Arbres de decisió

Martí Caixal Joaniquet: 1563587

Ricard López Olivares: 1571136

Sergi Bons Fuses: 1571359

Objectius

- Preprocessat de dades
- Tractament d’atributs continus
- Arbre de decisió
- cross validation

Dataset

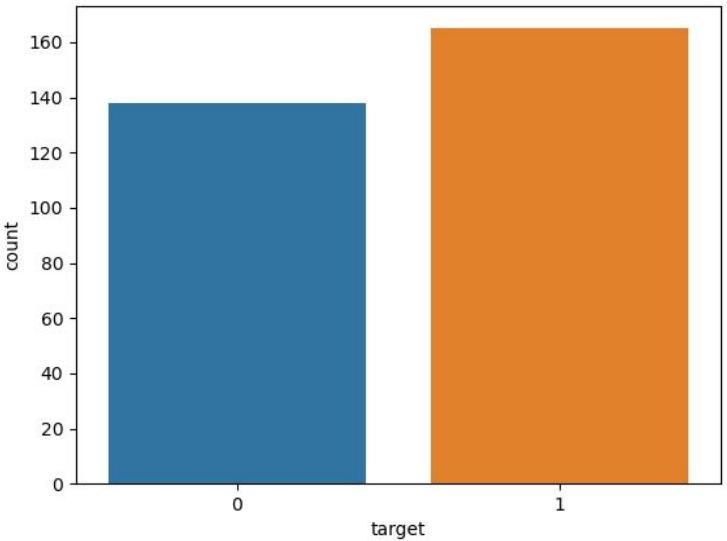
Heart Disease UCI

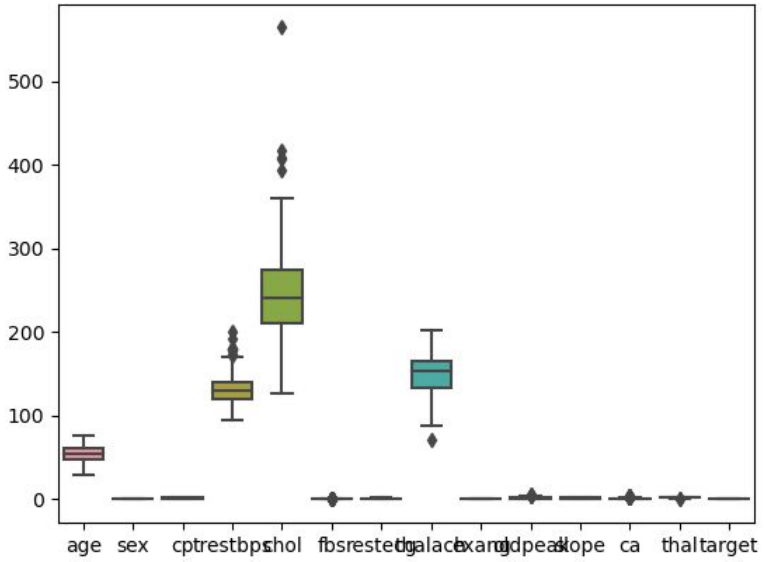
→ Predir enfermetat de cor

→ 303 mostres

→ Atributs continus i discrets

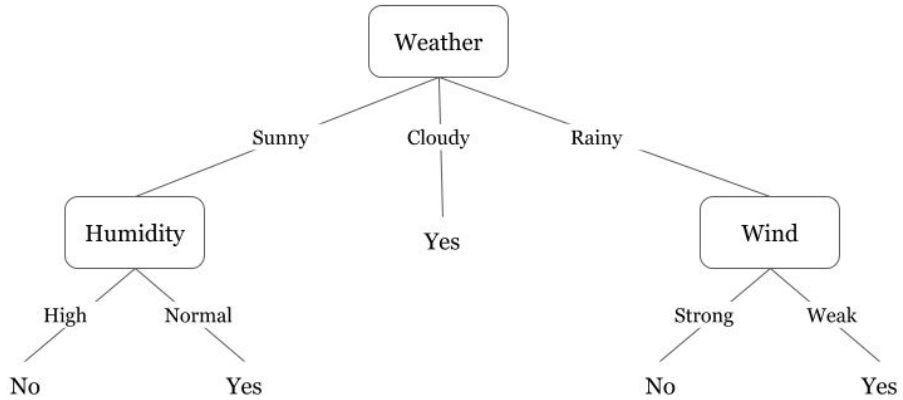
EDA


count
target
0
1
0
20
40
60
80
100
120
140
160


0
100
200
300
400
500
age
sex
cp
trestbps
chol
fbs
restecg
thalach
exang
oldpeak
slope
ca
thal
target

