a.a 2023/2024

PREDIZIONE PER RILASCIO DI CARTE DI CREDITO

Oltolini Edoardo – 869124

Pulcino Federico – 872491

Università degli Studi di Milano – Bicocca

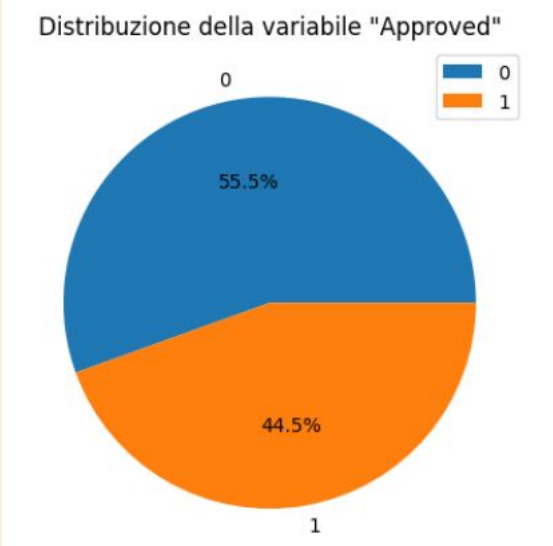
OBIETTIVO DELL'ELABORATO

L'obiettivo del progetto è, dato un dataset contenenti informazioni su degli individui (e.g. Genere, Età, Stato Occupazionale), di allenare modelli di machine learning che permettano di decidere l'esito, positivo o negativo, di una richiesta di rilascio di una carta di credito.

DESCRIZIONE DEL DATASET

Il Dataset è composto da **690** elementi, il **70%** di essi è stato utilizzato per il training dei modelli mentre il rimanente **30%** per la fase di testing, entrambi contengono un numero relativamente equo di esempi positivi e negativi.

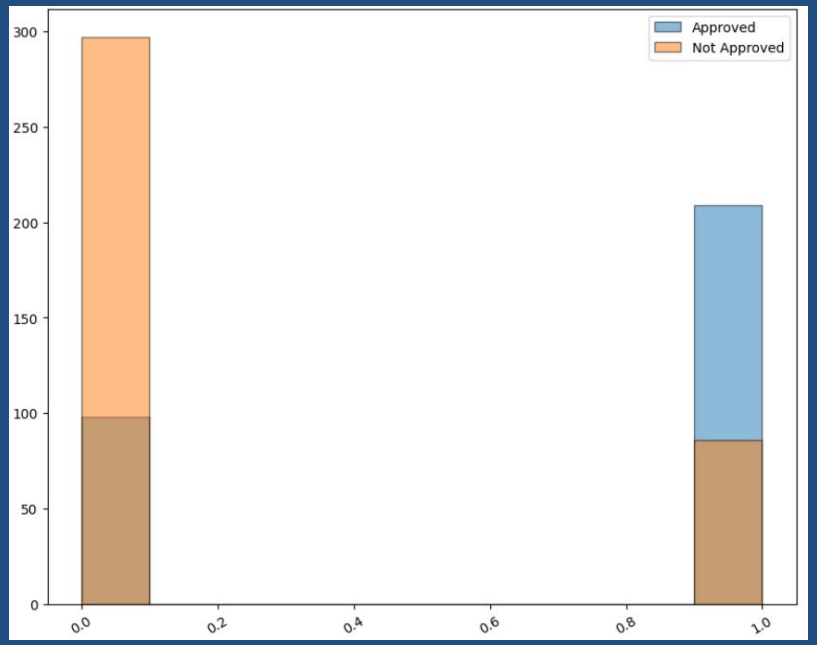
Il Dataset presenta **16** attributi, di cui uno, «Approved» è la variabile target che si vuole predire.



