UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

A.A. 2023/2024
Laurea Magistrale
Informatica

Progetto Wine Type

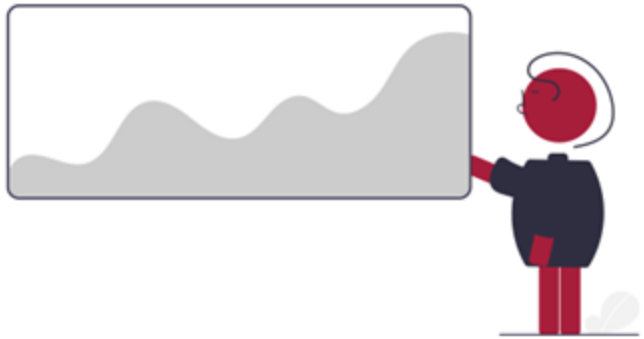
Machine Learning – Febbraio 2024

Realizzato da:

Cavaleri Matteo - 875050
Gargiulo Elio - 869184
Piacente Cristian - 866020

Introduzione del Progetto

- Svolgimento di un'analisi esplorativa su un dataset scelto
- Costruzione e allenamento di tre modelli di apprendimento
- Verifica delle prestazioni ed efficacia sul dataset
- Comparazione tra modelli e analisi dei risultati
- Considerazioni e conclusioni



Features Dataset e Target

	type	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	white	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6
1	white	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6
2	white	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6
3	white	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6
4	white	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6

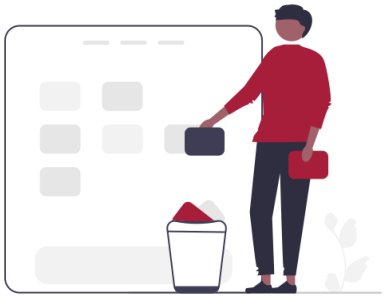
- Target: type (binario)
- 11 features continue e 1 categorica (quality intero tra 0 e 10)

La Scelta del Dataset

- Garantire coerenza e rilevanza nelle analisi successive
- Utilizzo di dati sensati anziché fittizi
- Dati principalmente numerici continui
- Non troppo complesso
- Compatibile con PCA

Pulizia e Casting del Dataset

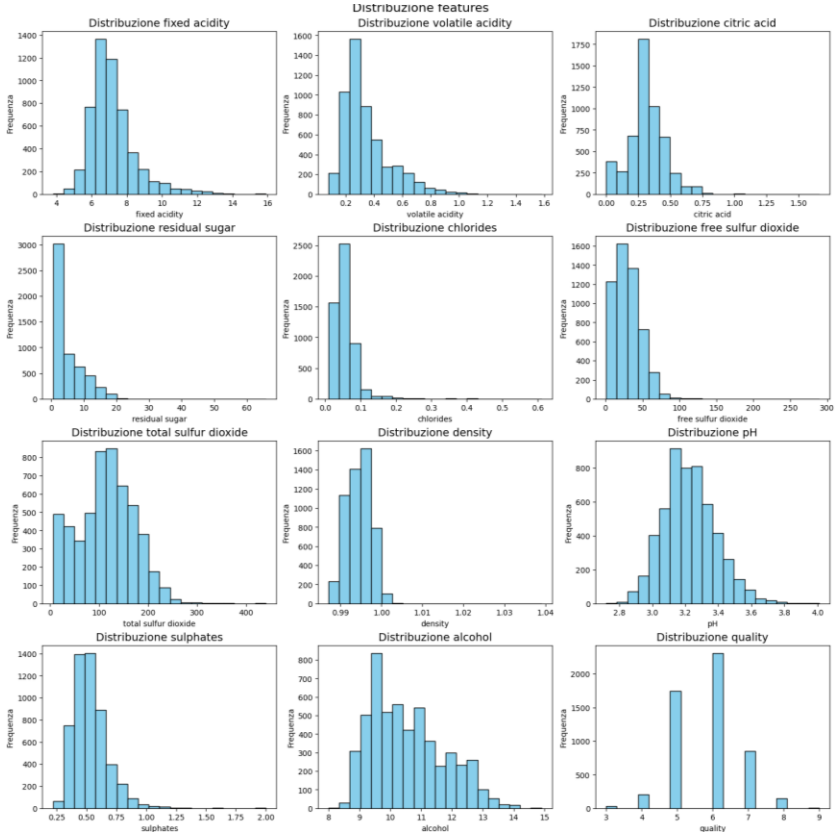
Prima di tutto è stato sistemato il dataset per evitare di riscontrare problemi su dati non conformi all'analisi esplorativa



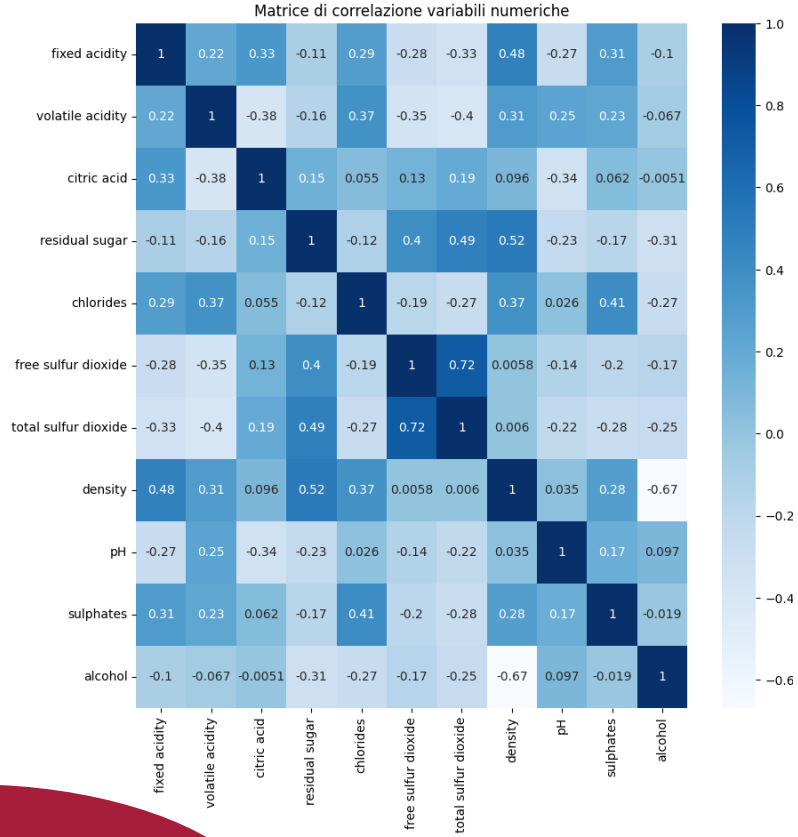
- Eliminazione di dati nulli
- Eliminazione di dati duplicati
- Casting al tipo booleano del target con Label Encoding (red -> False, white -> True)
- Casting al tipo categoria della quality.

