

# CENSUS INCOME PREDICTION

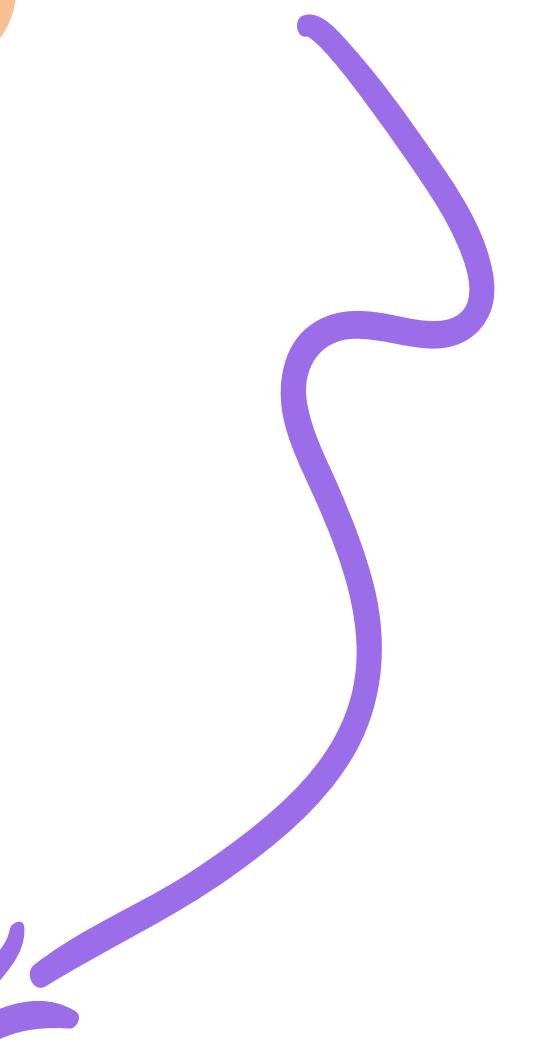


Presentazione

Nicolò Sansevrino 865889

# OVERVIEW

Introduzione  
Preprocessing  
Feature engineering  
Esperimenti sui modelli  
Tecniche di validazione  
Analisi dei risultati  
Conclusioni



# INTRODUZIONE

Il progetto mira a identificare dei classificatori in grado di predire se un individuo guadagna più o meno di 50k \$ l'anno

Il dataset proviene dal US Census Bureau risalente al 1994



# **ANALISI ESPLORATIVA**

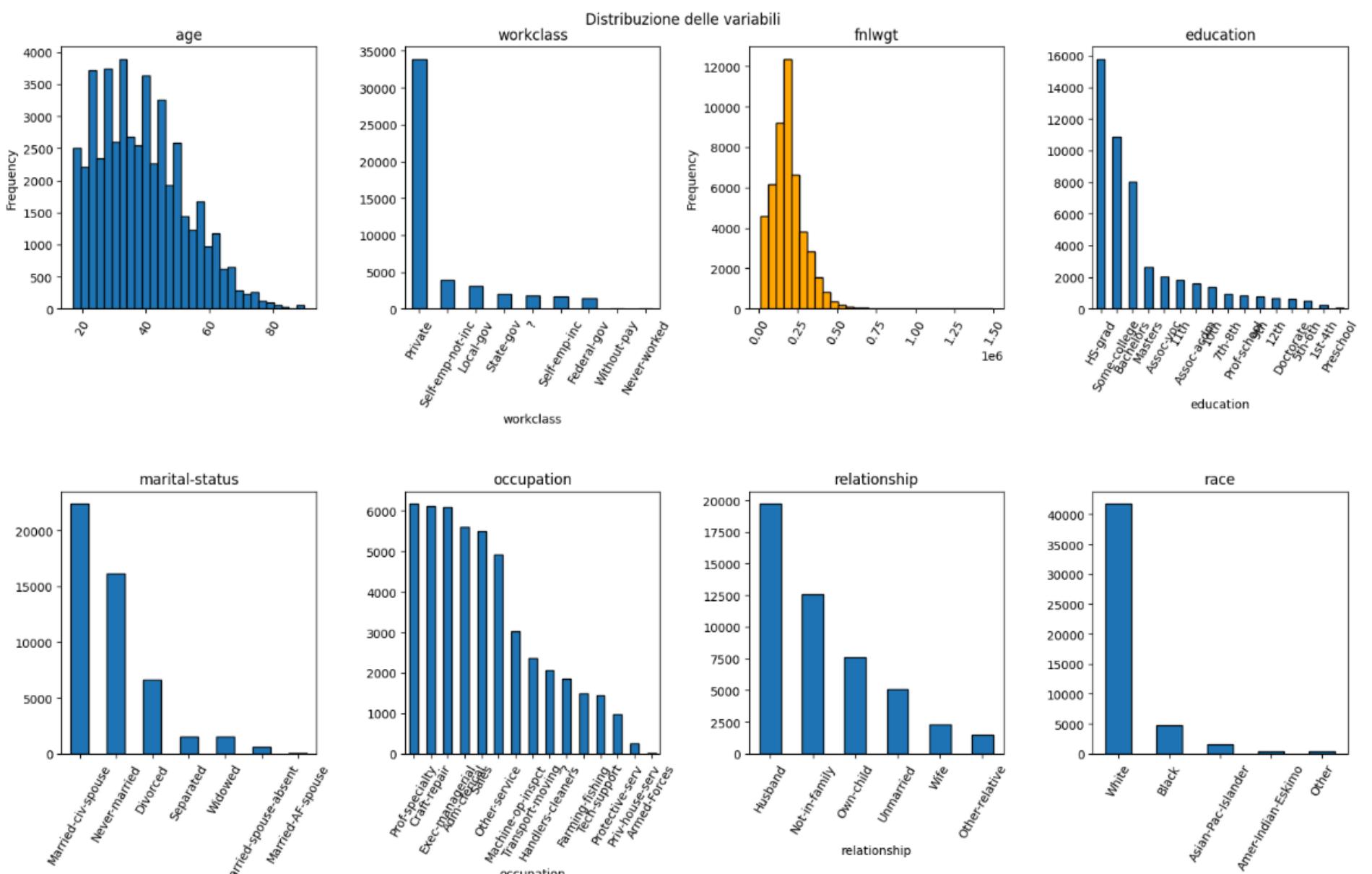
Il dataset contiene informazioni demografiche quali età, livello di educazione, provenienza, tipo di lavoro ecc

I tipi di dati sono di diversa natura (numerica, categorici)

Il dataset presenta imperfezioni quali valori null, duplicati e sbilanciamento



# DATA VISUALIZATION

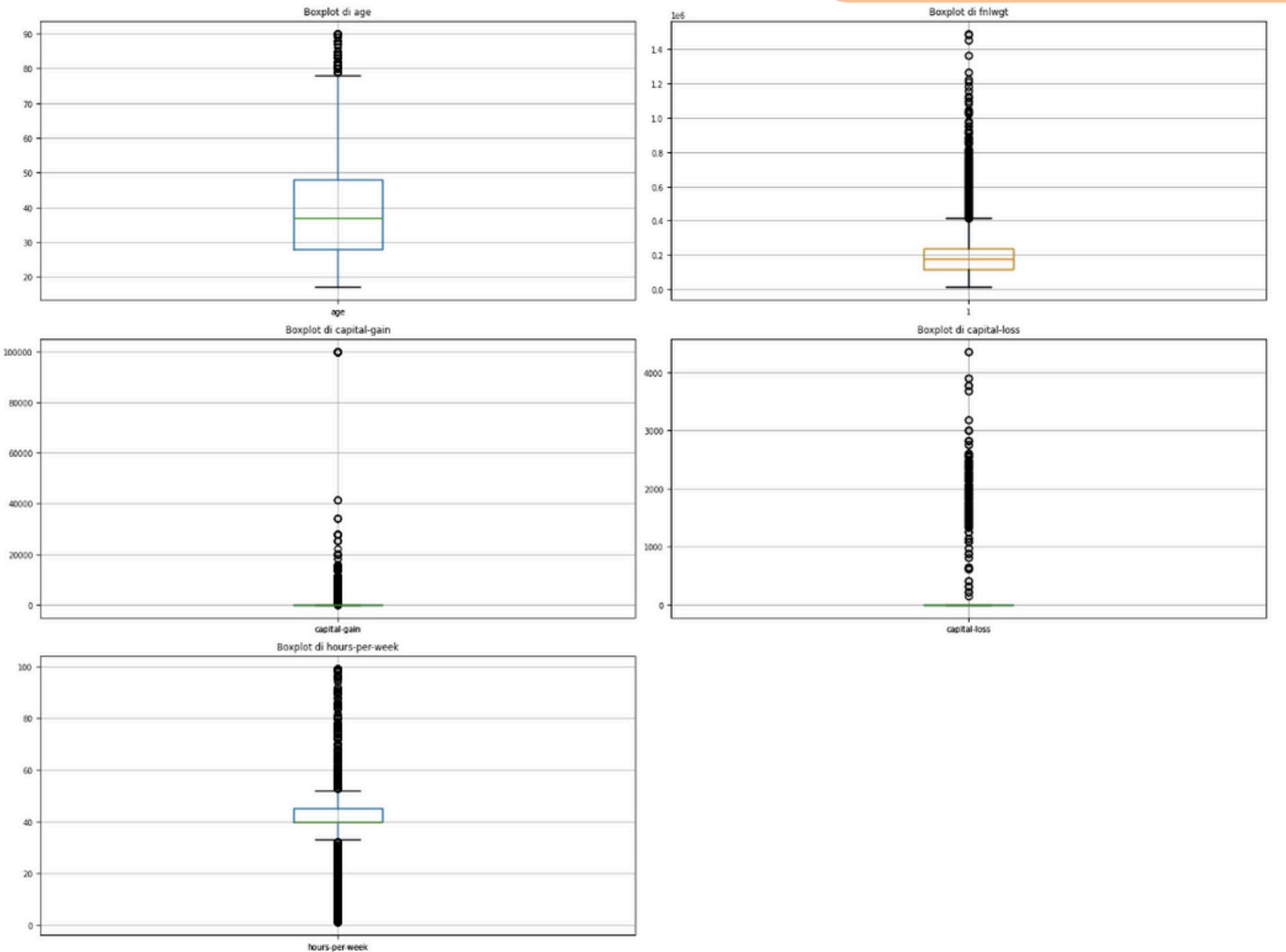


## CHECK FOR UNDERSTANDING

- la fascia di età più rappresentata: 18-50 anni
- i contratti più diffusi sono quelli di tipo privato
- il titolo di studio più comune è il diploma
- lo stato civile più comune è sposato/a
- l'etnia più comune è quella bianca
- il genere più comune è quello maschile
- il paese di provenienza più rappresentato è gli Stati Uniti

Attenzione: fnlwgt non è una feature!

