Classificazione dei vini basata su algoritmi di Machine Learning

Davide Locci AA 2021-2022
Prof.Giuliano Armano

Obiettivi

L'analisi si pone come obiettivo quello di risolvere un problema di classificazione immedesimandosi in un produttore di vino, al fine di prevedere, conoscendo le specifiche di un vino al momento della produzione, se questo verrà valutato dagli esperti in maniera sufficiente o insufficiente, e quindi indirettamente capire come verrà accolto dal mercato e dal consumatore finale, prevedendone le vendite.

Dataset

Il dataset su cui si basa l'analisi è disponibile al seguente link:
https://www.kaggle.com/datasets/ruthgn/wine-quality-data-set-red-white-wine?resource=download
Si tratta di un dataset contenente più di 6mila osservazioni, ognuna riguardante uno specifico vino portoghese, e 13 feature.
Ciascun vino è contraddistinto da un voto in una scala da 1 a 10 che ne rispecchia la bontà, attribuito da un gruppo di esperti.

Analisi

Panoramica generale sul dataset e data cleaning

Come prima cosa si è proceduto ad effettuare una panoramica generale del dataset e di quelle che sono le feature che riguardano ciascuna osservazione e quindi ciascun vino.
Il dataset, come accennato, è composto da 6497 osservazioni e 13 colonne.

Le colonne sono:

-Fixed Acidity: l'acidità fissa (g/dm3)
-Volatile acidity: l'acidità volatile (g/dm3)
-Citric acid: l'acido citrico (g/dm3)
-Residual sugar: gli zuccheri residui (g/dm3)
-Chlorides: i cloruri (g/dm3)
-Free sulfur dioxide: l'anidride solforosa libera (mg/dm3)
-Total sulfur dioxide: l'anidride solforosa totale (mg/dm3)
-Density: la densità del vino(g/cm3)
-pH: il pH
-Sulphates: i solfati (g/dm3)
-Alcohol: la gradazione alcolica del vino (%)

La funzione *describe()* ha permesso, per ognuna delle feature, di ottenerne un riassunto contenente il valore medio, massimo e minimo:
In questo modo è stato possibile confrontare questi valori con quelli disposti dalla legislatura portoghese in tema di vini, per individuare eventuali possibili valori errati. Confrontando i valori minimi e massimi delle variabili con i limiti di legge è risultato che, in alcuni casi, le variabili acidità volatile ed acido citrico presentano valori che vanno oltre i limiti legislativi. In particolare: l'acidità volatile non può eccedere gli 1,5 g/l mentre la funzione describe() indica come massimo 1,58 g/l ; l'acido citrico invece non può essere superiore ad 1 g/l mentre la funzione describe() indica come massimo 1,66 g/l.

Successivamente sono state svolte le operazioni casting e data cleaning.
Per quanto riguarda il data cleaning si è provveduto a:
- individuare eventuali valori nulli (non presenti)
-rimuovere le righe duplicate
-individuare i potenziali valori errati sulla base dei limiti di legge descritti sopra: si è appurato che l' osservazione che supera il limite di legge per quanto riguarda l'acidità volatile è soltanto una, mentre le osservazioni che superano il limite di legge per quanto riguarda l'acido citrico sono due. Di conseguenza, essendo un numero non rilevante rispetto al totale delle osservazioni, si è deciso di eliminarle.

Analisi esplorativa

La seconda fase del lavoro riguarda l'analisi esplorativa. Nello specifico sono stati individuati 4 obiettivi:

1)Categorizzare il target;
2)Disegnare la matrice di correlazione delle feature per individuare eventuali collinearità;
3)Rappresentare graficamente le relazioni tra variabile target e le altre feature, per avere una prima idea di quali sono le più influenti;
4)Individuare ed eliminare gli outlier

1)Categorizzazione del target

Il target, la colonna quality, si presenta come una variabile quantitativa discreta.
È infatti la colonna che rappresenta il punteggio attribuito dagli esperti, che può assumere soltanto 10 valori, in una scala di punteggio da 1 a 10.
Per svolgere il problema di classificazione, si è deciso di categorizzare la variabile in una variabile qualitativa sia con 2 categorie sia con 3 categorie.
Per quanto riguarda la categorizzazione in 2 categorie, tutti i valori compresi tra 0 e 5 rientrano nella categoria 'bad' mentre tutti i valori compresi tra 6 e 10 rientrano nella categoria 'good'.
Per quanto riguarda invece la categorizzazione in 3 categorie, i valori compresi tra 0 e 4 rientrano nella categoria 'bad' ; i valori compresi tra 5 e 6 rientrano nella categoria 'medium'; i valori compresi tra 7 e 10 rientrano nella categoria 'top'.