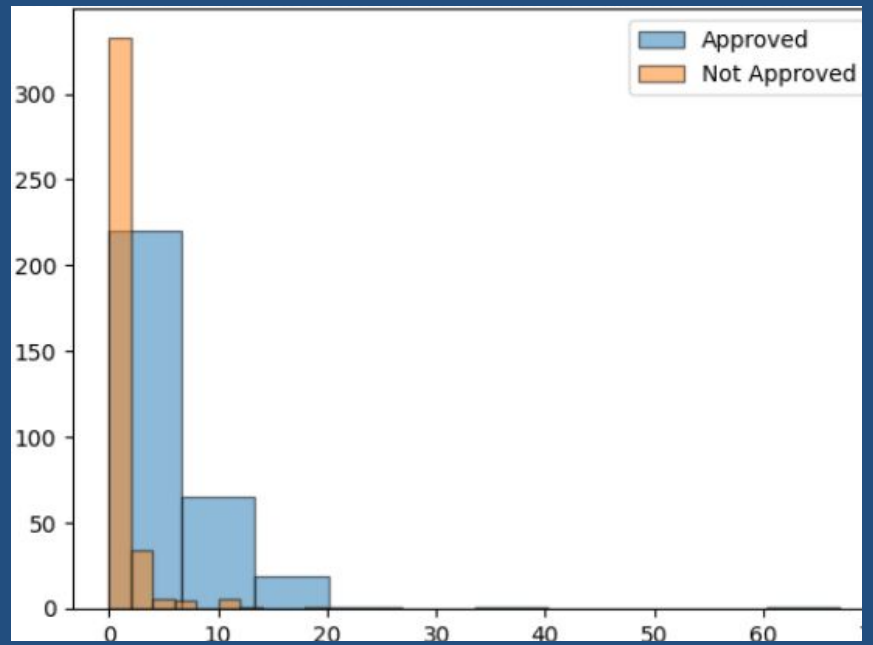
	Gender	Age	Debt	Married	BankCustomer	Industry	Ethnicity	YearsEmployed	PriorDefault	Employed	CreditScore	DriversLicense	Citizen	ZipCode	Income	Approved
0	1	30.83	0.000	1	1	Industrials	White	1.25	1	1	1	0	ByBirth	202	0	1
1	0	58.67	4.460	1	1	Materials	Black	3.04	1	1	6	0	ByBirth	43	560	1
2	0	24.50	0.500	1	1	Materials	Black	1.50	1	0	0	0	ByBirth	280	824	1
3	1	27.83	1.540	1	1	Industrials	White	3.75	1	1	5	1	ByBirth	100	3	1
4	1	20.17	5.625	1	1	Industrials	White	1.71	1	0	0	0	ByOtherMeans	120	0	1

ANALISI ESPLORATIVA

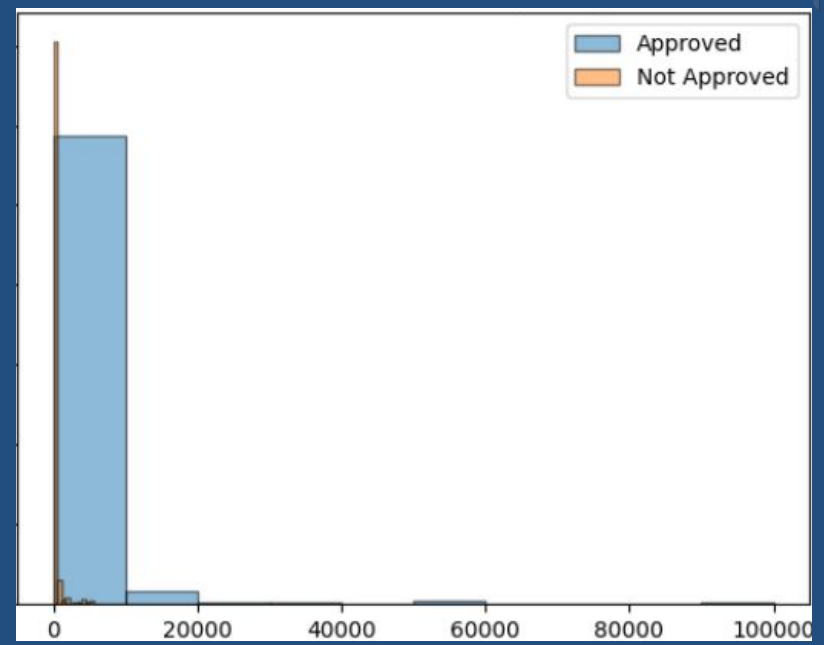
Dall'analisi del dataset gli attributi, «Employed», «Income» e «CreditScore» sembrano capaci di discriminare bene la variabile di interesse Approved.



