

WINE QUALITY PREDICTION

Raviotta Benedetto 816362
Salamone Fabio 816297

Obiettivo

L'obiettivo stabilito per questo progetto è lo sviluppo di due modelli di apprendimento che abbiano lo scopo di classificare la qualità del vino rosso.

I modelli di riferimento scelti:

- **Macchine a vettori di supporto**
- **Algoritmo random forest**



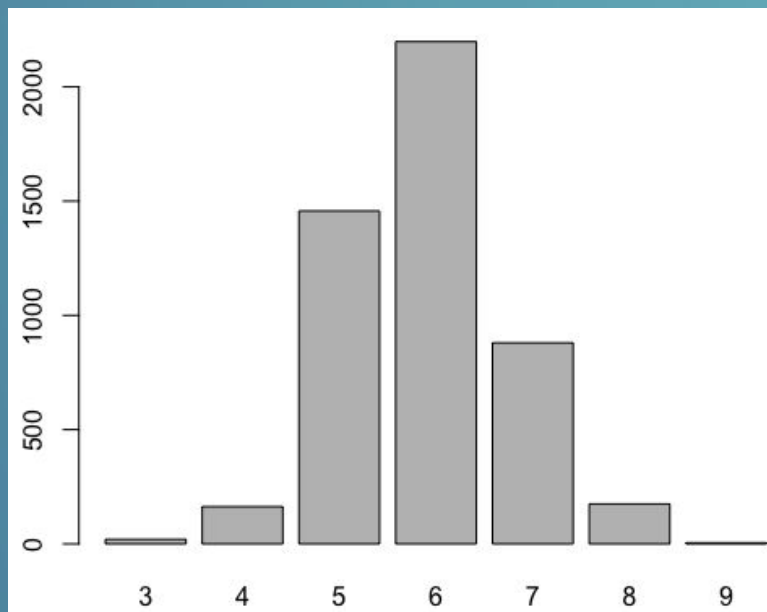
Descrizione dominio

Il dataset Red Wine Quality utilizzato contiene un elenco di 4898 istanze e 12 attributi di tipo reale. Gli attributi sono:

- Acidità fissa
 - Acidità volatile
 - Acido citrico
 - Zuccheri residui
 - Cloruri
 - Anidride solforosa libera
 - Anidride solforosa totale
 - Densità
 - pH
 - Solfati
 - Alcool
- La variabile target è : **qualità**.

Distribuzione delle etichette

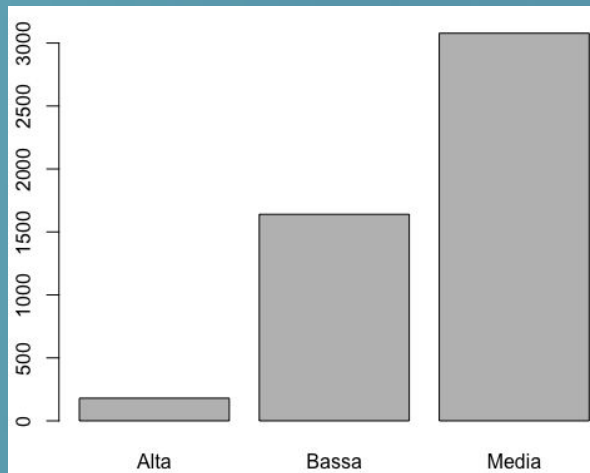
La qualità è rappresentata da un numero intero variabile da 1 a 10 (1=bassa, 10=alta)