Analisi delle Covariate e Target (1/4)

- Distribuzioni simili a gaussiane

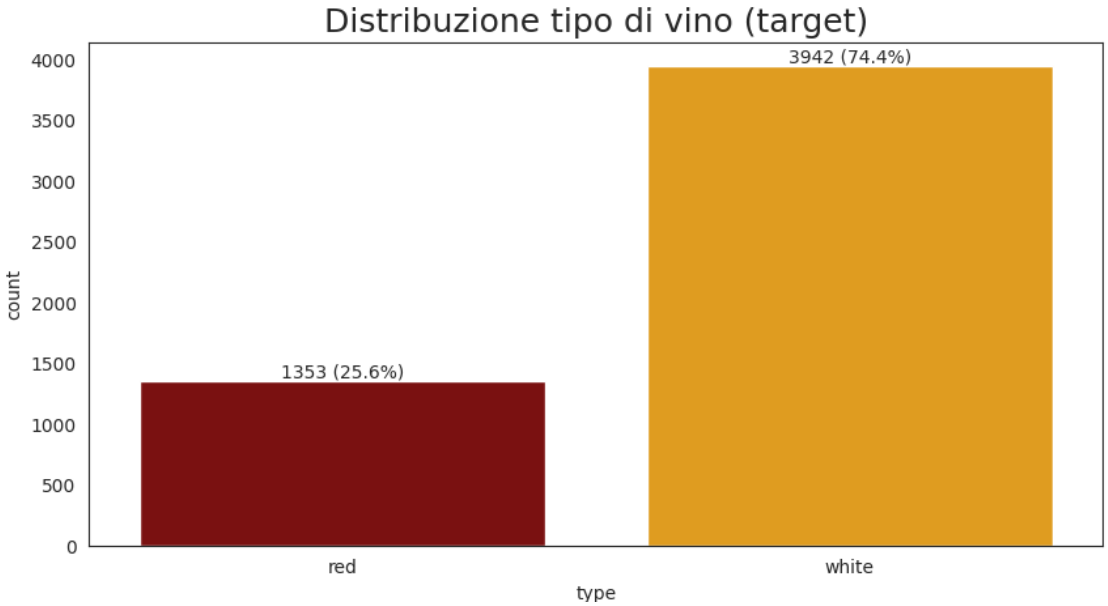


Analisi delle Covariate e Target (2/4)



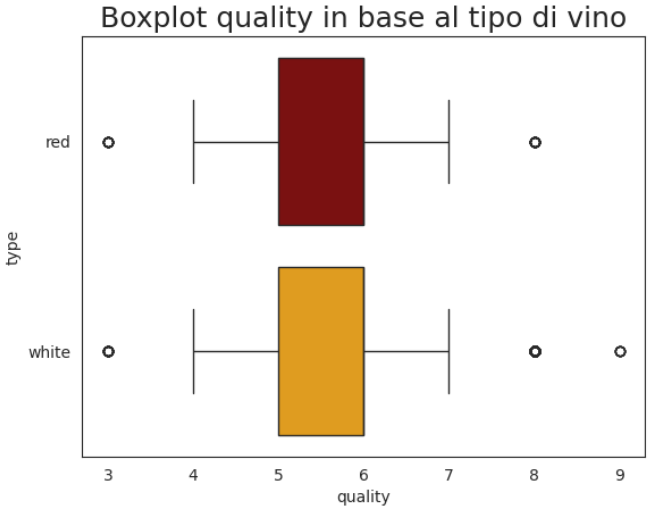
- L'unica correlazione rilevante è total sulfur dioxide con free sulfur dioxide

Analisi delle Covariate e Target (3/4)



- Target abbastanza sbilanciato

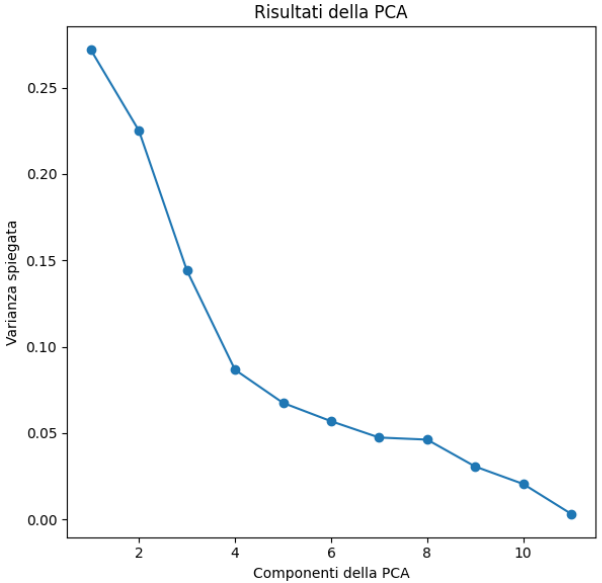
Analisi delle Covariate e Target (4/4)

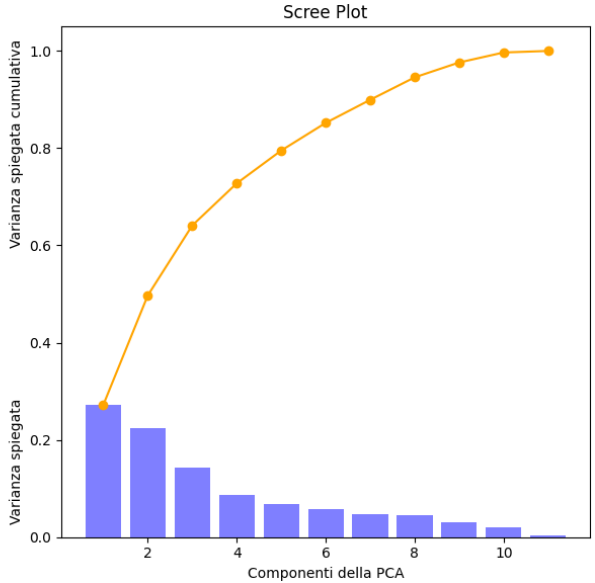


- Quality non aggiunge informazione sul target type
  → **Drop** della colonna

Principal Component Analysis (1/2)