Distribuzione di «Employed» rispetto alla variabile «Approved»



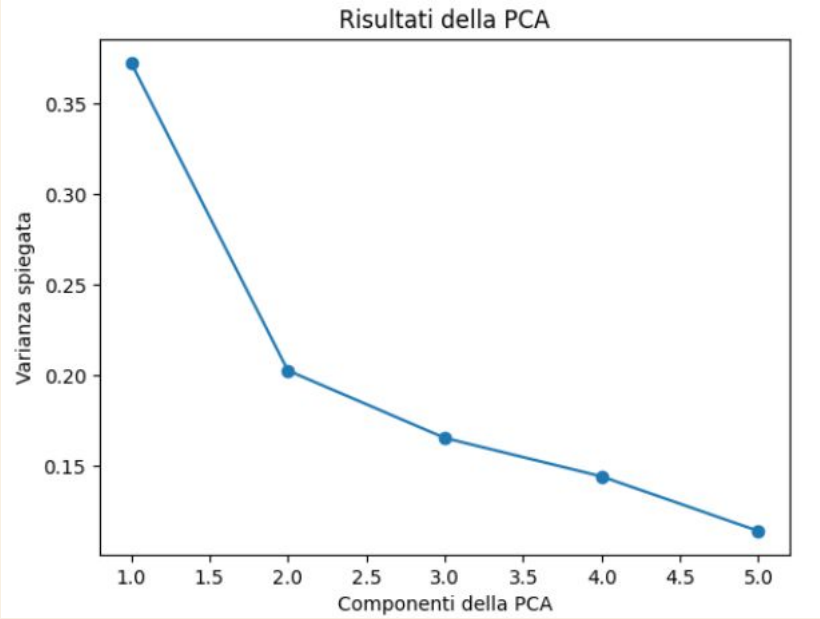
Distribuzione di «CreditScore» rispetto alla variabile «Approved»



Distribuzione di «Income» rispetto alla variabile «Approved»

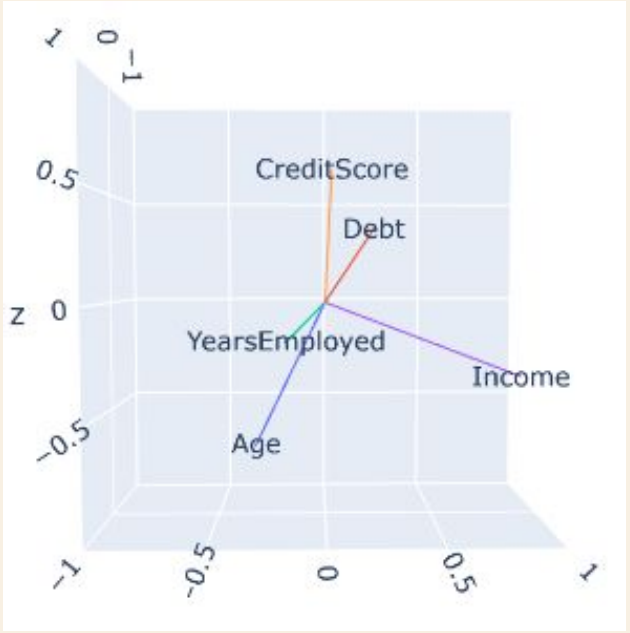
PRINCIPAL COMPONENTS ANALYSIS

QUANTE COMPONENTI PRINCIPALI MANTENERE




```
Varianza spiegata mantenendo 2 componenti 0.5756021217619177
Varianza spiegata mantenendo 3 componenti 0.7412778140532335
Varianza spiegata mantenendo 4 componenti 0.885713500766025
Varianza spiegata mantenendo tutte le componenti 1.0
```