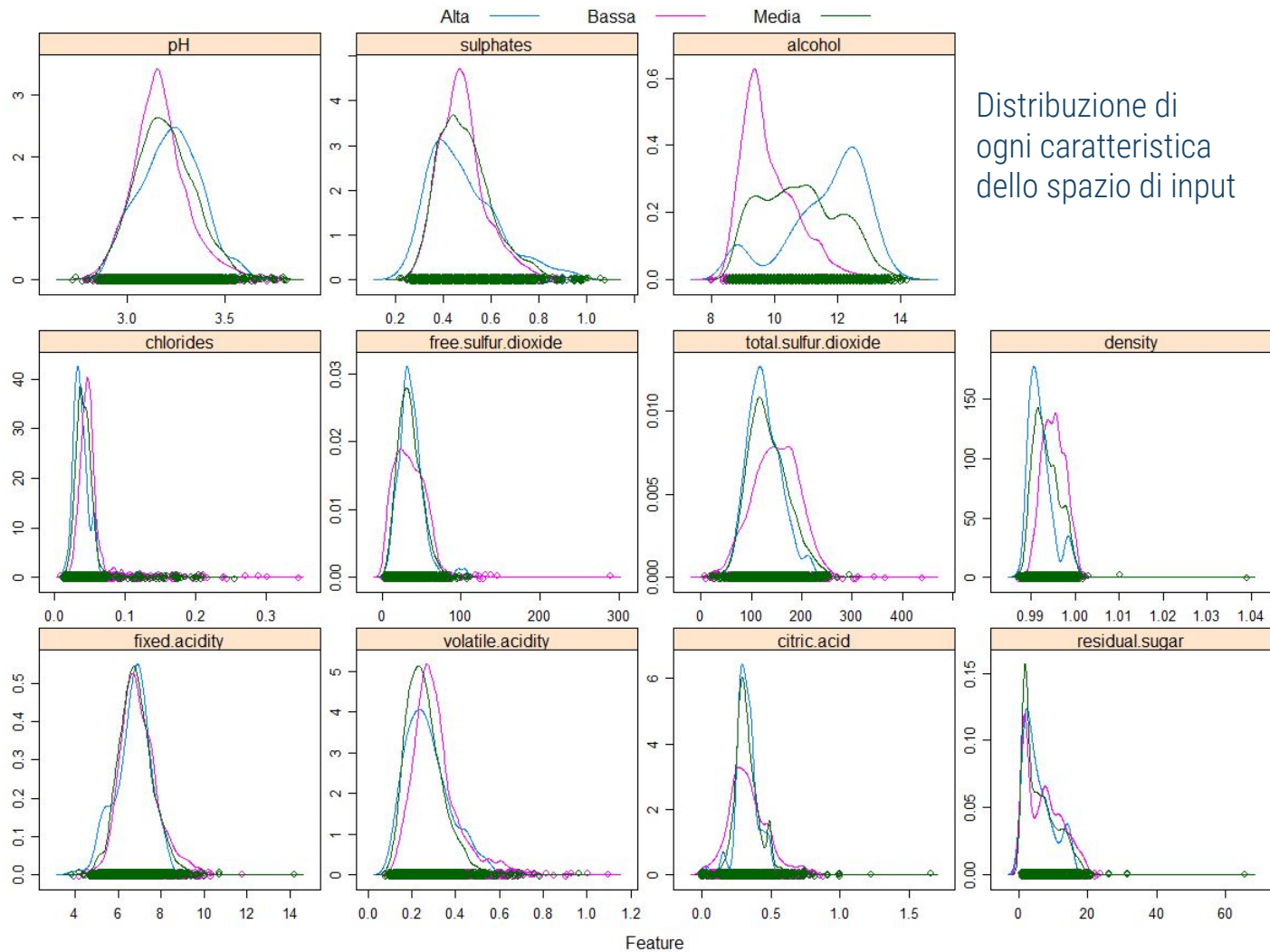


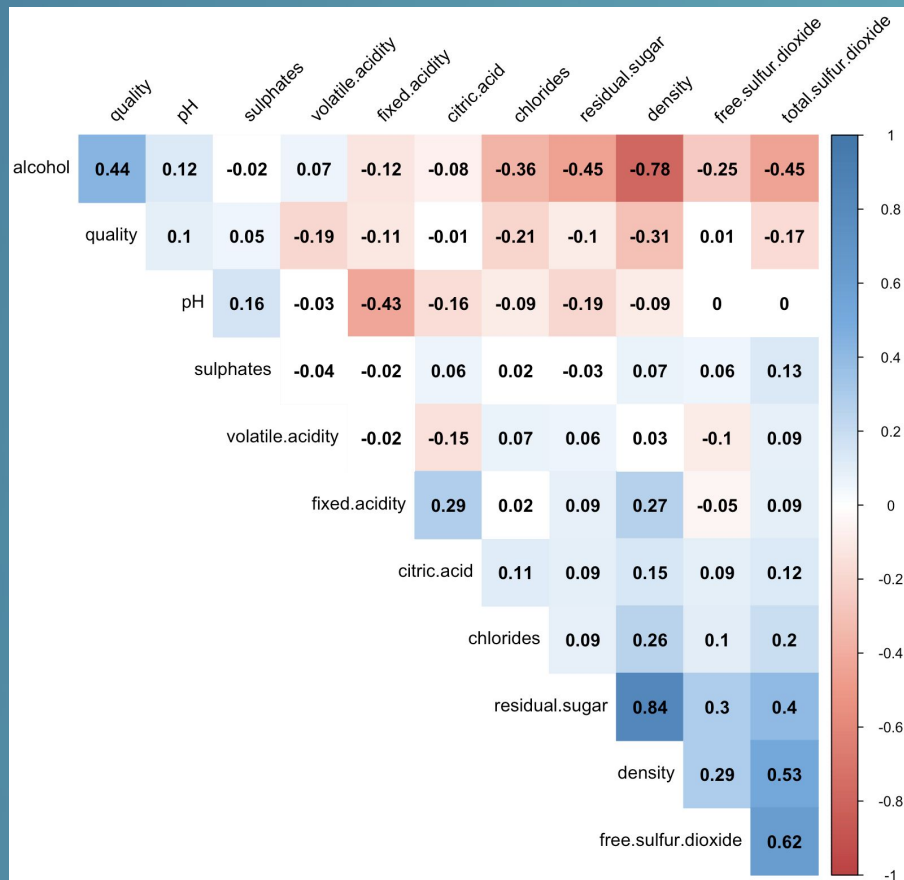
Etichette e codifica

Convertiamo i valori di qualità in 3 etichette:

- Qualità bassa (punteggio < 5)
- Qualità media (punteggio di 6 o 7)
- Qualità alta (punteggio > 7)







Modeling process

Abbiamo deciso di effettuare la classificazione su due diverse varianti del dataset:

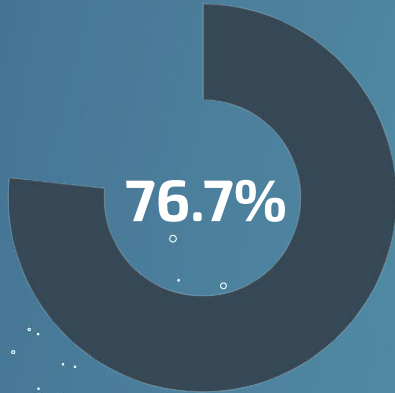
- Qualità suddivisa in alta, media, bassa