	Eigenvalue	Variance Percent	Cumulative Variance Percent
Comp 1	2.991077	27.186472	27.186472
Comp 2	2.476404	22.508515	49.694986
Comp 3	1.585096	14.407246	64.102232
Comp 4	0.953458	8.666165	72.768397
Comp 5	0.742378	6.747617	79.516014

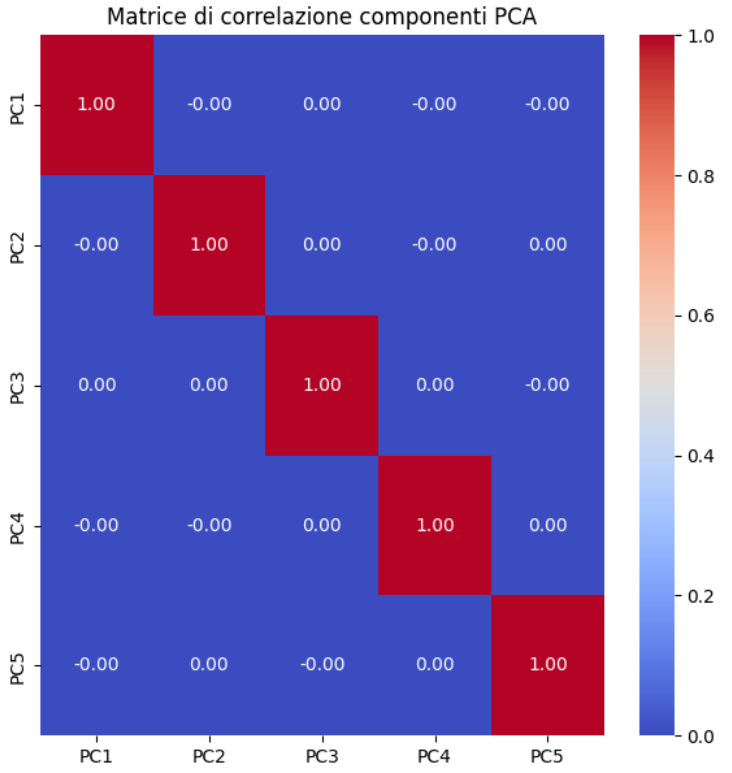




- 5 componenti spiegano ~ 80% varianza

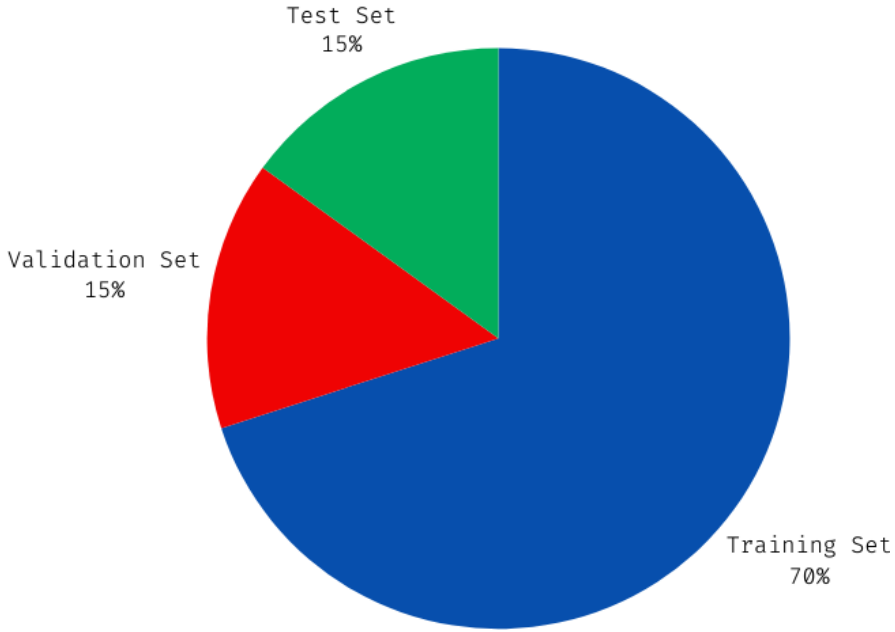
Principal Component Analysis (2/2)

- PCA efficace, componenti non correlate

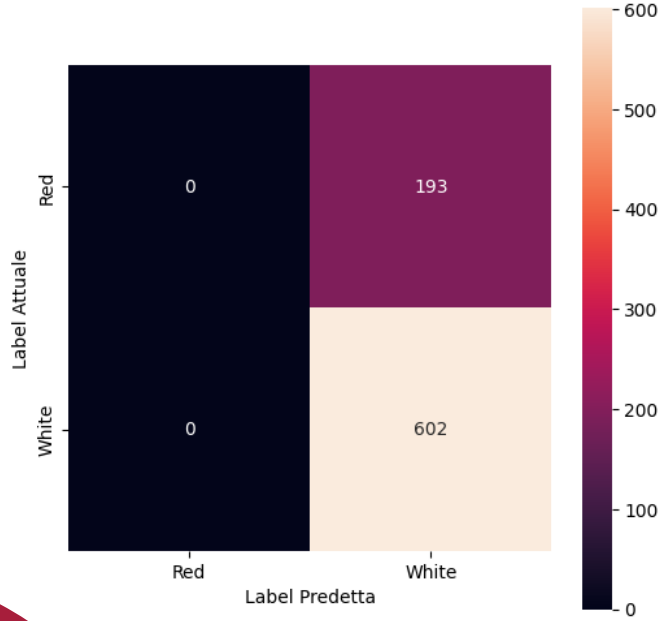


Operazioni Preliminari

- Suddivisione dei dati ottenuti dalla PCA con dimensione (5295, 5)