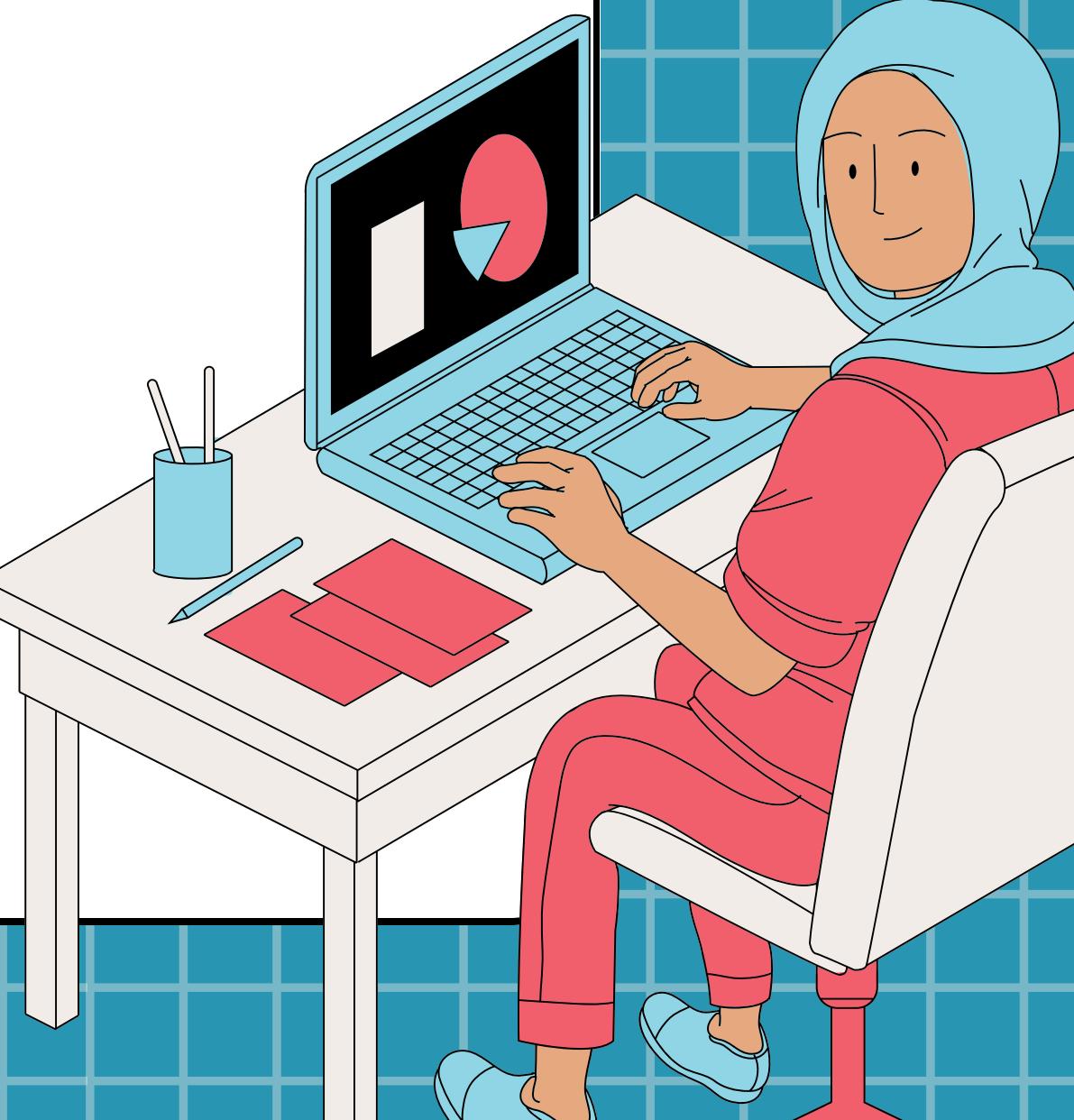
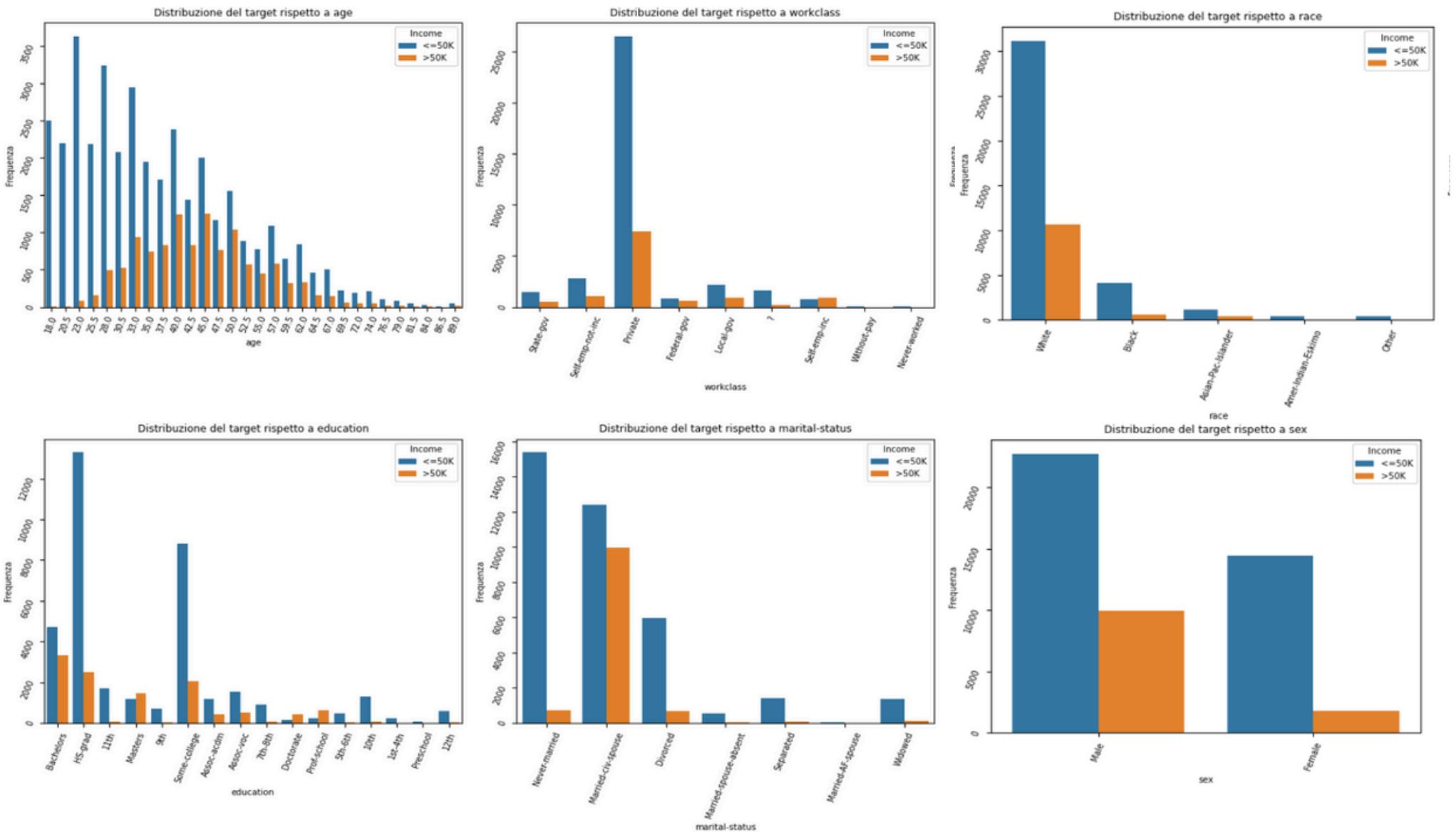
# BOXPLOT



- l'età media è 38 anni con outliers superiori
- capital loss e gain più comuni sono 0, ma esistono eccezioni molto alte
- le ore lavorative più comuni sono 40, ma esistono eccezioni sia superiori che inferiori
- fnlwgt presenta outliers superiori

Bisogna notare che il numero di outliers è molto elevato e non può essere ignorato

# DATA VISUALIZATION 2

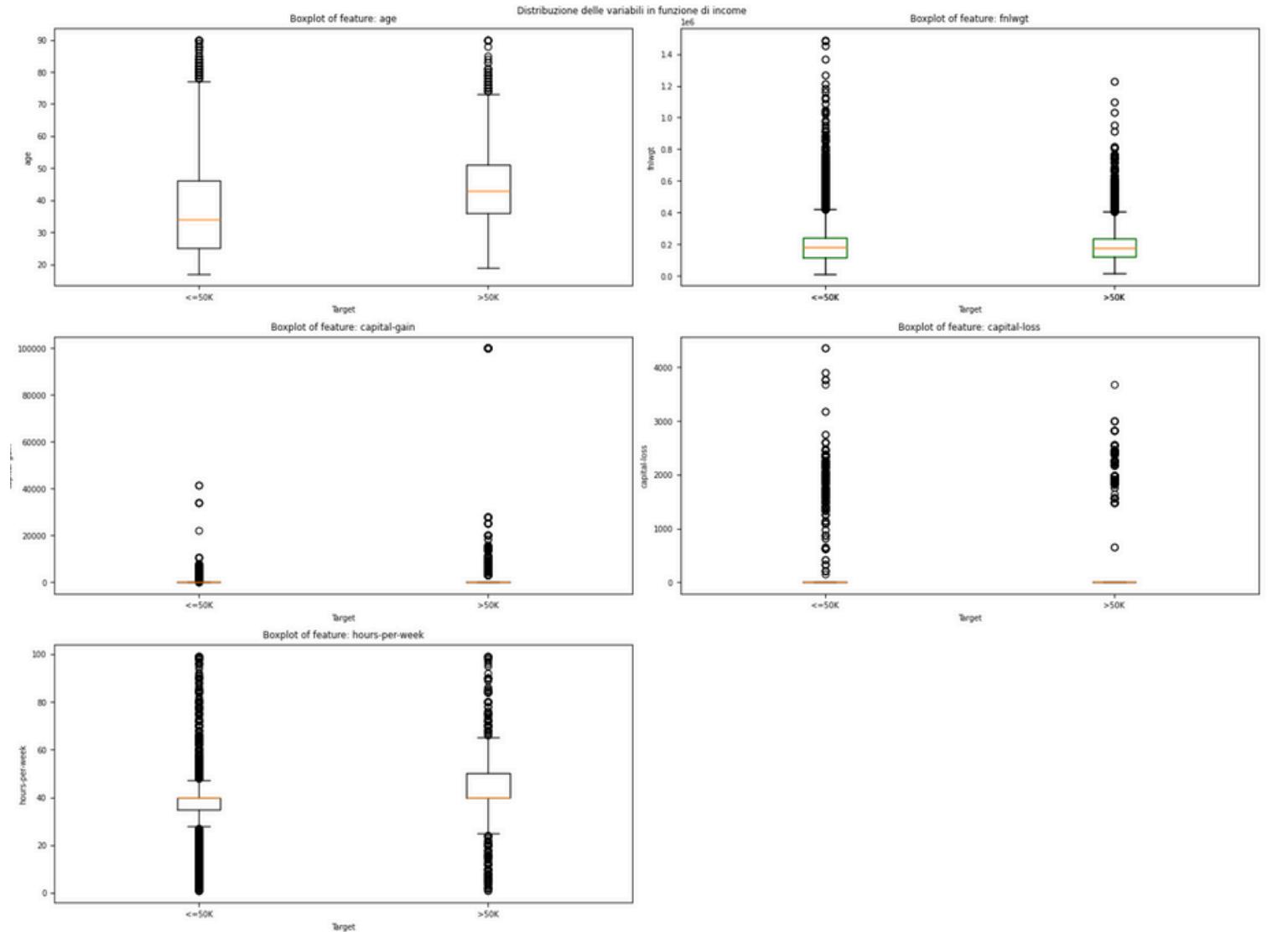


# CHECK FOR UNDERSTANDING

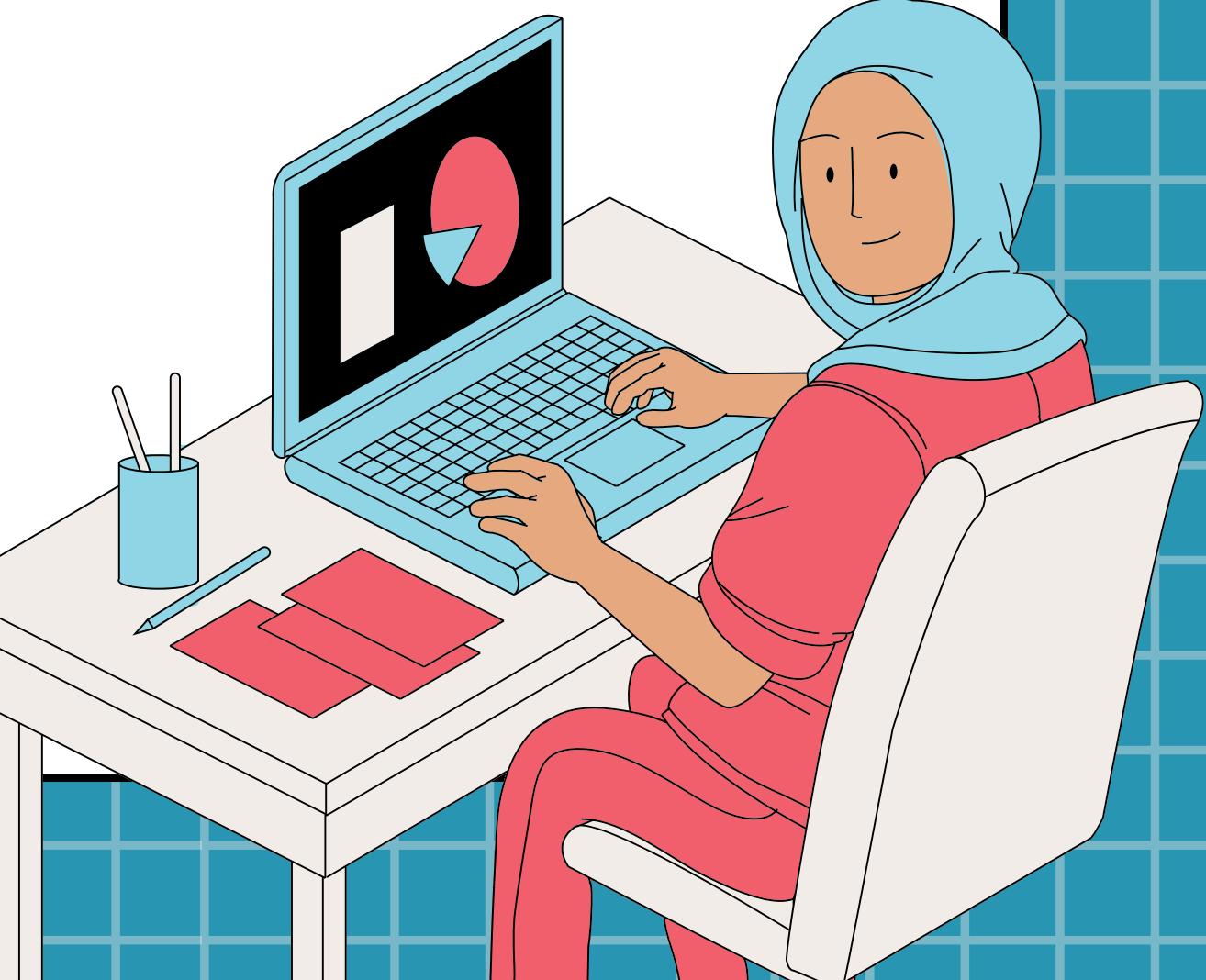
- le fasce d'età intermedie guadagnano di più
- chi lavora in proprio guadagna di più
- l'etnia contribuisce
- il titolo di studio ha un peso
- l'essere sposati è un fattore
- il genere sessuale determina il guadagno



# BOXPLOT 2



- chi guadagna di più ha un'età media più alta
- chi guadagna di più tende a lavorare più ore
- fnlwgt è equamente distribuito
- i valori più comuni di capital gain e loss è 0



# PREPROCESSING

Si sono susseguite molte fasi:

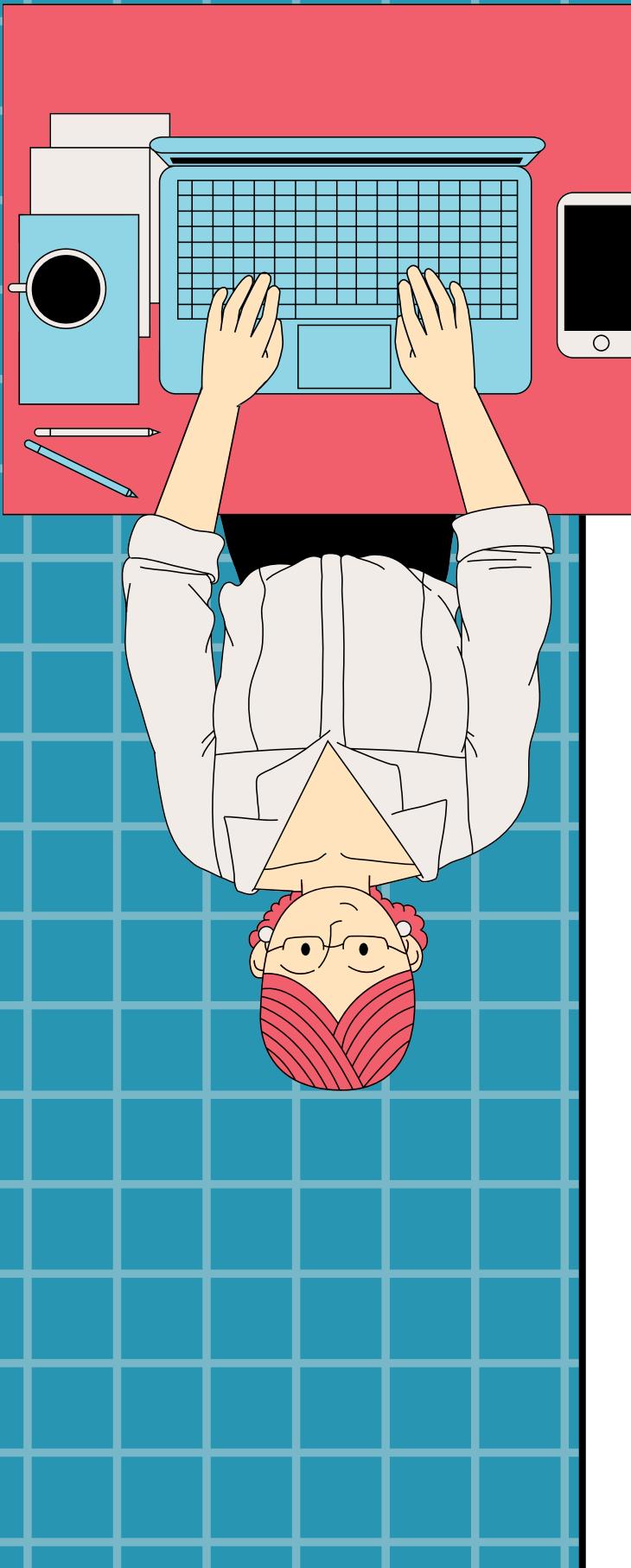
- valori duplicati
- valori null
- colonne 'education' e 'education-num'
- codifica dei generi
- accorpamento valori del target
- fusione di capital loss e capital gain
- gestione fnlwgt
- gestione outliers



# ALCUNI ASPETTI CHIAVE

- notazione differente dei valori mancanti all'interno del dataset individuata
- differente gestione dei valori null:
  - eliminazione per native-country
  - aggiunta dei valori 'never worked'
  - imputazione con unknown per workclass e occupation





## ALTRO SUL PREPROCESSING

Gli altri passaggi ridurre il più possibile il dataset senza di fatto perdere informazioni:

- education e education-num contengono le stesse informazioni
- capital loss e capital gain sono state fuse
- fnlwgt è stato estratto come peso, anzichè come feature
- outliers non eliminati

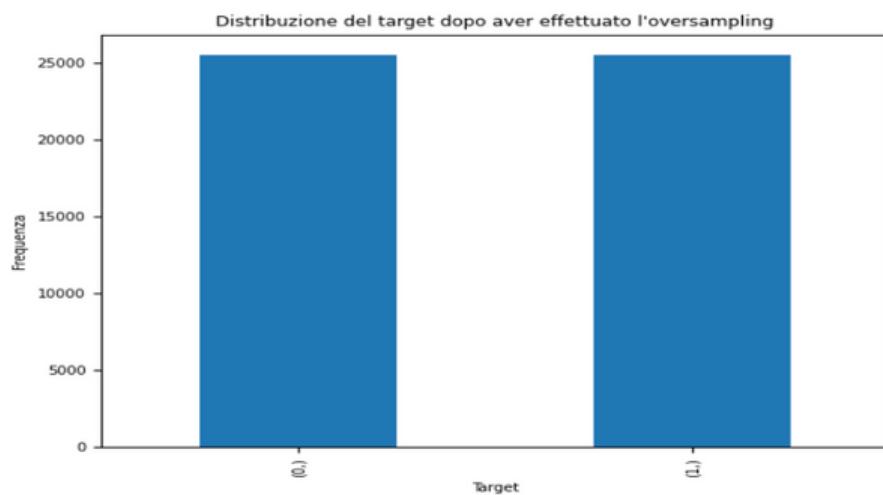
# FEATURE ENGINEERING

Fasi:

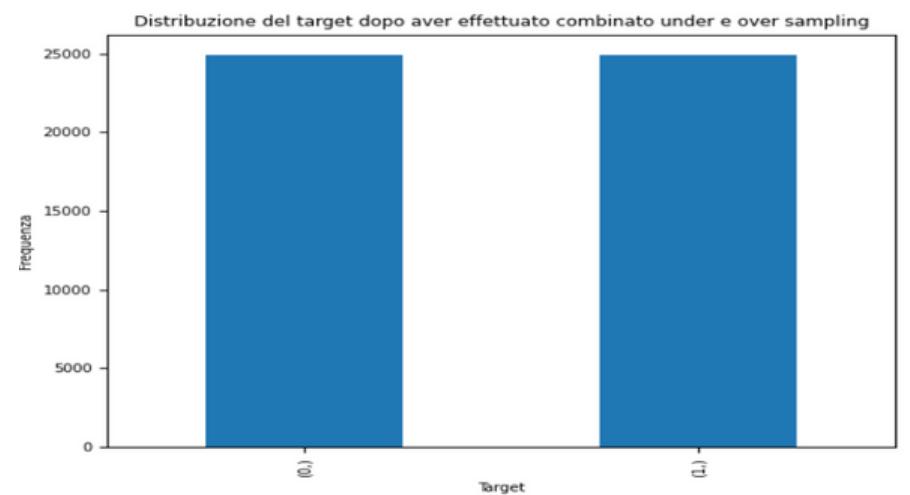
- one hot encoding
- standardizzazione features
- split in train e test
- bilanciamento dataset



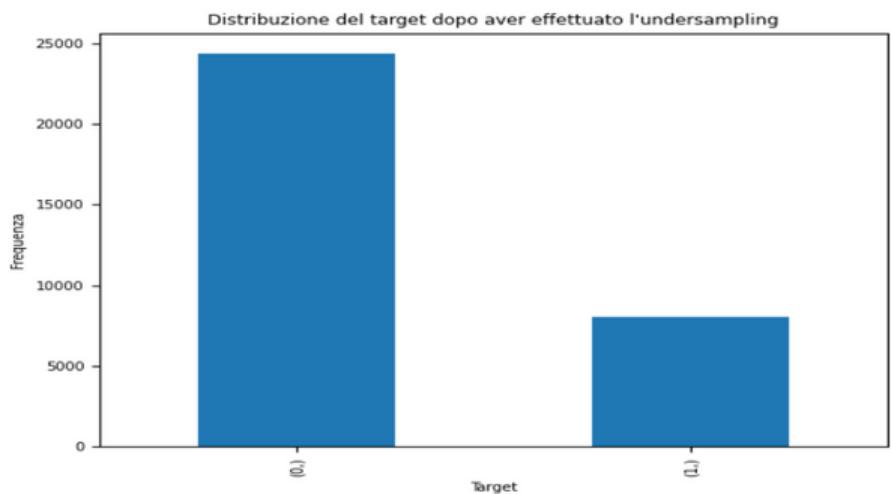
# BILANCIAMENTO DATASET



0: 25518  
1: 25518



0: 24907  
1: 24907



0: 24393  
1: 8038

# MODELLI

## Alberi decisionali

Perchè:

- facili da interpretare
- velocità di addestramento

Esperimenti:

1. dataset liscio, criterio Gini
2. dataset bilanciato, criterio Gini
3. dataset liscio, criterio Entropy
4. dataset bilanciato, criterio Entropy

## Support Vector Machines

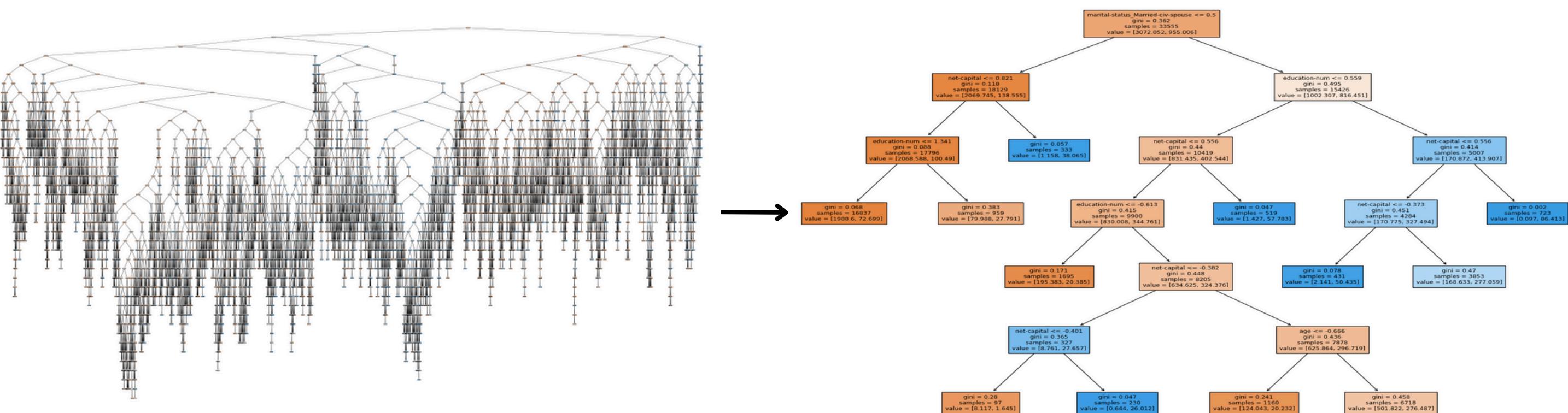
Perchè:

- adatti a problemi ad alta dimensionalità
- resistenti agli outliers

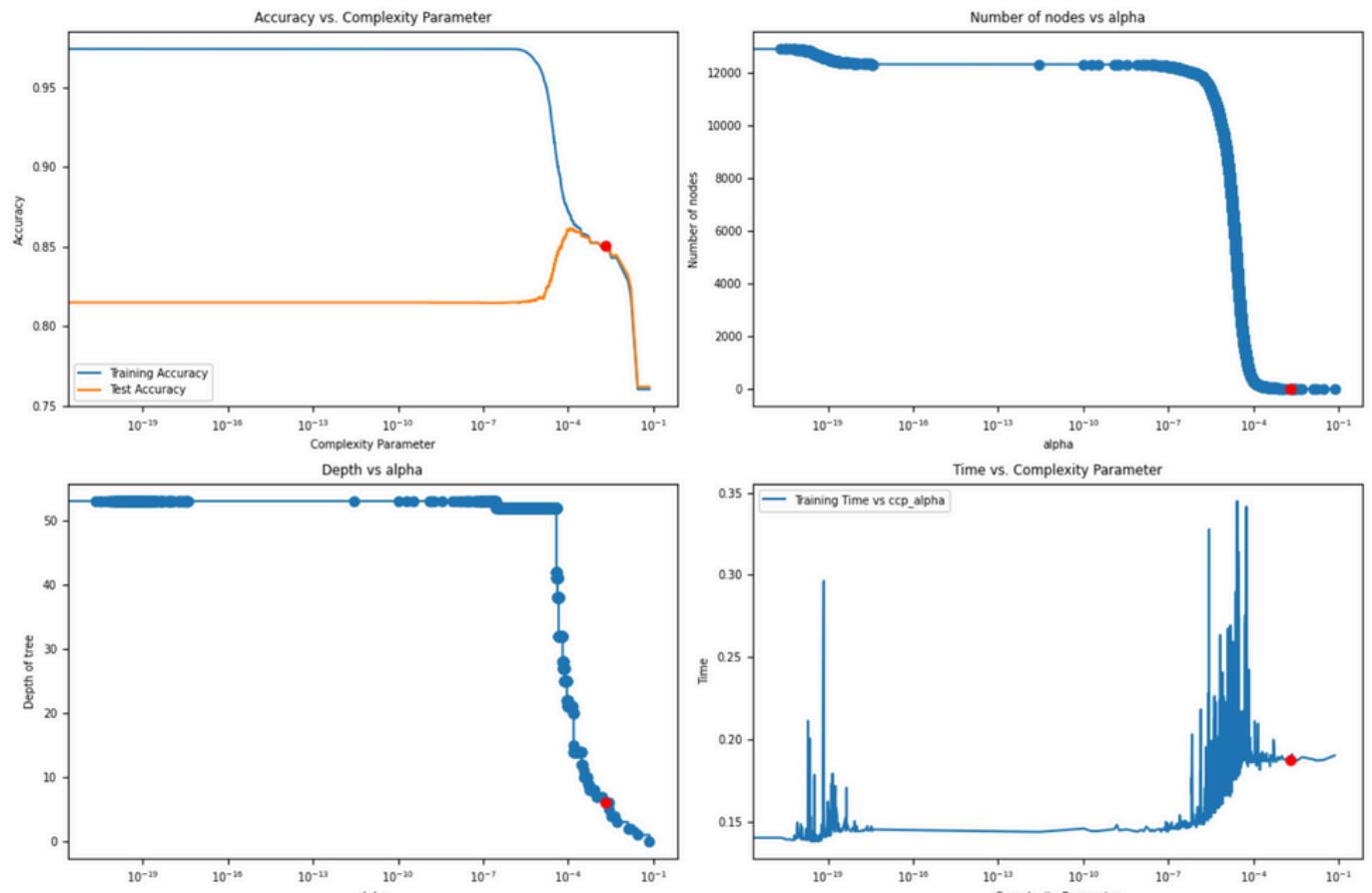
Esperimenti:

- $C=1$ ,  $\text{kernel}=\text{'linear'}$ , dataset liscio
- $C=1$ ,  $\text{kernel}=\text{'linear'}$ , dataset bilanciato
- $C=1$ ,  $\text{kernel}=\text{'rbf'}$ , dataset bilanciato

