Modellizzazione tramite tecniche di classificazione di data mining delle preferenze riguardo al vino bianco, a partire da caratteristiche fisico-chimiche

Bianchini Filippo, 871378 Maliwat Julius, 864520

Rabuzzi Giacomo, 864452 Robbiani Andrea, 865118

23 Novembre, 2022

Abstract

La nostra analisi si propone di identificare quale tra le differenti tecniche di classificazione di data mining riesca nel migliore dei modi a modellare le preferenze di gusto relative al vino bianco, basandosi su dati analitici che sono semplicemente accessibili durante la fase di certificazione di qualsiasi vino. Il modello K-NN allenato su set bilanciato fornisce risultati positivi, migliori che tutti gli altri testati.

Questo lavoro può essere utile per determinare quali prodotti possano essere ritenuti d'eccellenza, rispetto alla preferenze di gusto, in base a determinate caratteristiche fisico-chimiche.

1 Introduzione

Per condurre la nostra analisi abbiamo tratto ispirazione da un paper pubblicato nel 2009 da Paulo Cortez, Antònio Cerdeira, Fernando Almeida, Telmo Matos e Josè Reis, intitolato "Modeling wine preferences by data mining from physicochemical properties". Gli studiosi hanno deciso di sviluppare questa analisi considerando la

grande crescita che era in corso nel mercato del vino. Il Portogallo, nazione degli autori, era da poco entrato tra i 10 maggiori esportatori di vino a livello mondiale; in particolare l'esportazione del *vinho verde*, un vino originario della provincia storica del Minho, nell'estremo nord del paese, tra il 1997 e 2007, era aumentata del 36%.

In questo contesto di crescita hanno avuto un ruolo fondamentale le certificazioni ottenute dal prodotto e la sua qualità. Queste valutazioni si basano su dati fisico-chimici e analisi sensoriali (gusto e olfatto). Tra questi diversi tipi di dati è difficile individuare una relazione.

L'obiettivo del loro lavoro, come del nostro, era trovare la tecnica di Data Mining che meglio sapesse classificare le osservazioni rispetto alla relativa variabile target.

Il fine di questo lavoro era quello di poter migliorare le fasi di produzione e vendita del vino affinché il prodotto risultasse migliore al gusto e la sua commercializzazione maggiormente remunerativa.

Gli autori dell'analisi originale propongono di utilizzare a questo scopo la regressione multipla, le reti neurali e tecniche di support vector machines, ossia supervised learning models; inoltre utilizzano, oltre al dataset *white wine quality,* anche un dataset *red wine quality* che presenta le stesse identiche variabili.

In conclusione hanno dichiarato che il modello SVM è quello preferibile, avendo un'accuracy pari a 86.8% (per quanto riguarda il nostro stesso dataset).

Diversamente dal nostro lavoro, il paper di Cortez et al. non riclassifica la variabile risposta rendendola binaria, ma piuttosto la stima in un intervallo di confidenza.

Il risultato della suddetta analisi è stato importante per l'industria del vino in quanto questo approccio è basato su test oggettivi e può essere integrato nei processi decisionali soggettivi dai quali tendenzialmente dipendono le certificazioni.

Il report che segue si dividerà in diverse sezioni oltre a questa introduzione: materiali e metodi (2.1 e 2.2), che consiste in una spiegazione dei dati, delle tecniche e dei modelli di data mining utilizzati; risultati (3), dove vengono presentati tutti i risultati ottenuti durante l'analisi attraverso rappresentazioni grafiche e tabelle; discussioni (4), sezione dove vengono fornite le conclusioni del lavoro svolto. Infine è presente una bibliografia (5) dove sono presenti i rimandi ai documenti citati.

2 Materiali e metodi

2.1 Materiali

Il dataset da noi scelto ha come titolo “White Wine Quality”, proviene da una ricerca condotta nel 2009 dall’*Università di Minho, Guimaraes, Portogallo* e dalla *Commissione della viticoltura della regione del Vinho Verde*.

Le osservazioni sono 4989 e riguardano la variante del bianco del Vinho Verde; sono state raccolte tra il maggio del 2004 e il febbraio del 2007 e riguardano prodotti DOP che sono stati testati da un ente preposto.

Le dodici variabili chimico-fisiche sono state calcolate in laboratorio. Per quanto riguarda invece la variabile che contiene le preferenze di gusto, sono stati effettuati assaggi “al buio”. Ogni assaggiatore restituiva almeno tre pareri sensoriali con un voto tra 0 e 10. Il valore riportato è il valore mediano.

Le variabili Fisico-chimiche sono l'input, quella sensoriale è l’output.

Queste sono le variabili:

- fixed acidity: l’acidità fissa è costituita dalle sostanze acide presenti in un vino, che non sono portate a volatilizzare, ma al contrario restano all’interno del vino per tutta la sua vita.
- volatile acidity: l’acidità volatile è costituita da molecole che tendono a disperdersi nell’aria, la quantità totale al fine di essere un pregio, secondo esperti di settore, dev’essere inferiore allo 0,7%. Varcato questo limite, diventa un difetto.
- citric acid: è un acidificante utilizzato per correggere l'acidità in mosti e vini, che svolge inoltre un'azione stabilizzante come antiossidante. Se in quantità non eccessive può donare freschezza e sapore al vino.
- residual sugar: conosciuta anche come dolcezza residua (RS), la quantità di zucchero nel vino che si ottiene con la fine naturale della fermentazione, più i vini sono fermentati minore è questa quantità (min 0.6-0.7g/L).
- chlorides: la quantità di cloruri (sali) presenti nel vino, sono presenti in quantità limitata (mediamente 0,05-0,2 g/L).
- free sulfur dioxide: L’aggiunta di biossido di zolfo (anidride solforosa) costituisce una diffusa pratica enologica giacché esso agisce da conservante, antiossidante. Esso previene l’instabilità microbiologica durante il processo di vinificazione e le fermentazioni secondarie nei vini dolci.
- total sulfur dioxide: biossido di zolfo totale.