- Qualità suddivisa in alta (≥ 7) e bassa (< 7)

In entrambi i casi il dataset è stato diviso casualmente in due sottoinsiemi :
70% delle istanze per il training del modello e 30% delle istanze per il test del modello

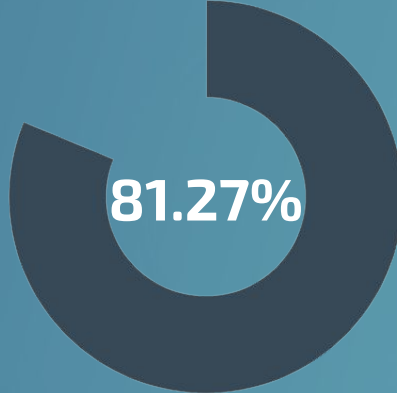
Confronto accuratezza

Dataset 3 etichette
(alta, media, bassa)

SVM

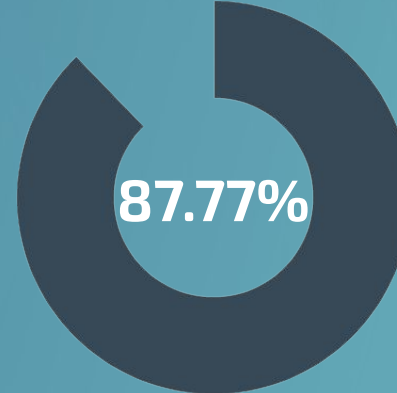


Random Forest

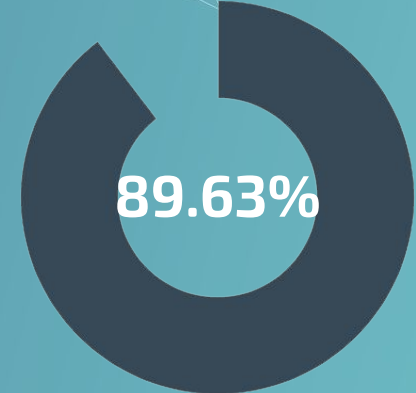


Dataset 2 etichette
(alta, bassa)