Modello Baseline



- Punto di partenza per la costruzione dei modelli

- Classificazione per tutti white
  → Test accuracy ~ 0.7572

Modelli di Apprendimento

- Rete Neurale
- Support Vector Machines
- Albero di Decisione

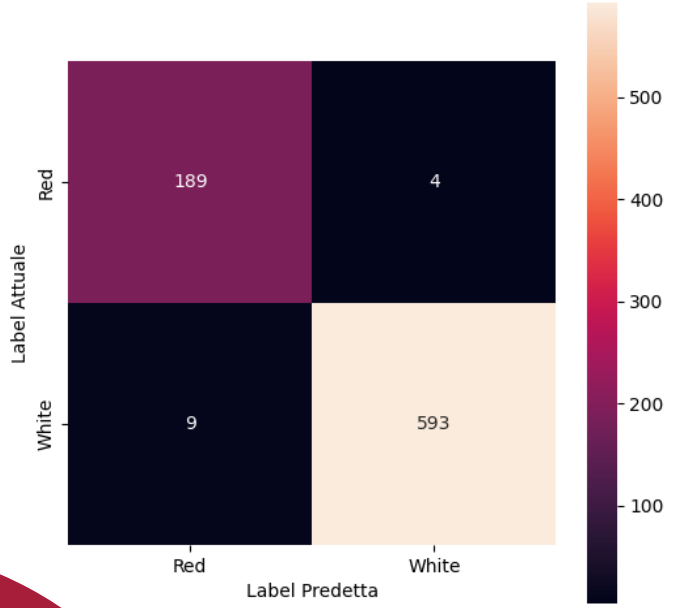
- Approccio Naive: immediato

→ Iperparametri di default/comuni

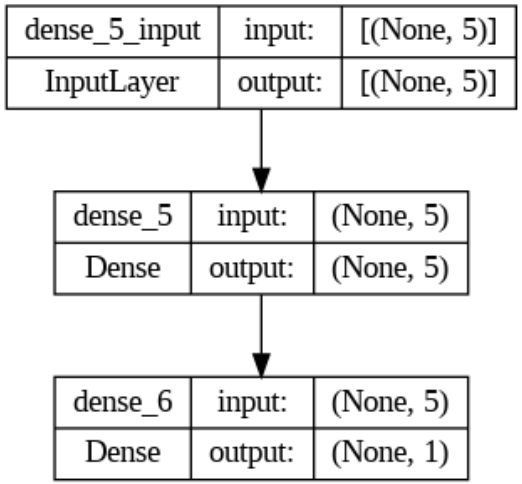
- Approccio Ottimale: complesso

→ Ricerca degli Iperparametri migliori

Rete Neurale Naive



Training Accuracy	Test Accuracy	Training Loss	Test Loss
~ 0.9833	~ 0.9836	~ 0.1030	~ 0.1073



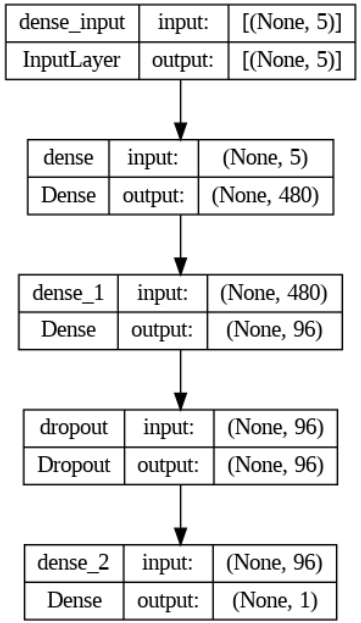
- Training Standard con 10 epoche

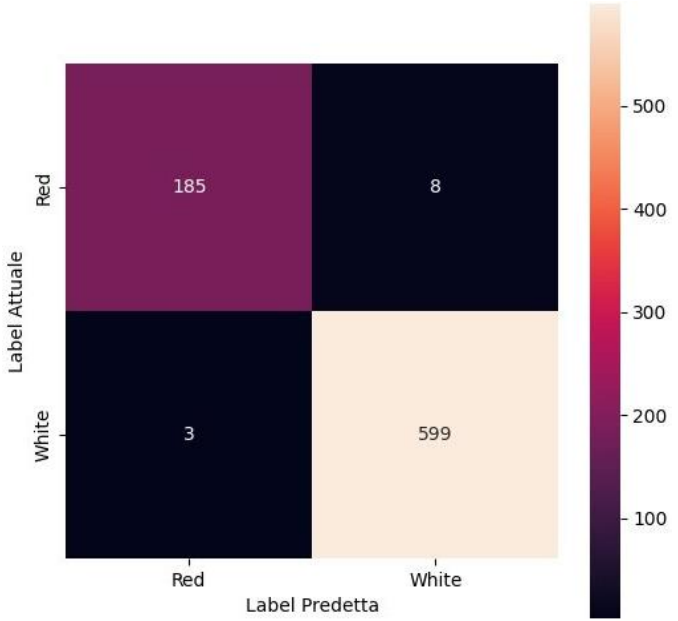
Rete Neurale Ottimale

Units	Learning Rate	Layers
480	0.001	1

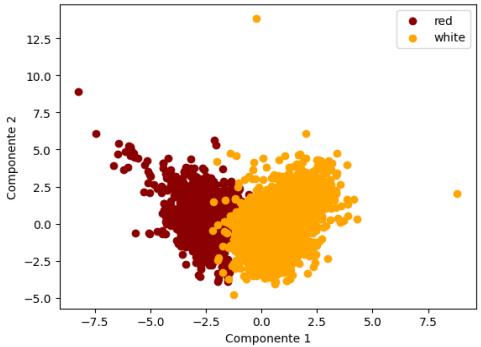
Training Accuracy	Test Accuracy	Training Loss	Test Loss
~ 0.9906	~ 0.9862	~ 0.0321	~ 0.0591