2)Matrice di correlazione

Dopo aver categorizzato il target, si è provveduto a disegnare la matrice di correlazione.
La matrice di correlazione misura e rappresenta l'indice di correlazione per ciascuna coppia di variabili del dataset. L'indice di correlazione è una misura specifica usata nell'analisi della correlazione per quantificare la forza della relazione lineare tra due variabili.
Quest'ultimo è compreso in un intervallo che va da -1 a 1:
-un valore di -1 indica la massima correlazione inversa;
-un valore di 1 indica la massima correlazione positiva;
-0 indica assenza di correlazione lineare e quindi assenza di effetto lineare tra le due variabili.

La correlazione può essere misurata soltanto tra due variabili numeriche, per questo prima occorre filtrare tutte le variabili non numeriche.

Se la correlazione tra due variabili (positiva o negativa) è molto vicina a 1 (oppure a -1) significa che c'è collinearità e che in sintesi quelle due variabili forniscono lo stesso contenuto informativo e stanno misurando la stessa cosa, potendo procedere all'eliminazione di una delle due in modo da non inserire poi nel modello un'informazione ridondante.

La matrice di correlazione disegnata è risultata essere questa:

	fixedAcidity	volatileAcidity	citricAcid	residualSugar	chlorides	freeDioxide	totDioxide	density	pH	sulphates	alcohol
fixedAcidity	1.000000	0.215524	0.333857	-0.104390	0.289023	-0.281678	-0.327964	0.478430	-0.271334	0.304970	-0.102810
volatileAcidity	0.215524	1.000000	-0.385664	-0.163769	0.366509	-0.348857	-0.400410	0.309733	0.245599	0.230168	-0.065752
citricAcid	0.333857	-0.385664	1.000000	0.148952	0.059122	0.130118	0.192969	0.099382	-0.347290	0.060160	-0.010398
residualSugar	-0.104390	-0.163769	0.148952	1.000000	-0.123032	0.398754	0.488006	0.520987	-0.234424	-0.174861	-0.305210
chlorides	0.289023	0.366509	0.059122	-0.123032	1.000000	-0.186033	-0.269204	0.371819	0.025049	0.405636	-0.269558
freeDioxide	-0.281678	-0.348857	0.130118	0.398754	-0.186033	1.000000	0.720442	0.006556	-0.141139	-0.198419	-0.170630
totDioxide	-0.327964	-0.400410	0.192969	0.488006	-0.269204	0.720442	1.000000	0.007621	-0.222010	-0.276118	-0.250652
density	0.478430	0.309733	0.099382	0.520987	0.371819	0.006556	0.007621	1.000000	0.033977	0.282649	-0.667580
pH	-0.271334	0.245599	-0.347290	-0.234424	0.025049	-0.141139	-0.222010	0.033977	1.000000	0.168387	0.097754
sulphates	0.304970	0.230168	0.060160	-0.174861	0.405636	-0.198419	-0.276118	0.282649	0.168387	1.000000	-0.016945
alcohol	-0.102810	-0.065752	-0.010398	-0.305210	-0.269558	-0.170630	-0.250652	-0.667580	0.097754	-0.016945	1.000000