- density: la densità del vino è simile a quella dell'acqua e dipende dalla percentuale di alcool e zucchero presente.
- pH: descrive l'acidità o la basicità del vino in valori comprese tra 0 e 14. Il vino ha tendenzialmente un pH che varia tra 2.8 e 4, i produttori di bianco cercano di mantenerlo tra 3 e 3.5.
- sulphates: i solfiti vengono aggiunti al vino per integrare la funzione disinfettante, antiossidante e stabilizzante dell'anidride solforosa.
- alcohol: percentuale di alcool presente nel vino.
- quality: variabile output (basato su dati sensoriali, ha un punteggio tra 0 e 10).

Non sono presenti valori mancanti.

2.2 Metodi

2.2.1 Analisi Preliminare dei Dati

Il primo passaggio svolto è stato l'analisi preliminare dei dati.

- analisi dei missing values nel dataset.
- controllo dei valori anomali.
- riclassificazione della variabile target come *eccellente* (1) se il suo valore è compreso tra 7 e 10 e come *non eccellente* (0) se il suo valore è compreso tra 0 e 6. La variabile è dicotomica.
- divisione delle osservazioni, tramite campionamento casuale, in training set, validation set e test set.
- Training set (small training set): è un insieme di osservazioni usate per la fase di apprendimento del metodo di Data Mining. La variabile dipendente è nota.
- Validation set: set di osservazioni utilizzato per ottimizzare i parametri del modello e per valutarne le performance.
- Test set: parte del dataset utilizzata solo per l'assessment finale del metodo di classificazione o regressione. Le osservazioni contenute in questo dataset non sono mai state usate durante la fase di training e validation.

Continuiamo l'analisi esplorativa dei dati sullo small training set.

- Verifica della presenza di valori anomali
- Verifica e interpretazione di zeri e eventuali valori negativi nel dataset
- Verifiche grafiche per l'analisi delle distribuzioni delle variabili, in modo da ottenere ulteriori informazioni e testarne la normalità.
- Trasformazione delle variabili per ridurre gli outliers e rendere normali le distribuzioni.
- Analisi della correlazione tra variabili.

- Standardizzazione per media e varianza dei dati.
- Verifica della distribuzione delle variabili condizionatamente al target.

Eseguiamo le stesse trasformazioni fatte sul training set sul validation set.

2.2.2 Modelli di classificazione

Classificatore K-nearest neighbors

Il K-NN è un algoritmo non parametrico utilizzato per prevedere la classe di appartenenza di una nuova osservazione avendo calcolato la distanza con le osservazioni già note.

L'unico paramento presente nel K-NN è *k,* detto parametro di tuning. Rappresenta il numero di vicini da considerare dall'osservazione scelta.

L'algoritmo calcola la distanza tra ogni osservazione.

Dopo aver individuato la classe più presente nelle *k* osservazioni più vicine, l'algoritmo la assegna alla nuova osservazione con probabilità proporzionale al numero di osservazioni di quella classe nell'intorno più vicino di dimensione *k*.

Essendo l'unico parametro del modello il suo valore modifica la complessità. Maggiore è il suo valore, maggiormente saranno lisce le linee di confine dei gruppi. Se *k* è troppo piccolo si può incorrere in un problema di overfitting.

Il modello non ha ipotesi a priori. Fondamentale è determinare il valore di *k* che meglio sappia classificare le osservazioni.

Regressione logistica

La Regressione Logistica è un'espansione di un processo di regressione lineare e fa parte dei generalized linear models. Questi modelli vengono utilizzati ad esempio quando la regressione lineare non è funzionante in quanto la variabile target è dicotomica.

È usata per descrivere i dati, fare analisi predittiva e per spiegare la relazione tra una variabile target e le covariate.

La Regressione Logistica stima il logaritmo dell'odds di un evento, quindi della sua ragione di scommessa ($\pi(x)$) (*logit* è il suo logaritmo). Infatti il modello si presenta come:

$$logit(\pi(x)) = \beta_0 + \sum_{k=1}^{p} \beta_j x_j \qquad logit(\pi(x)) = ln[\frac{\pi(x)}{1-\pi(x)}]$$

Tramite la regressione logistica è possibile individuare quali variabili siano maggiormente esplicative e quali invece possano essere omesse. Si opera una selezione stepwise rispetto al Criterio di Akaike, che deve essere minimizzato.

Analisi discriminante

Il metodo successivo è l'analisi discriminante, che è costituita da due diversi metodi: analisi discriminante lineare (LDA) e analisi discriminante quadratica (QDA). La differenza tra questi due metodi consiste nel fatto che LDA costruisce linee di confine lineari, mentre QDA quadratiche.

Si tratta di una tecnica multivariata di classificazione che separa gli oggetti in uno o più gruppi basandosi sulle caratteristiche misurabili di tali oggetti. Queste caratteristiche misurabili sono chiamate predittori o variabili indipendenti, i gruppi invece sono la variabile risposta o dipendente.

Lo scopo è creare un modello che possa essere usato per fare previsione in modo da comprendere la relazione tra variabili indipendenti e dipendente. Vengono usati dati con classi già note. Si basa sul Teorema di Bayes.

