PRÀCTICA CRI
ARBRES DE DECISIÓ

RESOLEM UN PROBLEMA DE CLASSIFICACIÓ
AMB ARBRES DE DECISIÓ

Dataset: Mobile Price Classification (KAGGLE)

Coneixement, Raonament i incertesa

Jan Gras Serra 1636517

Lucía Sánchez Guillén 1633311

Youssef Cahouach Guella Ikhlaf 1638618

UAB

Índex

Introducció

En aquesta pràctica, el nostre objectiu és desenvolupar un classificador d'aprenentatge supervisat. Implementarem els algoritmes d'arbres de decisió ID3 i C4.5. També farem ús de tècniques per gestionar valors NaN i per a la discretització. Finalment, avaluarem a partir d'una mètrica específica el funcionament del nostre classificador.

ID3

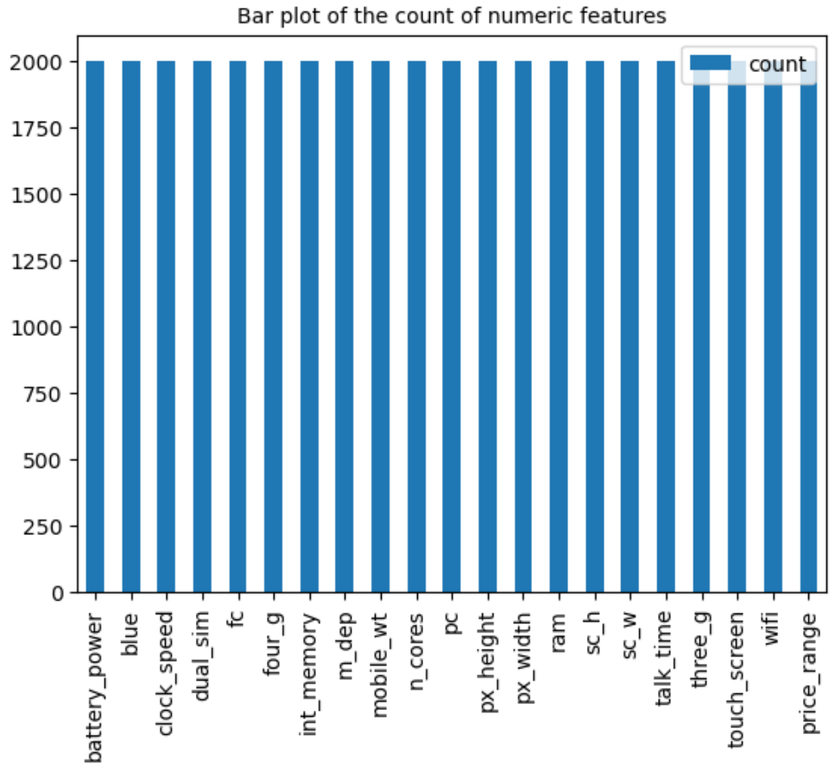
Aquest algorisme només pot gestionar atributs categòrics, per tant, les dades s'han de discretitzar.

C4.5

Aquest algoritme pot gestionar tant atributs categòrics com continus.

Anàlisi de dades

No tenim NaNs.




```
Data columns (total 21 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   battery_power  2000 non-null   int64
 1   blue           2000 non-null   int64
 2   clock_speed    2000 non-null   float64
 3   dual_sim       2000 non-null   int64
 4   fc             2000 non-null   int64
 5   four_g         2000 non-null   int64
 6   int_memory     2000 non-null   int64
 7   m_dep          2000 non-null   float64
 8   mobile_wt      2000 non-null   int64
 9   n_cores        2000 non-null   int64
 10  pc             2000 non-null   int64
 11  px_height      2000 non-null   int64
 12  px_width       2000 non-null   int64
 13  ram            2000 non-null   int64
 14  sc_h           2000 non-null   int64
 15  sc_w           2000 non-null   int64
 16  talk_time      2000 non-null   int64
 17  three_g        2000 non-null   int64
 18  touch_screen   2000 non-null   int64
 19  wifi           2000 non-null   int64
 20  price_range    2000 non-null   int64
dtypes: float64(2), int64(19)
```

