Explainable AI

Wine Quality

Hanan Francesca Anderloni (mat. 889079)
Gaia Fratesi (mat. 882225)
Nicolò Melchioretto (mat. 893145)

Introduzione

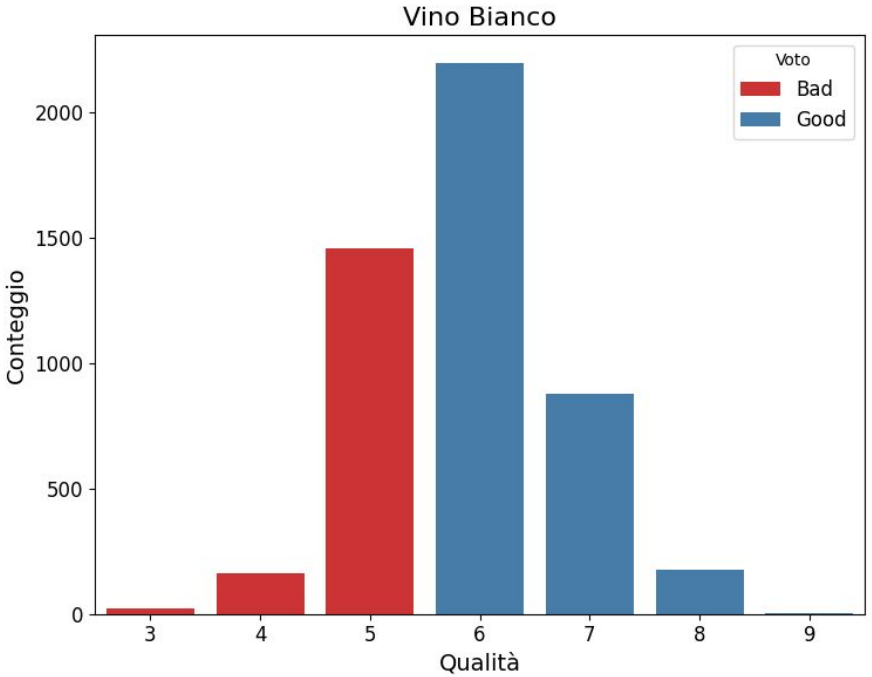
L'obiettivo del progetto consiste nel creare spiegazioni dettagliate e comprensibili riguardo gli output di un modello Random Forest impiegato per un task di classificazione.

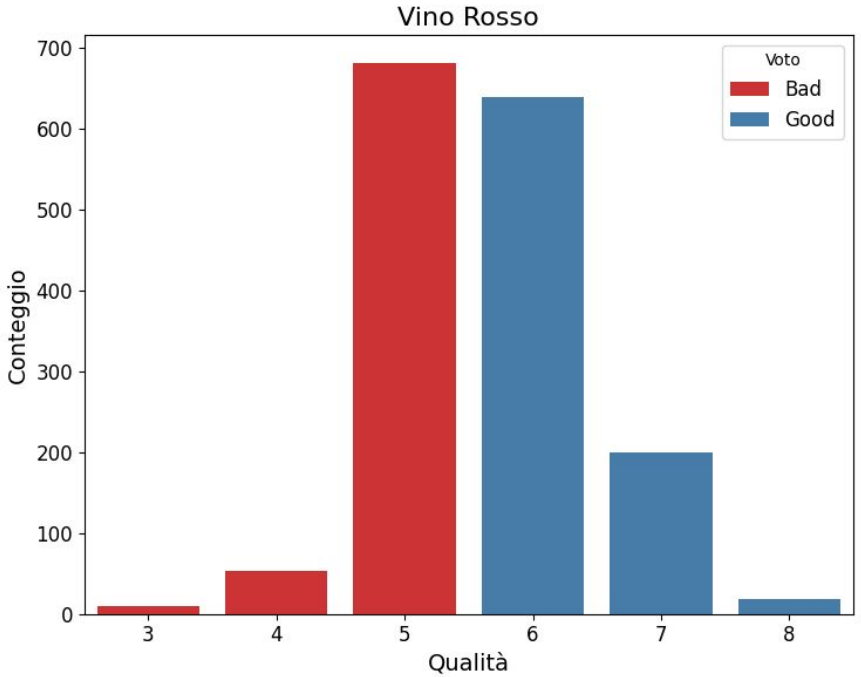
Queste spiegazioni possono essere sia a livello globale, che forniscono una visione d'insieme del contributo delle variabili, e a livello locale, che si concentra su singole istanze.