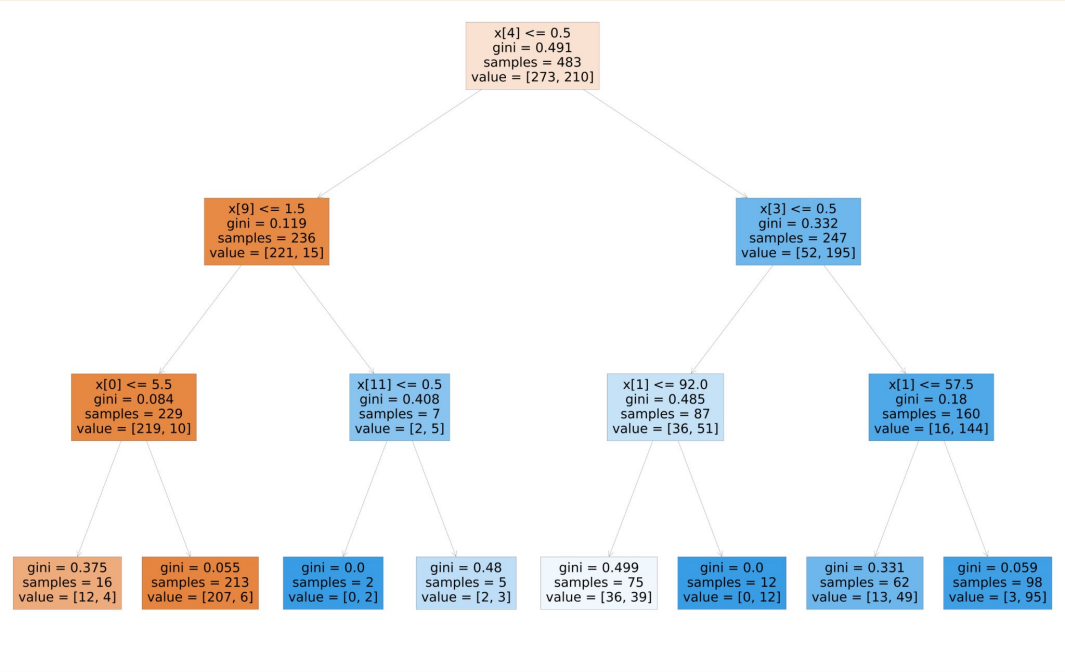
ANALIZZARE LA CORRELAZIONE TRA LE VARIABILI NUMERICHE

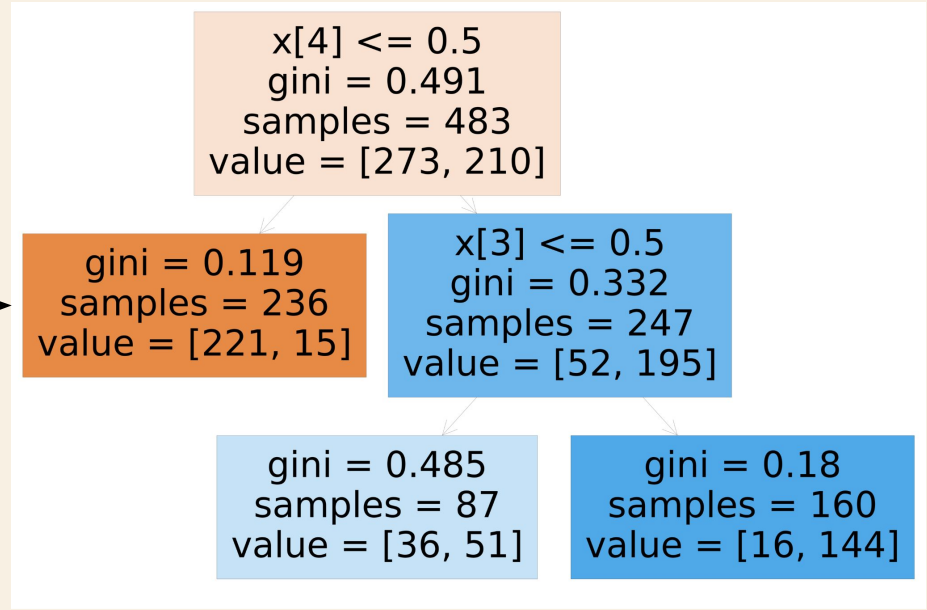



```
Distanza di Age dall'origine: 0.8421435676709103
Distanza di Debt dall'origine: 0.9932475757026835
Distanza di YearsEmployed dall'origine: 0.6195067100242775
Distanza di Income dall'origine: 0.9999887533511271
Distanza di CreditScore dall'origine: 0.9594203428098699
```

PRUNED DECISION TREE

Il CART, Classification & Regression (Decision) Tree, è uno strumento di supporto alle decisioni strutturato a forma di albero. Ogni nodo corrisponde ad un attributo ed ogni ramo a un valore del corrispondente attributo. L'accuratezza iniziale è del 78%.





max_depth = 3, Accuracy = 83%

ccp_alpha = 0.012…, Accuracy = 84%

Motivazioni Decision Tree

- L’albero decisionale può gestire sia variabili categoriche che numeriche e non richiede ipotesi sulla distribuzione dei dati.

