Decision Tree

- → D.T Intution
- → Working D.T
- → Entropy / Info. gain
- → Code

→ Data Scientist Aintel

↳ Employee attrition

1/0

1. Find the chances of attrition of a Employee

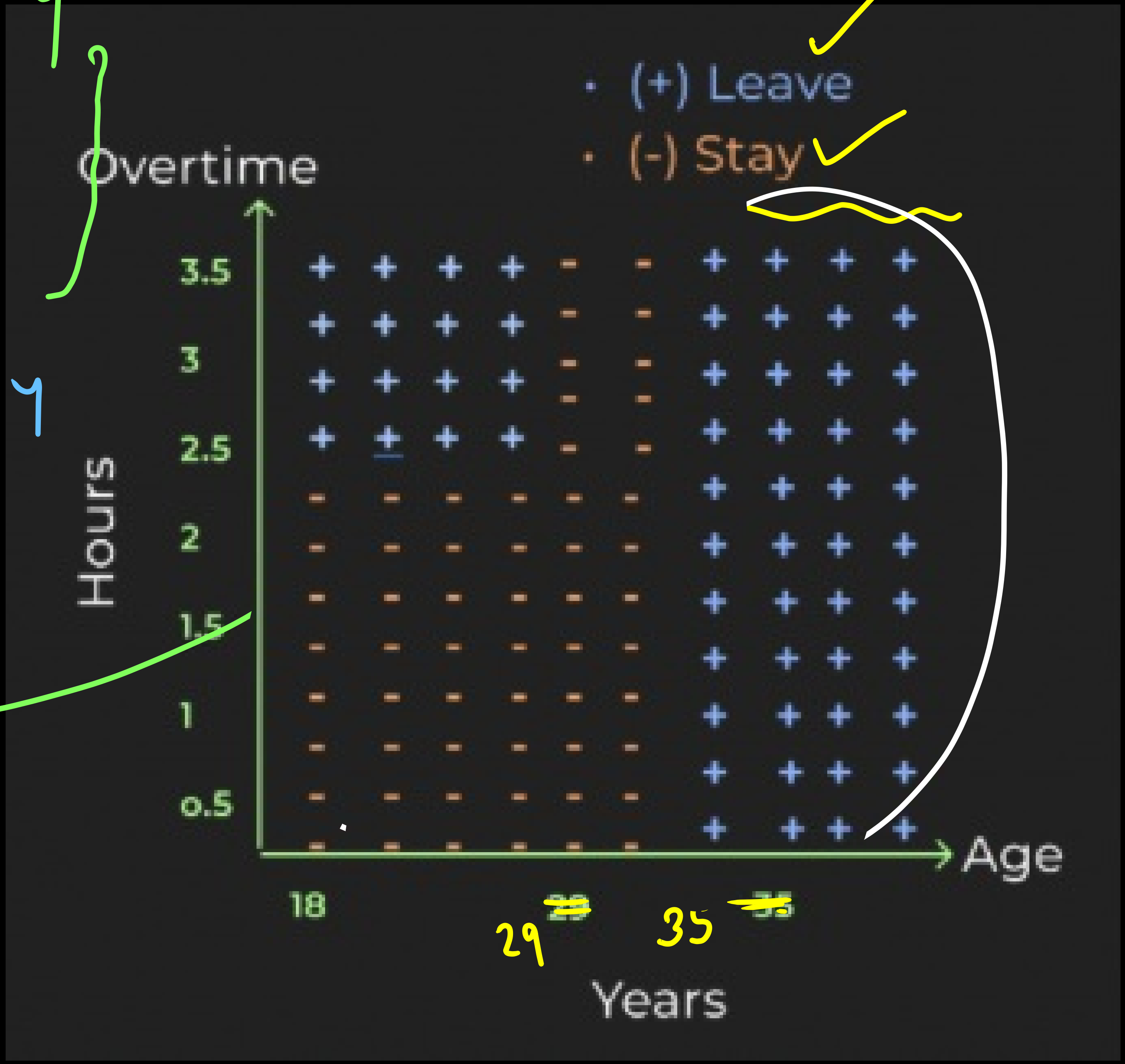
2. What are the key factor responsible for Attrition

