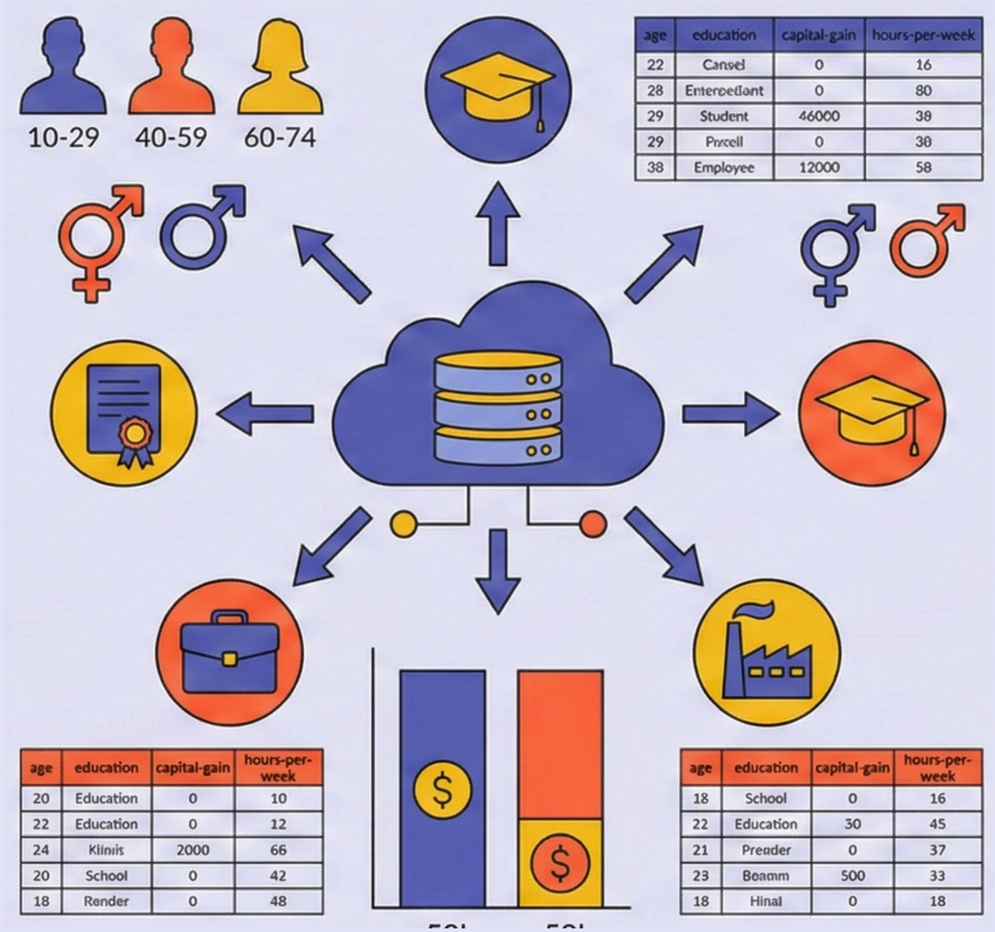
PREDICTIA VENITULUI UNEI PERSOANE

Andrei Barbu
Matei Caragea



SETUL DE DATE

1. UCI Adult Income
2. 48842 inregistrari
3. 14 feature-uri

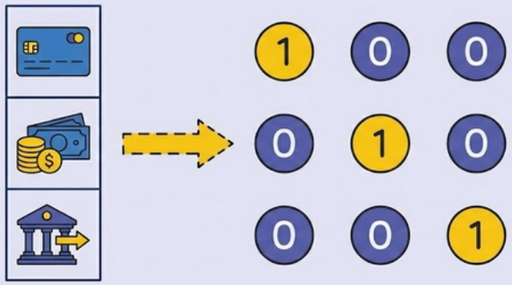
PREPROCESARE

Completarea valorilor lipsa

Simplificarea atributelor

Scalare cu StandardScaler

Transformare cu One-Hot Encoding



SETUL FINAL

1. 48842 inregistrari
2. 43 atribute

	age	education-num	sex	capital-gain	capital-loss	hours-per-week
0	0.025996	1.136512	1	2.844559	-0.221264	-0.034087
1	0.828308	1.136512	1	-0.297918	-0.221264	-2.213032
2	-0.046942	-0.419335	1	-0.297918	-0.221264	-0.034087
3	1.047121	-1.197259	1	-0.297918	-0.221264	-0.034087
4	-0.776316	1.136512	0	-0.297918	-0.221264	-0.034087

	native-country	income	workclass_Local-gov	workclass_Never-worked	...
0		1	0	False	False ...
1		1	0	False	False ...
2		1	0	False	False ...
3		1	0	False	False ...
4		0	0	False	False ...