# DECISON TREE - EXP 1

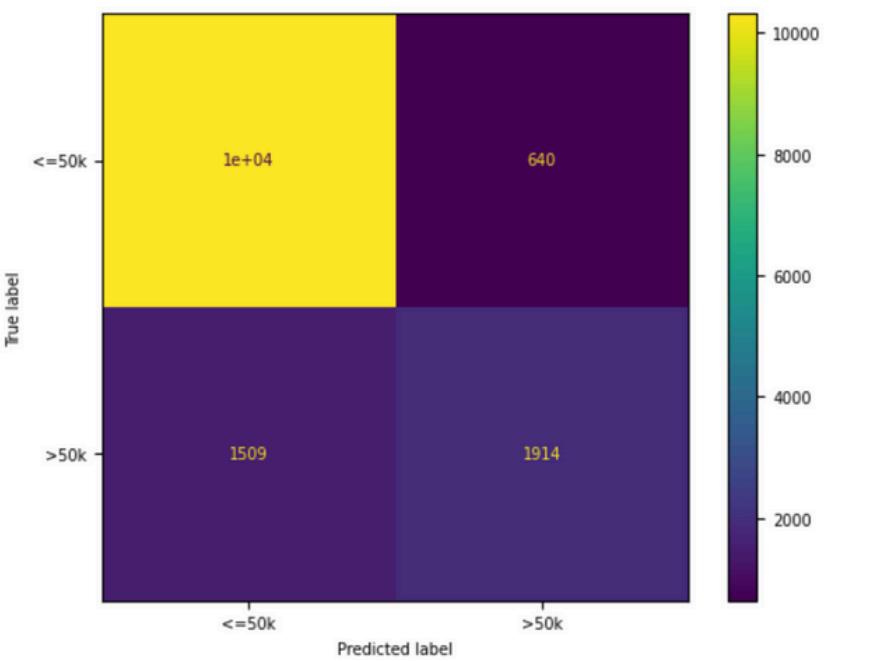


# DECISON TREE - EXP 1

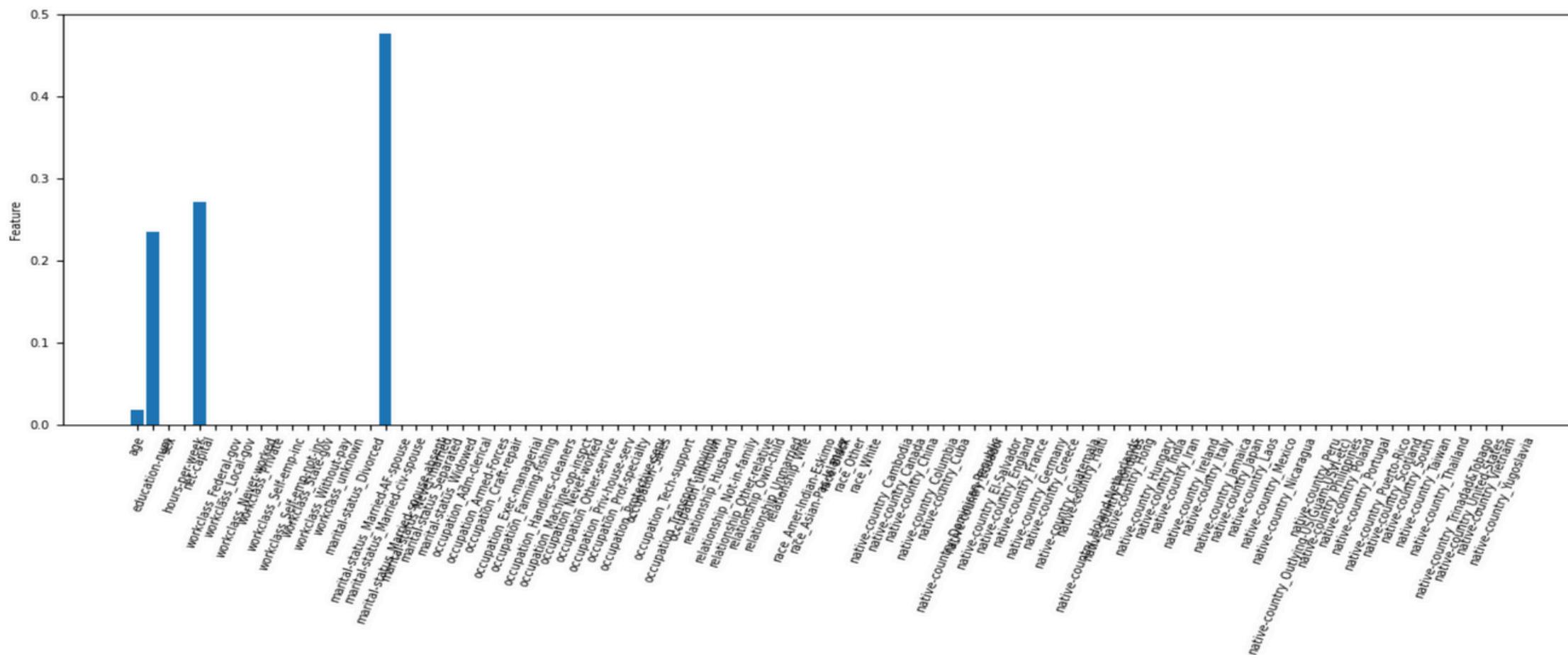


Training accuracy: 0,84    Test accuracy: 0,85  
 Nodi: 23    Profondità: 6    Tempo di addestramento: 0,19s

	precision	recall	f1-score	support
$\leq 50k$	0.87	0.94	0.91	10959
$> 50K$	0.75	0.56	0.64	3423
accuracy			0.85	14382
macro avg	0.81	0.75	0.77	14382
weighted avg	0.84	0.85	0.84	14382



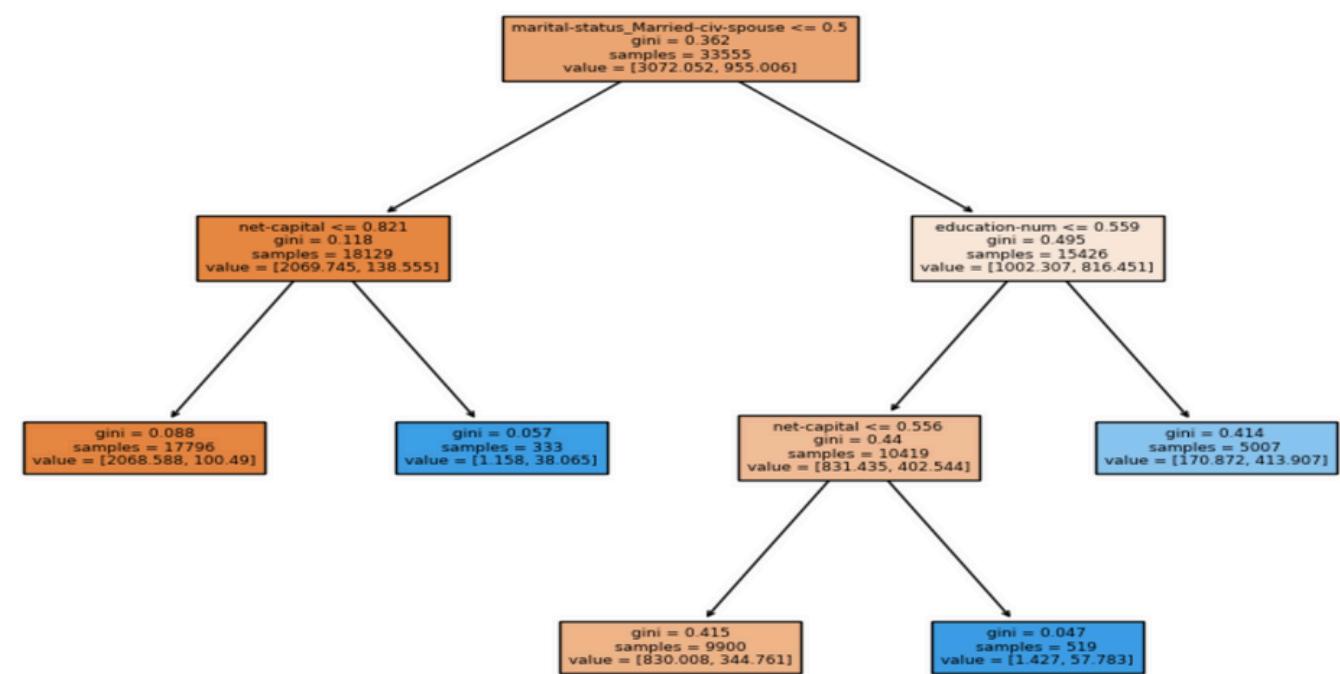
# DECISON TREE - EXP 1



Feature più importanti:

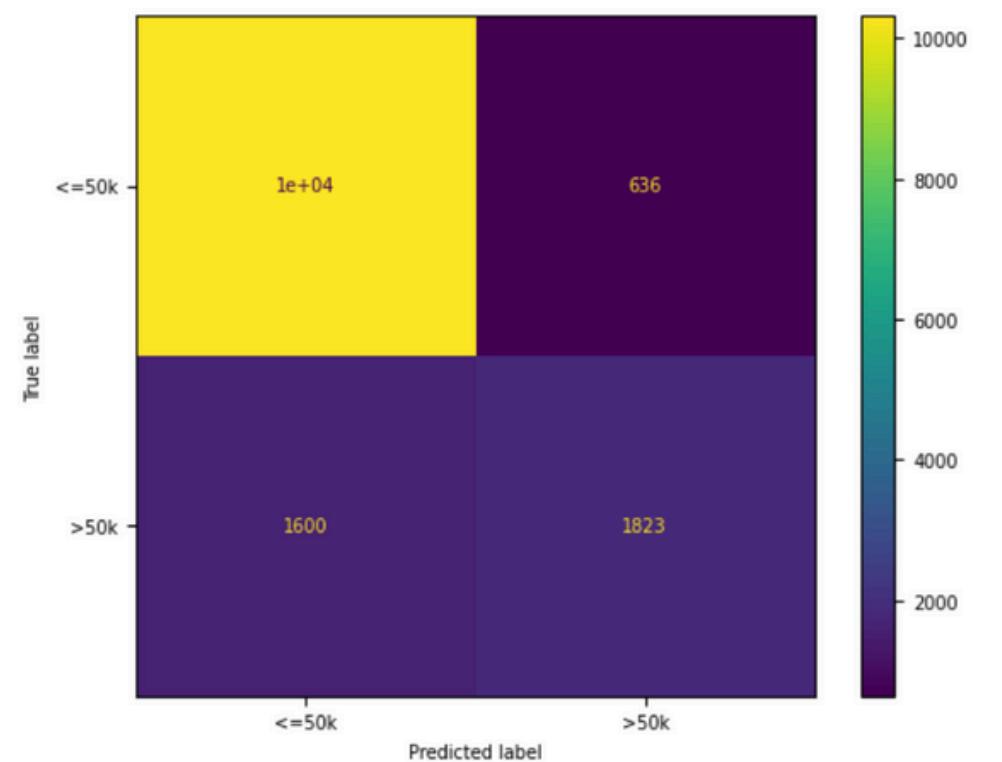
- marital status married civ spouse: 0,47
- net capital: 0,27
- education num: 0,23
- age: 0,01
- native country France: 0,00

# EXTRA



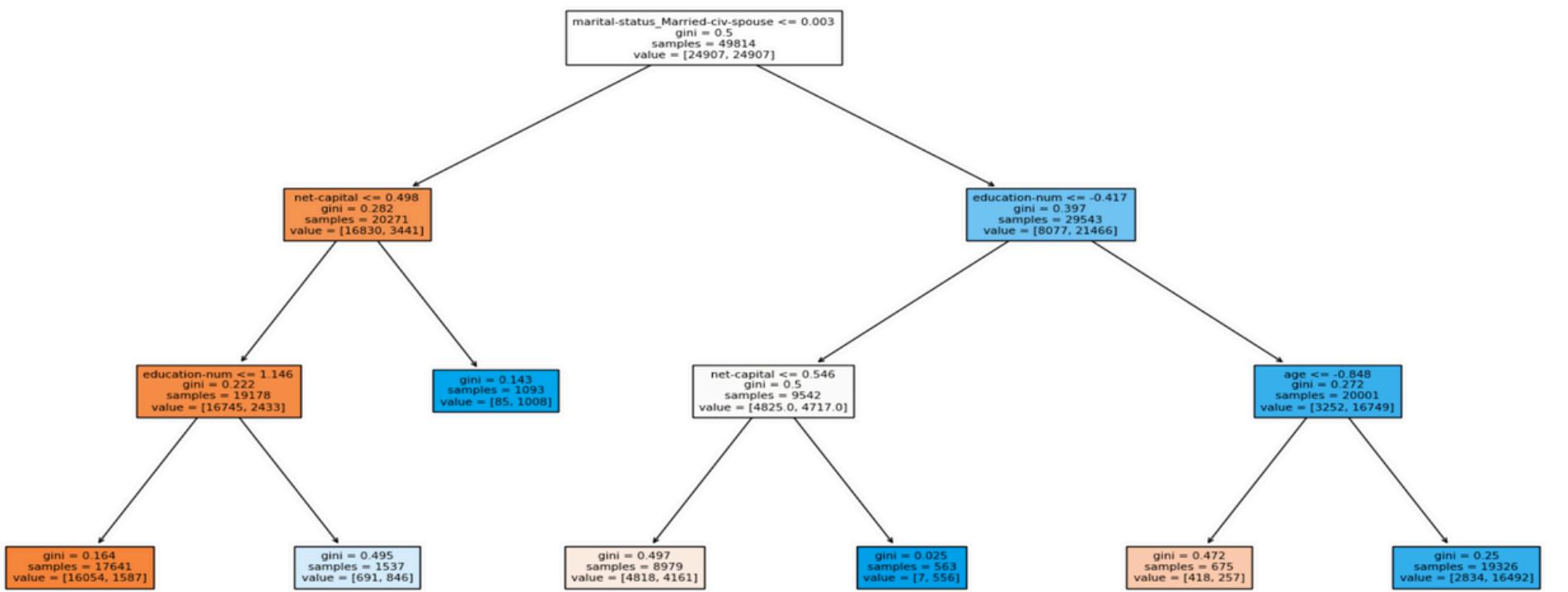
Feature più importanti:  
 marital-status-Married-civ-spouse: 0.56,  
 net-capital: 0.22,  
 education-num: 0.21,  
 age: 0, native-country-France: 0