↳ Feature Importance

Age Overtime | y

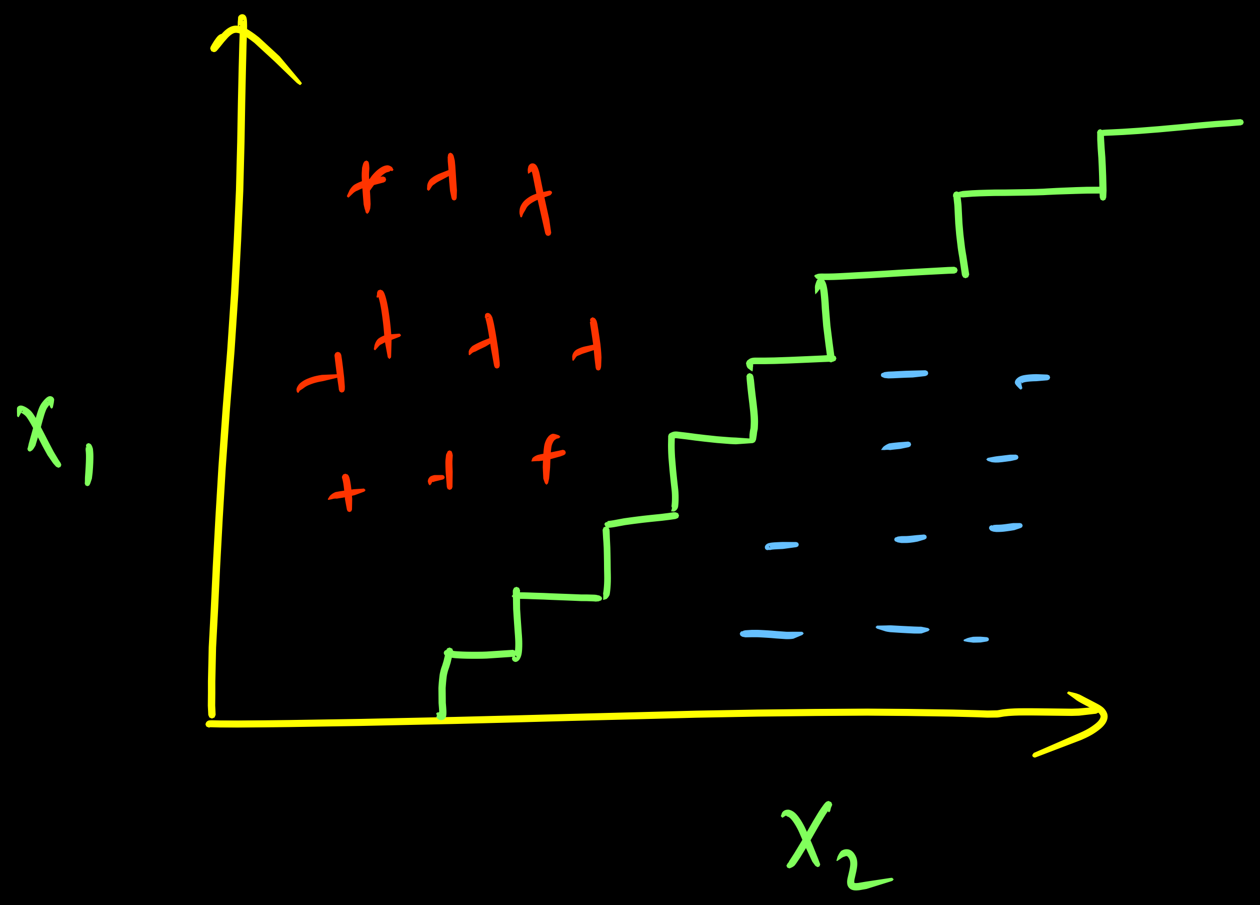
Age Overtime | y

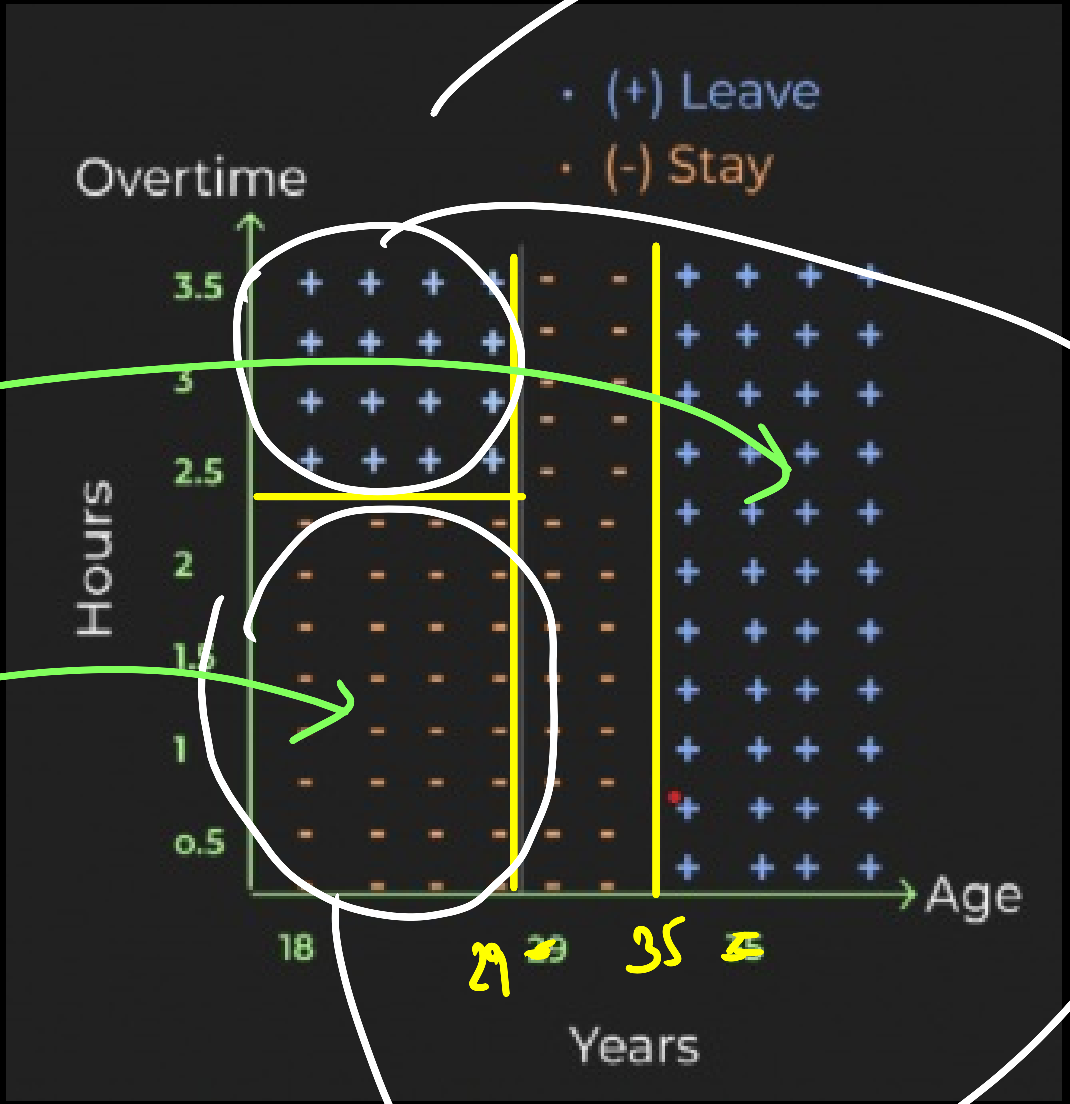
27 2



→ Axis parallel hyperplane

→ Non-linear model

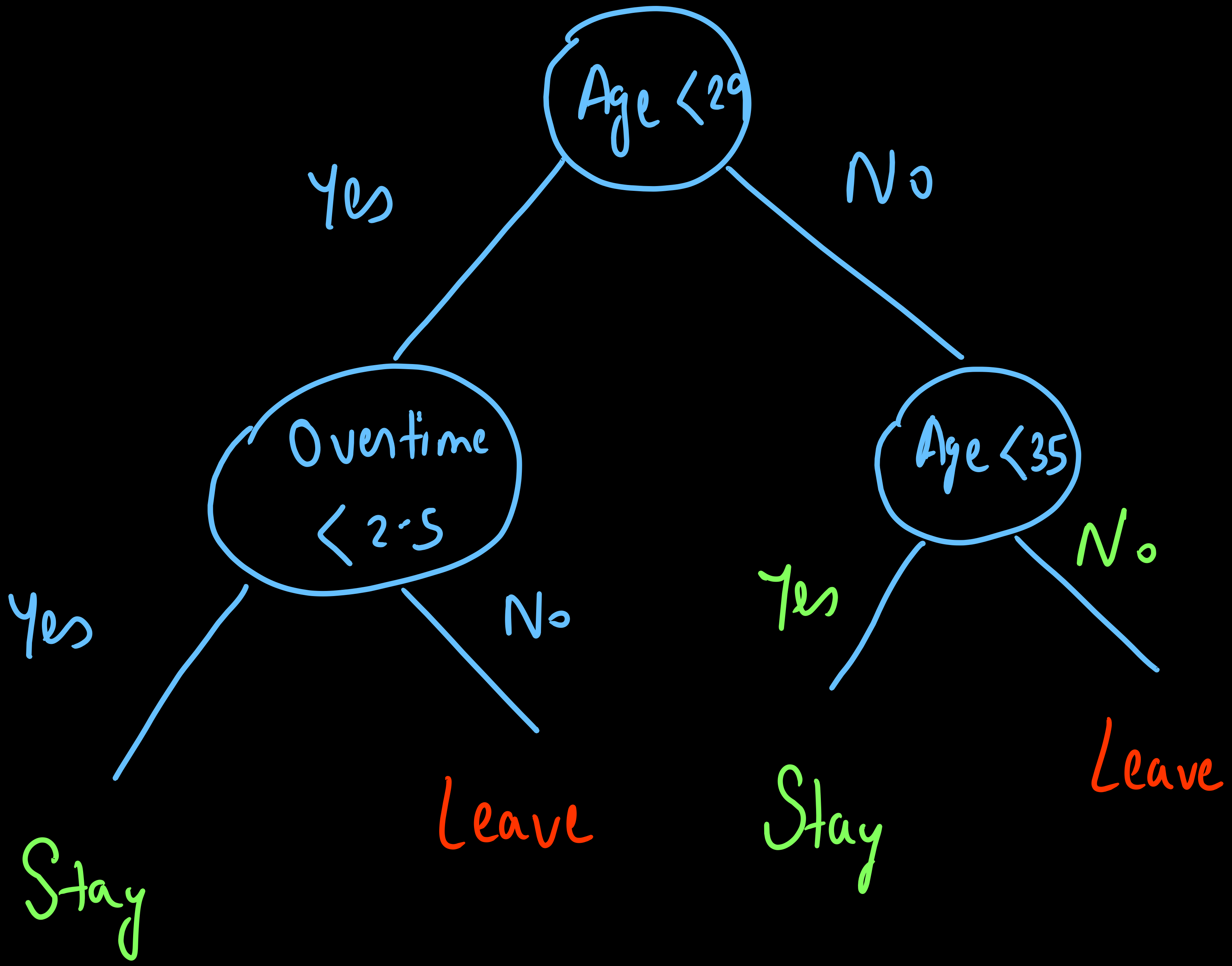

x_1
x_2
+
−

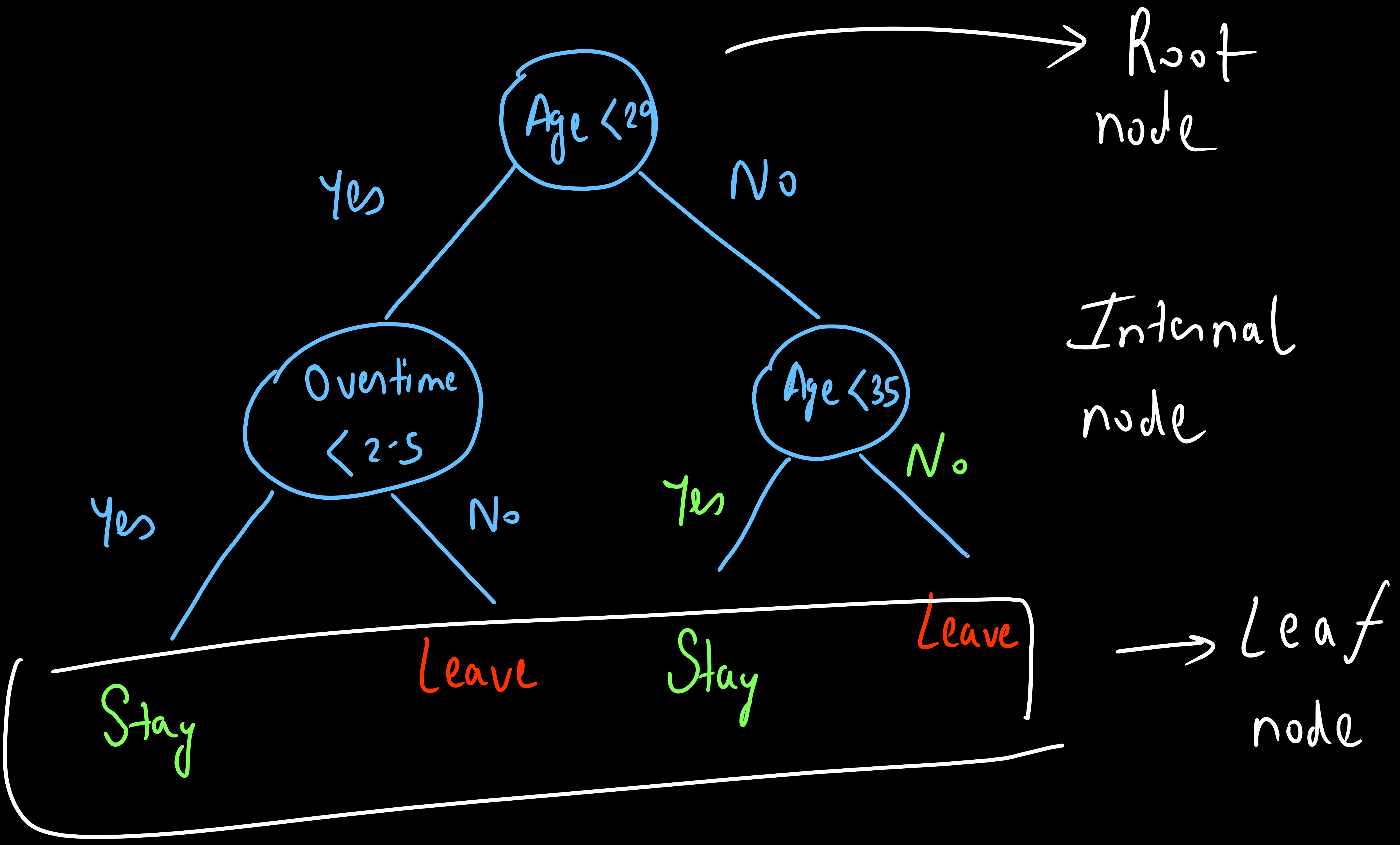