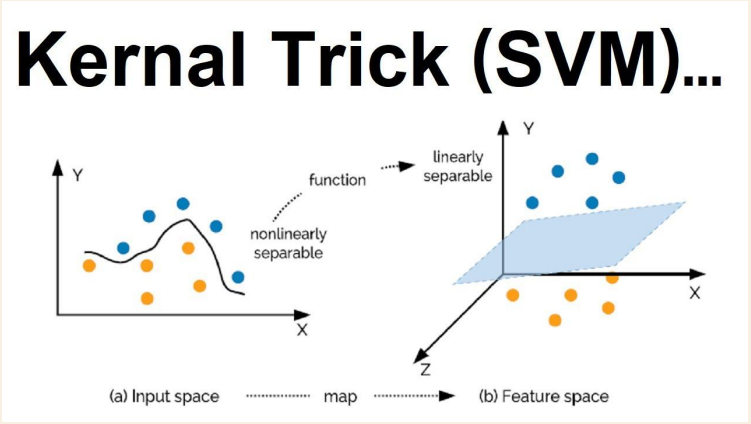
- Sono modelli intuitivi e facilmente interpretabili; possono essere quindi visualizzati graficamente, il che aiuta a comprendere meglio le loro decisioni.

SVM

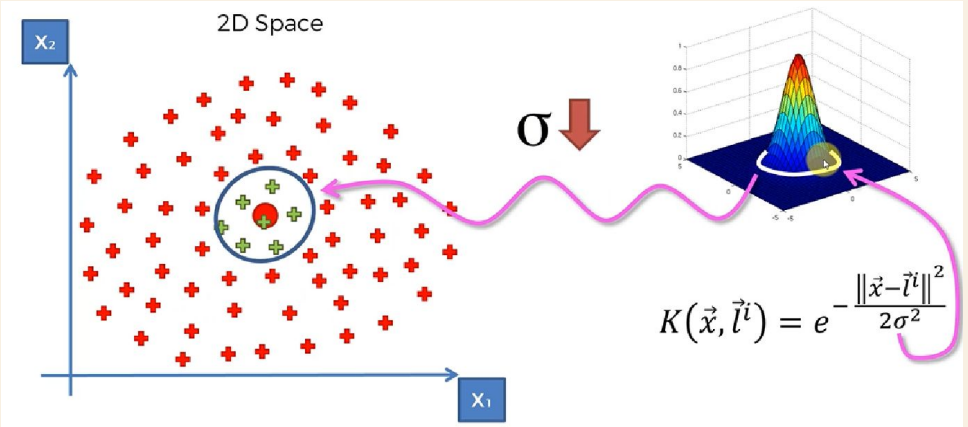
01 Algoritmo di Machine Learning supervisionato

02 Adattabilità con funzioni Kernel

03 Applicazione del modello al dataset



Rappresentazione Kernel Trick



Rappresentazione Kernel RBF

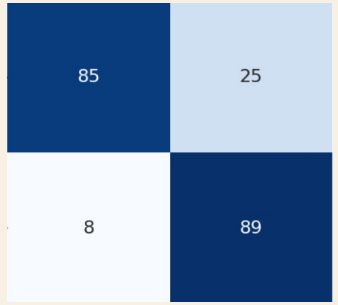
Motivazioni SVM

- Classificazione binaria sulla variabile target
- Alta dimensionalità
- Presenza di parametri regolarizzatori per prevenire l'overfitting
- Diverse tipologie di Kernel (Lineare e RBF) che consentono di effettuare un migliore studio sui dati

