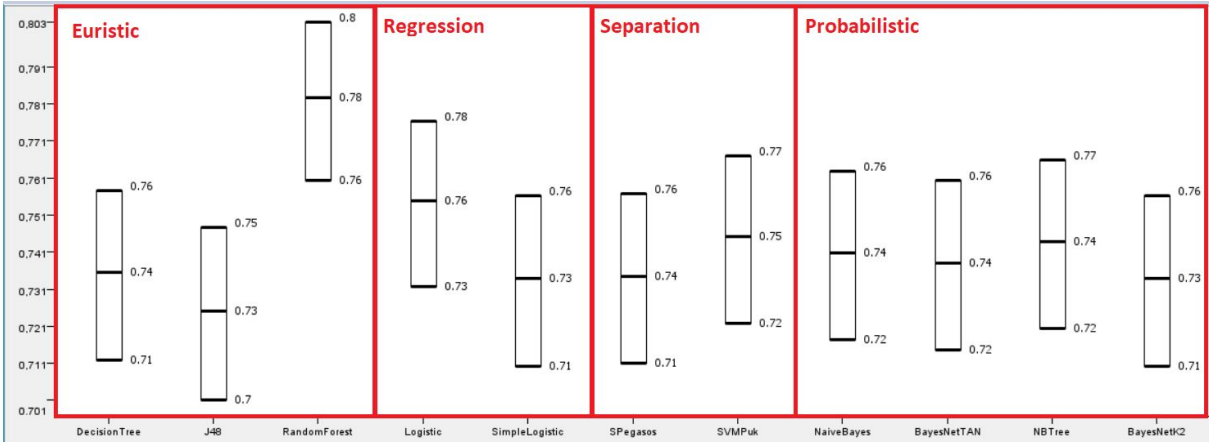

Figura 3: Box plot accuratezza

Dal grafico si può dedurre che non vi siano differenze significative tra le macro categorie dei classificatori. Per quanto riguarda i modelli euristici si possono notare notevoli differenze tra i modelli, con il Random Forest che riesce a ottenere un intervallo di accuratezza migliore, il più alto tra tutti i modelli considerati globalmente. Per quanto riguarda le altre categorie di classificatori non risultano particolari tendenze.

ROC Curve

Un ulteriore metodo utilizzato per valutare il modello di classificazione è la **ROC curve (Receiver operating characteristic)** che consente di rappresentare, sull'asse delle ordinate, la percentuale del numero totale di veri positivi (classe positiva effettivamente prevista come tale, TP) e, sull'asse delle ascisse, la percentuale di falsi positivi (classe negativa erroneamente prevista come positiva, FP).

La ROC Curve ci permette di quantificare, per ogni data percentuale di veri positivi (TP) che il classificatore è in grado di prevedere, qual è la percentuale di falsi positivi (FP), ossia di errori, che il nostro classificatore commette.

Dalla ROC curve è inoltre possibile ricavare l'**AUC (Area Under the Curve)** che ne è una sua rappresentazione numerica e permette quindi di confrontare in modo più preciso i diversi modelli.

Valutazione ROC curve

Valutando le performance in termini di ROC Curve andiamo a confrontare le diverse curve ottenute. Nella Fig.4 vengono mostrate le tre curve migliori, relative a Random Forest, SVM Puk e NaiveBayes.

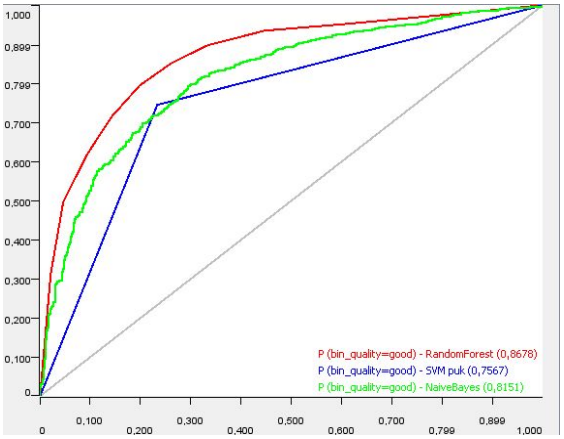

Figura 4: ROC curve

Come si può notare dalla figura 4 in corrispondenza di ~ 55% di True Positive il modello Random Forest restituisce all'incirca il 10% di False Positive mentre, per esempio, il modello SMOPuk in corrispondenza del 20% di TP, restituisce il 10% di istanze classificate erroneamente come positive, risultando quindi meno performante. E' quindi chiaro come il modello Random Forest sia il più performante in confronto agli altri modelli.