Training

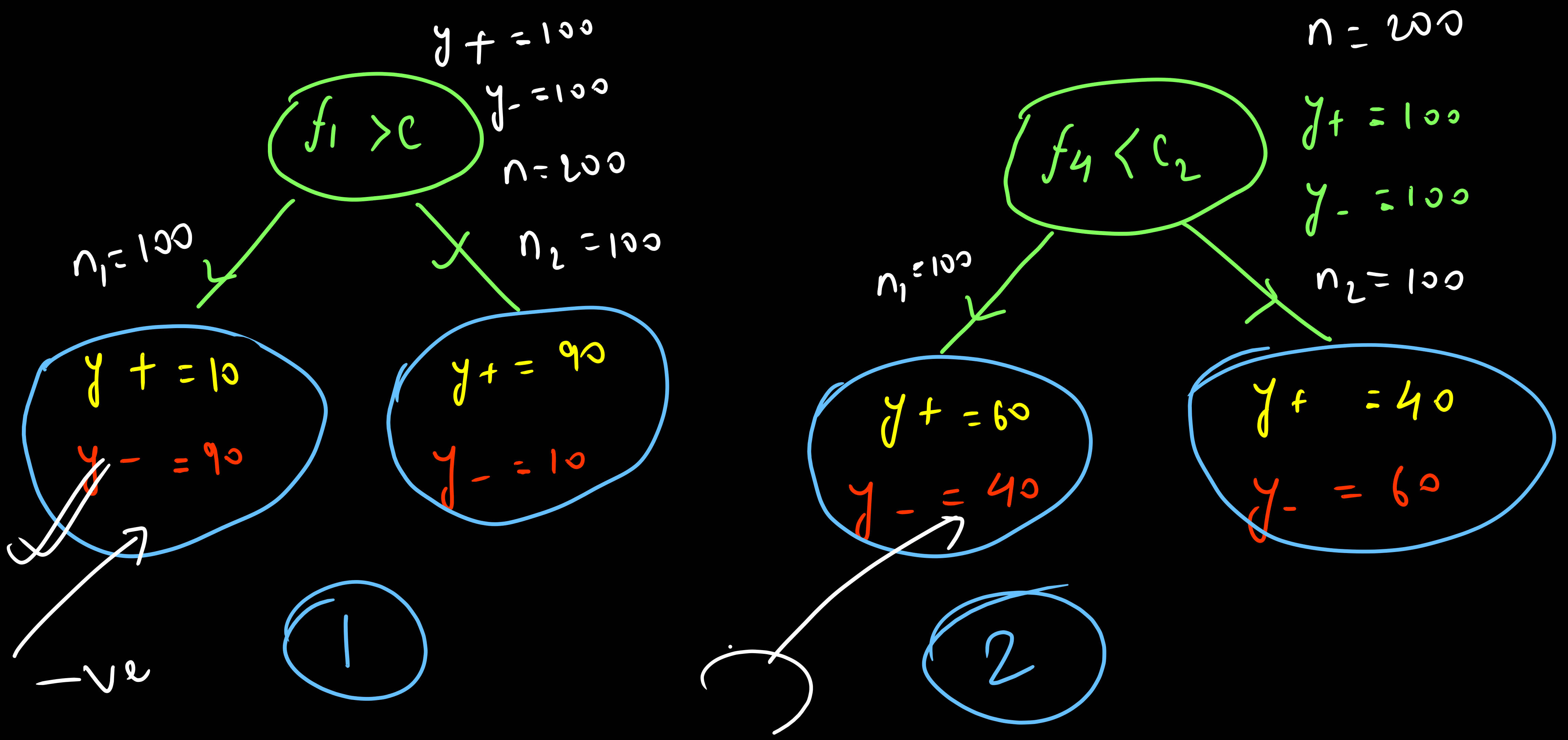
```
If (Age <29):
    If(Overtime <2.5):
        Stay
    else:
        Leave
else
    If(Age <35):
        Stay
    else:
        Leave
```

```
If (Age <29):
    If(Overtime <2·5):
        Stay
    else:
        Leave
else
    If(Age <35):
        Stay
    else:
        Leave
```




Age <29
Yes
No
Root node
Overtime <2.5
Age <35
Internal node
Yes
No
Yes
No
Stay
Leave
Stay
Leave
Leaf node

→ Which options to choose



→ More homogenous = More confidence in prediction

→ Entropy

↳ Measures Randomness

↳ Impurity (Heterogenity)

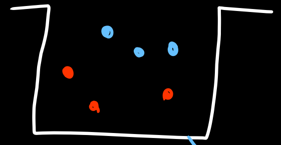
$$H(y) = -\sum_{i=1}^{K} p_i \log p_i$$

$$\boxed{H(y) = -\sum_{i=1}^{\overset{\text{No of classes}}{K}} \underbrace{p_i}_{\text{Probab. of each class}} \log_2 p_i}$$

No of classes

Probab. of each class

Ranges (0 - 1)