Apartat 1

Estructura de dades




```
tree = \
{'Weather': {
    'Sunny': {'Humidity': {
                    'High' : 'NO',
                    'Normal': 'YES'
                }
    },
    'Cloudy' : 'YES',
    'Rainy' : { 'Wind': {
                    'Strong' : 'NO',
                    'Weak' : 'YES'
                }
    }
}}
```

Tipus d'atributs

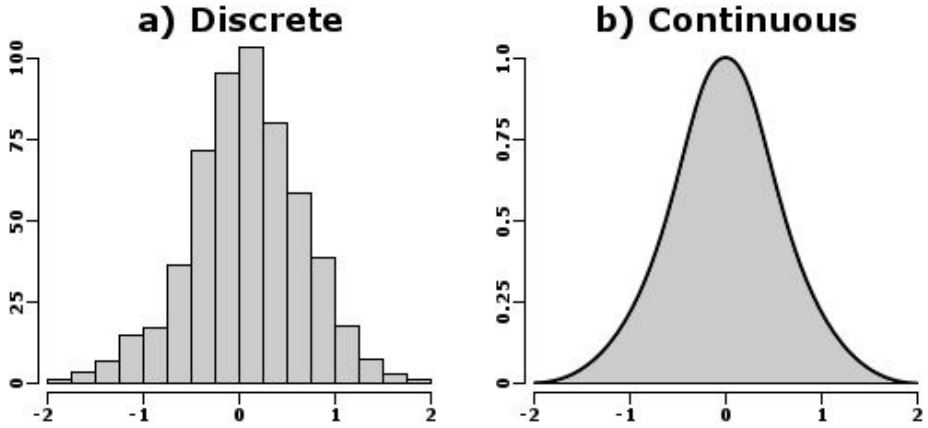
Discrets:

- Funcionen bé
- Generen proques branques

Continus:

- Generen masses branques
- Cal tractar-los

→ Classificar en N intervals



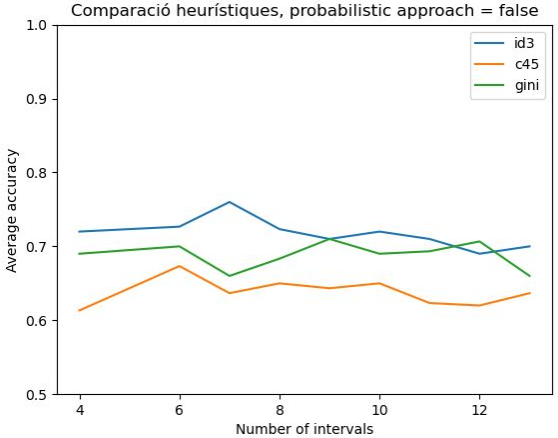
Heuristiques

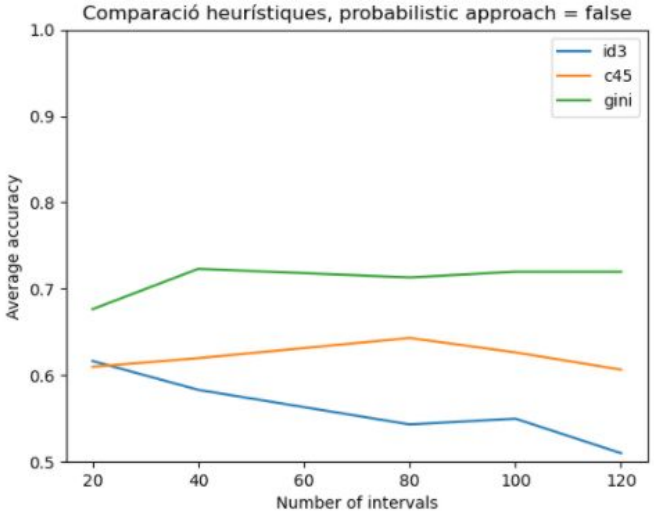
$$Gain(S,A) = -\sum_{x=1}^{m}\left(p(x) * log_2 p(x)\right) - \sum_{v \epsilon A}\frac{|S_v|}{|S|} * \sum_{x \epsilon v}\left(p(x) * log_2 p(x)\right)$$