Tractament de dades

	battery_power	blue	clock_speed	dual_sim	fc	four_g	int_memory	m_dep	mobile_wt	n_cores	...	px_height	px_width	ram	sc_h	sc_w	talk_time	three_g	touch_screen	wifi	price_range
0	842	0	2.2	0	1	0	7	0.6	188	2	...	20	756	2549	9	7	19	0	0	1	1
1	1021	1	0.5	1	0	1	53	0.7	136	3	...	905	1988	2631	17	3	7	1	1	0	2
2	563	1	0.5	1	2	1	41	0.9	145	5	...	1263	1716	2603	11	2	9	1	1	0	2
3	615	1	2.5	0	0	0	10	0.8	131	6	...	1216	1786	2769	16	8	11	1	0	0	2
4	1821	1	1.2	0	13	1	44	0.6	141	2	...	1208	1212	1411	8	2	15	1	1	0	1

5 rows × 21 columns

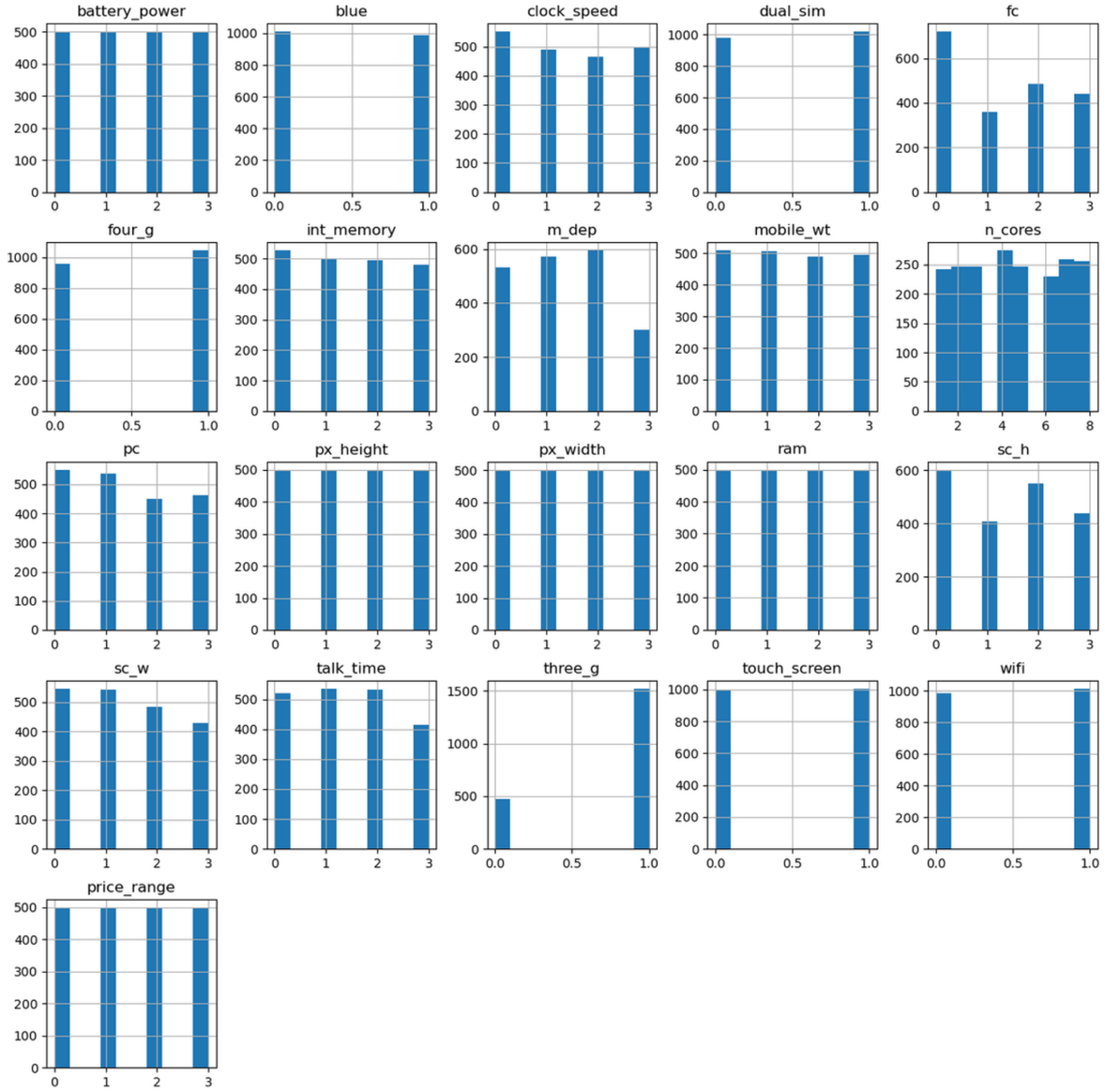
Omplim NaNs a partir de la mitjana de cada columna (Apartat B).