PANORAMICA DELLE PERFORMANCE

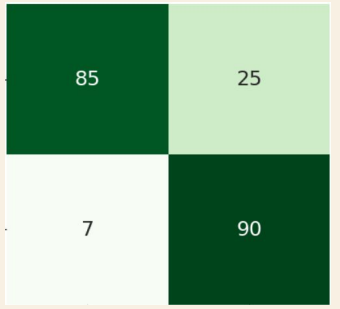
Pruned Decision Tree
Tempo di training: 0.008 secondi

	precision	recall	f1-score	support
0	0.91	0.77	0.84	110
1	0.78	0.92	0.84	97
accuracy			0.84	207
macro avg	0.85	0.85	0.84	207
weighted avg	0.85	0.84	0.84	207



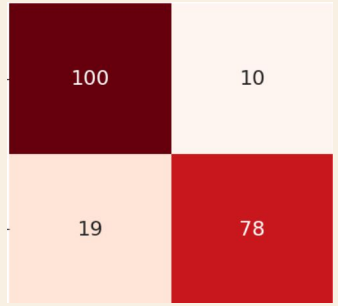
SVM - rbf kernel
Tempo di training: 0.139 secondi

	precision	recall	f1-score	support
0	0.92	0.77	0.84	110
1	0.78	0.93	0.85	97
accuracy			0.85	207
macro avg	0.85	0.85	0.85	207
weighted avg	0.86	0.85	0.85	207



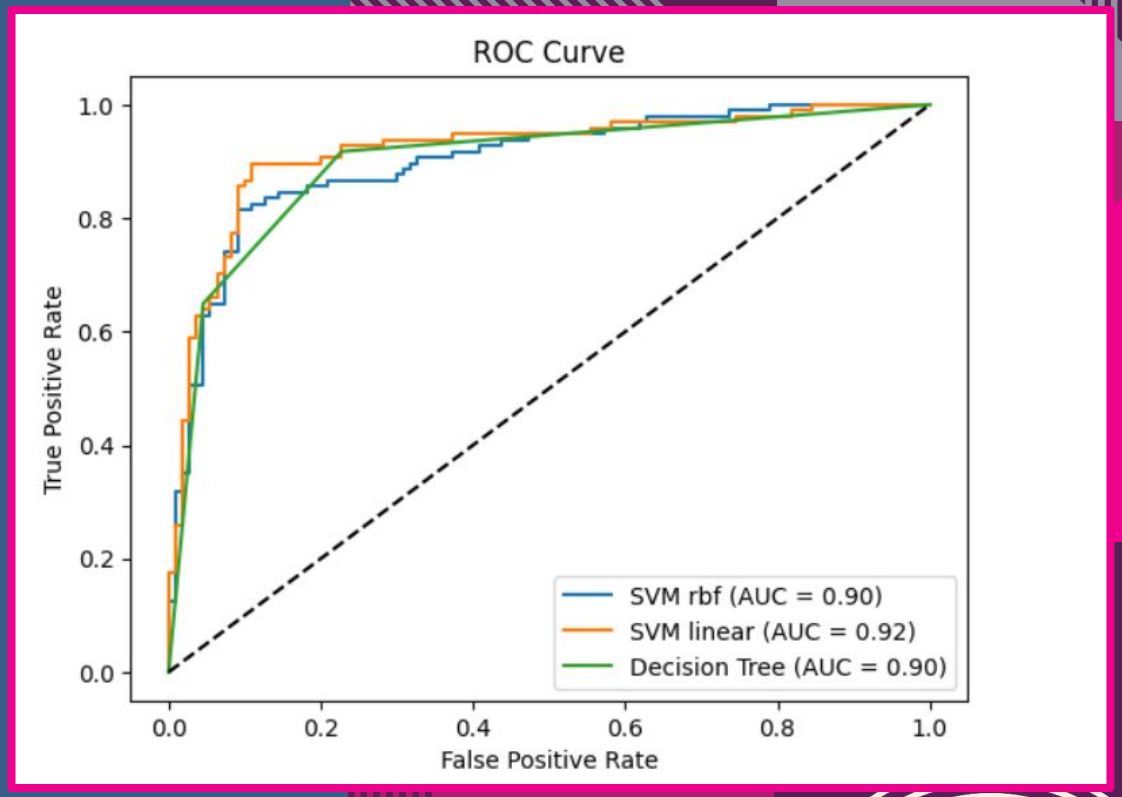
SVM - linear kernel
Tempo di training: 2.822 secondi

	precision	recall	f1-score	support
0	0.84	0.91	0.87	110
1	0.89	0.80	0.84	97
accuracy			0.86	207
macro avg	0.86	0.86	0.86	207
weighted avg	0.86	0.86	0.86	207