- Keras Tuner Library
- Obiettivo riduzione dell'overfitting
- Layers di Dropout

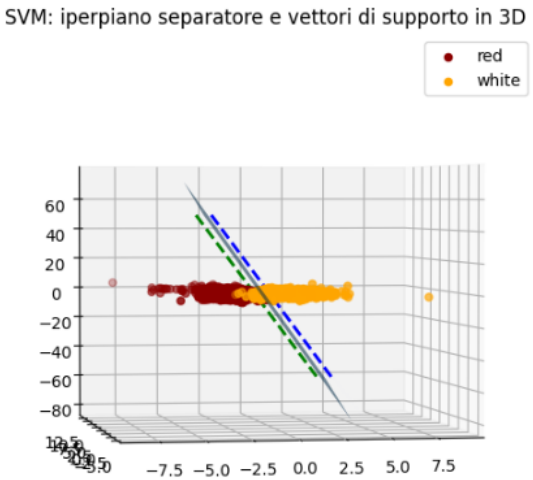




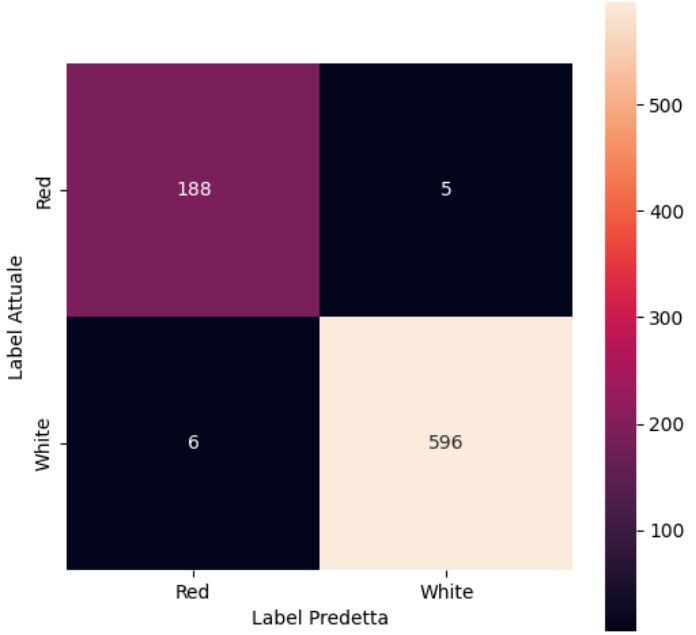
Support Vector Machines Naive



- Intuizione da PCA



Training Accuracy	Test Accuracy
~ 0.9854	~ 0.9862

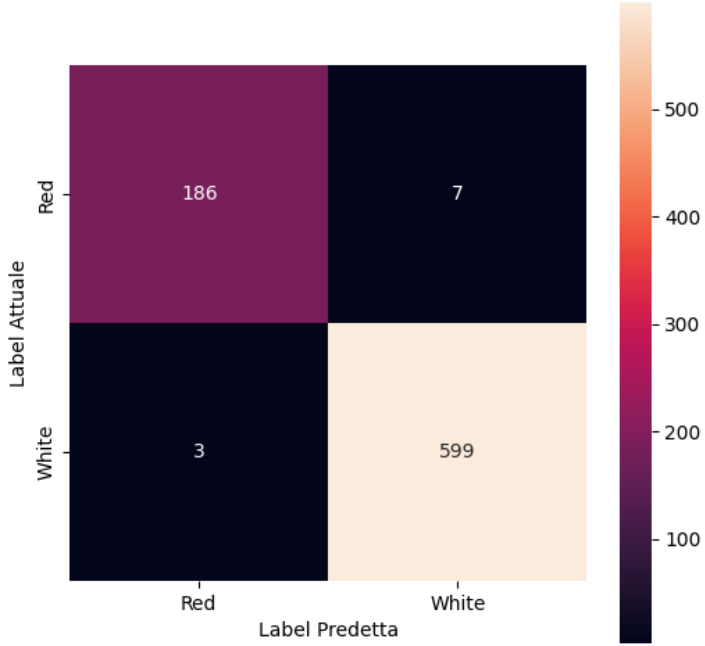


Support Vector Machines Ottimale

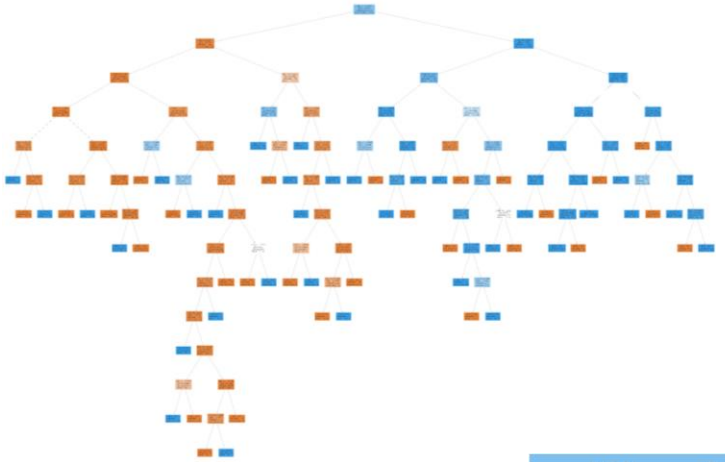
Kernel	C	Gamma
rbf	100	0.01

Training Accuracy	Test Accuracy
~ 0.9895	~ 0.9874

- Hyperparameter Tuning con Grid Search

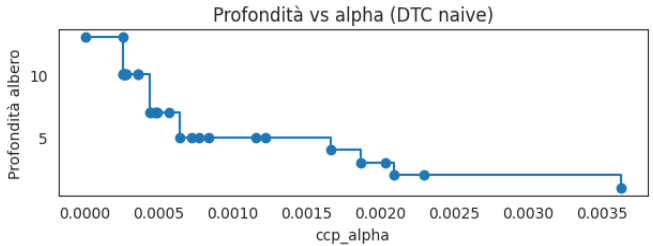


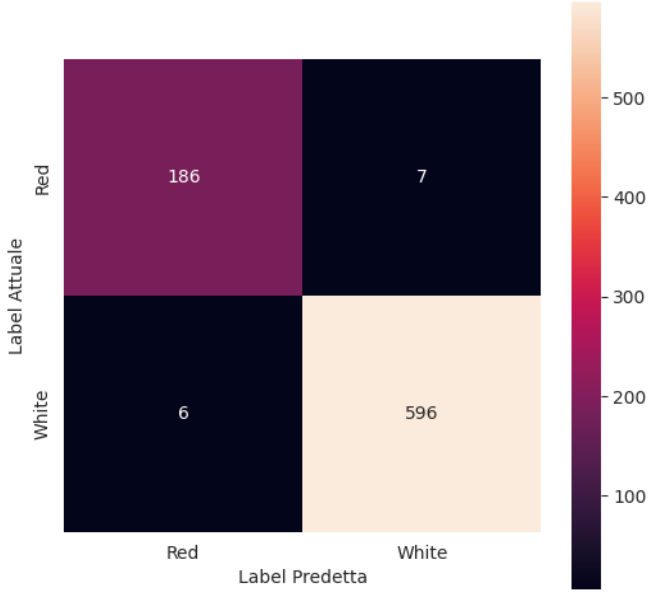
Albero Decisionale Naive



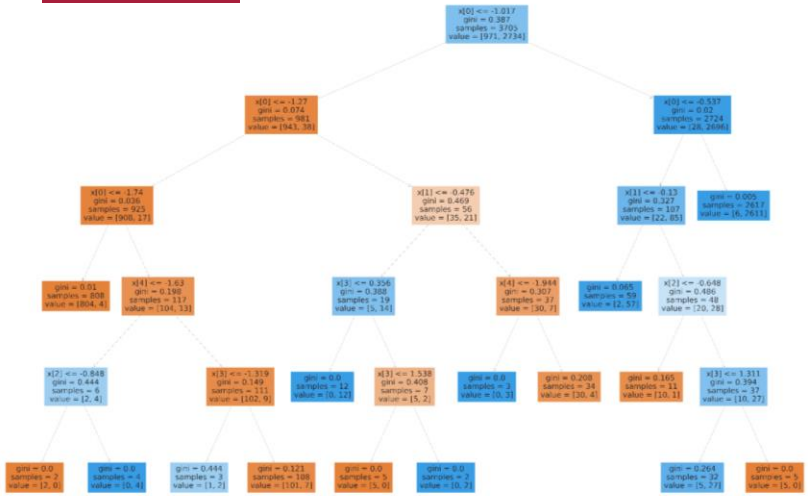
Training Accuracy	Test Accuracy
1.0	~ 0.9836