Esperimento extra: albero extra small



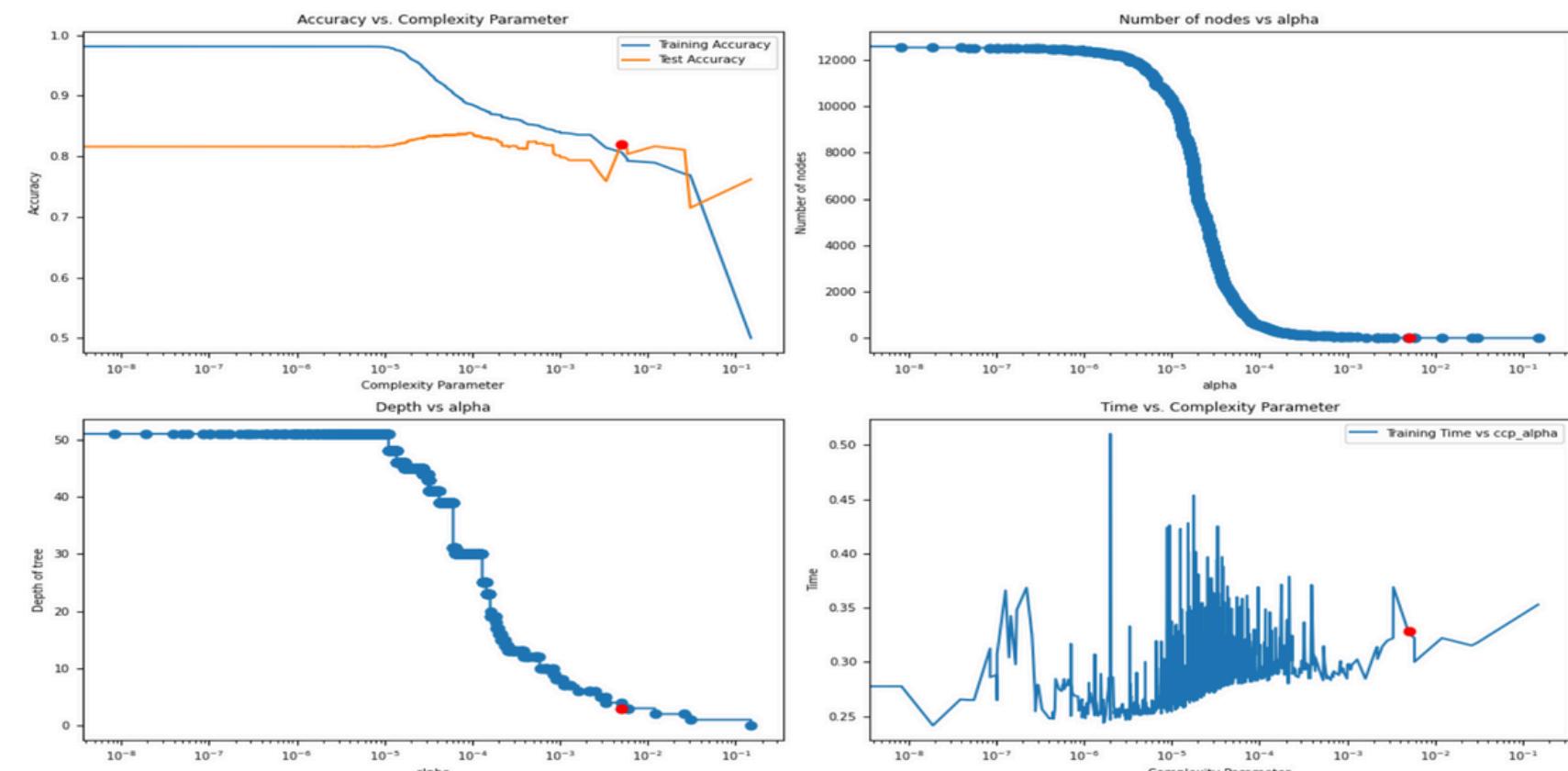
Vengono estremizzati i risultati dell'albero precedente!

# DECISON TREE - EXP 2

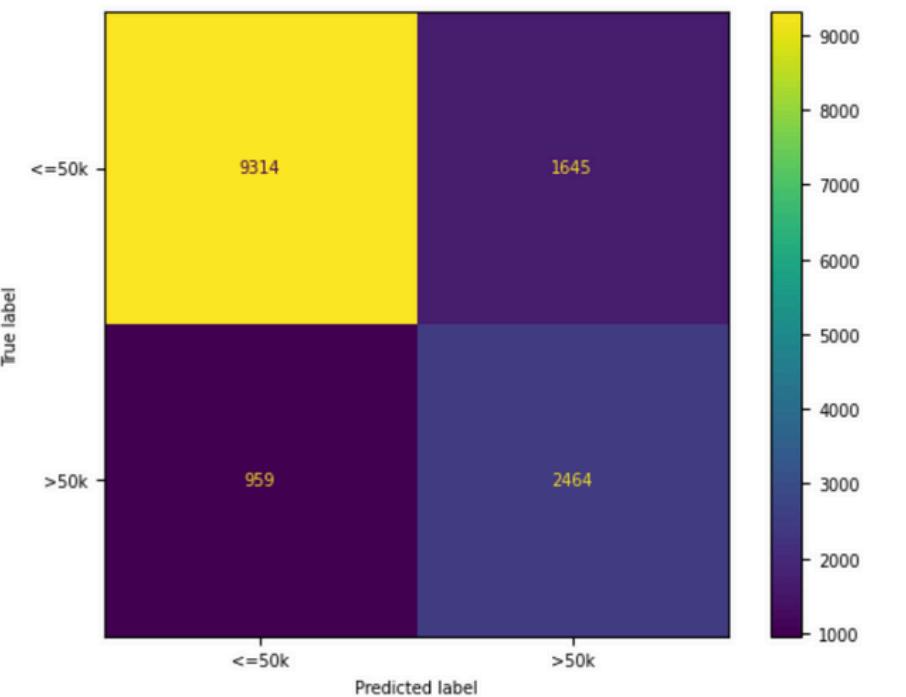


Esperimento 2: dataset bilanciato e criterio Gini

# DECISON TREE - EXP 2

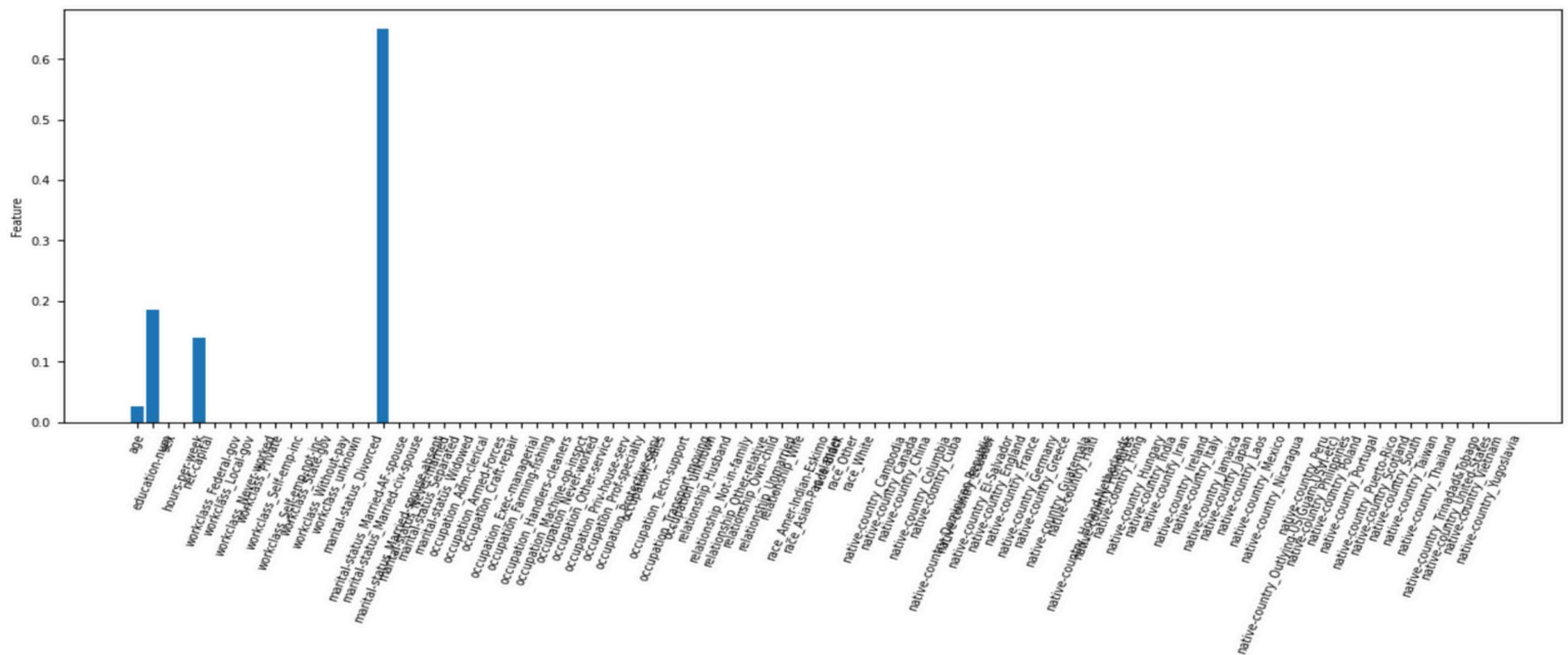


	precision	recall	f1-score	support
$\leq 50k$	0.91	0.85	0.88	10959
$> 50K$	0.60	0.72	0.65	3423
accuracy			0.82	14382
macro avg	0.75	0.78	0.77	14382
weighted avg	0.83	0.82	0.82	14382



Training accuracy: 0,806    Test accuracy: 0,818  
 Nodi: 13    Profondità: 3    Tempo di addestramento: 0,28s

# DECISON TREE - EXP 2

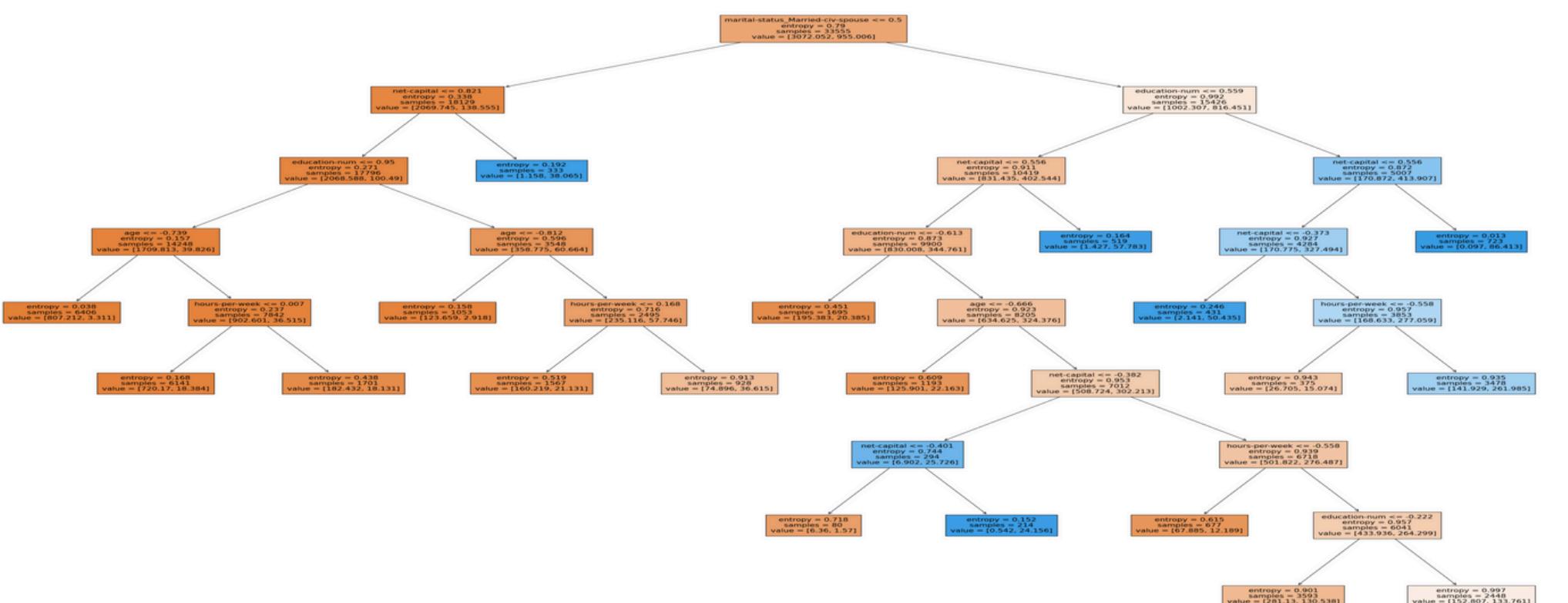


Features:

- marital-status-Married-civ-spouse: 0.65
- education-num: 0.18
- net-capital: 0.14
- age: 0.02
- native-country-France 0.0