In relazione all'AUC, le curve migliori risultano essere quelle dei modelli Random Forest e NaiveBayes. L'AUC più alto viene ottenuto dal modello Random Forest, con il valore di 0.867.

Featured vs Full Dataset

In parallelo al dataset a cui è stata applicata la Feature Selection abbiamo preso in considerazione anche il dataset completo con tutti i suoi attributi. Questa scelta è stata fatta per poter quantificare l'effettiva perdita e/o incremento dell'accuratezza, e quindi valutare l'efficacia di una Feature Selection rispetto ai diversi modelli.

Una volta ottenute le due accuratezze (Featured e Full DS) abbiamo calcolato la seguente grandezza:

$$Delta\ ACC = \frac{ACC\ Full\ DS\ -\ ACC\ Feature\ DS}{ACC\ Full\ DS} \cdot 100$$

Ottenendo così la percentuale di accuratezza persa passando dal dataset Full a quello Featured. Per ogni modello sono state ottenute le seguenti stime:

Row ID	D Accuracy_Diff
SVMPuk	2.489
Decision Tree	2.122
SimpleLogistic	1.668
MultilayerPerc...	1.573
J48	1.563
sPegasos	1.287
NBTree	0.86
Random Forest	0.749
Logistic	0.025
BayesNetTAN	-0.141
NaiveBayes	-1.134
BayesNetK2	-2.192

Figura 5: Delta tra ACC full DS e ACC feature DS

E' facile notare (Figura 5) quanto lo scarto tra le due accuracy sia minimo con valori che oscillano intorno all' |1%| e tocca il suo massimo con il modello SVMPuk: il 2.5%. Per questo modello si passa infatti da uno **0.751** Full DS ad un **0.746** di accuratezza del Featured DS, una perdita del tutto trascurabile. In generale sono perdite che rientrano pienamente nell'intervallo di confidenza delle rispettive accuratezze, suggerendo quindi che non ci sia un significativo discostamento.

Dato importante per l'implementazione di modelli attraverso la feature selection e che ci permette di coglierne la sua efficacia. Passando infatti da ben 11 attributi a 5, e quindi una riduzione delle dimensioni del dataset di oltre il 50%, abbiamo un'accuratezza sostanzialmente invariata.

Conclusioni

In definitiva possiamo dire di aver raggiunto buoni risultati per la classificazione della qualità del vino sulla base dei parametri alcohol, total sulfur dioxide, density, volatile acidity e sulphates.

Utilizzando un modello RandomForest si raggiunge infatti un' accuracy dell' 80% che ci permette di identificare la buona qualità del vino nella maggior parte dei casi. Il risultato non è certamente ideale poiché una volta su cinque il nostro classificatore commetterà un errore, ma risulta essere comunque soddisfacente data la natura dei nostri dati.
La variabile quality infatti concentra la maggior parte dei suoi record tra i valori 5 e 6, non vi è quindi una netta differenziazione delle qualità.

A questo proposito è interessante notare come gli zuccheri residui siano del tutto ininfluenti per l'algoritmo, con un GainRatio pari a 0. La dolcezza del vino sembra quindi essere un parametro del tutto trascurabile nella valutazione della sua qualità.

Non stupisce invece l'influenza dell'acido acetico, che in quantità elevate porta ad uno sgradevole sapore di aceto.

E' importante sottolineare la difficoltà di questo tipo di classificazione rappresentata dalla natura soggettiva della nostra variabile target. La qualità del vino è infatti basata su valori sensoriali che pur avvalendosi di alcuni metri di giudizio saranno sempre influenzati dalla soggettività dell'individuo.

Riferimenti

[1] Cappellini, M. Il vino italiano fattura 11 miliardi. Italia primo produttore mondiale (2019) https://www.ilsole24ore.com/art/il-vino-italiano-fattura-11-miliardi-italia-primo-produttore-mondiale-ABAdqbkB

[2] Red Wine Quality, Kaggle Dataset https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009

[3] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009