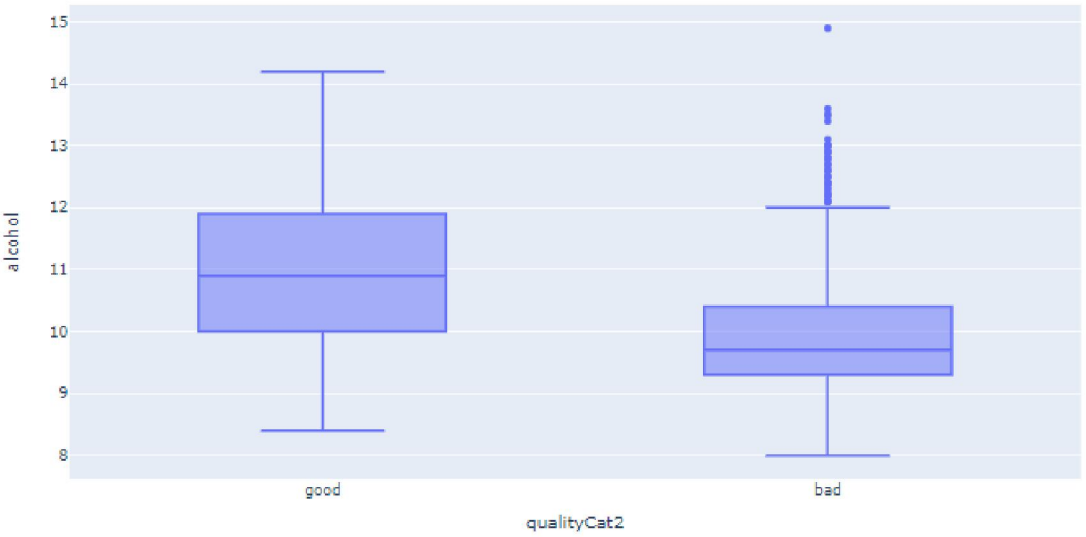
Si può notare come la correlazione più elevata (messa in risalto anche dal colore acceso più vicino al rosso) sia quella tra le feature freeDioxide e totDioxide, pari a 0.72. Tuttavia, essendo alta ma non altissima, si è deciso di non procedere all'eliminazione di una delle due.

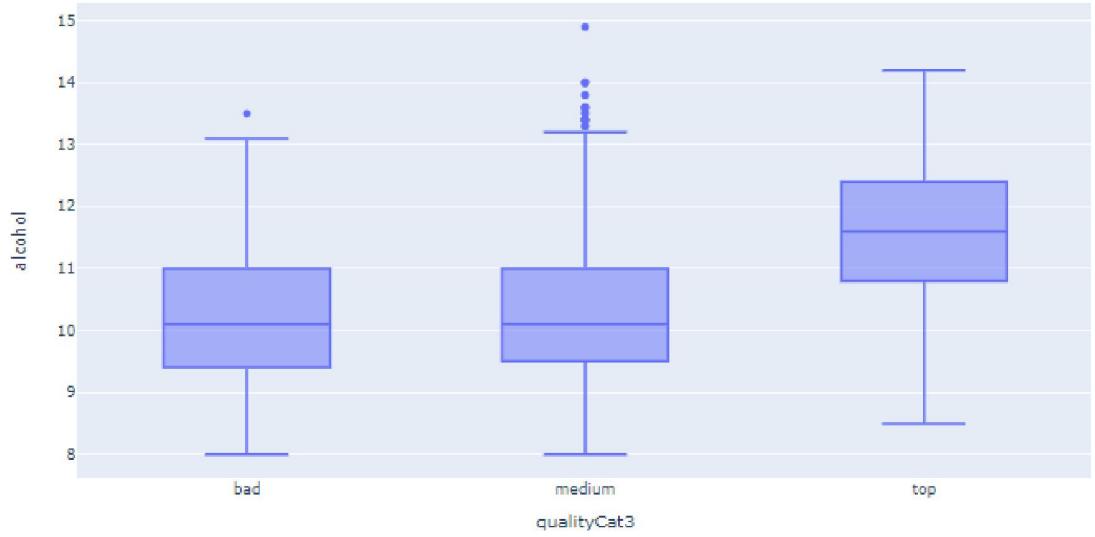
3)Rappresentazioni grafiche delle relazioni tra variabile target e feature.

Il terzo obiettivo dell'analisi esplorativa è quello di visualizzare in maniera grafica le relazioni del target con ciascuna delle feature del dataset.
Lo scopo è quello di ottenere una prima idea su quelle che sono le feature maggiormente in grado di poter influenzare la categoria di appartenenza di ciascuna osservazione.
Per ciascuna feature, è stato disegnato sia il grafico che mette in relazione la feature con il target suddiviso in 2 categorie, sia il grafico che mette in relazione la feature con il target suddiviso in 3 categorie.