# DECISON TREE - EXP 3

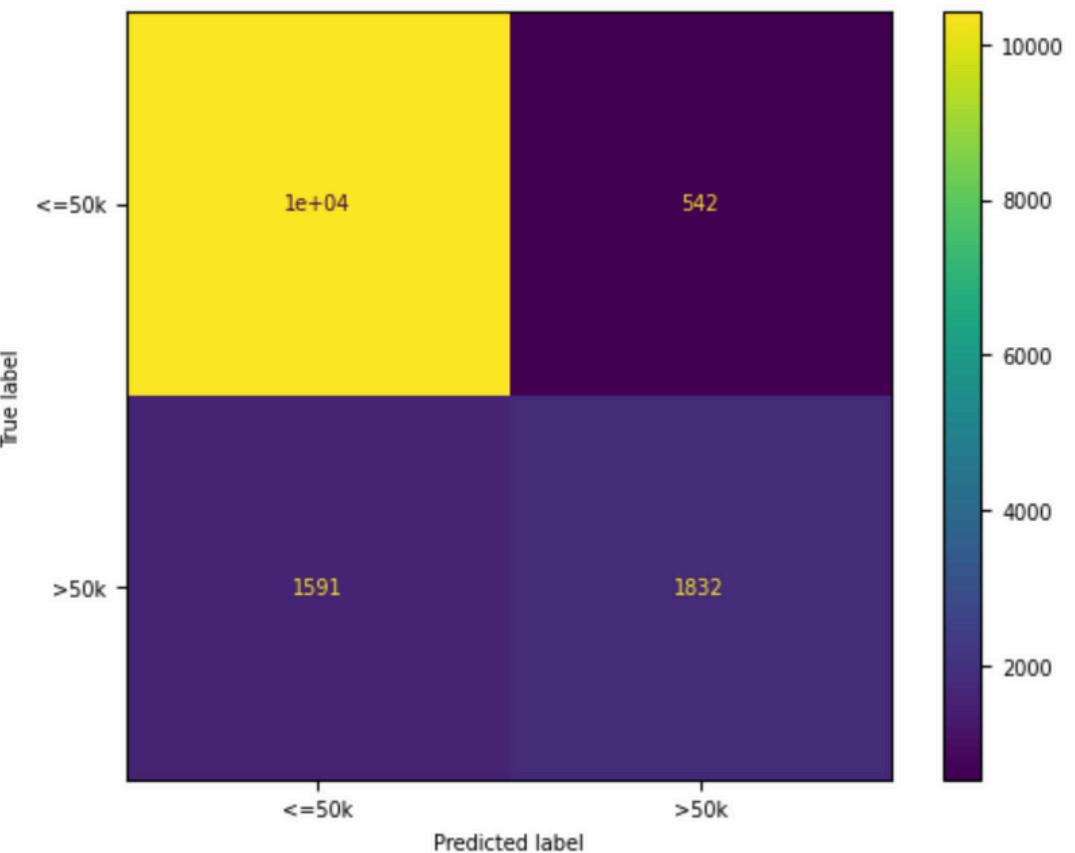
Esperimento 3: dataset liscio e criterio Entropy



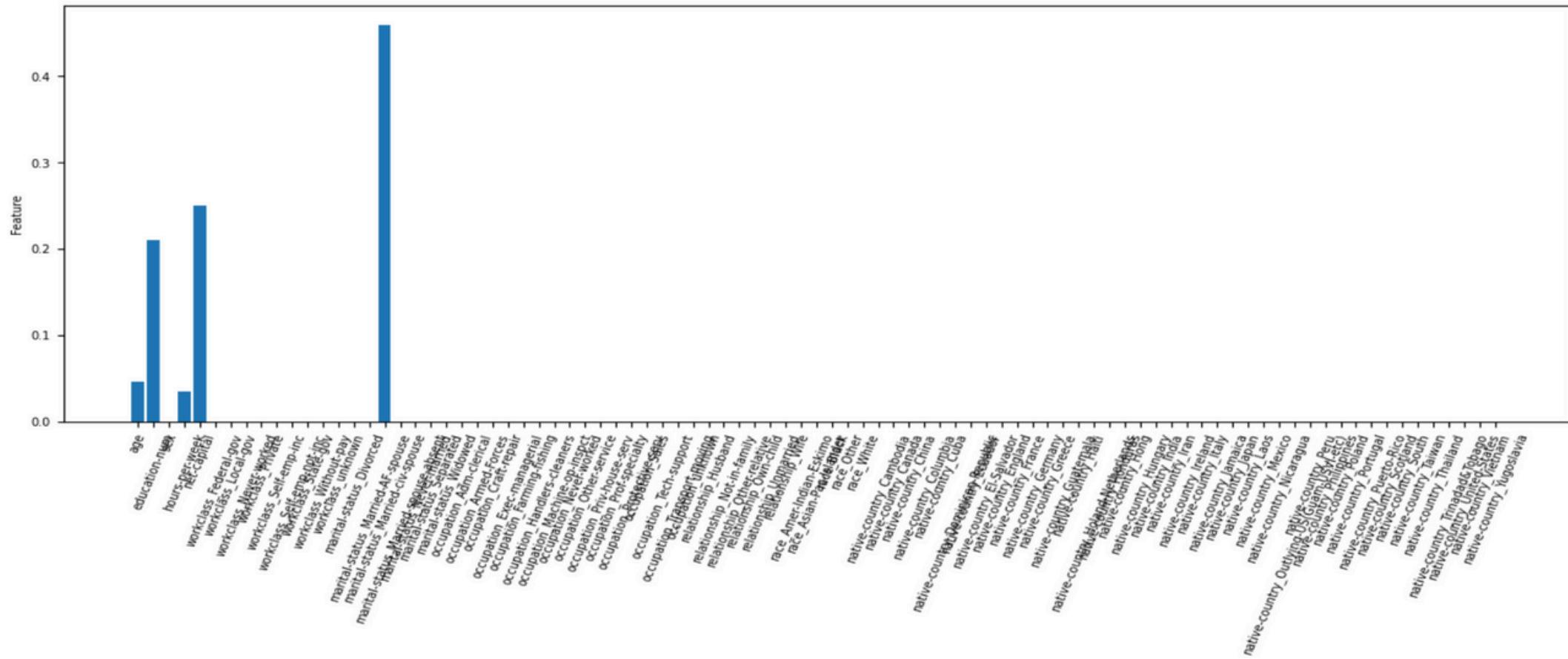
# DECISON TREE - EXP 3

	precision	recall	f1-score	support
$\leq 50k$	0.87	0.95	0.91	10959
$> 50K$	0.77	0.54	0.63	342
accuracy			0.85	14382
macro avg	0.82	0.74	0.77	14382
weighted avg	0.84	0.85	0.84	14382

(I dettagli quali tempo di addestramento, numero nodi, profondità ecc sono gli stessi dell'esperimento 1)



# DECISON TREE - EXP 3

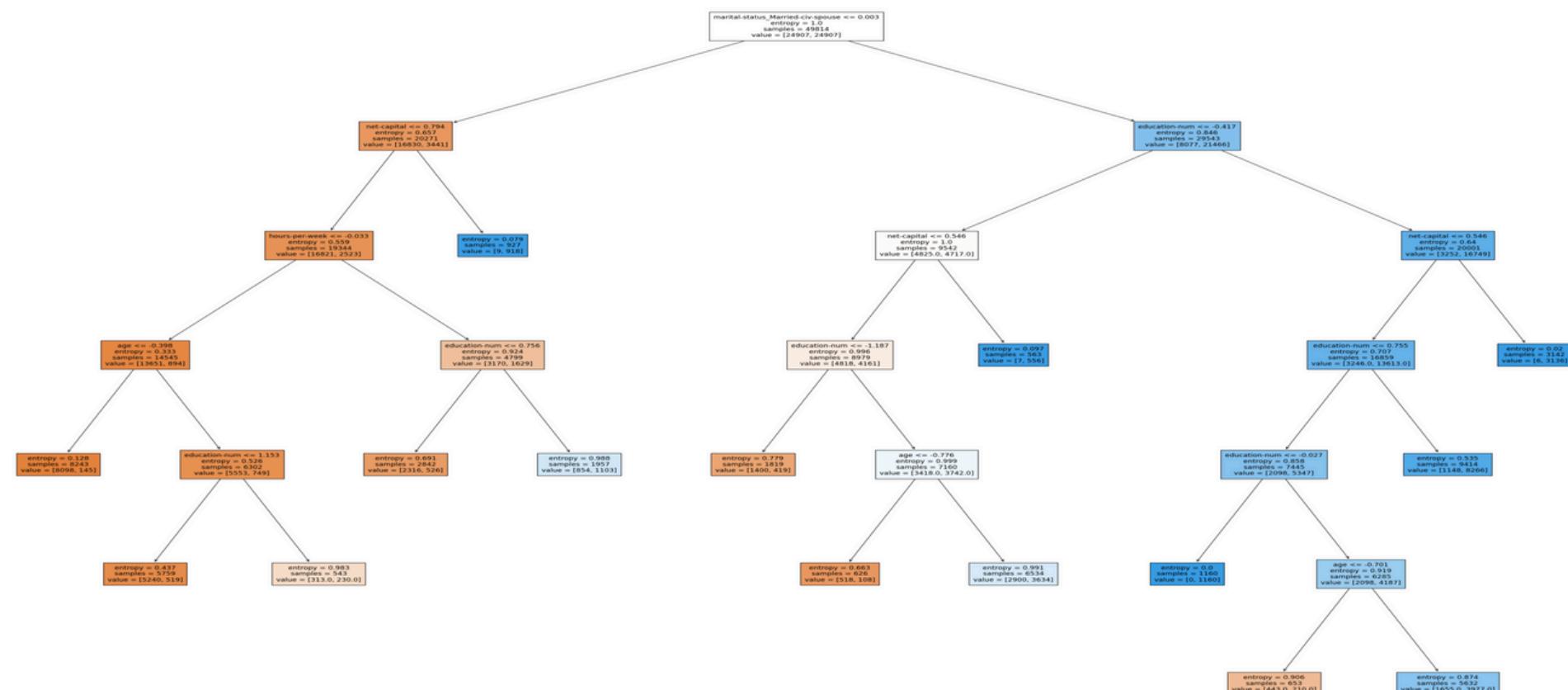


Features:

- marital-status-Married-civ-spouse: 0.46
  - net-capital: 0.25
- education-num: 0.21
  - age: 0.05
- hours-per-week 0.03

# DECISON TREE - EXP 4

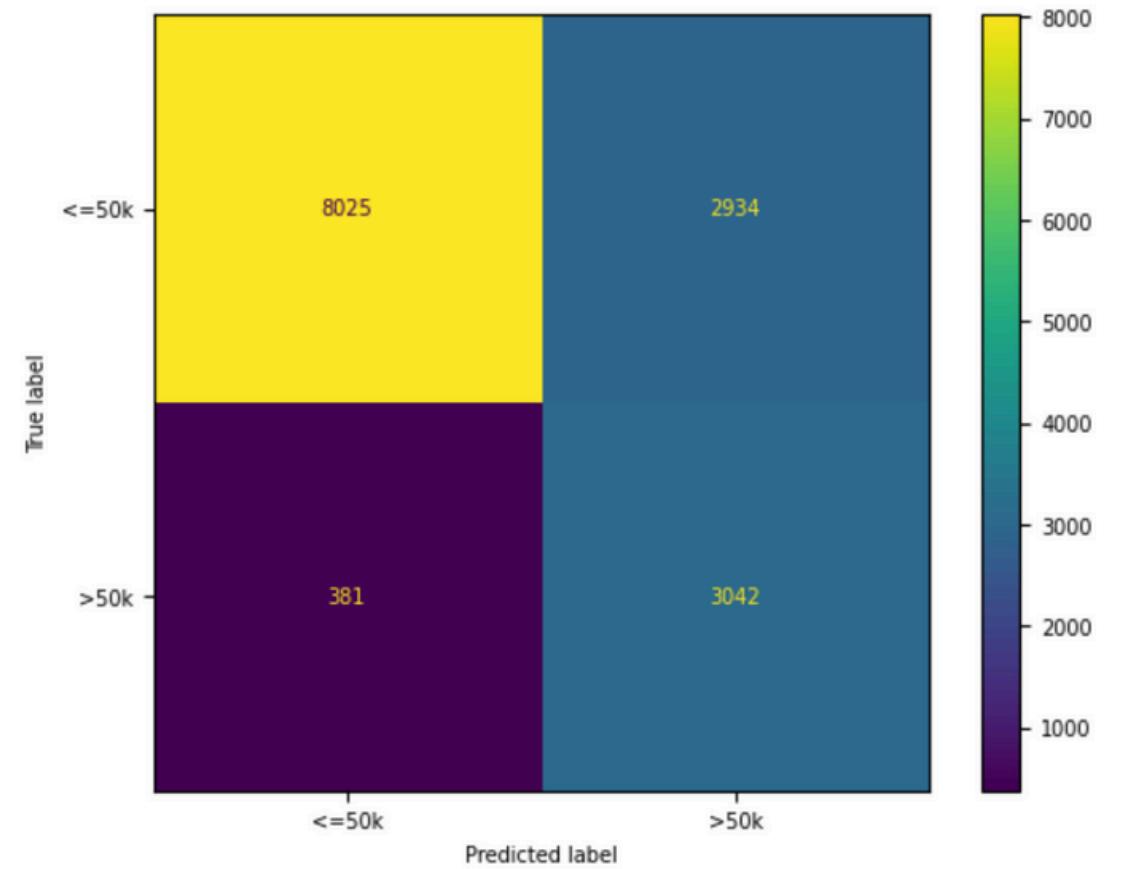
Esperimento 4: dataset bilanciato e criterio Entropy