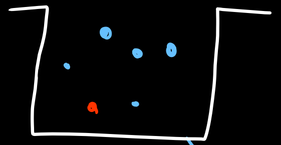
$P(\text{blue}) = 3/6 = \boxed{1/2}$

$P(\text{red}) = 3/6 = \boxed{1/2}$

$$H(y) = -\left[P_{blue} \log_2 P_{blue} + P_{red} \log_2 P_{red}\right]$$

$$= -\left[\tfrac{1}{2} \log_2 (\tfrac{1}{2}) + \tfrac{1}{2} \log_2 (\tfrac{1}{2})\right]$$

$$\boxed{= 1}$$

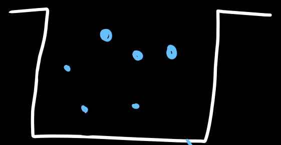


$P_{blue} = 5/6$

$P_{red} = 1/6$

$$H(y) = -\left[P_b \log_2 P_b + P_r \log_2 P_r\right]$$

$$= -\left[\tfrac{5}{6} \log_2 (5/6) + \tfrac{1}{6} \log_2 (1/6)\right]$$

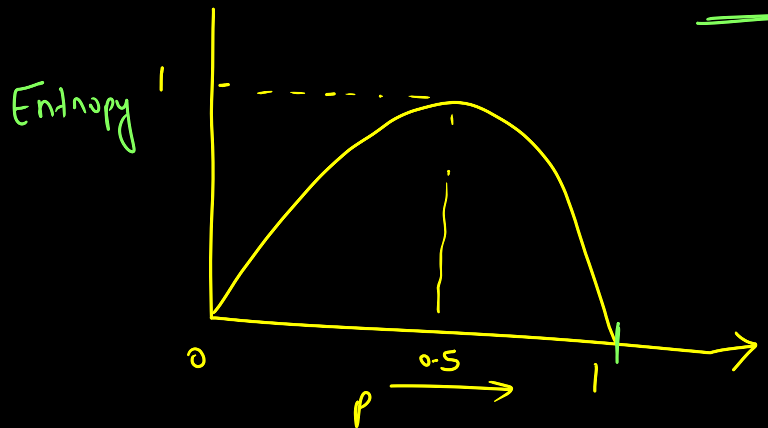


$P_b = 1$

$P_r = 0$

$$H(y) = -\left[1 \underbrace{\log_2 (1)}_{0} + \underbrace{0 \log_2 (0)}_{0}\right]$$

$$= 0$$



→ Break until 22:19

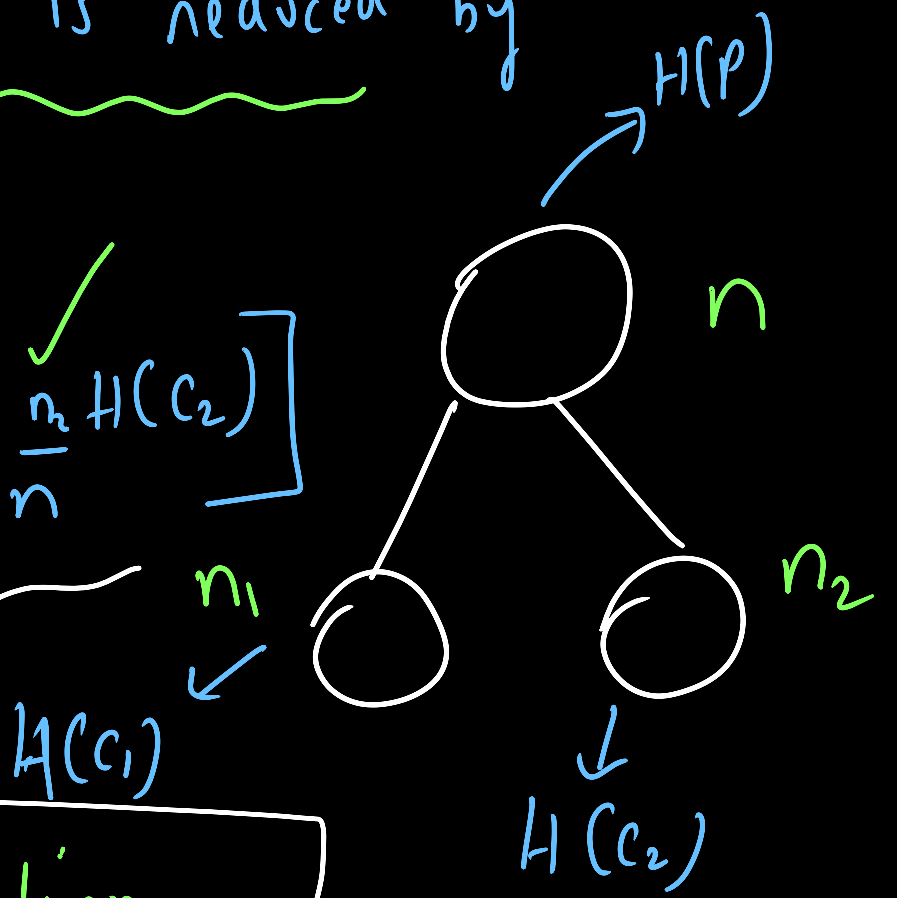
Information Gain

(Age, Gender)