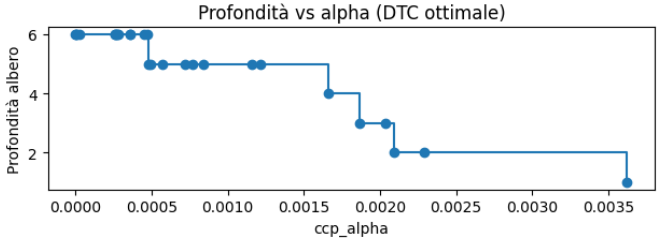
x[0] <= -1.017
gini = 0.387
samples = 3705
value = [971, 2734]





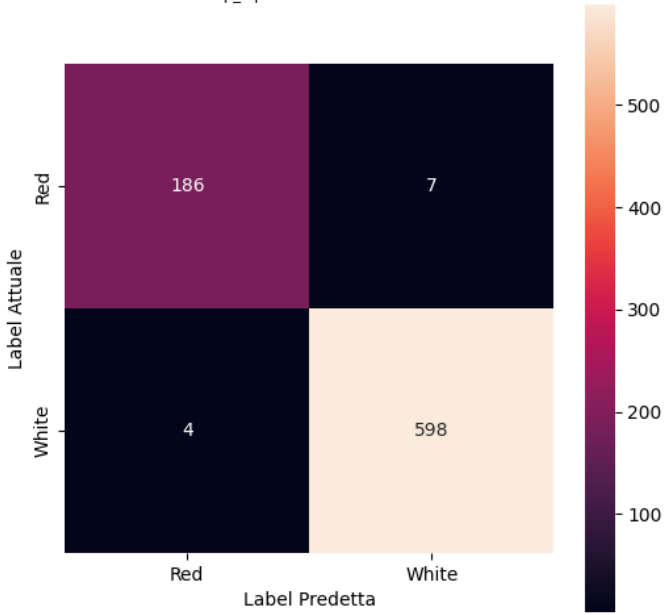
Albero Decisionale Ottimale



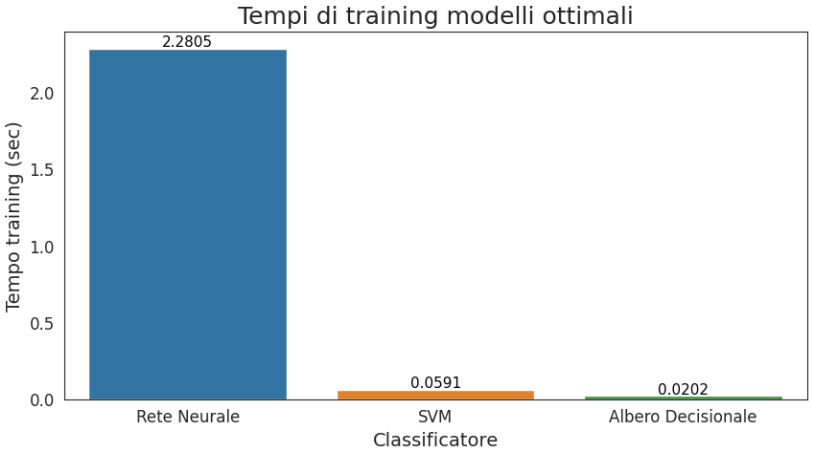


Alpha	Criterion	Max Depth	Max Features	Splitter
0.0005	Gini	6	None	Best

Training Accuracy	Test Accuracy
~ 0.9919	~ 0.9862

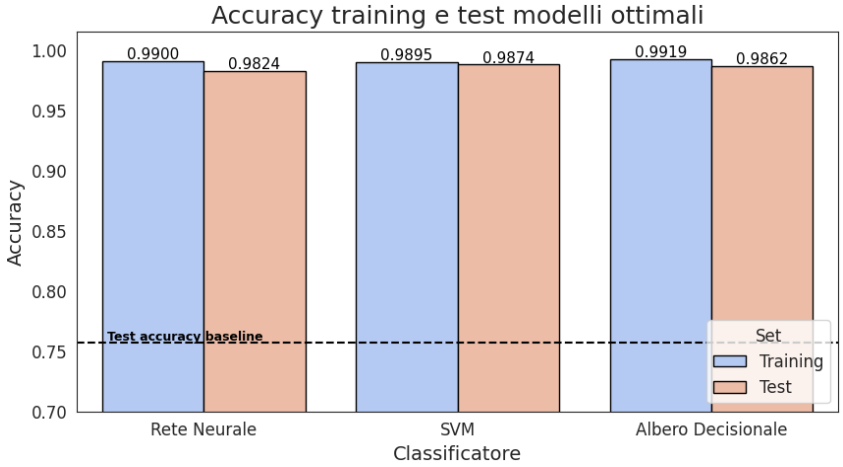


Valutazione Performance Ottimali (1/6)