Discretitzem les variables contínues dividint les dades en 4 quartils.

	battery_power	blue	clock_speed	dual_sim	fc	four_g	int_memory	m_dep	mobile_wt	n_cores	...	px_height	px_width	ram	sc_h	sc_w	talk_time	three_g	touch_screen	wifi	price_range
0	0	0	2	0	0	0	0	2	3	2	...	0	0	2	0	2	3	0	0	1	1
1	1	1	0	1	0	1	3	2	1	3	...	2	3	2	3	1	1	1	1	0	2
2	0	1	0	1	1	1	2	3	2	5	...	3	3	2	1	0	1	1	1	0	2
3	0	1	3	0	0	0	0	2	1	6	...	3	3	2	2	2	1	1	0	0	2
4	3	1	1	0	3	1	2	2	1	2	...	3	1	1	0	0	2	1	1	0	1

5 rows × 21 columns

Mètrica seleccionada



Atributs desbalancejats:

- fc
- m_dep
- pc
- sc_h
- three_g

Per garantir que no ignorem aquests valors que apareixen tan poc freqüentment en comparació amb la resta, fem servir la mètrica **F1-score**.

La F1-score és la mitjana harmònica de la precisió i la recall.

Classificador: resultats

Mobile Price Classification

Model	F1-score
ID3	0.7428
C4.5 (Guany Informació)	0.7388
C4.5 (Gini)	0.7404

Titanic

Model	F1-score
ID3	0.6576
C4.5 (Guany Informació)	0.6701
C4.5 (Gini)	0.6576

Conclusions

Mobile Price Classification

ID3	
Accuracy	0.7425
Precision	0.7505
Recall	0.7425
F1-score	0.7449

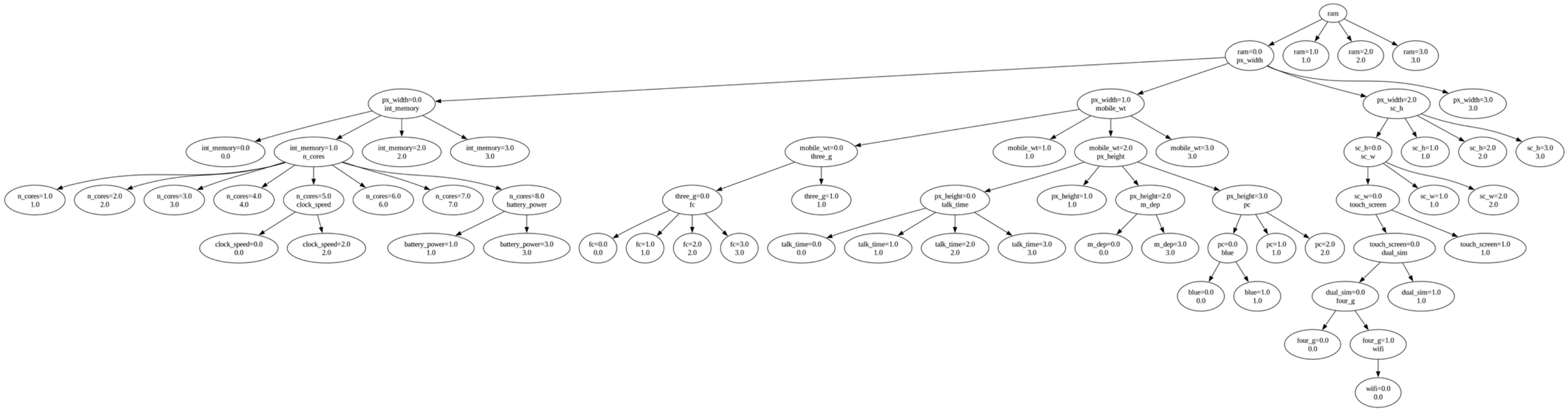
Arbre Resultant

Titanic

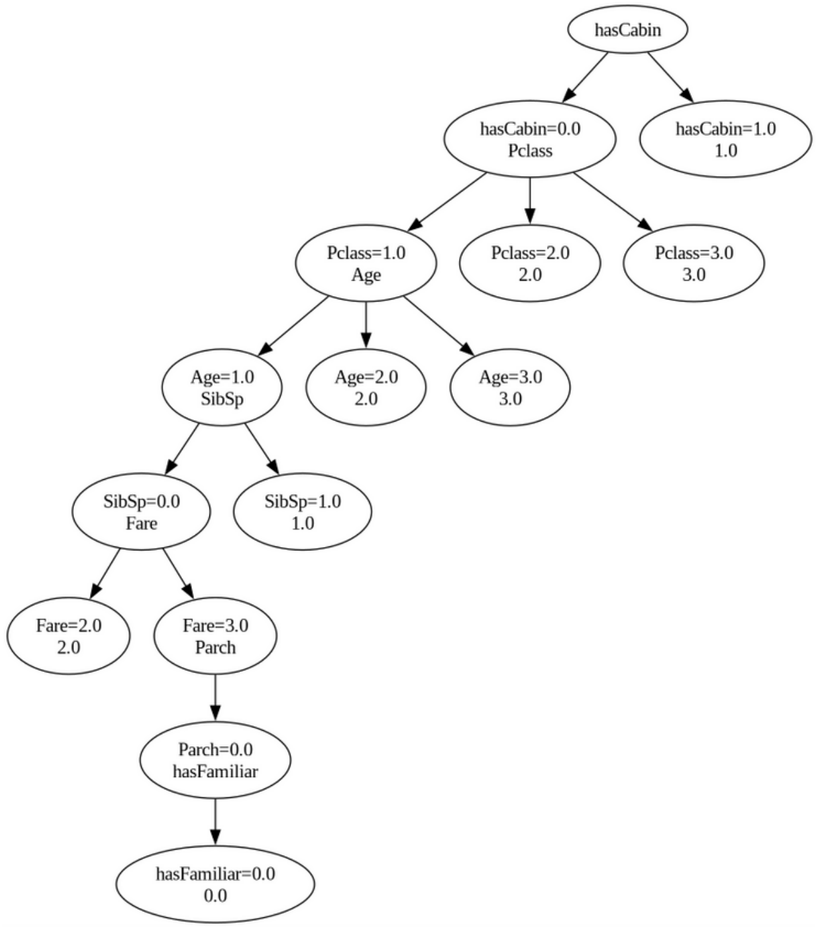
C4.5 (Guany Informació)	
Accuracy	0.6434
Precision	0.6420
Recall	0.6434
F1-score	0.6306

Arbre Resultant

Mobile Price Classification



Titanic


hasCabin
hasCabin=0.0
Pclass
hasCabin=1.0
1.0
Pclass=1.0
Age
Pclass=2.0
2.0
Pclass=3.0
3.0
Age=1.0
SibSp
Age=2.0
2.0
Age=3.0
3.0
SibSp=0.0
Fare
SibSp=1.0
1.0
Fare=2.0
2.0
Fare=3.0
Parch
Parch=0.0
hasFamiliar
hasFamiliar=0.0
0.0

GRÀCIES

GM08:30_3

UAB