→ Used to decide which feature to split on

→ Measure how much entropy is reduced by making a split

→ $IG = H(P) - \left[\frac{n_1}{n} H(C1) + \frac{n_2}{n} H(C_2) \right]$



→ High IG → high reduction in entropy (Heterogenous) → More Homogenous

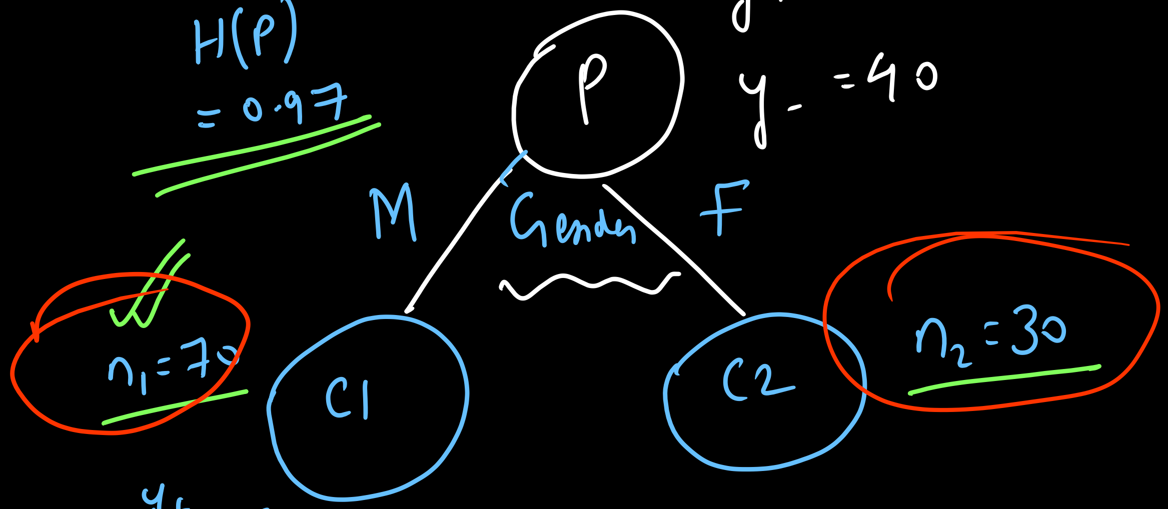
✓ $n = 100$

$y_+ = 60$
$y_- = 40$

Age	Gender	y
20	M	y_+
→ 30	F	⋮
5	F	y_-

$H(P) = 0.97$

$n = 100$
$y_+ = 60$
$y_- = 40$



$y_+ = 50$
$y_- = 20$
$H(c_1) = 0.863$

$\frac{n_1}{n}$

$y_+ = 10$
$y_- = 20$
$p = \frac{10}{30}$
$H(c_2) = 0.918$

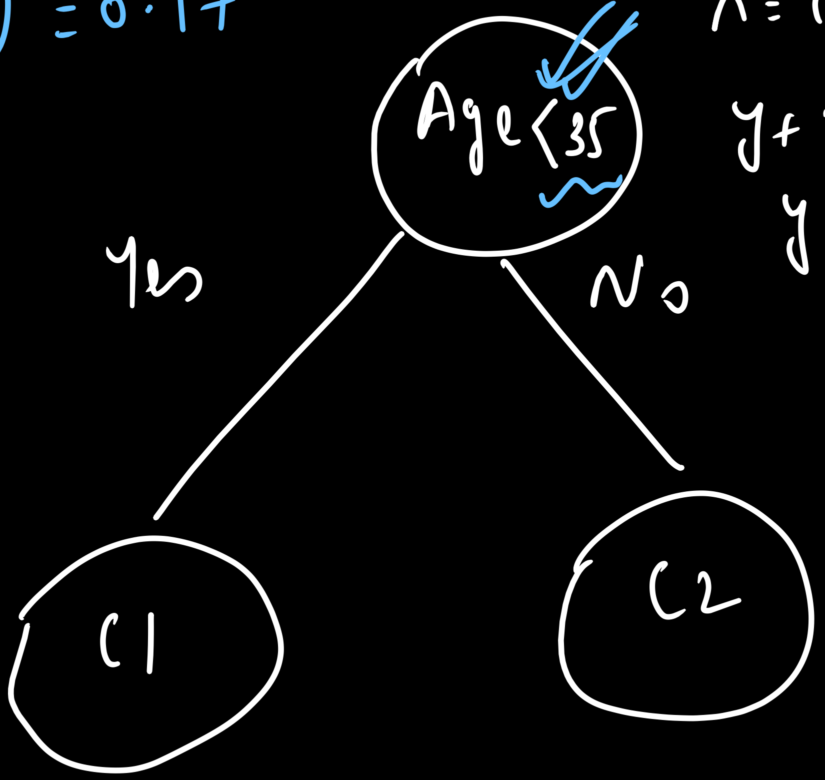
$\frac{n_2}{n}$

$$IG^{(Gender)} = H(P) - \left[\frac{n_1}{n} H(c_1) + \frac{n_2}{n} H(c_2)\right]$$

