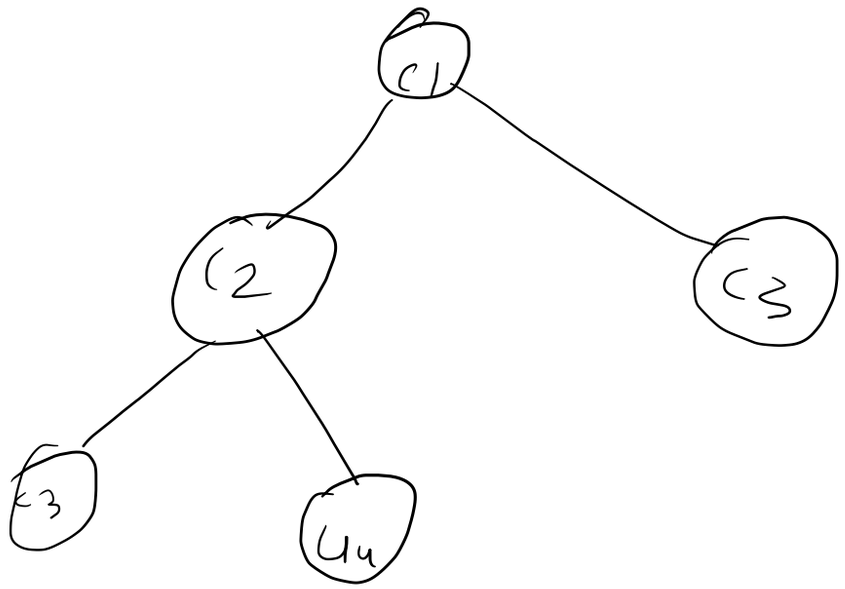


feature → C1, C2, C3, C4

1. pruning ← overfitting
2. feature importance
3. gini vs entropy ← [data bias]