Verifica assunzioni analisi discriminante

Entrambi i metodi assumono che le variabili indipendenti all'interno delle classi si distribuiscono normalmente e necessitano che i dati soddisfino determinate condizioni:

- LDA: matrici di varianza e covarianza per le variabili indipendenti siano uguali in tutte le classi.
- QDA: matrici di varianza e covarianza non necessariamente uguali per tutte le classi.

2.2.3 Downsampling

Tecnica di ricampionamento utilizzabile se il dataset oggetto di analisi ha proporzioni di classe asimmetriche (nel nostro caso 78% excellent vs 22% poor). La classe che costituisce gran parte dei dati è chiamata classe maggioritaria, l'altra minoritaria.

Il downsampling comporta la rimozione casuale delle osservazioni dalla classe maggioritaria per evitare che lo sbilanciamento dovuto dalla sua presenza domini nella fase di training dei modelli.

Ecco i passaggi del processo di downsampling:

1 In primo luogo, separeremo le osservazioni presenti nel training in due data frames differenti in base al valore assunto dalla variabile target, ossia per classe.

2 Successivamente, ri-campioneremo la classe maggioritaria senza sostituzione, impostando una numerosità campionaria pari a quella della classe minoritaria.

3 Infine, combineremo il data frame della classe di maggioranza sottocampionato con il data frame originale della classe minoritaria.

2.2.4 Metriche di valutazione per la classificazione binaria

Accuracy: rapporto tra numero di stime corrette e il numero totale di campioni in input.

AUC: rappresenta l'area sotto la curva ROC (traccia la probabilità di un risultato vero positivo (sensibilità) in funzione della probabilità di un risultato falso positivo per una serie di punti di cut-off)

Balanced Accuracy: media aritmetica tra Sensitivity e Specificity.

3 Risultati

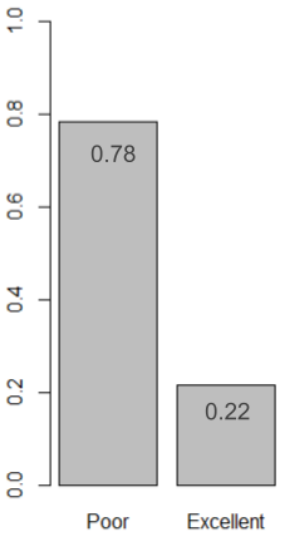
Per prima cosa abbiamo controllato l'assenza di valori mancanti e di valori anomali (per esempio numeri negativi). Abbiamo allora riclassificato la variabile quality dando ad ogni osservazione con valore pari o maggiore di 7 il valore 1 e alle osservazioni con punteggio quality inferiore a 7 abbiamo attribuito il valore 0.

Abbiamo diviso il data set in training set (80% delle osservazioni totali) e test set (20% delle osservazioni). Il training set è stato nuovamente diviso in uno small training set (75% delle osservazioni di training set) e un validation set (25% delle osservazioni di training set).

L'analisi esplorativa successivamente condotta è stata effettuata su small training set.

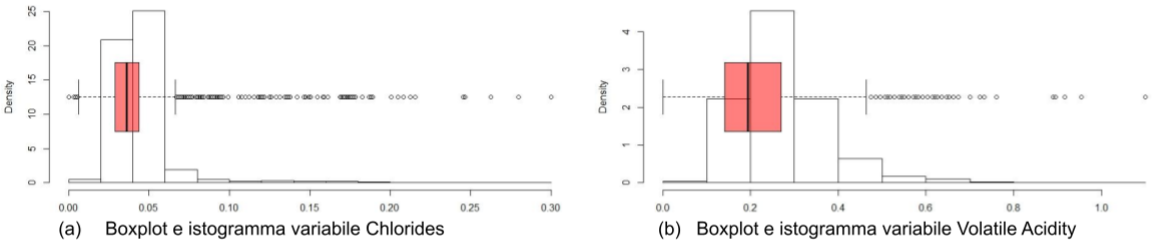
La variabile target quality risulta essere sbilanciata, come possiamo vedere in figura 1.

Figura 1: Distribuzione variabile quality



Abbiamo analizzato gli istogrammi e i boxplot delle variabili per iniziare a studiarne la distribuzione e per verificare la presenza di outliers. Tutte le variabili tranne alcohol presentano outliers. Le variabili con code più pesanti e che necessitano di una trasformazione risultano essere volatile.acidity, residual.sugar, chlorides, free.sulfure.dioxide e sulphates (figura 2).

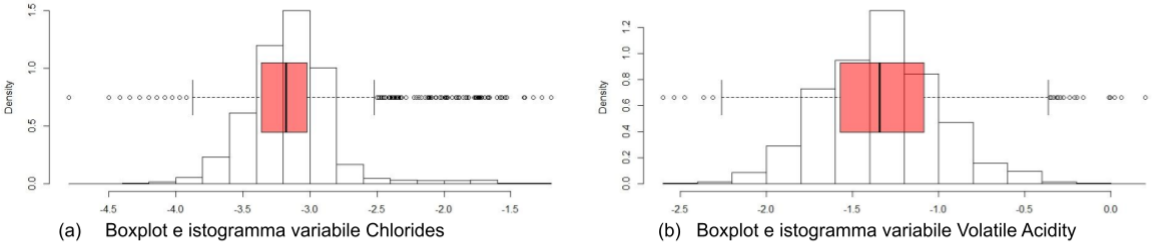
Figura 2: Rappresentazione grafica variabili Chlorides e Volatile Acidity



(a) Boxplot e istogramma variabile Chlorides (b) Boxplot e istogramma variabile Volatile Acidity