$$= 0.97 - [0.7 \times 0.863 + 0.3 \times 0.918]$$

$$= 0.0914$$

$H(P) = 0.97$



$n = 100$

$y_+ = 60$

$y_- = 40$

$n_1 = 60$

$y_+ = 50$

$y_- = 10$

$n_2 = 40$

$y_+ = 10$

$y_- = 30$

$H(C_2) = 0.8112$

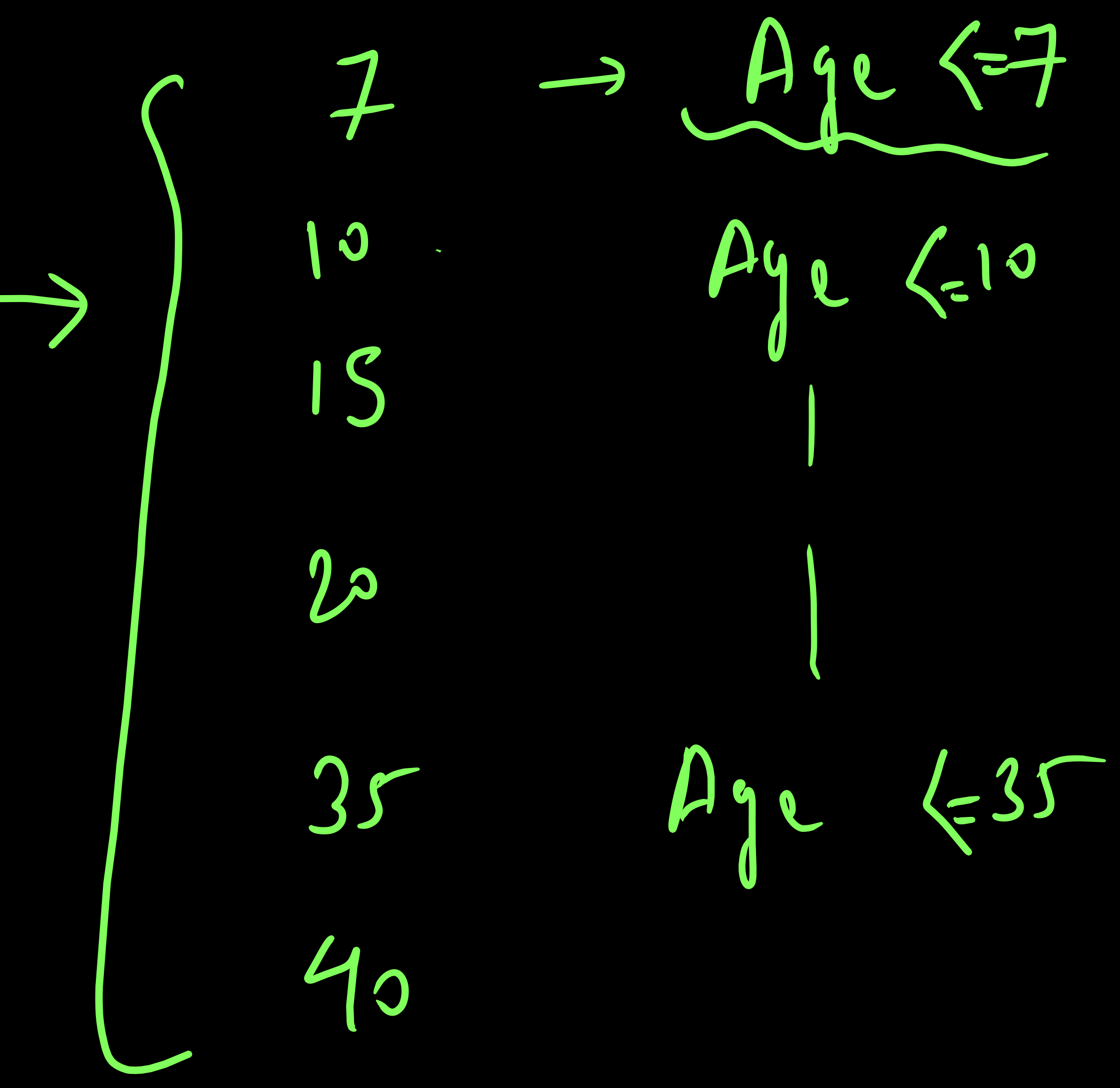
✓ $H(C_1) = 0.65$

$$IG(Age<35) = H(P) - \left[\frac{n_1}{n} H(C_1) + \frac{n_2}{n} H(C_2)\right]$$

$$= 0.97 - \left[\frac{60}{100} \times 0.65 + \frac{40}{100} \times 0.8112\right]$$

$$= 0.2565$$

Age	Gender	y
10	M	y+
20	F	
15		y-
7		
35		
40		



0 - 10

10 - 20

20 - 30

30 - 40