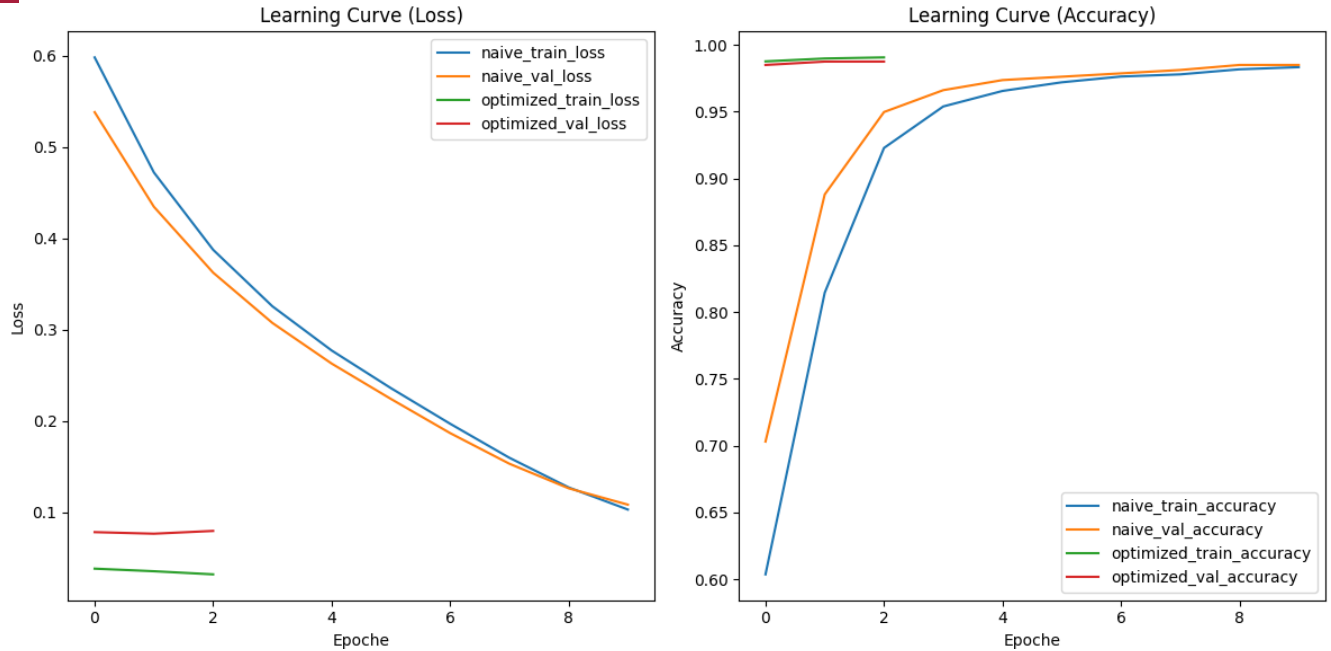


- Tempo elevato per la Rete Neurale

- Accuracy simili, siamo in presenza di un leggero overfitting

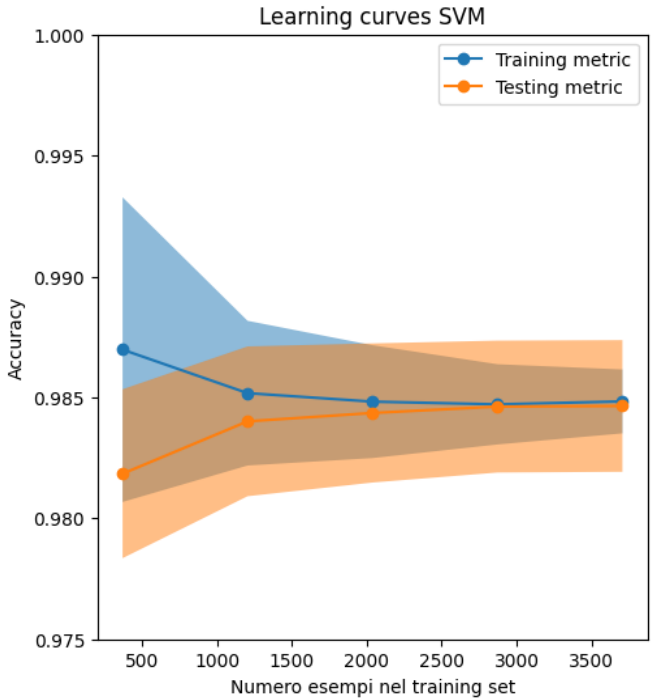


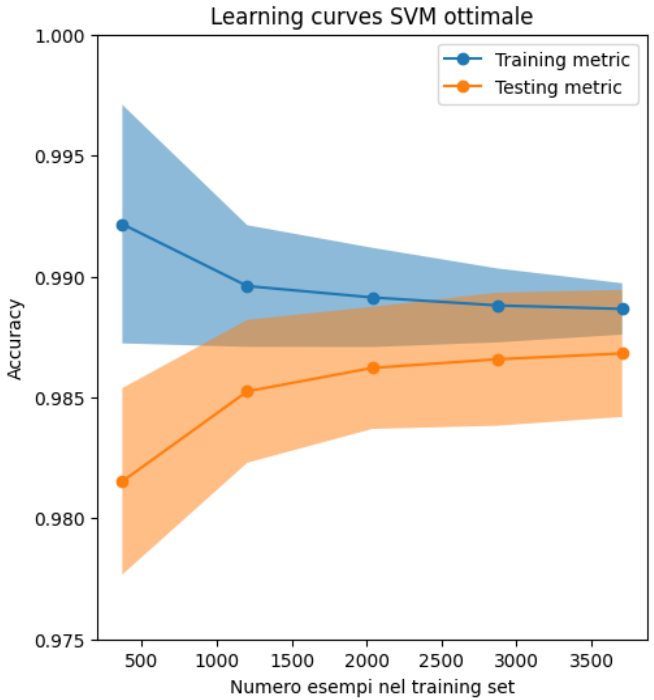
Valutazione Performance Ottimali (2/6)



- Rete Neurale Ottimale converge subito con migliori valori, ma maggior overfitting

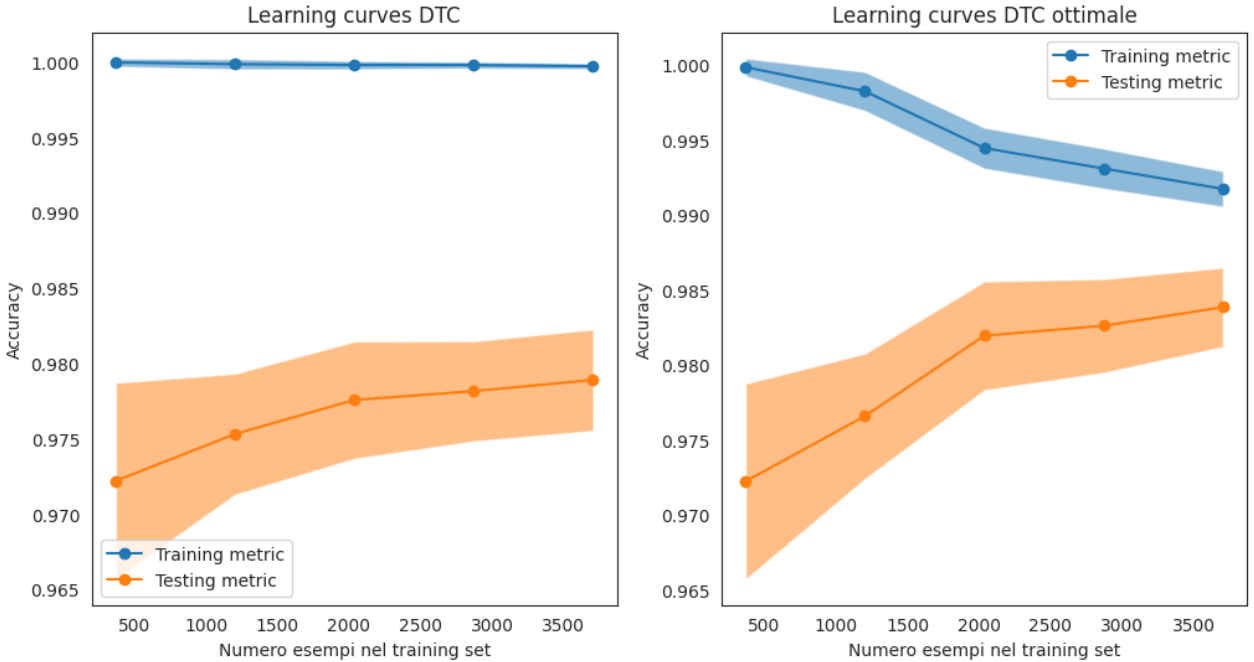
Valutazione Performance Ottimali (3/6)