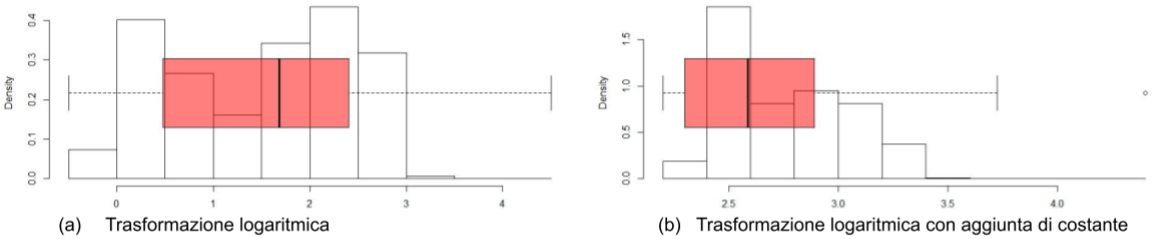
Attraverso un'analisi grafica, valutiamo come performi la trasformazione logaritmica sulle variabili individuate, prima di applicarla definitivamente. Essa sembra funzionare bene con 4 variabili (figura 3), mentre con residual sugar non sembra andare molto bene. Decidiamo quindi di applicare sempre questa trasformazione, ma questa volta aggiungendo una costante pari a 10 (figura 4). Ora la trasformazione sembra performare meglio.

Figura 3: Rappresentazione grafica variabili Chlorides e Volatile Acidity dopo trasformazione logaritmica



(a) Boxplot e istogramma variabile Chlorides (b) Boxplot e istogramma variabile Volatile Acidity

Figura 4: Confronto trasformazioni di Residual Sugar

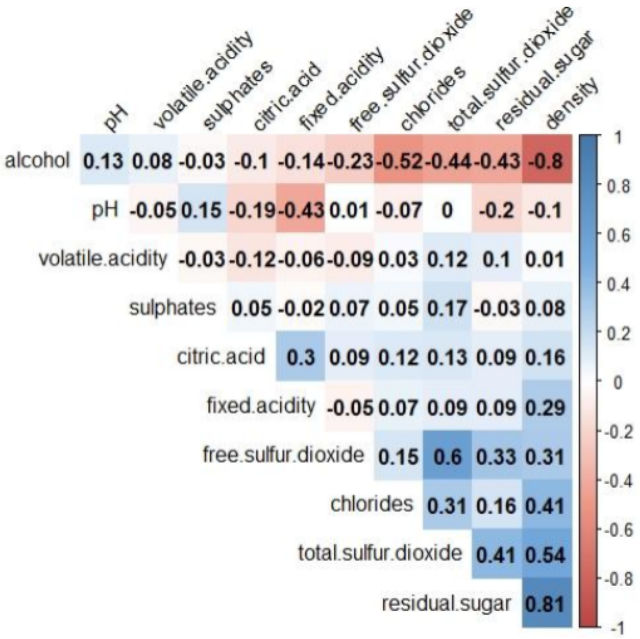


(a) Trasformazione logaritmica (b) Trasformazione logaritmica con aggiunta di costante

Tre variabili (density, total.sulfur.dioxide, residual sugar) presentano outliers potenzialmente problematici. Una volta individuati, decidiamo di eliminarli.

A questo punto valutiamo la correlazione tra le variabili. Quelle con correlazione più alta risultano essere residual.sugar - density (0.81) e alcohol - density (-0.8) (figura 5). Risultato prevedibile poiché la densità del vino dipende proprio dalla concentrazione di alcohol e di zuccheri. Decidiamo lo stesso di mantenere density all'interno dell'analisi e della successiva classificazione.

Figura 5: Heatmap basata sui valori di correlazione



Giunti a questo punto, abbiamo standardizzato per media e varianza, siccome abbiamo trattato gli outliers più influenti.

Abbiamo proseguito questa parte di analisi verificando le assunzioni per l'applicazione dell'analisi discriminante. Le variabili, condizionatamente alla classe, non risultano avere una matrice di varianze e covarianze comune (figura 6). Per quanto riguarda l'analisi della normalità, da un punto di vista grafico alcune sembrano avere una distribuzione approssimabile ad una normale (figura 7), tuttavia attraverso il test di Shapiro-Wilk per la normalità (figura 8), nessuna variabile risulta avere una distribuzione normale.