SVM



Random Forest



Confronto dei risultati

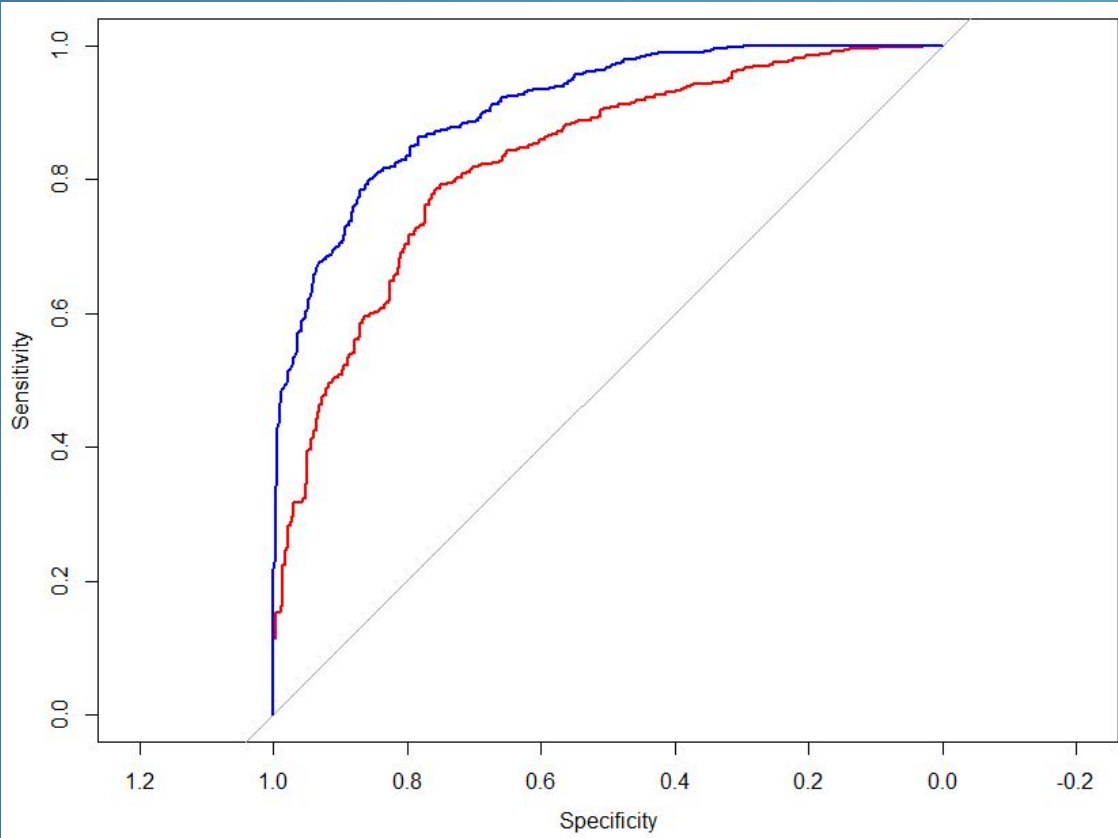
	Precision	Recall	F1
SVM	0.9573	0.8956	0.9254
RF	0.9025	0.9686	0.9344