CURVE ROC

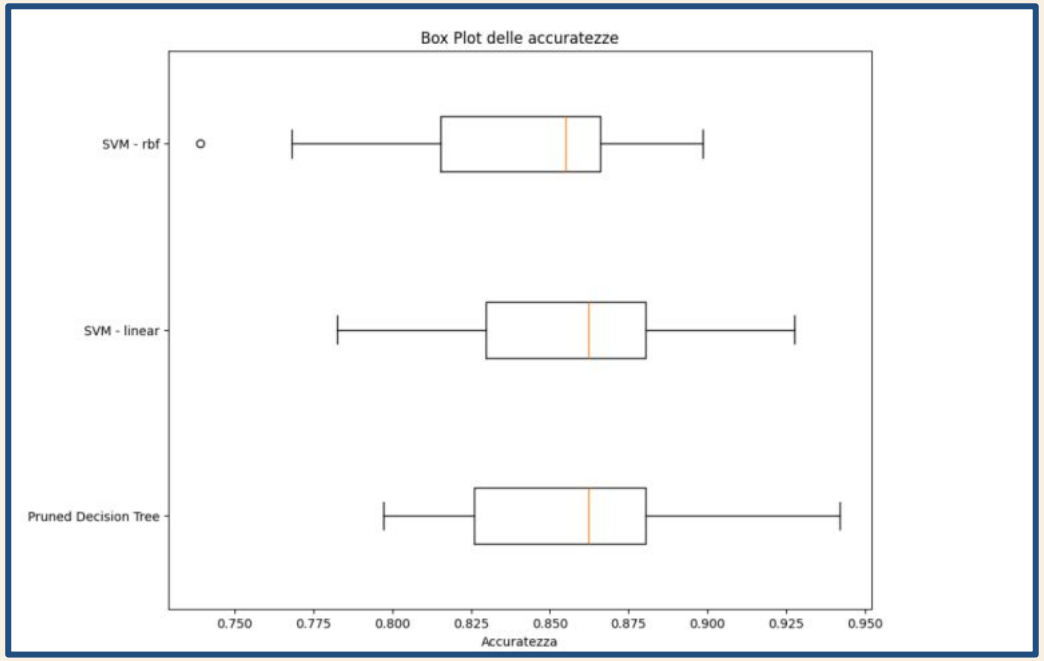
Le curve ROC (Receiver Operating Characteristic) sono una misura di performance per valutare i modelli, non solo rispetto ad una specifica classe ma anche rispetto a diversi modelli. Mettono in rapporto il **True Positive Rate** contro il **False Positive Rate**.



10-FOLD CROSS-VALIDATION

Suddivisione del dataset in 10 parti di uguale numerosità e, a ogni passo k, la k-esima parte del dataset diventa quella di convalida, mentre la restante costituisce l'insieme di addestramento.

```
Intervalli di confidenza Pruned Decision Tree:  (0.8275114310743121, 0.8710392935633691)
Intervalli di confidenza SVM - linear:  (0.8283874177986328, 0.8875546111868746)
Intervalli di confidenza SVM - rbf:  (0.83394632501335, 0.8819957039721574)
```



CONCLUSIONI

Tutti i modelli si sono rivelati più che sufficientemente adeguati sia in termini di accuratezza che in termini di performance, mostrando un'accuratezza tra l'**84%** e l'**86%** e performance con un ROC AUC da **0.90** a **0.92**, tuttavia Il modello SVM con kernel lineare però è quello che si distingue maggiormente per quanto riguarda l'area sotto la curva ROC, in quanto ottiene quella maggiore. Tuttavia, ulteriori ricerche potrebbero essere necessarie per ottimizzare ulteriormente questi modelli e migliorare le loro prestazioni.