- Leggero overfitting in SVM ottimale

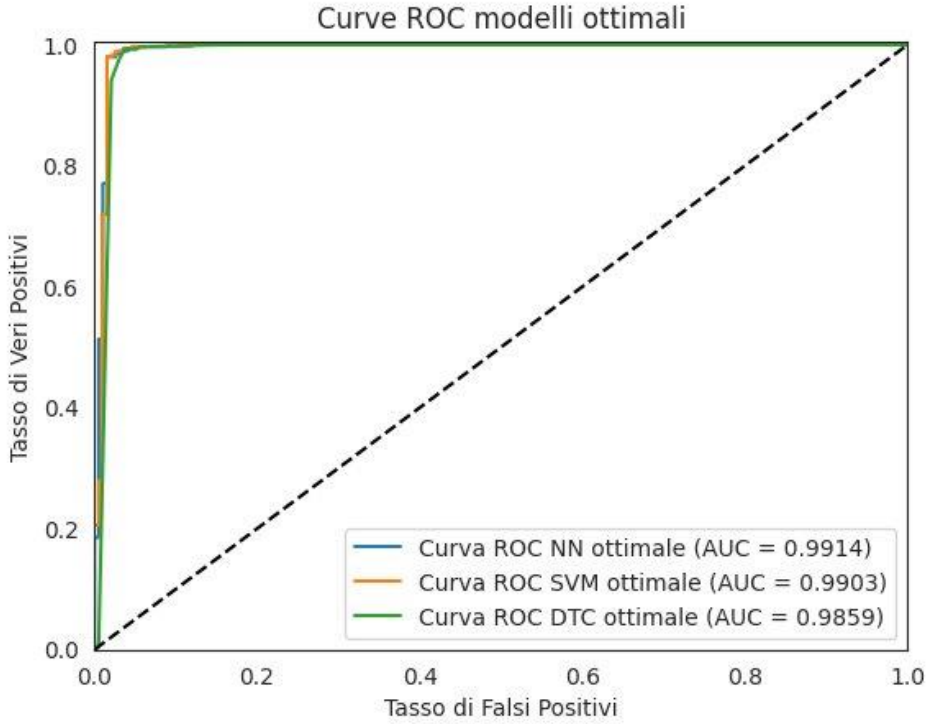
Valutazione Performance Ottimali (4/6)



- Minor overfitting in Albero Decisionale Ottimale

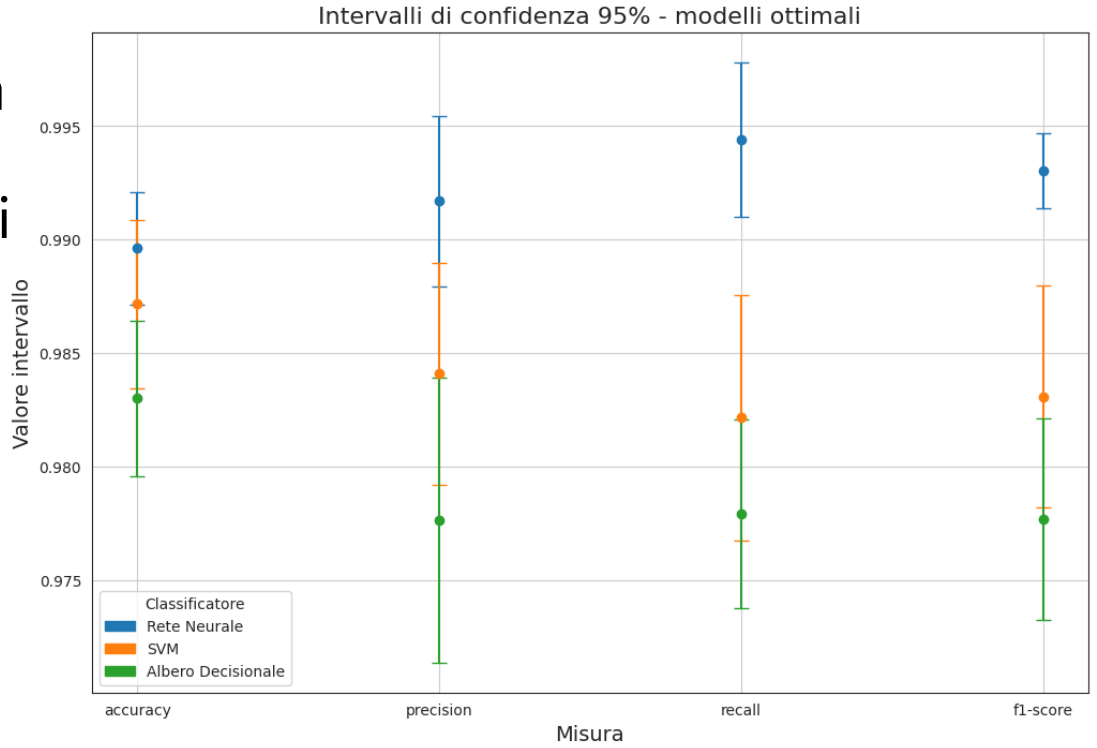
Valutazione Performance Ottimali (5/6)

- Curve ROC ottimali sovrapposte
- AUC vicine ad 1



Valutazione Performance Ottimali (6/6)

- Stratified 10-Fold Cross Validation
- La Rete Neurale ha valori maggiori degli altri due modelli



Considerazioni e Conclusioni

- I modelli ottimali ottenuti riportano buone metriche di performance anche se sono in leggero overfitting
- La Rete Neurale ha valori maggiori degli altri due modelli in cross validation richiedendo però maggior tempo per l'addestramento
- La SVM risulta essere il miglior compromesso tra risultati e complessità
- L'Albero Decisionale risulta essere il modello più veloce e semplice da utilizzare