↓

tree

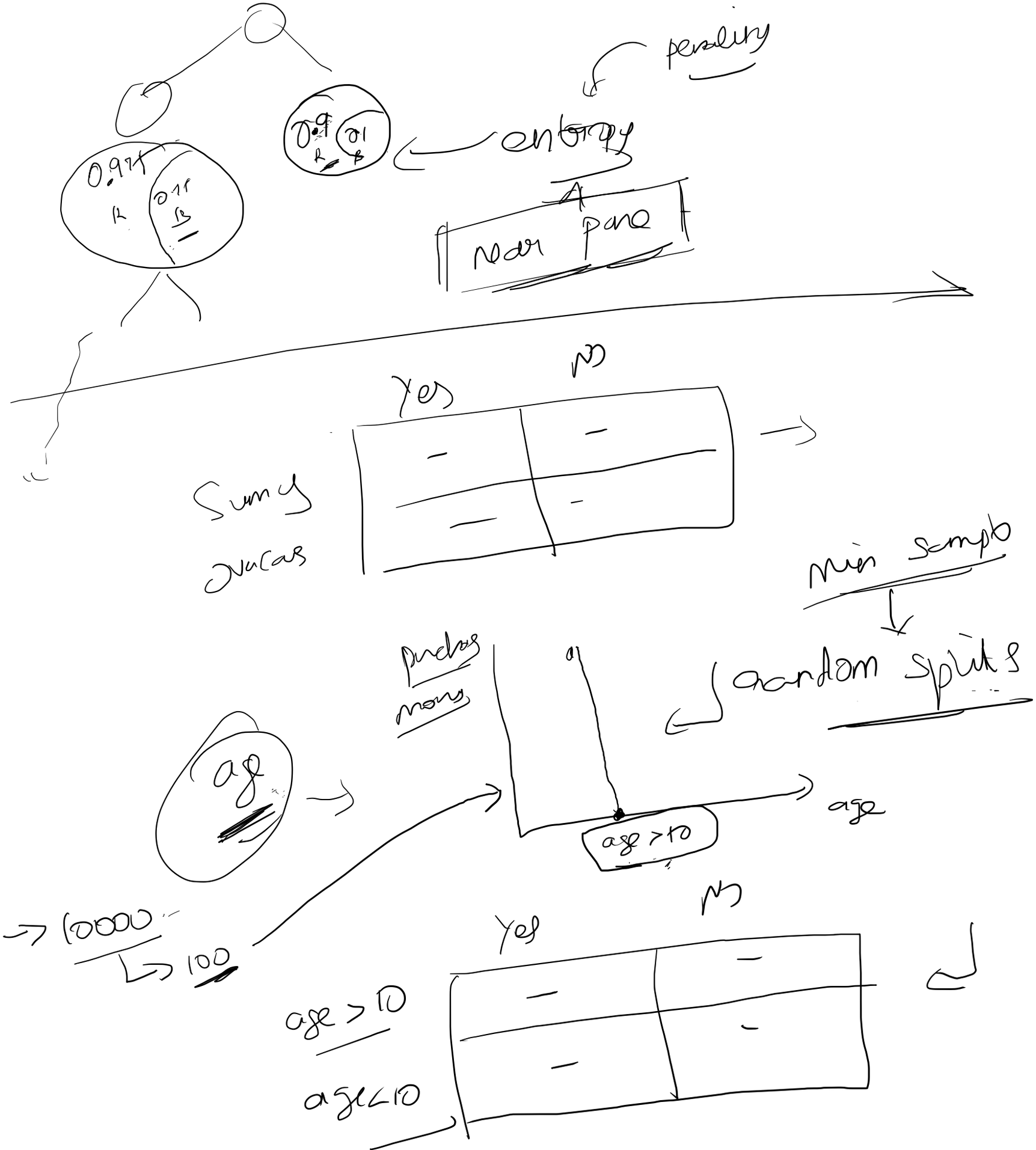
feature importance

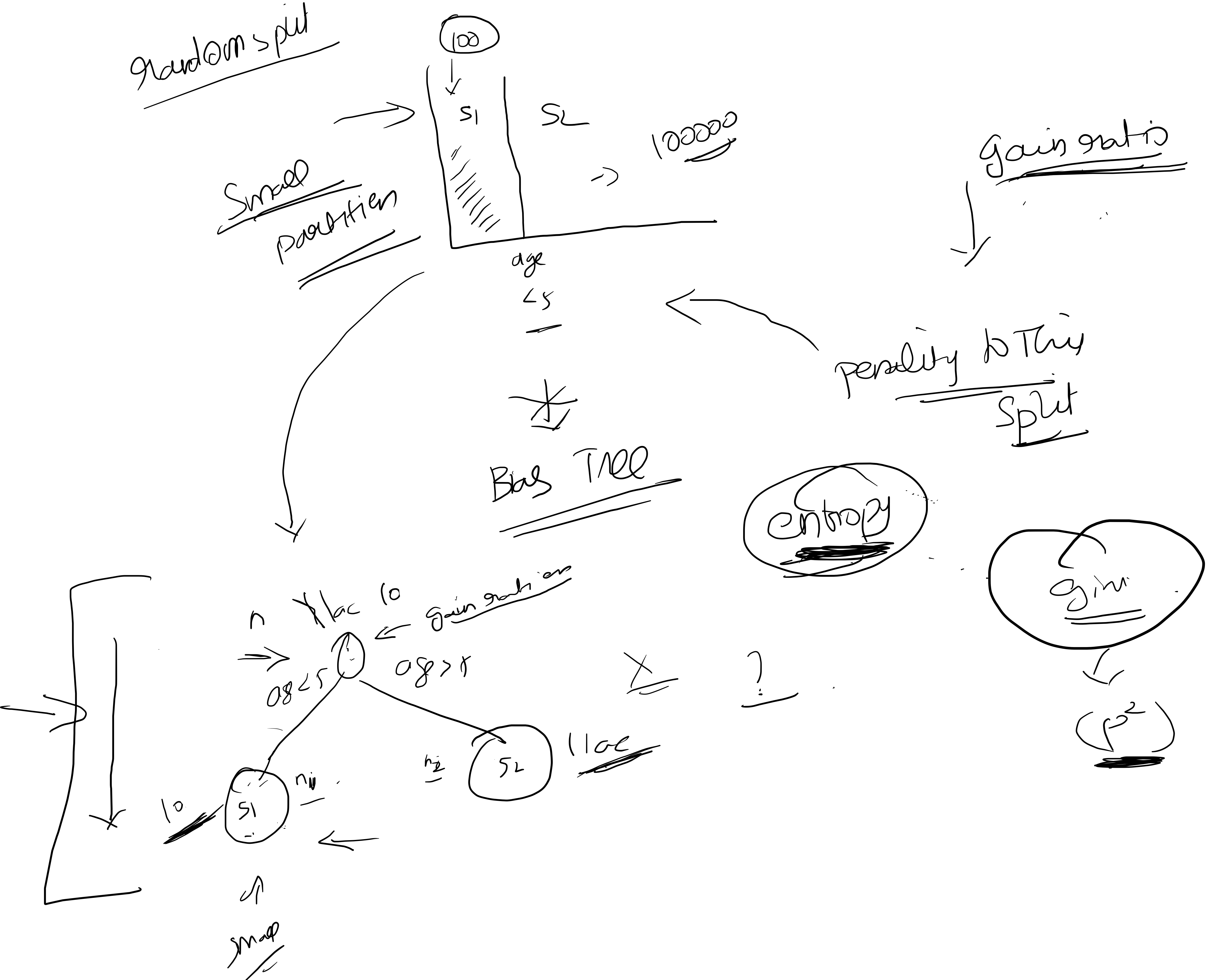
→ MDA → Mean decreasing accuracy

→ MDI (gini) → gini importance

↳ Mean decreasing Impurity

permutation based


0.97
R
0.11
B
0.9
R
0.1
B
entropy
node pure
Sunny
Yes
No
Min sample
Min bucket
random splits
Previous node
age
age > 10
age
age -> 10000
100
age > 10
age < 10
Yes
No


random split
100
S1
S2
100000