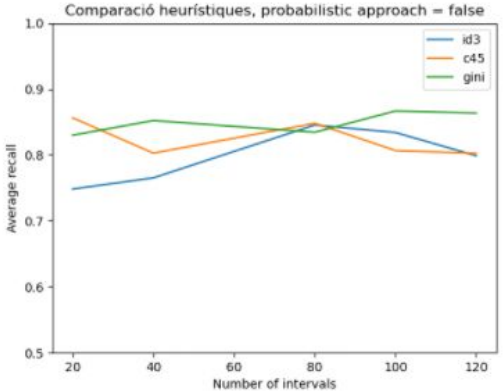
$$GainRatio(S,A) = \frac{Entropia(S) - Entropia(S,A)}{-\sum_{v \epsilon A}\left(\frac{|S_v|}{|S|} log_2\left(\frac{|S_v|}{|S|}\right)\right)}$$

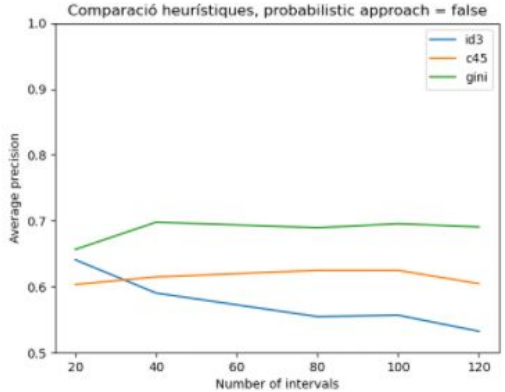
$$Gini(S,A) = 1 - \sum_{v \epsilon A}\left(\frac{|S_v|}{|S|}^2\right)$$

Resultats


Comparació heurístiques, probabilistic approach = false
id3
c45
gini
Average accuracy
Number of intervals


Comparació heurístiques, probabilistic approach = false
id3
c45
gini
Average accuracy
Number of intervals


Comparació heurístiques, probabilistic approach = false
id3
c45
gini
Average recall
Number of intervals


Comparació heurístiques, probabilistic approach = false
id3
c45
gini
Average precision
Number of intervals

Mètrica utilitzada

Malaltia?	Actual YES	Actual NO	Total
Predicted YES	TP = 5	FP = 0	5
Predicted NO	FN = 25	TN = 270	295
Total	30	270	300

Accuracy	92%
Precision	100%
Recall	16%
F1 Score	27%

Apartat 2

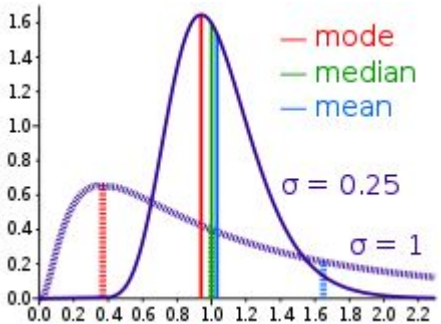
Valors nuls i erronis

Atributs continus:

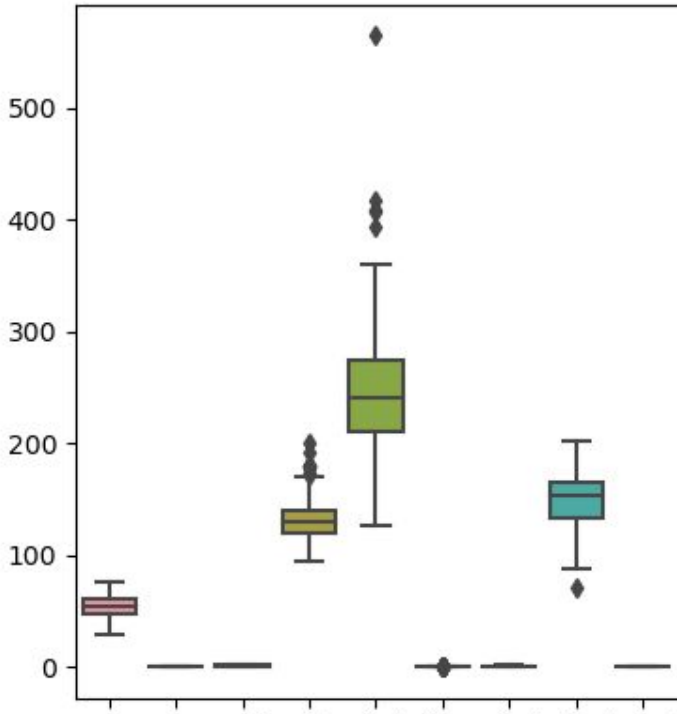
- Mitjana

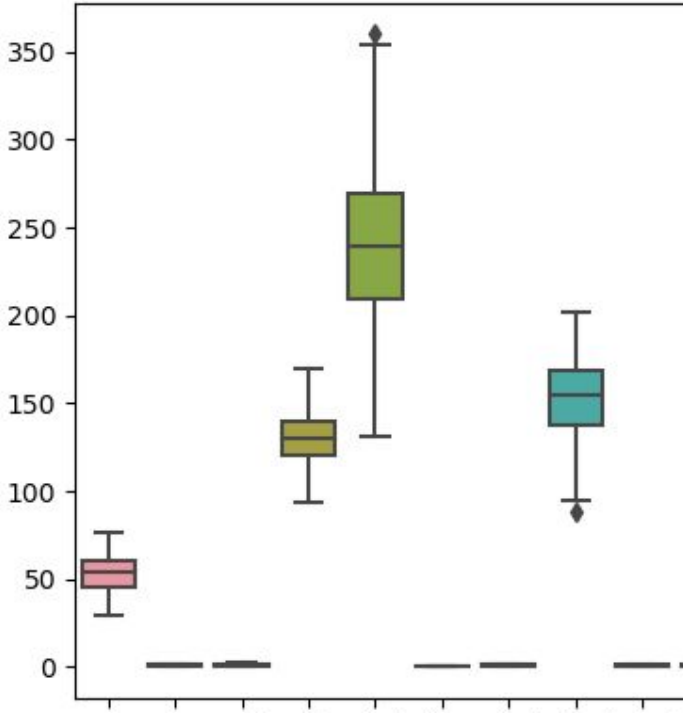
Atributs discrets:

- Moda

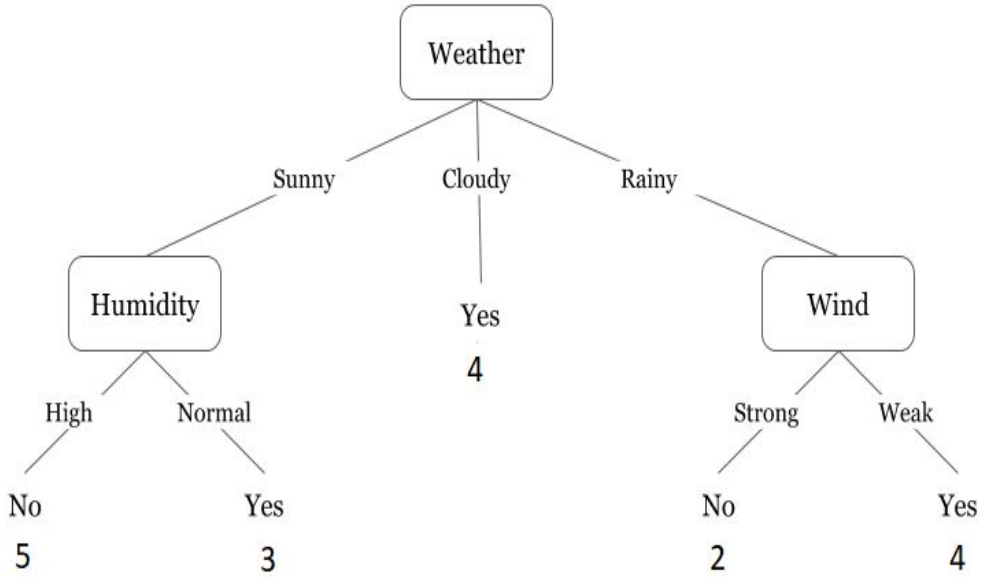


Outliers



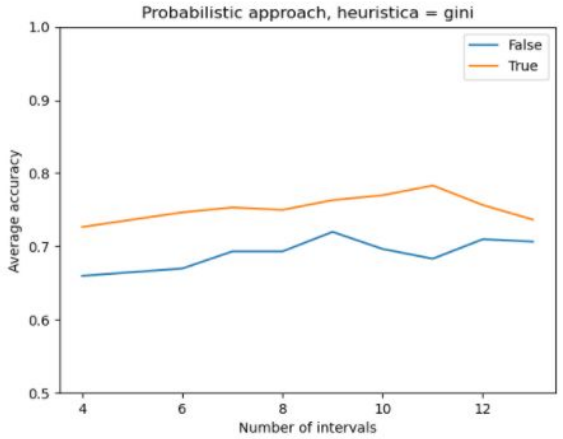


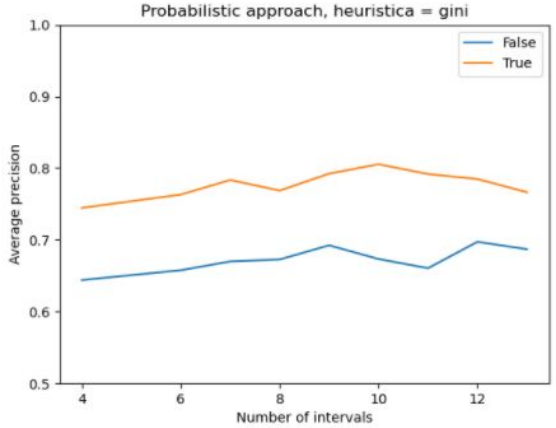
Probabilistic approach

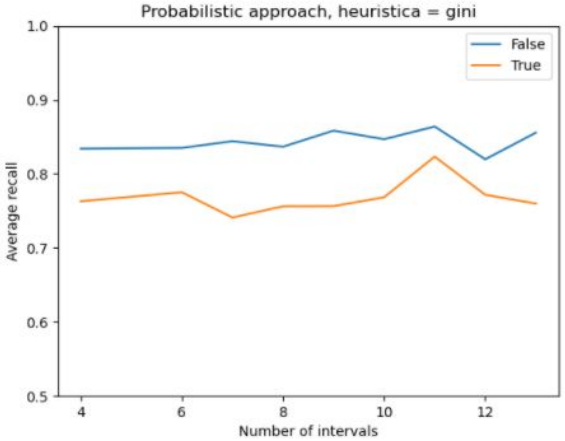


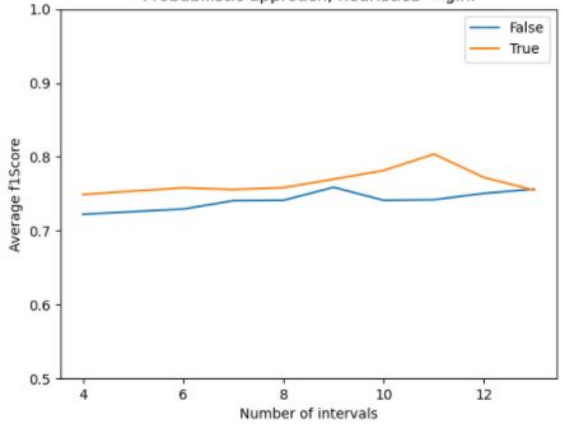

```
tree = \
{'Weather': {
    'Sunny': {'Humidity': {
                    'High' : ('NO', 5),
                    'Normal': ('YES', 3)
                }
    },
    'Cloudy' : ('YES', 4),
    'Rainy' : { 'Wind': {
                    'Strong' : ('NO', 2),
                    'Weak' : ('YES', 4)
                }
    }
}}
```