Vengono impiegati due strumenti: LIME (Local Interpretable Model-agnostic Explanations) e SHAP (SHapley Additive exPlanations).

Exploration & Visualization

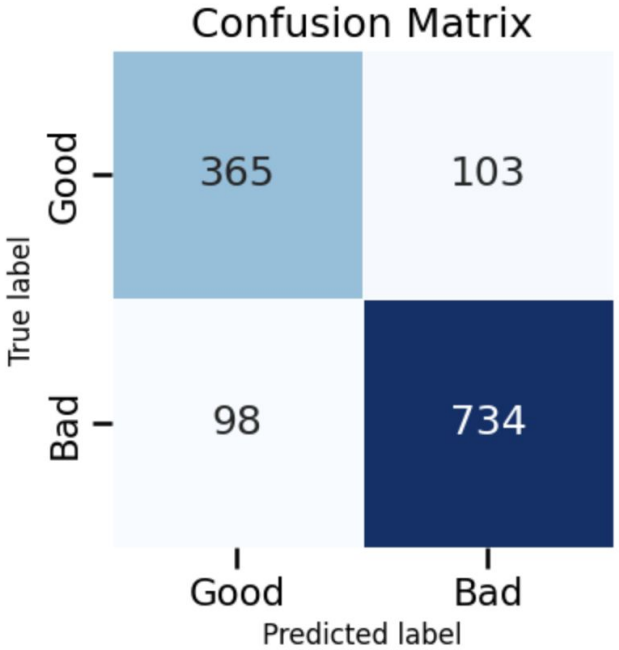
Si riportano le distribuzioni grafiche della *Qualità* del vino raggruppate per tipo di vino (*bianco*, *rosso*). Si mette in evidenza la ricodifica della qualità in *Good* (sufficiente, > 6) e *Bad* (insufficiente, ≤ 5).





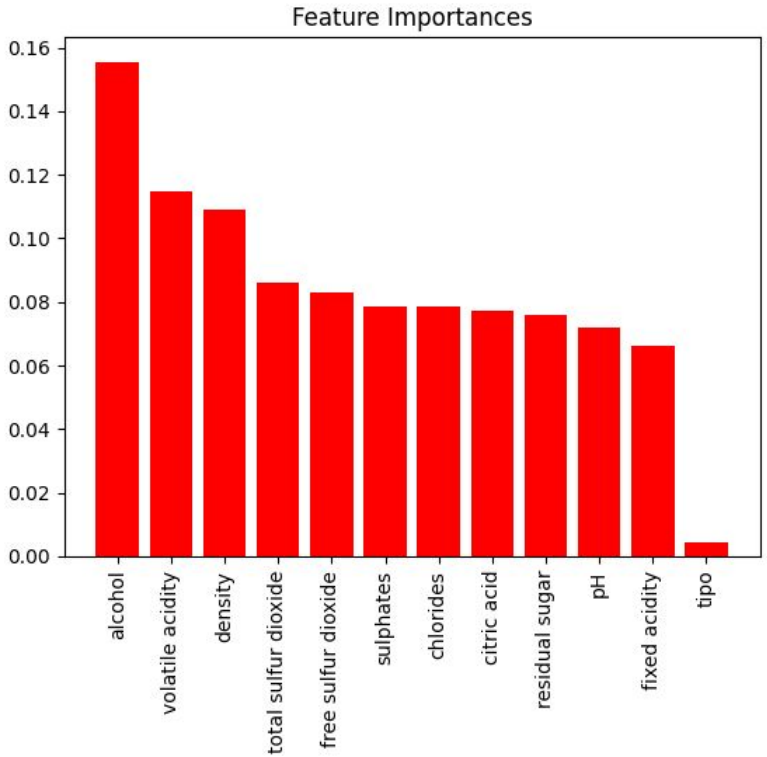
Random Forest

Si riportano le performance classificative del modello tramite confusion matrix e metriche di valutazione.