Figura 6: Ellissi condizionate alla variabile target

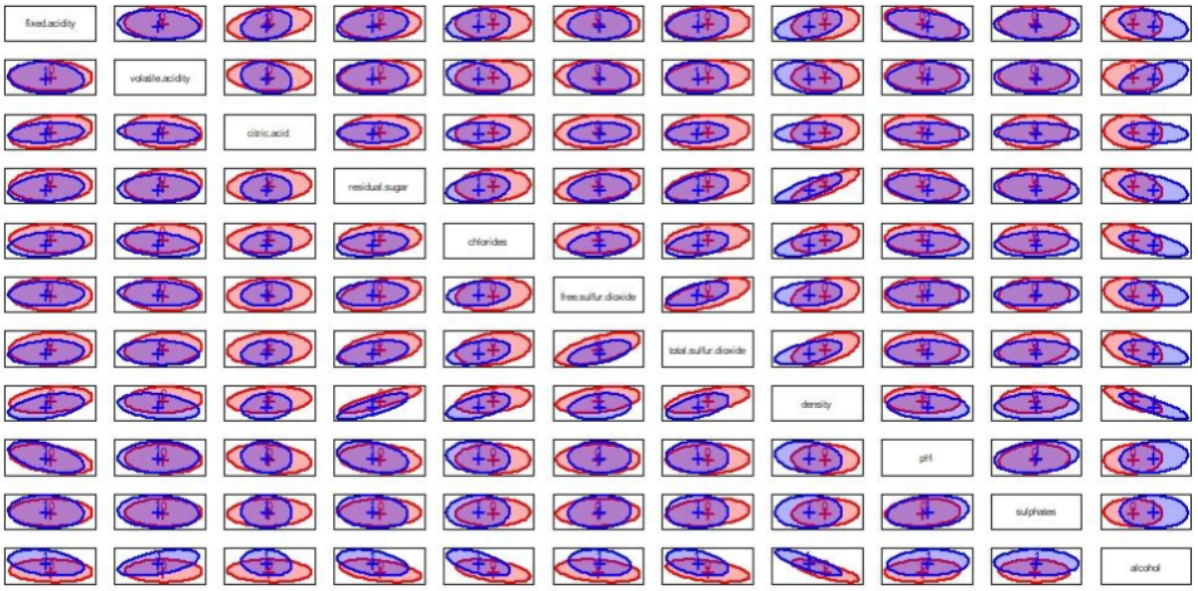
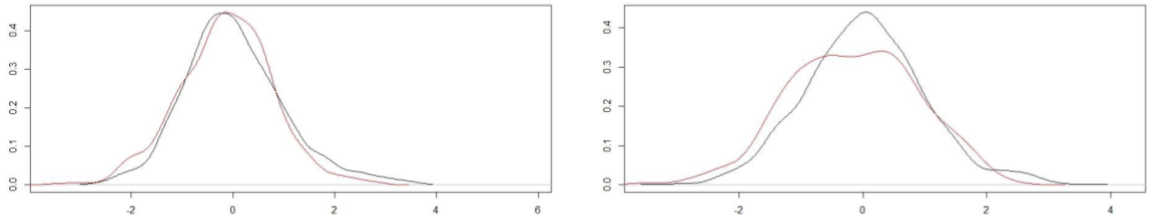


Figura 7: Rappresentazione grafica distribuzione delle variabili condizionata alla classe



(a) Distribuzione di Fixed Acidity

(b) Distribuzione di Volatile Acidity

Figura 8: Risultati del test di Shapiro per la verifica della normalità

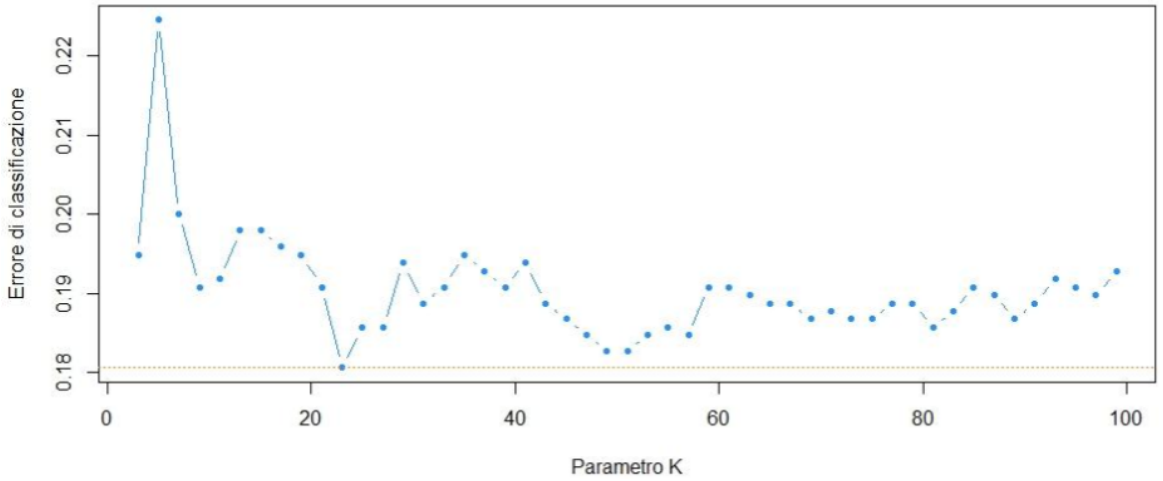
	Excellent	Poor
fixed.acidity	0.01614	0.00000
volatile.acidity	0.01193	0.00000
citric.acid	0.00000	0.00000
residual.sugar	0.00000	0.00000
chlorides	0.00000	0.00000
free.sulfur.dioxide	0.00000	0.00000
total.sulfur.dioxide	0.00001	0.00005
density	0.00000	0.00000
pH	0.00027	0.00000
sulphates	0.00038	0.00000
alcohol	0.00000	0.00000

Prima di passare alla parte di implementazione dei modelli abbiamo effettuato le stesse trasformazioni operate sul training set anche sul validation set.

K-NN

Il primo metodo che siamo andati a testare è stato il k-nn. Siamo partiti dall'individuazione del valore k ottimale, ovvero quello che presenta l'errore di classificazione nel validation più basso. Il valore che risponde a questo criterio è 23 (figura 9). Applichiamo il metodo sullo small training per verificare che distingua bene le classi all'interno del set di allenamento, abbiamo ottenuto un'accuracy del 84.16%. Andiamo ora a testarlo sul validation set ottenendo un'accuracy del 81.94%.