Classe positiva : Bassa

Classe positiva : Alta

	Precision	Recall	F1
SVM	0.5733	0.77883	0.6603
RF	0.8265	0.6033	0.6975

Confronto curve ROC



SVM

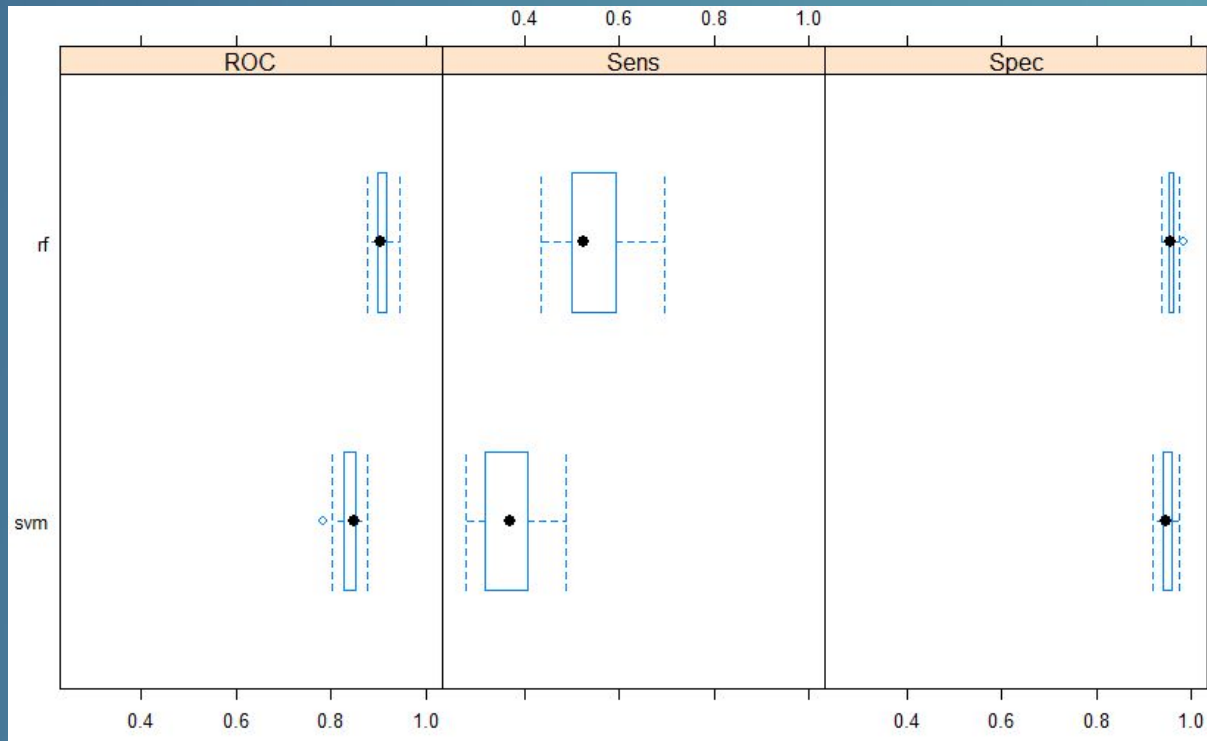
AUC: 0.8279

Random Forest

AUC 0.909

Il valore di AUC 0.8279, indica che c'è una probabilità del 82.79% che il modello sarà in grado di distinguere tra classe positiva e classe negativa.

Confronto tra i due modelli - 1



Notiamo la stessa specificità, quindi non è un elemento discriminante di scelta tra un modello e l'altro. Possiamo notare delle differenze per quanto riguarda la ROC e la sensibilità, ritenendole sufficientemente distinte.


Confronto tra i due modelli - 2

I modelli non presentano tempi computazionali molto differenti, dunque il tempo computazionale non influenza la scelta del modello da utilizzare.

(s)	Everything	FinalModel
SVM	132.47	1.58
RF	185.21	1.72

A decorative network diagram in the top-left corner, consisting of white dots connected by thin white lines, forming a complex web-like structure.

Conclusioni

- Accuratezza predizione dei modelli simile
 - Differenze sulla precisione date dalla classe:
 - Classe “Bassa”: entrambi i modelli presentano un buon livello di precisione
 - Classe “Alta”: notevole differenza di precisione
 - Limitazioni
 - Dataset sbilanciato
 - Numero ridotto di attributi
 - Soluzioni proposte
 - Aggiunta di ulteriori informazioni rilevanti al dataset
- 
- A decorative network diagram in the bottom-right corner, consisting of small white dots and circles connected by thin white lines, forming a sparse, scattered network.



GRAZIE!

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**.

Please keep this slide for attribution.