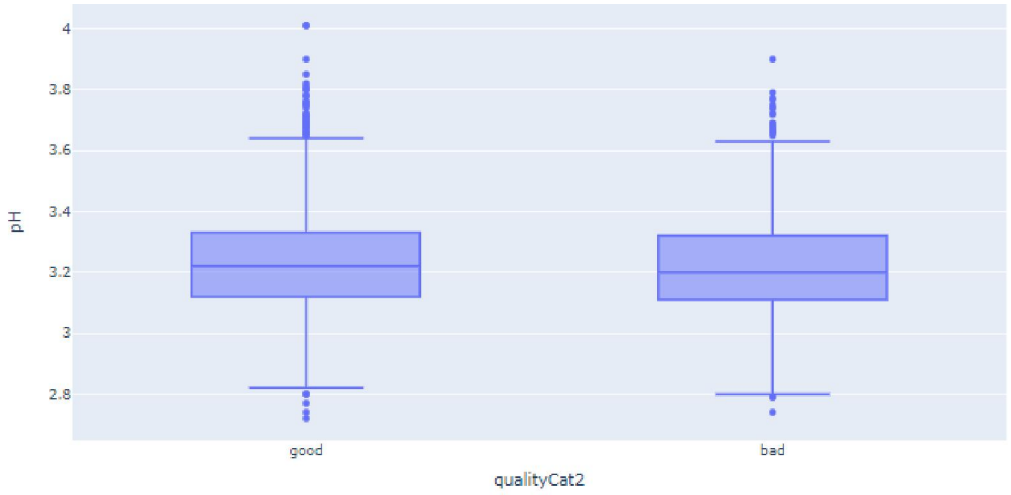
Di seguito, a titolo di esempio, i grafici che rappresentano la relazione tra target e 2 delle 12 feature del dataset: rispettivamente alcohol e pH.

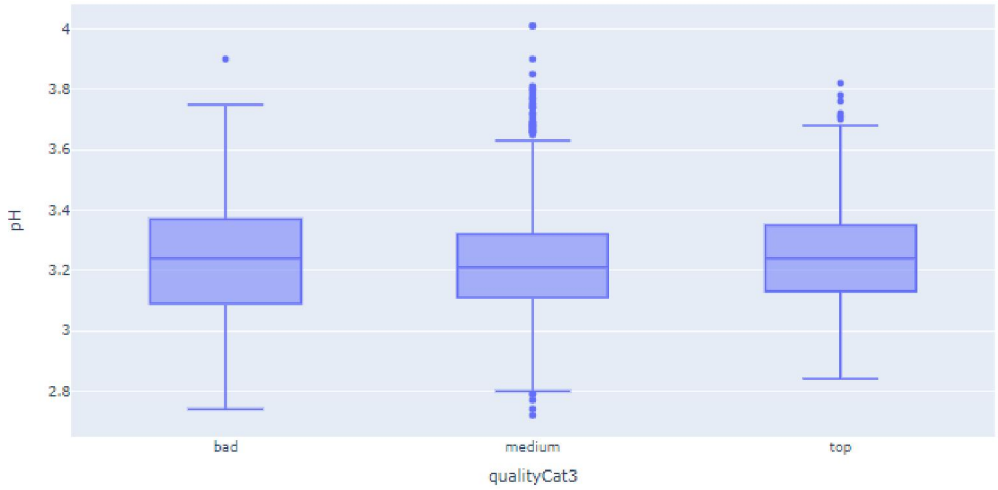
-Target & Alcohol





-Target & pH





4)Outlier detection

L'ultimo passo dell'analisi esplorativa riguarda l'individuazione e la gestione di eventuali outlier, ovvero valori anomali presenti nelle feature.
Per individuare e gestire gli outlier è stata creata una funzione *find_outliers()*.
La funzione individua, per ciascuna variabile, gli outlier basandosi sull' Interquartile Range Rule, ovvero sul calcolo, per ogni variabile, dell'Interquartile Range (IQR).
L'IQR di una variabile è la differenza tra il terzo quartile e il primo quartile di quella stessa variabile:

IQR = Q3-Q1

Una volta calcolato l'IQR questo va moltiplicato per 1,5 ed il prodotto ottenuto va:
-sommato al terzo quartile Q3 ottenendo così la soglia superiore, per la quale ogni valore che supera questa soglia viene considerato un outlier;
-sottratto al primo quartile Q1 ottenendo così la soglia inferiore, per la quale ogni valore

inferiore ad essa viene considerato un outlier.