Small partition
age
<5
gain ratio
penalty to this split
Bias Tree
entropy
gini
p^2
n 1lac 10
gain ratio
age<5
age>5
S1
S2
n1
n2
10
1lac
small

original accuracy

C_1 → gain ↑ → remove → accuracy ↓

C_2 → gain ↑ → remove → accuracy ↓

permutation of feature : (C_1, C_2, C_3, C_4, C_5)

Boruta

↳ permuting feature import using RF

→ C_1

C_1, C_2

C_1, C_2, C_3

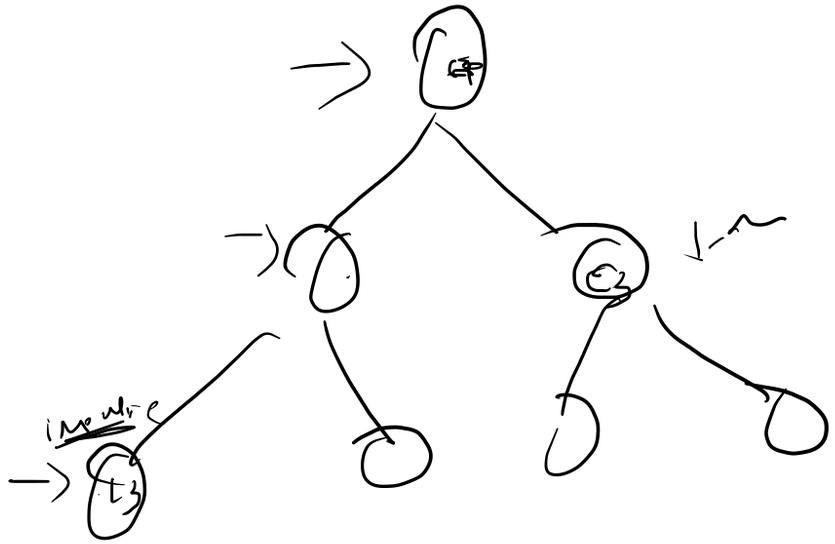
C_1, C_2, C_3, C_4 ...