Figura 9: Rappresentazione dei valori k testati e relativi errori di classificazione



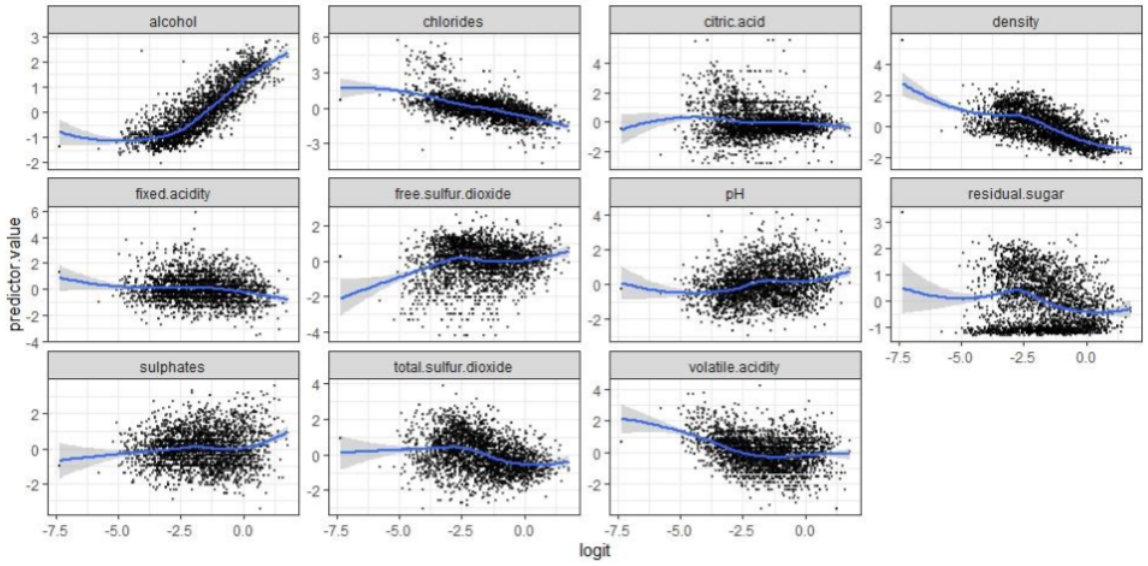
Regressione logistica

Implementiamo su small training il modello di regressione logistica al fine di spiegare la variabili quality mediante l'uso di tutte le covariate o di un sottoinsieme di esse. Procediamo con una procedura di data selection step basata sull'AIC per vedere se si possono selezionare solo un sottoinsieme di covariate. I risultati ottenuti con tutte e tre le possibili direzioni di selezione ci consigliano di mantenere tutte le covariate. Stimiamo nuovamente il modello eliminando i punti influenti, ma, sebbene il valore dei parametri cambi leggermente, il modello finale contiene le stesse variabili.

Come per il K-NN testiamo il modello prima sullo small training stesso ottenendo un accuracy pari 80.86%, poi sul validation con un'accuracy dell'80.71%.

Verifichiamo poi che la relazione tra covariate e i valori previsti sia lineare, ciò però non accade (figura 10).

Figura 10: Rappresentazione relazione singole covariate Logit



Analisi Discriminante

LDA

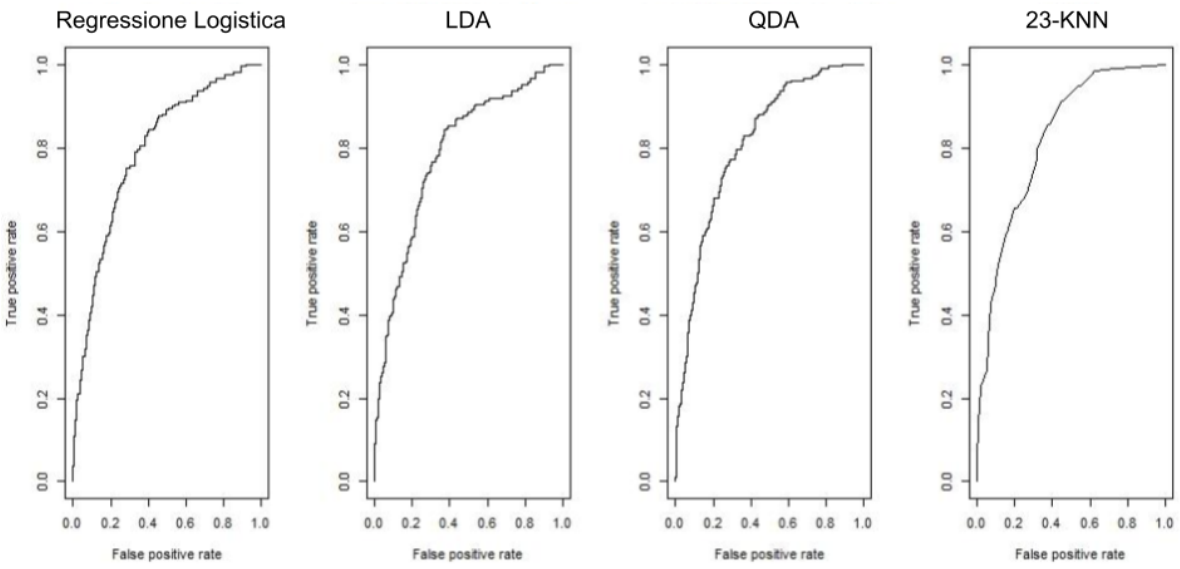
Anche in questo caso applichiamo LDA per provare a spiegare la variabile target attraverso l'uso delle covariate a disposizione. Nonostante le assunzioni non siano rispettate, applichiamo lo stesso questo metodo. Otteniamo un'accuracy del 80.52% sullo small training e del 80.71% su validation set.

QDA

Applichiamo anche il metodo QDA poiché le variabili presentano matrice di varianze e covarianza non comune. Dall'applicazione sullo small training ricaviamo un'accuracy del 79.53% e del 78.27% sul validation set. Nonostante un'accuracy più bassa degli altri metodi, siamo qui in presenza di una sensitivity del 60.48%, valore piuttosto alto se confrontato agli altri metodi.