	occupation_Transport-moving	relationship_Not-in-family
0	False	True
1	False	False
2	False	True
3	False	False
4	False	False

	relationship_Other-relative	relationship_Own-child
0	False	False
1	False	False
2	False	False
3	False	False
4	False	False

	relationship_Unmarried	relationship_Wife	race_Asian-Pac-Islander
0	False	False	False
1	False	False	False
2	False	False	False
3	False	False	False
4	False	True	False

	race_Black	race_Other	race_White
0	False	False	True
1	False	False	True
2	False	False	True
3	True	False	False
4	True	False	False

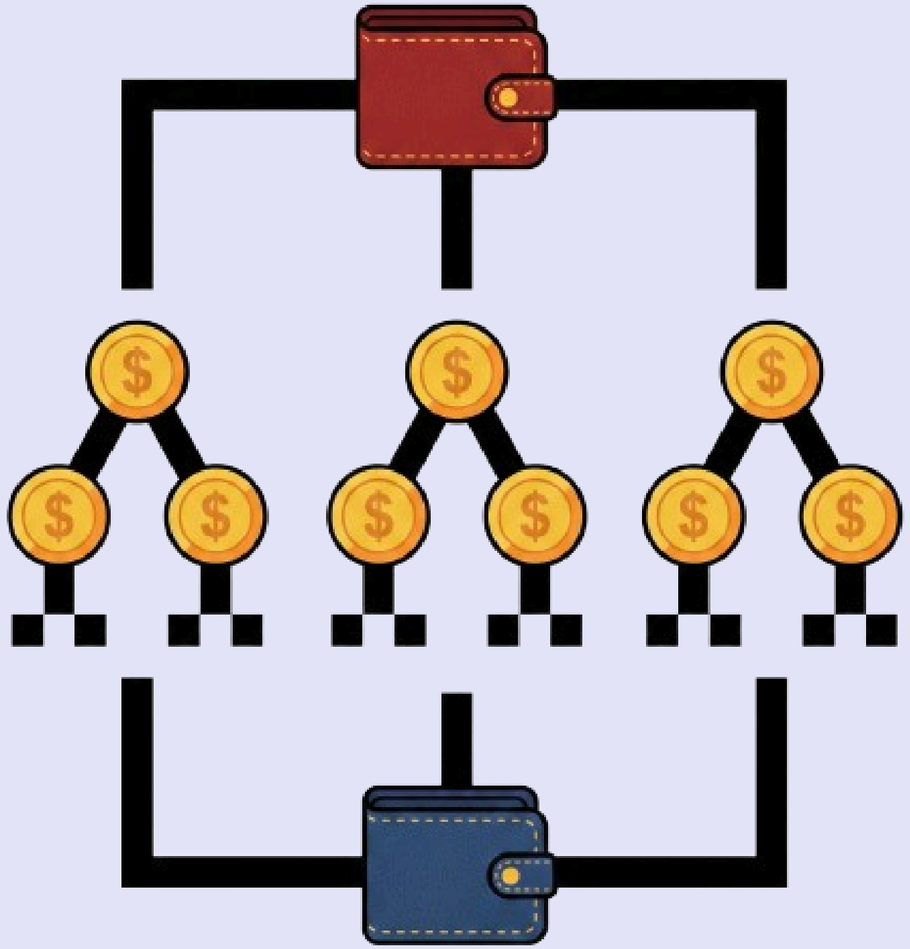
[5 rows x 43 columns]

RANDOM FOREST

- antrenare: 80%
- testare: 20%

- class_weight = balanced → penalizeaza erorile pe clasa minoritara pentru a combate dezechilibrul

- max_depth = 15 → previne overfitting-ul



REZULTATE

sex	accuracy	selection_rate	FNR	FPR
female	0.91	0.13	0.3	0.06
male	0.76	0.48	0.08	0.30

race	accuracy	selection_rate	FNR	FPR
not black	0.8	0.39	0.12	0.22
black	0.88	0.18	0.21	0.1

DIRECTII VIITOARE

Exponentiated Gradient Reduction

modificarea ponderilor

SMOTE

modificarea setului de antrenament prin generarea de date artificiale

MULTUMIM!