Accuracy	0.845
Precision	0.876
Recall	0.882
F1 Score	0.879

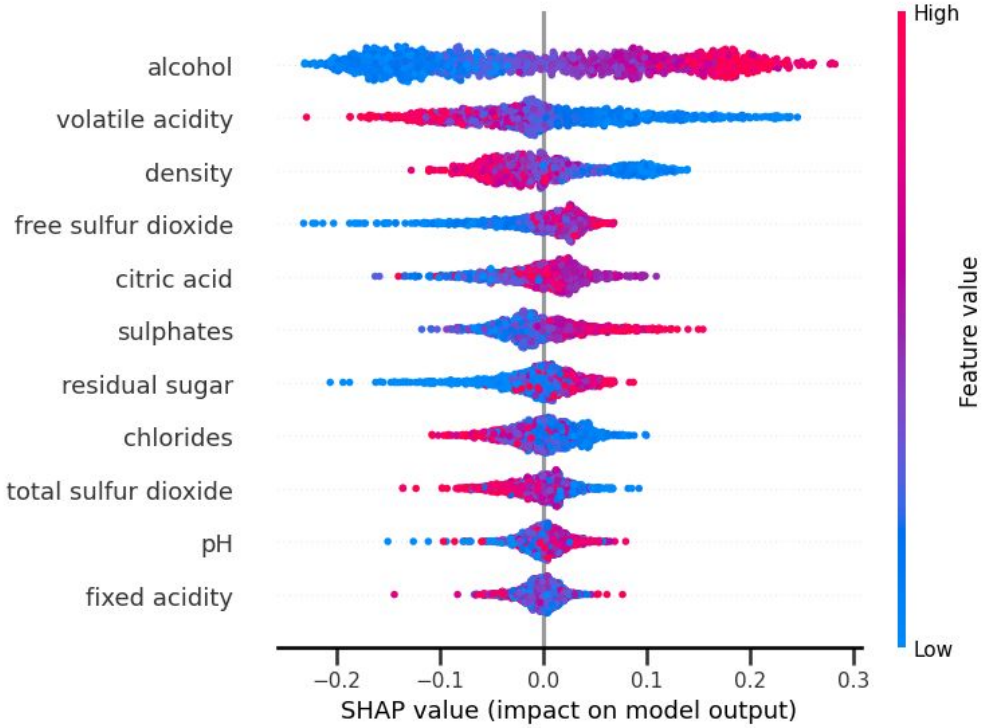
Global Explanation



Un primo tool di spiegazione globale è grafico di *Feature Importance* riferito a tutte le variabile impiegate per allenare il modello.

Si noti che la variabile *tipo* è quasi ininfluente: si sceglie di escluderla subito.

SHAP Summary Plot



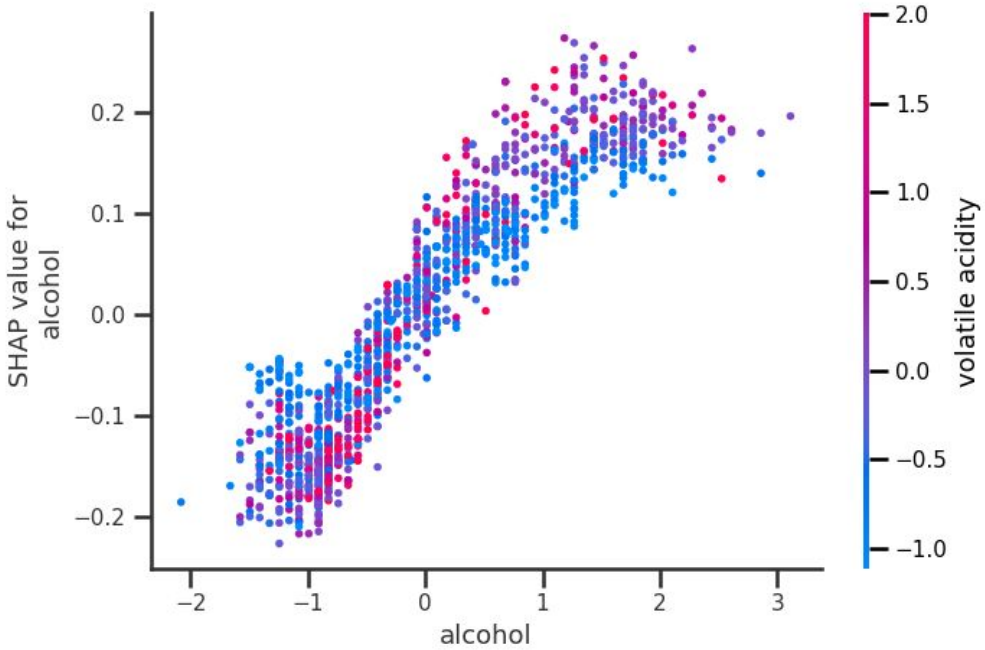
Un vino con alto valore di *alcohol* avrà una probabilità di essere classificato come *Good* più alta di un vino con basso valore di *alcohol*; un valore di *alcohol* intermedio ha scarsa influenza sulla posterior (quindi sulla prediction).