→ C_1, C_2, C_3, C_4, C_5 ←

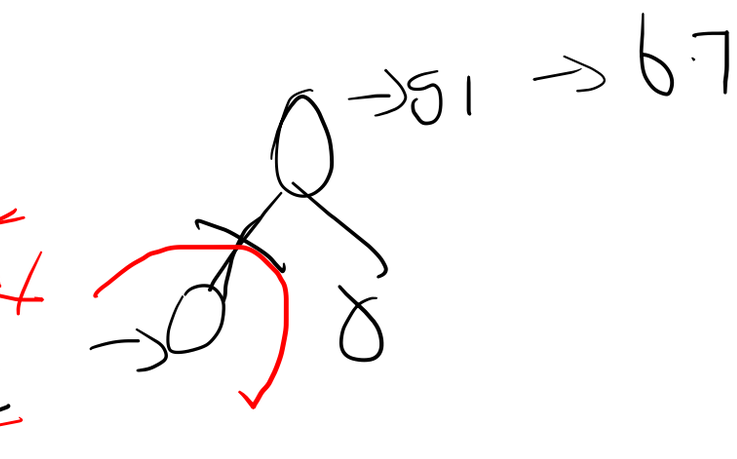
C_2
C_3
C_4
C_5

(2, C

Gain



impurity

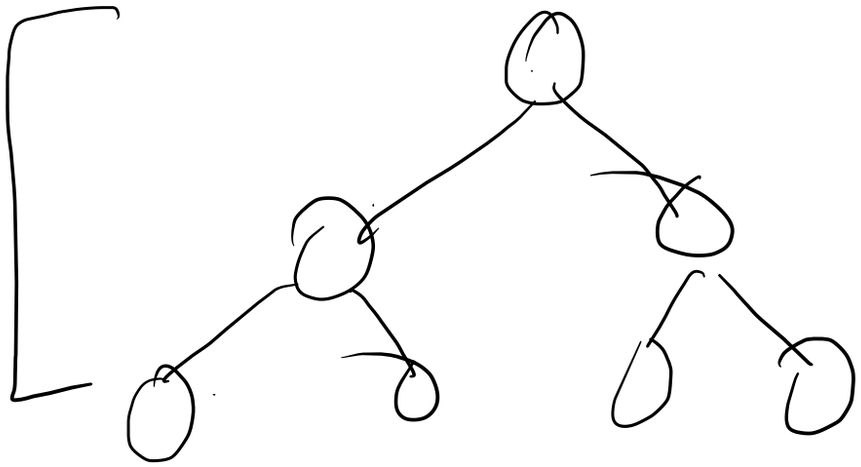


Drop

$g = 0.2$

How my impurity

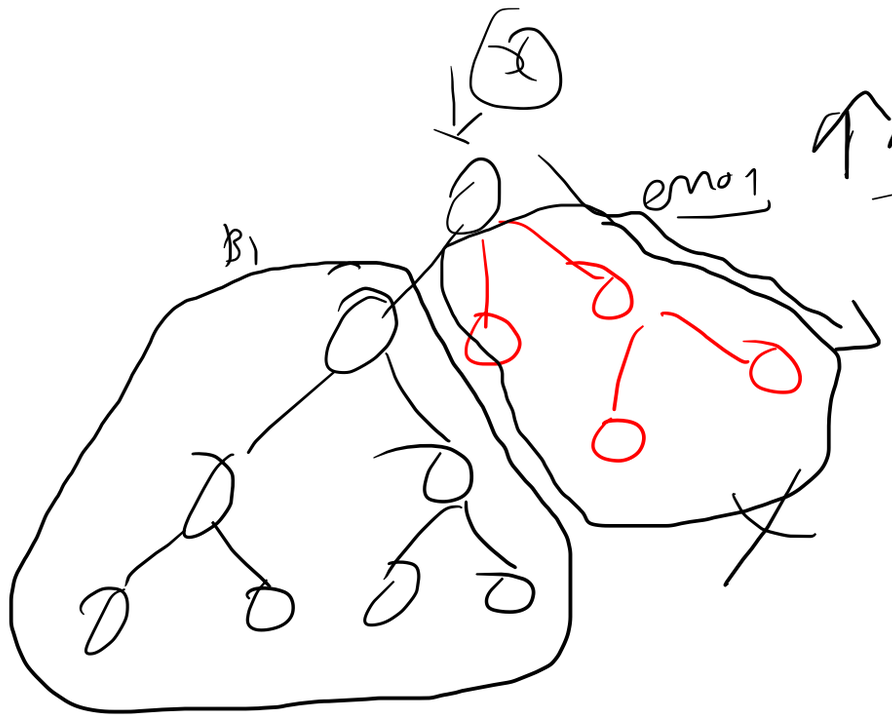
Average



Overfitting

1. Max depth → 3
2. pruning

↳ at branch



error

C → NL

error

misclassification

error

pruning

pre pruning

post pruning

1. Al-a tireal Tree
2. pruning

1. Tree
2. pruning