Probabilistic approach



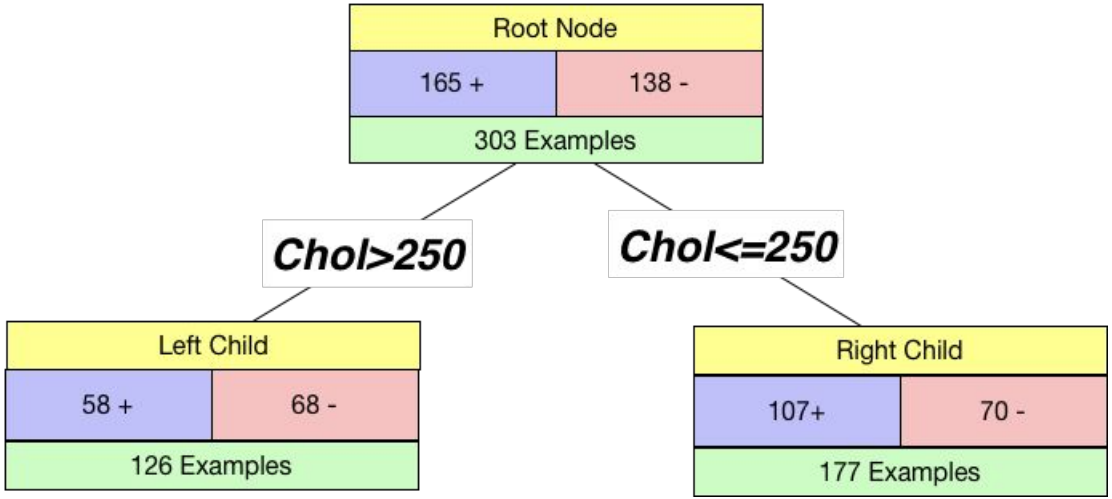




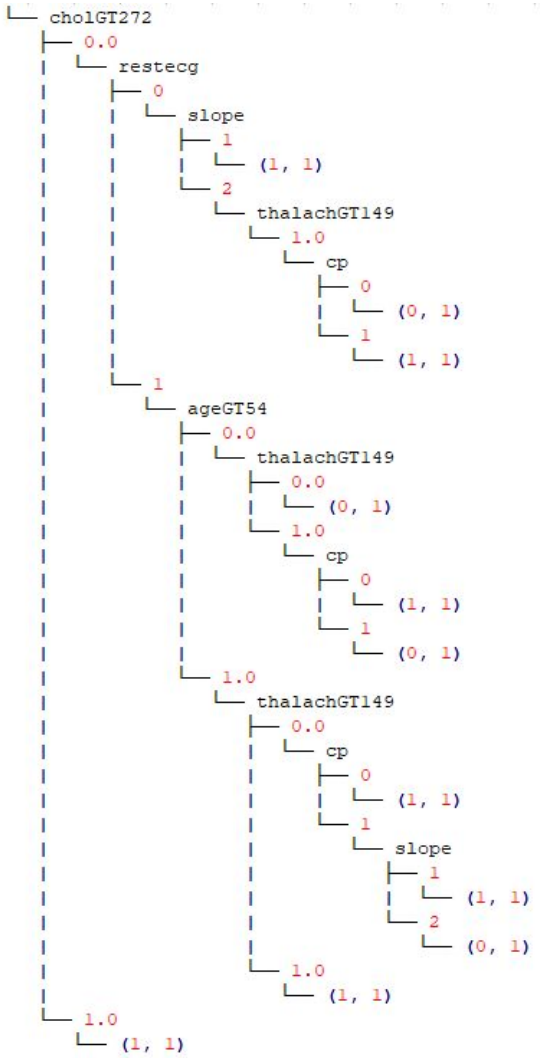


Apartat 3

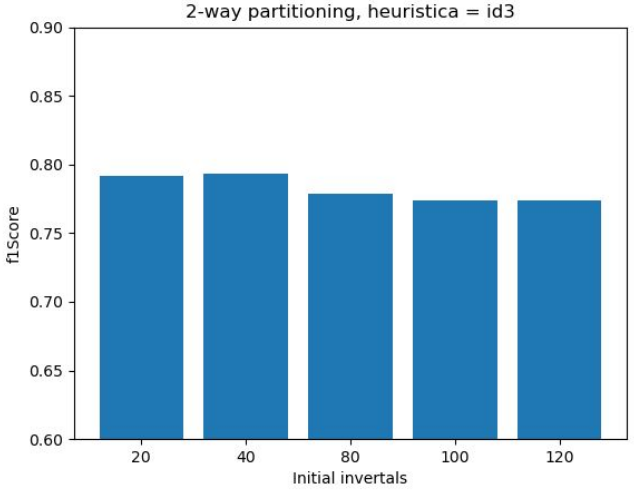
2-Way Partitioning

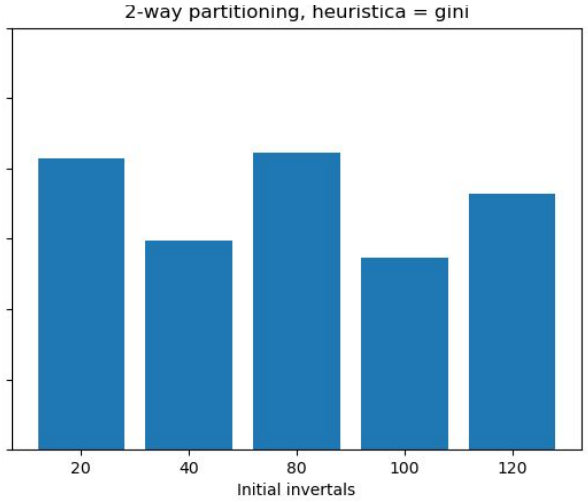


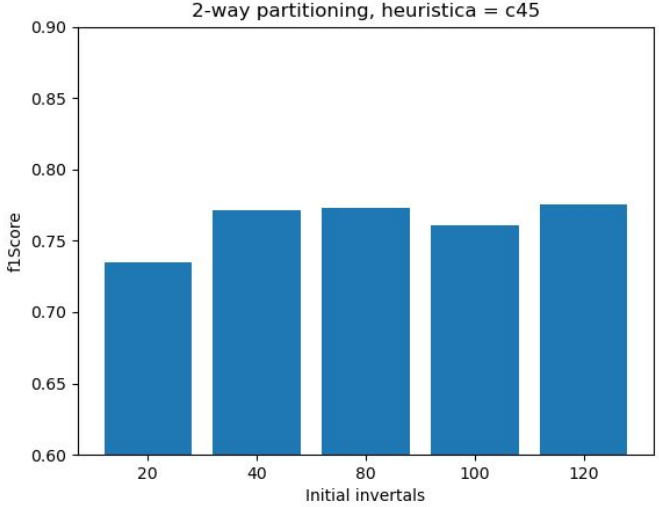