Valori alti di *volatile acidity* corrispondono a *P(vote = Bad)* maggiori.

Il valore di *fixed acidity* ha scarsa influenza sull'output.

Un vino con valori alti di *sulphates* aumentano la *P(vino = Good)*.

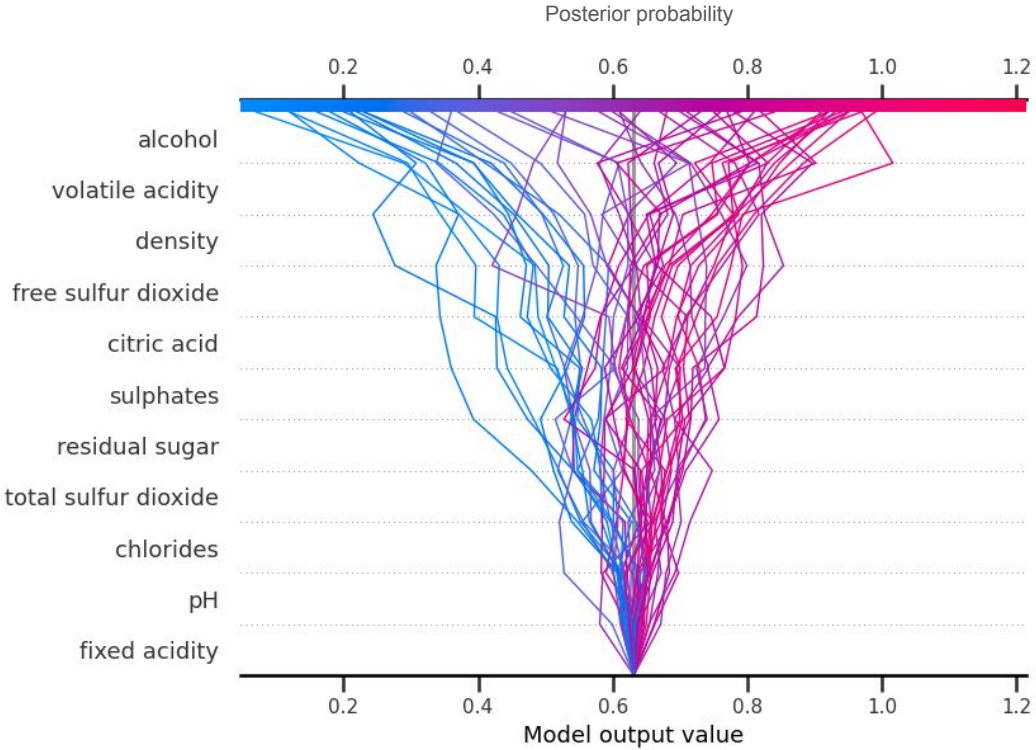
SHAP Dependence Plot



All'aumentare del valore di *alcohol* del vino, il suo SHAP value (quindi, il suo impatto sulla determinazione della prediction del modello) aumenta.