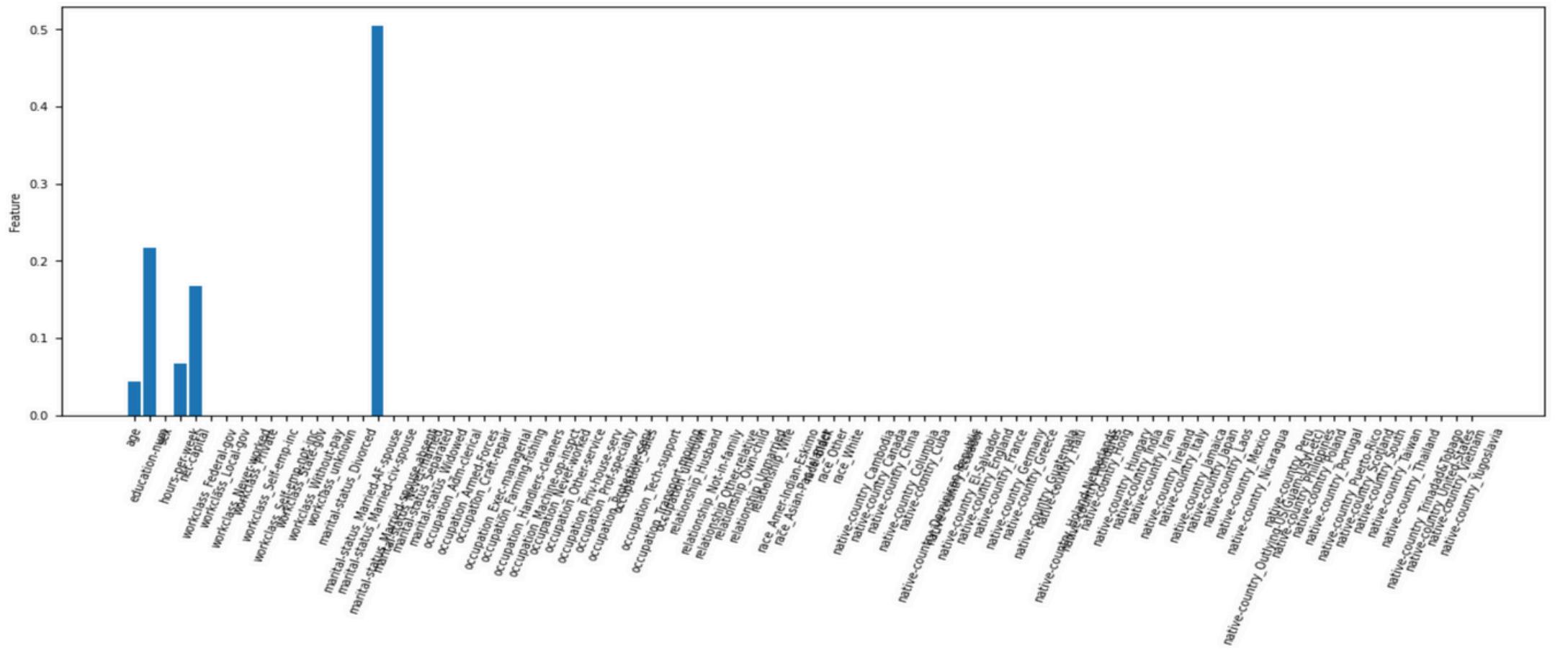
# DECISION TREE - EXP 4

	precision	recall	f1-score	support
$\leq 50k$	0.95	0.73	0.83	10959
$> 50K$	0.51	0.89	0.65	3423
accuracy			0.77	14382
macro avg	0.73	0.81	0.74	14382
weighted avg	0.85	0.77	0.79	14382

(I dettagli quali tempo di addestramento, numero nodi, profondità ecc sono gli stessi dell'esperimento 2)



# DECISION TREE - EXP 4



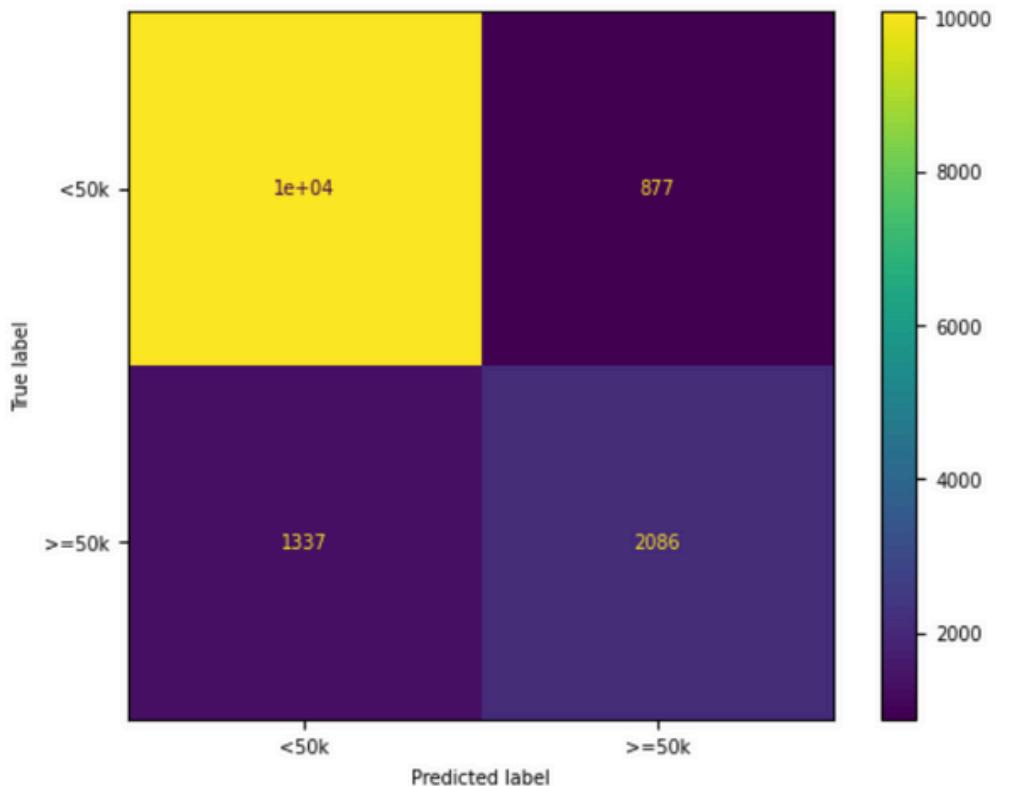
features:

- marital-status-Married-civ-spouse: 0.50
    - education-num: 0.21
      - net-capital: 0.17
    - hours-per-week 0.06
      - age: 0.04

# SVM - EXP 1

	precision	recall	f1-score	support
$\leq 50k$	0.88	0.92	0.90	10959
$> 50K$	0.71	0.61	0.65	3423
accuracy			0.85	14382
macro avg	0.79	0.76	0.78	14382
weighted avg	0.84	0.85	0.84	14382

Esperimento 1 con kernel lineare, C=1 e dataset liscio

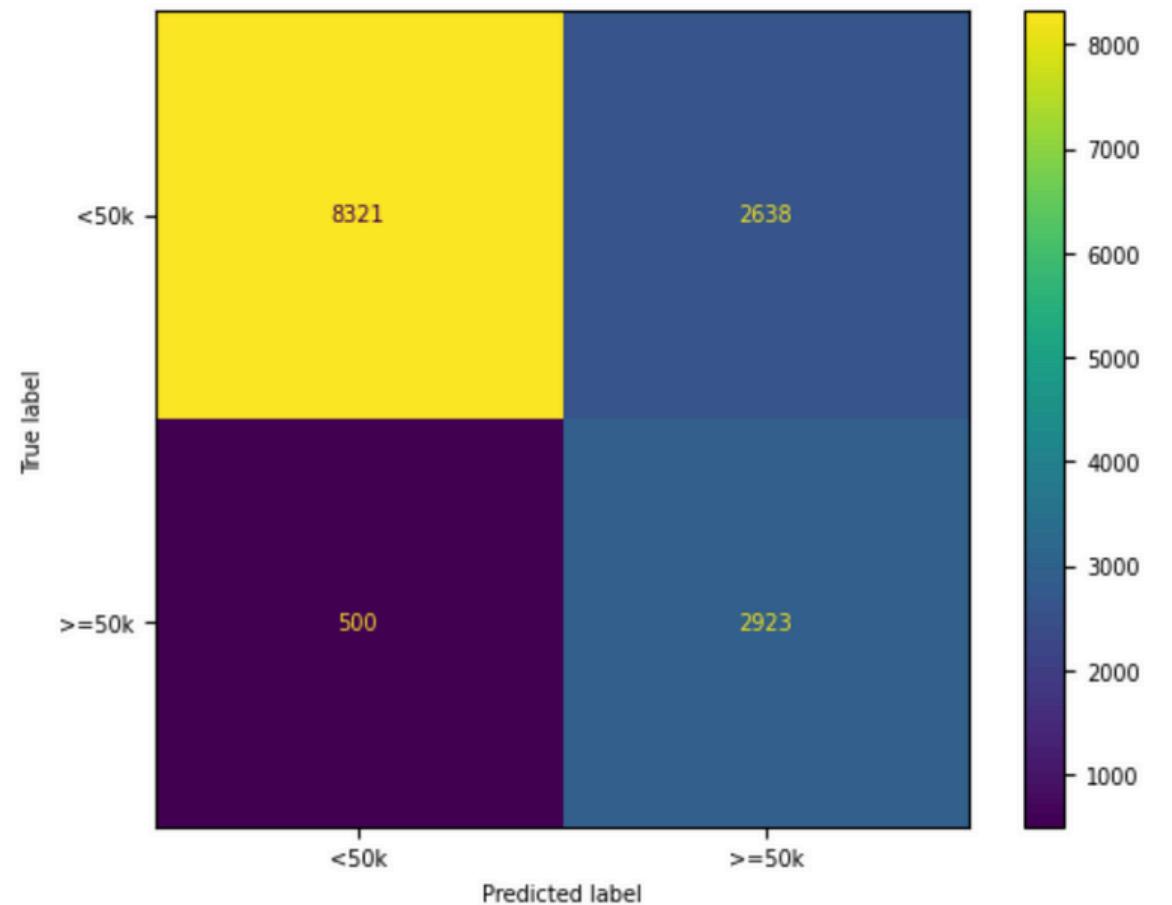


circa 30s di addestramento

# SVM - EXP 2

	precision	recall	f1-score	support
$\leq 50k$	0.94	0.76	0.84	10959
$> 50K$	0.53	0.85	0.65	3423
accuracy			0.78	14382
macro avg	0.73	0.81	0.75	14382
weighted avg	0.84	0.78	0.80	14382

Esperimento 2 con kernel lineare, C=1 e dataset bilanciato

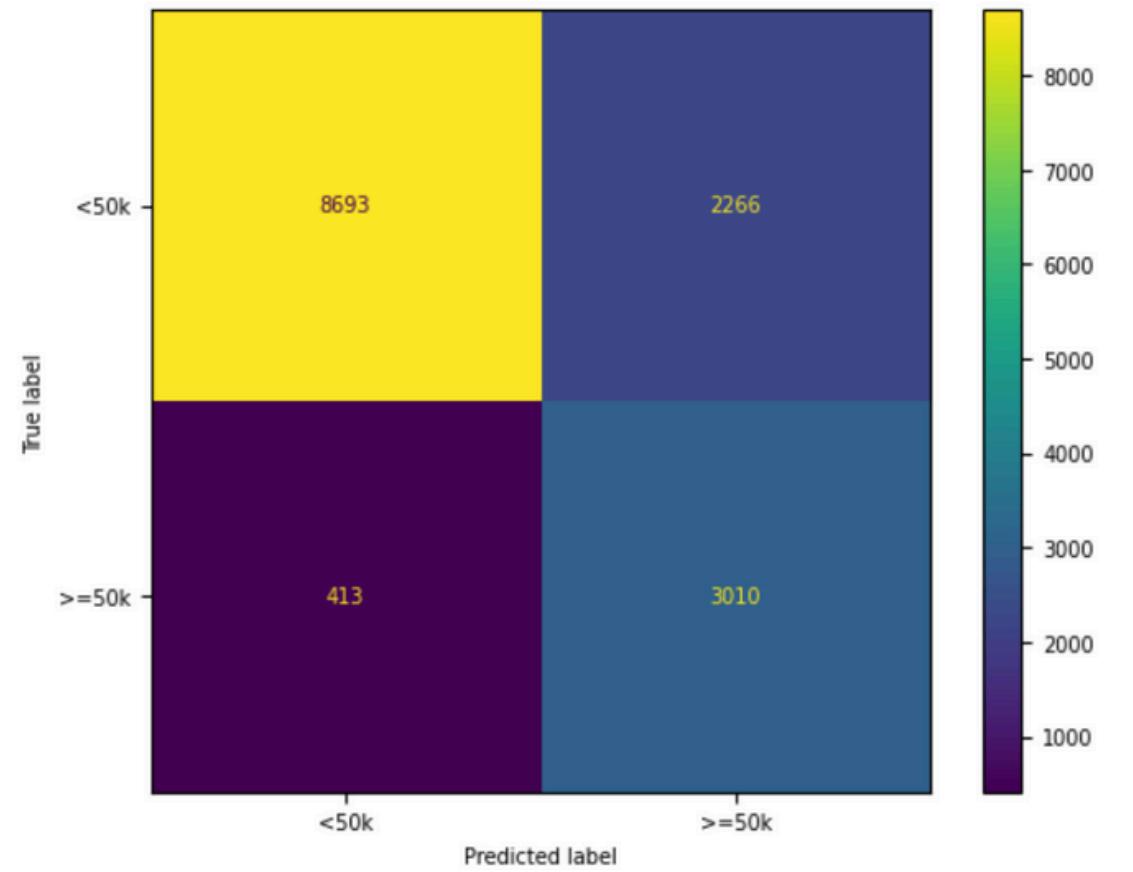


circa 1min e 30s di addestramento

# SVM - EXP 3

	precision	recall	f1-score	support
$\leq 50k$	0.95	0.79	0.87	10959
$> 50K$	0.57	0.88	0.69	3423
accuracy				14382
macro avg	0.76	0.84	0.78	14382
weighted avg	0.86	0.81	0.82	14382