Curve ROC e AUC

Figura 11: Curve ROC



Valutazione Performance dei modelli su Validation set

Figura 12: Risultato modelli su validation set

	Accuracy	AUC	Balanced Accuracy
KNN	81.94	82.17	67.73
Regressione logistica	80.71	79.23	61.75
LDA	80.71	78.63	62.79
QDA	78.27	81.26	71.80

In figura 12 sono riportati in tabella i risultati dei modelli su validation set. Per valutare i nostri modelli di classificazione, diamo priorità alle metriche AUC (figura 11) e Balanced Accuracy, dato che stiamo trattando un problema di classificazione con target sbilanciato e vogliamo dare lo stesso peso a entrambe le classi. Valutiamo comunque l'accuracy, in quanto siamo interessati a vedere se i modelli superano la soglia minima di accuracy del 78%, che è l'accuracy del modello di classificazione che prevede solo classe 0.

- KNN: ha performato meglio rispetto agli altri modelli in termini di Accuracy e AUC, soffrendo però relativamente dal punto di vista di Balanced Accuracy. Quest'ultimo risultato è dovuto al fatto che in questo caso il modello presenta una sensitivity molto bassa, in quanto le previsioni del modello erano quasi tutte 0.
- Regressione logistica e LDA: i modelli hanno performato male in termini delle metriche più importanti(AUC e Balanced Accuracy), risultato prevedibile dato che le assunzioni previste dai modelli non sono state rispettate.
- QDA: ha performato meglio rispetto agli altri modelli dal punto di vista del Balanced Accuracy, grazie alla sensitivity alta (figura 13). Ha l'accuracy peggiore, risultato anche dal fatto che le assunzioni di normalità delle variabili condizionate dalla classe non sono state soddisfatte.

Figura 13: Sensitivity e specificity modelli su validation set

	Sensitivity	Specificity
KNN	42.86	92.60
Regressione logistica	28.57	94.94
LDA	31.43	94.16
QDA	60.48	83.12

Da queste osservazioni decidiamo di scartare la regressione logistica e il modello LDA, dato che hanno ottenuto risultati peggiori nelle 2 metriche più importanti. Inoltre da questi risultati possiamo dire che la classificazione performa meglio con linee di decisione non lineari.

Procediamo quindi con il modello KNN e QDA.

Modelli implementati su training e test set

Siamo ora pronti per testare i due metodi scelti su training set e test set.

Applichiamo a questi due set la trasformazione logaritmica come abbiamo adoperato precedentemente su small training set. Segue poi il controllo della distribuzione delle variabili e degli outliers. Individuiamo tre outliers potenzialmente problematici nelle variabili citric acid, residual sugar, total sulfur dioxide, density e li rimuoviamo.

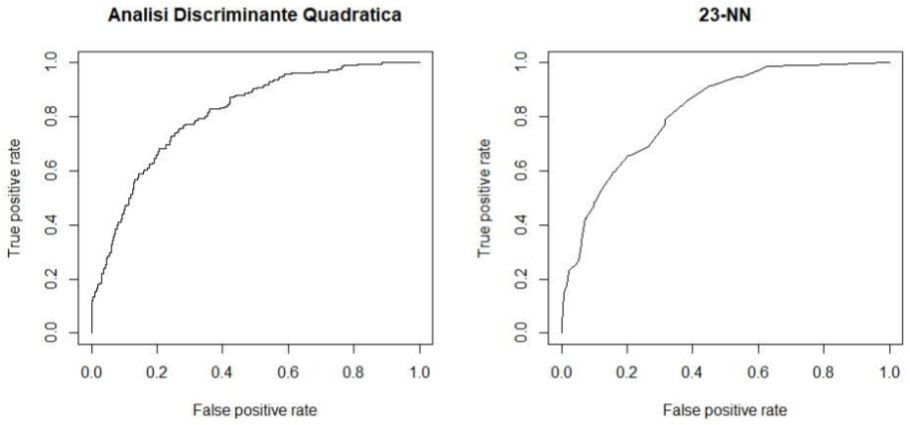
A questo punto standardizziamo i due set per media e varianza del training set..

Implementiamo ora i due metodi scelti: 23-NN e QDA.

Applichiamo il metodo K-NN con k=23, prima sul training per controllare se il modello fitti bene ai dati di allenamento, ottenendo un'accuracy del 83.60%, poi testiamo il metodo sul test set e ricaviamo un'accuracy del 82.35%.

Ricalcoliamo il modello QDA sul training set e lo applichiamo sullo stesso, otteniamo un accuracy del 79.62%. Applicando il modello sul test set riportiamo un'accuracy del 79.39%.

Figura 14: Curve di ROC dei modelli applicati al test set



Valutazione Performance dei modelli su Test set

Figura 15: Risultato modelli finali su test set

	Accuracy	AUC	Balanced Accuracy
KNN	82.35	83.92	66.37
QDA	79.39	83.09	71.82

Come nel validation set, KNN performa meglio in termini di Accuracy e AUC (figura 14) rispetto a QDA (figura 15). Entrambi mostrano risultati deludenti per quanto riguarda Balanced Accuracy, con QDA che performa leggermente meglio (figura 16).

Figura 16: Sensitivity e Specificity modelli finali su test set

	Sensitivity	Specificity
KNN	38.20	94.53
QDA	58.49	85.16

Downsampling

Siccome le classi nel dataset soffrono di un moderato sbilanciamento (quasi 80-20) e poiché i modelli stimati su dati non bilanciati non ci soddisfano soprattutto per ciò che riguarda la sensitivity, e quindi la balanced accuracy, abbiamo deciso di operare una forma di sottocampionamento per equiparare le classi all'interno del training set e allenare quindi i modelli scelti (KNN e QDA) su set di dati bilanciati.