L'interazione di *volatile acidity* non ha invece un impatto sullo SHAP value di *alcohol*.

SHAP Decision Plot



Si noti che la prior probability di *Good* è pari a ≃0.6 (base value di questo Decision Plot).

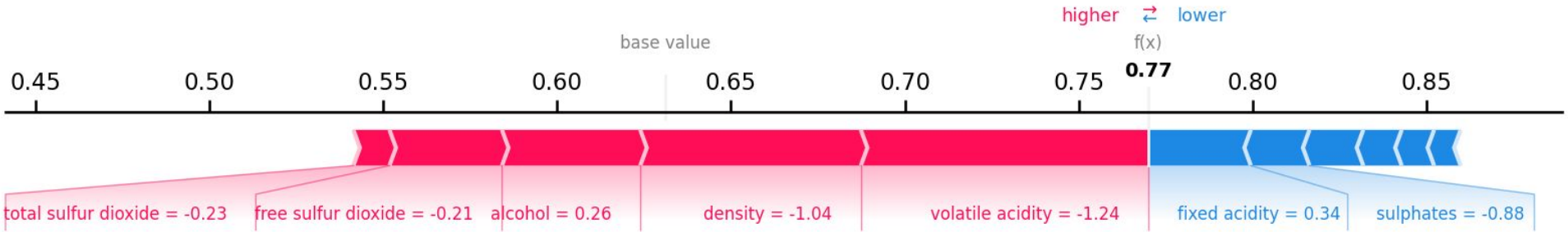
Vengono evidenziate le traiettorie del calcolo delle posterior per singola osservazione determinate dai valori assunti dalle variabili in ciascuna di esse.

Le traiettorie sono abbastanza omogenee con alcuni casi di deviazioni significative.

Local Explanation

SHAP Force Plot

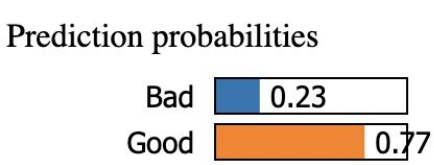
instance : 1

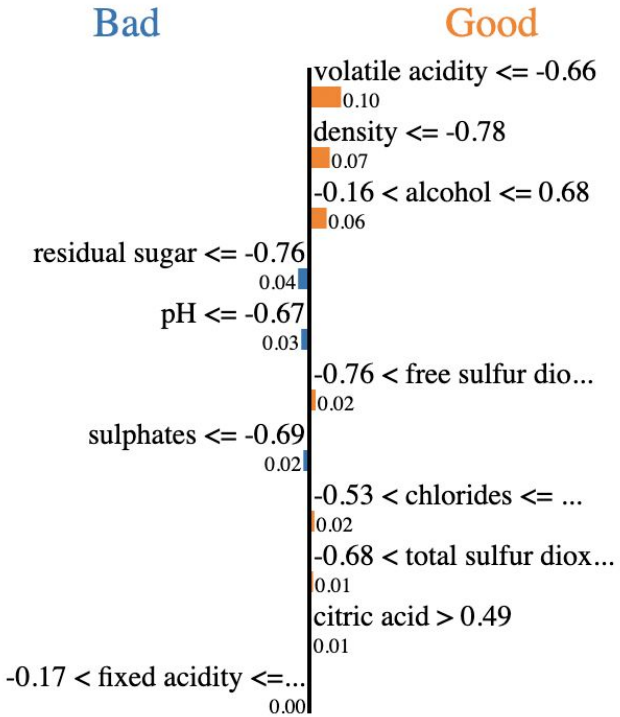


In questo caso si nota che, ad esempio, volatile acidity e fixed acidity spingono in direzioni opposte; tuttavia, l'intensità del contributo della prima è molto superiore rispetto a quello della seconda.

LIME Explanation

```
Instance: 1
True class: 1
```





Feature	Value
volatile acidity	-1.24
density	-1.04
alcohol	0.26
residual sugar	-0.84
pH	-0.93
free sulfur dioxide	-0.21
sulphates	-0.88
chlorides	-0.43
total sulfur dioxide	-0.23
citric acid	2.90
fixed acidity	0.34