Esperimento 3 con kernel radiale, C=1 e dataset bilanciato



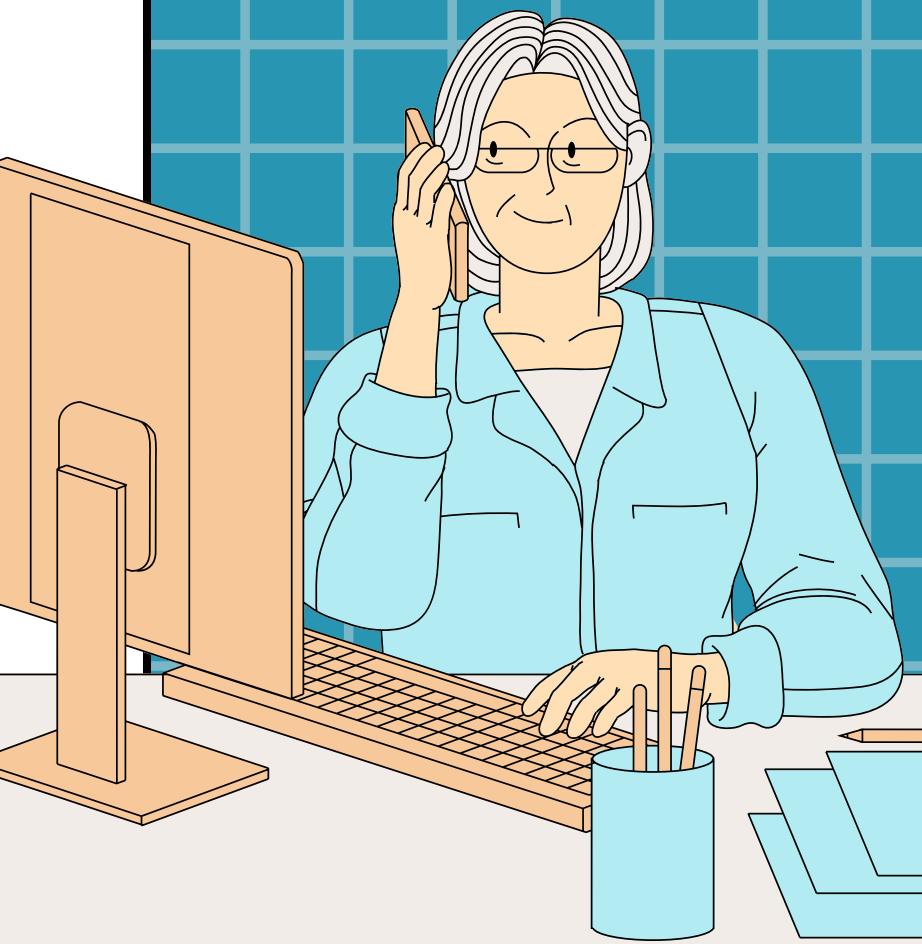
circa 1min e 30s di addestramento

# VALIDAZIONE ESPERIMENTI

La tecnica usata è Repeated K-fold perchè più adatta per dataset sbilanciati.

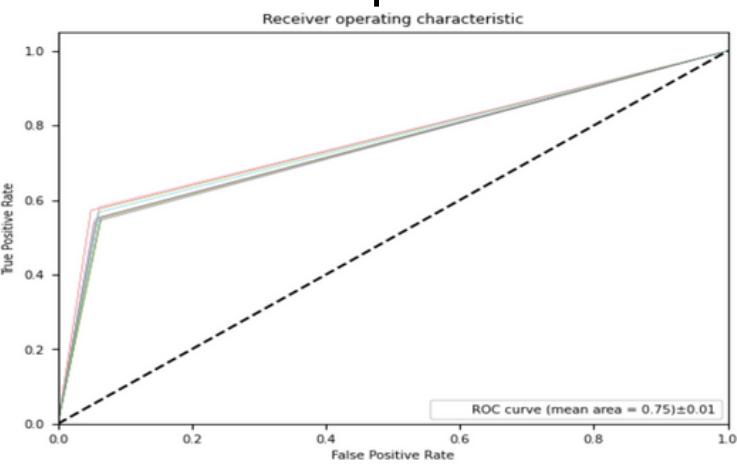
Sono stati creati intervalli di confidenza per ogni fold e su quello sono stati calcolati nuovi report.

Infine per ogni fold di ogni esperimento è stata creata la curva ROC.



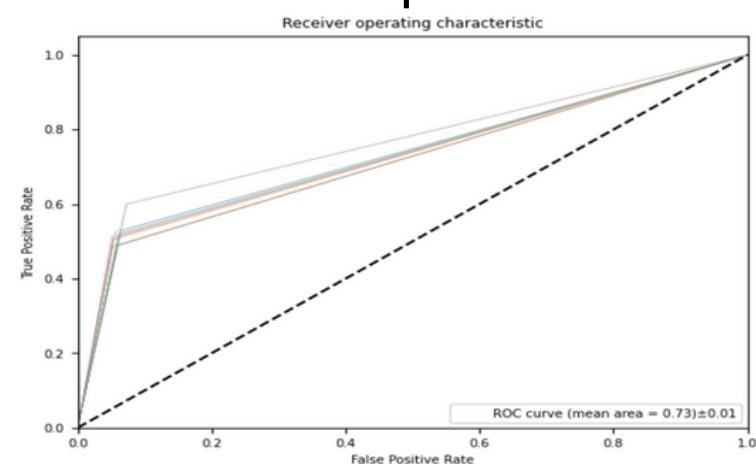
# TECNICHE DI VALIDAZIONE DECISION TREES

Exp 1



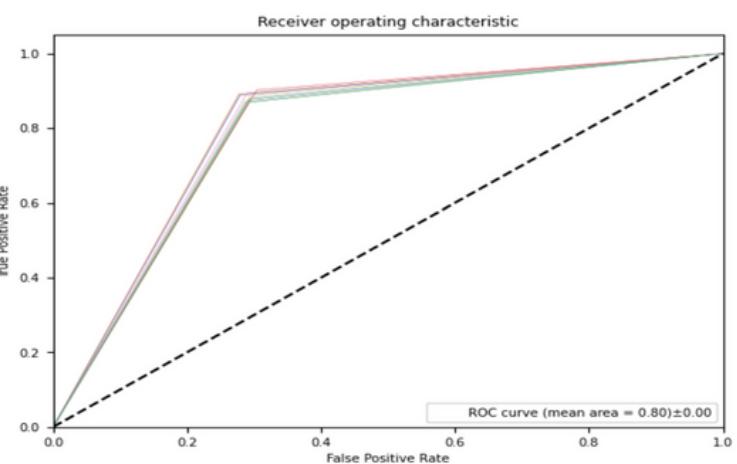
Accuracy: (0.846, 0.853)  
AUC score:  $0.75 \pm 0.01$

Exp 3



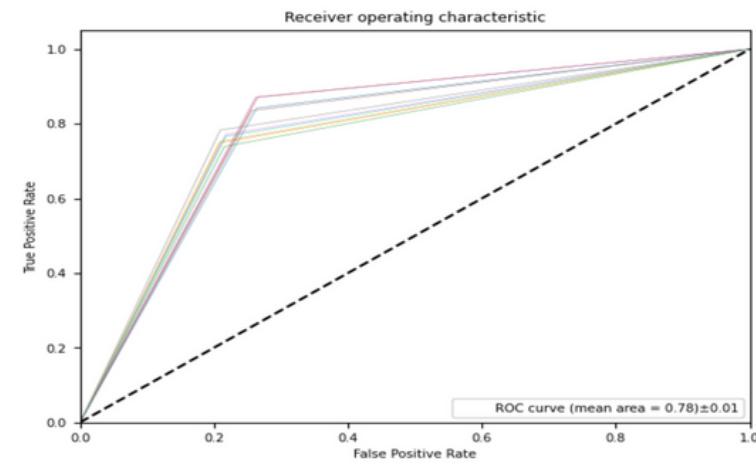
Accuracy: (0.847, 0.855)  
AUC score:  $0.73 \pm 0.01$

Exp 2



Accuracy: (0.743, 0.775)  
AUC score:  $0.75 \pm 0.01$

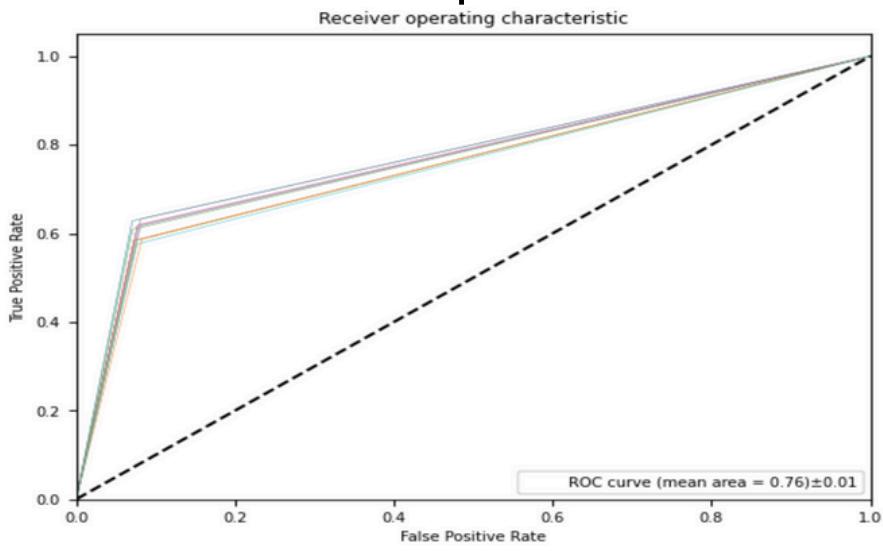
Exp 4



Accuracy: (0.753, 0.769)  
AUC score:  $0.78 \pm 0.01$

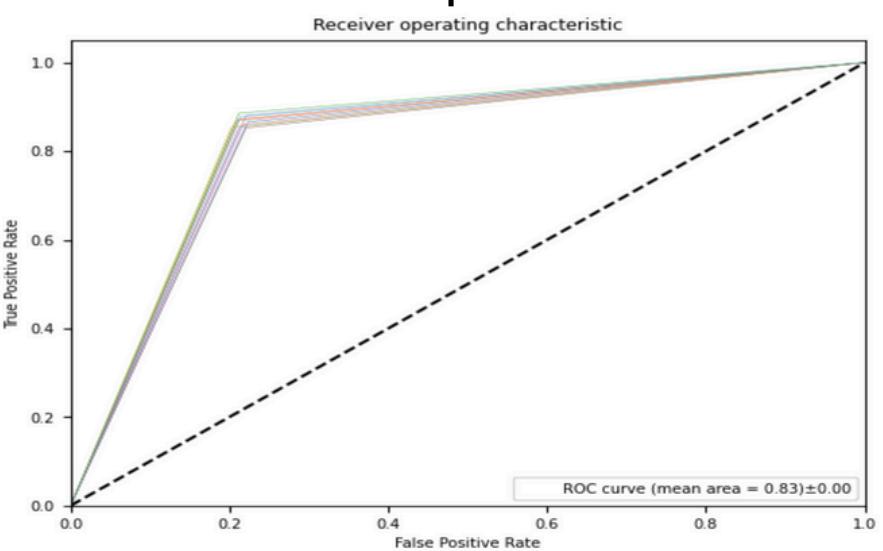
# TECNICHE DI VALIDAZIONE SVM

Exp 1



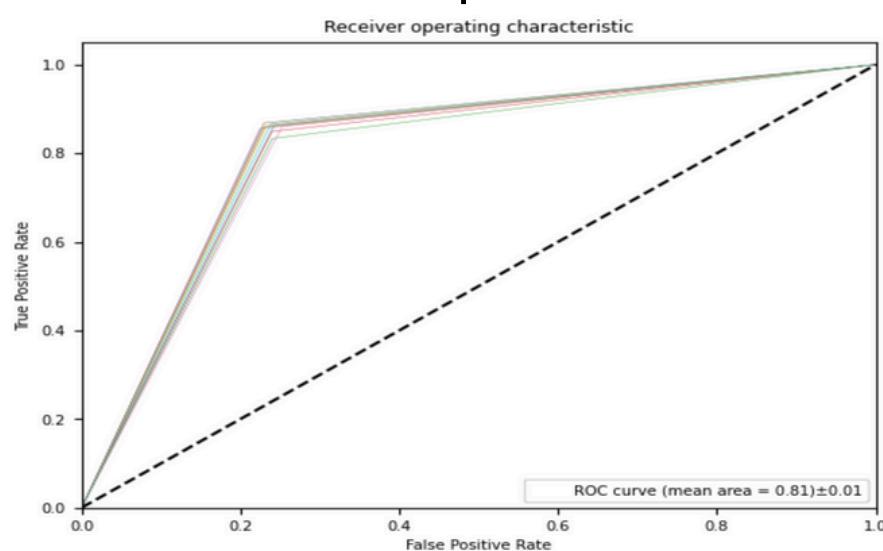
Accuracy: (0.843, 0.852)  
AUC score:  $0.76 \pm 0.01$

Exp 3



Accuracy: (0.801, 0.809)  
AUC score:  $0.83 \pm 0.00$

Exp 2



Accuracy: (0.781, 0.791)  
AUC score:  $0.81 \pm 0.01$

# ANALISI DEI RISULTATI

Possiamo dire che:

- i risultati ottenuti con validazione sono in linea con gli esperimenti
- introduciamo l'AUC score come metrica
- negli alberi pesa di più il bilanciamento che il criterio di splitting
- nelle SVM il bilanciamento del dataset è importante, ma anche il kernel
- fnlwgt non pare rilevante
- tempi di addestramento molto diversi
- difficoltà a sperimentare con le SVM



# CONCLUSIONI

La natura del problema è complessa.

Strade percorse:

- bilanciamento dataset,
- standardizzazione
- opportuna scelta di classificatori.

Limiti:

- non è stato possibile effettuare tutti gli esperimenti previsti
- difficoltà di predizione per un task che coinvolge fattori personali
- 

Possibilità future:

- Naïve Bayes (feature categoriche One Hot Encoded)
- Altri esperimento con SVM (grid search + varianti)



**CI SONO  
DOMANDE?**



**Nicolò Sansevrino**  
**865889**