Abbiamo costruito lo small training bilanciato a partire dallo small training originale già con le trasformazioni logaritmiche applicate, sottocampionando le osservazione con classe 0, e a partire da questo set abbiamo operato la standardizzazione su small training e validation.

Una volta ottenuto il set bilanciato, abbiamo allenato nuovamente il modello knn sullo small training e abbiamo identificato il k-ottimale, che è risultato pari a 3 (valore basso, quindi linee di decisione piuttosto complesse, come ci saremmo potuti aspettare). Abbiamo fittato anche il modello qda e ottenuto quindi il training e il validation error per entrambi i modelli allenati su set di dati bilanciati.

A questo punto abbiamo ricreato il training totale accertandoci che fosse bilanciato e contenesse le osservazioni presenti nello small training bilanciato e nel validation. Anche in questo caso i dati avevano già subito le trasformazioni logaritmiche ed è stato sufficiente standardizzare training e test set sulla base delle osservazioni del training.

Abbiamo allora allenato i modelli 3NN e QDA sul training set complessivo e abbiamo operato l'opera di classificazione sul test set, in modo da poter quindi valutare le performance dei nostri modelli su set bilanciati.

Qui in seguito (Figura 17) troviamo lo schema riassuntivo delle metriche di valutazione dei 2 modelli che abbiamo deciso di considerare allenati differentemente su training set bilanciato o meno:

Figura 17: Risultato modelli finali su test set

	Accuracy	AUC	Balanced Accuracy
KNN	82.35	83.92	66.37
$KNN_{Balanced}$	72.65	82.93	74.87
QDA	79.39	83.09	71.82
$QDA_{Balanced}$	69.49	82.63	74.39

Balanced - modello allenato su training set con target bilanciato

Valutazione Performance dei modelli su Test set

Scartiamo subito il modello $QDA_{Balanced}$ dato che performa peggio di $KNN_{Balanced}$ in tutte le metriche di classificazione (figura 17).

Figura 18: Sensitivity e specificity modelli finali su test set

	Sensitivity	Specificity
KNN	38.20	94.53
$KNN_{Balanced}$	78.77	70.96
QDA	58.49	85.16
$QDA_{Balanced}$	83.02	65.76

Balanced - modello allenato su training set con target bilanciato

Come possiamo notare in figura 18, decidiamo di scartare KNN perché ha una sensitivity troppo bassa, e quindi classifica male i prodotti eccellenti. Tra i 2 modelli rimasti scegliamo il modello KNN allenato sul training set bilanciato, dato che presenta un miglior equilibrio tra sensitivity e specificity, e quindi una miglior Balanced Accuracy, a discapito di pochi punti percentuali in AUC.

4 Discussioni

L'analisi che abbiamo svolto ci ha permesso di identificare vari modelli atti alla classificazione delle preferenze di gusto relative al vino bianco, basandoci su dati analitici che sono semplicemente accessibili durante la fase di certificazione di qualsiasi vino. Dopo aver allenato vari modelli su training set sbilanciati, e quindi con una classe maggioritaria, abbiamo deciso di provare ad operare un sottocampionamento per allenare i modelli risultati più corretti nella prima fase su training set bilanciati. I risultati ottenuti sono stati promettenti e grazie all'utilizzo di varie metriche con peculiarità differenti siamo giunti alla conclusione che il modello KNN allenato su set bilanciato è il migliore. Questo infatti è l'unico che, pur mantenendo buoni livelli di AUC, riesce a classificare in modo sufficientemente corretto entrambe le classi di interesse.

Un possibile sviluppo futuro di questa analisi potrebbe essere la valutazione del modello KNN semplice, che risulta avere ottimi livelli di accuracy e AUC, ma pessime performance nella classificazione della classe eccellente, su soglie differenti in modo da correggere quest'ultima problematica.

5 Bibliografia

1) Paulo Cortez, Antònio Cerdeira, Fernando Almeida, Telmo Matos e Josè Reis. Modeling wine preferences by data mining from physicochemical properties. (http://www3.dsi.uminho.pt/pcortez/wine5.pdf)
2) https://www.wineilvino.it/di-vino-e-daltre-facezie/lacidita-nel-vino-approfondimenti/#:~:text=L'acidit%C3%A0%20fissa%20%C3%A8%20costituita,appartenenti%20alla%20famiglia%20delle%20acetiche.
3) https://www.randoxfood.com/wine-analysis/citric-acid/
4) https://glossario.wein.plus/zucchero-residuo#:~:text=Di%20norma%2C%20un%20vino%20contiene,a%20450%20g%2Fl).
5) https://online.scuola.zanichelli.it/industriagroalimentare2ed-files/laboratorio/p86_Adulterazioni_sofisticazioni.pdf
6) https://www.rivistadiagraria.org/articoli/anno-2015/i-solfiti-in-enologia/#:~:text=Il%20diossido%20di%20zolfo%20
7) https://www.ravazzi.it/I/wineblog/i-solfiti-nel-vinocosa-sono-davvero-e-a-cosa-servono-1531#:~:text=solfiti%20nel%20vino-,Cosa%20sono%20i%20solfiti%3F,azione%20disinfettante%2C%20antiossidante%20e%20stabilizzante.
8) https://www.aeb-group.com/it/determinazione-del-ph-del-vino#:~:text=I%20valori%20di%20pH%20del,8%20per%20i%20vini%20rossi.