Per ogni colonna poi, si è deciso di eliminare “brutalmente” gli outlier soltanto se non eccedenti il 3% de totale delle osservazioni del dataset.
Ne è risultato che:
-le colonne citricAcid, residualSugar, freeDioxide, totDioxide, density, pH, sulphates, alcohol, sono state interamente ripulite dagli outlier;
-le colonne fixedAcidity, volatileAcidity, chlorides, hanno mantenuto gli outlier, che sono stati gestiti nella parte successiva di data transformation.

Data standardization

Prima di dare in pasto i dati agli algoritmi di Machine Learning, questi sono stati ulteriormente processati.

a)La prima operazione eseguita è stata l’encoding del target che ha permesso di passare, per la variabile target, da modalità categoriche a modalità numeriche (es. good→1.0)

b)Successivamente è stato svolto l’encoding dell’unica variabile categorica presente nel dataset, la variabile type, tramite l’utilizzo del OneHot Encoder, il più adatto per variabili categoriche non ordinate.

c)Infine sono state gestite le variabili quantitative, scalandole.
In questo caso si è optato per l’utilizzo di due scaler diversi:

-Il MinMax Scaler per le variabili i cui outlier sono stati eliminati dalla funzione *find_outliers()*
Il MinMax Scaler è lo scaler più utilizzato, ed esegue lo scaling in un range compreso tra un valore minimo ed un valore massimo, che di default sono 0 e 1.

-Il Robust Scaler per le variabili i cui outlier non sono stati gestiti dalla funzione *find_outliers()*
Il RobustScaler è lo scaler robusto, che gestisce la presenza degli outlier nella variabile che si sta scalando, e si utilizza appunto quando si hanno outlier nei nostri dati.

Analisi mediante algoritmi di classificazione

L'ultima parte dell'analisi prevede l'applicazione di algoritmi di Machine Learning per la classificazione dei vini.
Prima di procedere alla vera e propria applicazione degli algoritmi si sono suddivisi entrambi i dataset (quello con 2 e quello con 3 categorie per il target) in training set e test set in maniera randomica, attribuendo al training set il 70% delle osservazioni ed al test set il restante 30%. In questo modo è stato possibile utilizzare le osservazioni appartenenti al training set per addestrare gli algoritmi, e le osservazioni appartenenti al test set per testarne le performance.

Si è deciso di utilizzare tre diversi algoritmi di classificazione: il Naive Bayes classifier, la Regressione logistica, ed il Random forest classifier.
Ciascuno dei tre algoritmi è stato applicato sia per la classificazione binaria (applicata al dataset in cui la variabile target è suddivisa in due sole categorie) sia per la classificazione multiclasse (applicata al dateset in cui la variabile target è suddivisa in tre categorie).
Relativamente alla classificazione multiclasse, nel caso del Naive Bayes e della Regressione logistica, è stato opportuno settare dei parametri appositi che ne consentissero l'esecuzione.

Comparazione dei risultati ottenuti dagli algoritmi

Una volta applicati tutti e tre gli algoritmi, è stata realizzata una tabella che consentisse di comparare le performance ottenute da ciascuno di essi. La metrica calcolata per la misurazione delle performance è stata l'accuracy, ovvero il rapporto tra osservazioni predette in maniera corretta dal classificatore ed il numero totale di osservazioni da predire.

Di seguito le tabelle ottenute, rispettivamente una per la classificazione binaria ed una per la classificazione multiclasse:

Algorithm	Performance(accuracy)
Naive Bayes	0.690821256038647 3
Logistic Regression	0.759144237405107
Random Forest	0.7750172532781229

Algorithm	Performance(accuracy)
Naive Bayes	0.7556935817805382
Logistic Regression	0.7853692201518289
Random Forest	0.7943409247757074

Dalle tabelle è possibile notare come sia nel caso della classificazione binaria, sia nel caso della classificazione multiclasse, i risultati migliori sono quelli ottenuti dal Random forest classifier, che per la classificazione multiclasse sfiora l'80% di accuracy. A seguire, rispettivamente la Regressione logistica ed il Naive Bayes classifier.