Arbre resultant



2-Way Partitioning

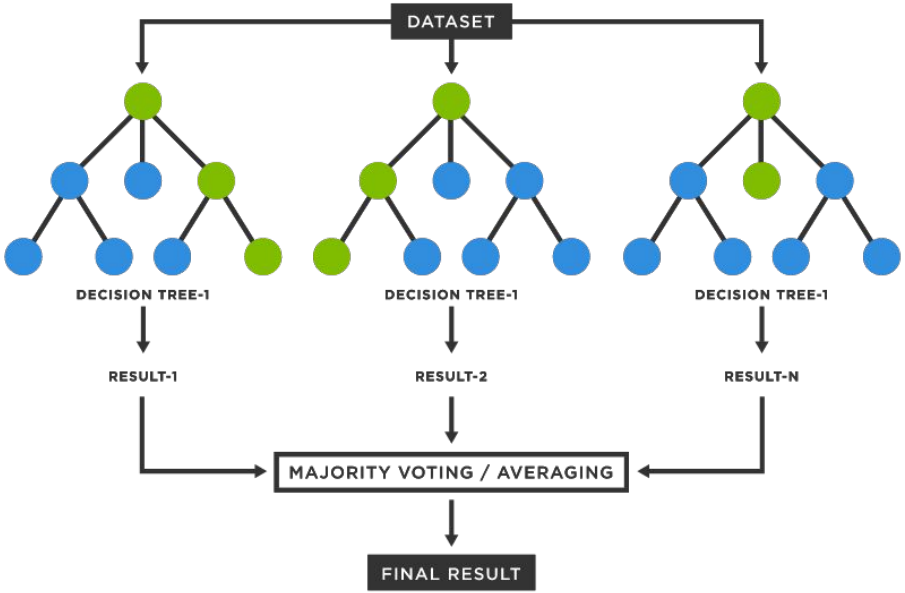






Apartat 4

Random Forest



TRAINING

Original = [1, 2, 3, 4, 5, 6]

Arbre 1 = [1, 2, 2, 3, 5, 5]

Arbre 2 = [1, 1, 4, 4, 5, 6]

Arbre 3 = [1, 4, 4, 4, 6, 6]

Random Forest