Previous Class - July 1

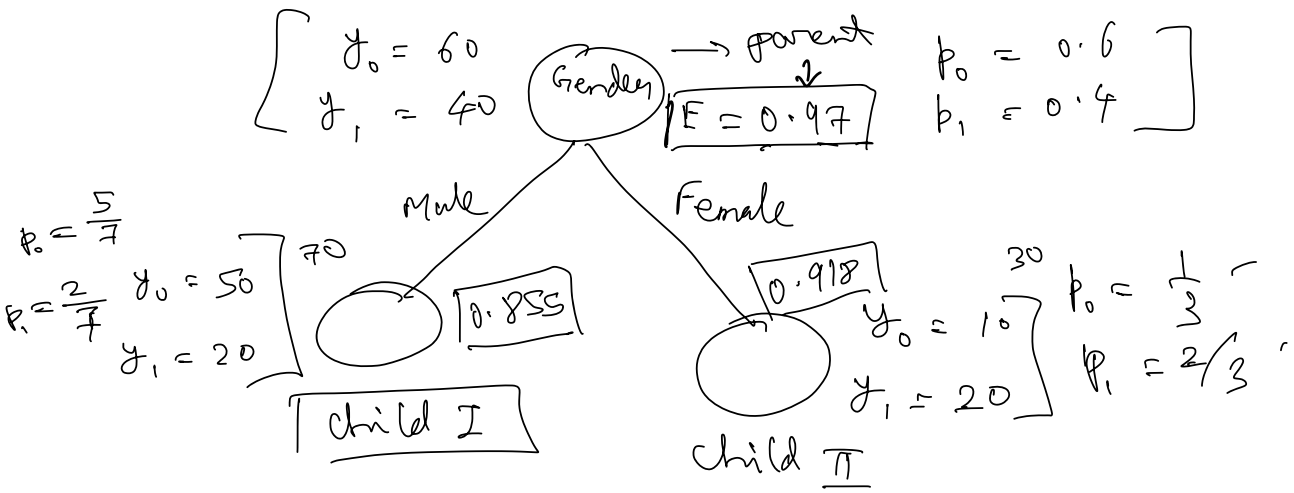
1) Recap - Quizz
2) Employee Attrition Dataset - graph
3) Purity of Nodes & Entropy
4) Plot for entropy
5) Weighted Entropy
6) Gini Impurity
7) Comparing Gini Impurity with Entropy
8) Code walkthrough (Time Permits)

Today's agenda

1) Recap
2) Splitting on numerical features. ✓
3) Overfit vs Underfit. ✓
4) Hyperparameters ✓
5) Visualizing a DT. ✓
6) Impact of outliers - decrease the depth
7) Feature Scaling -

8) Encoding of categorical features
9) DT for high-dim data
10) Data Imbalance
11) Feature Importance ✓
12) DT Regression

Recap

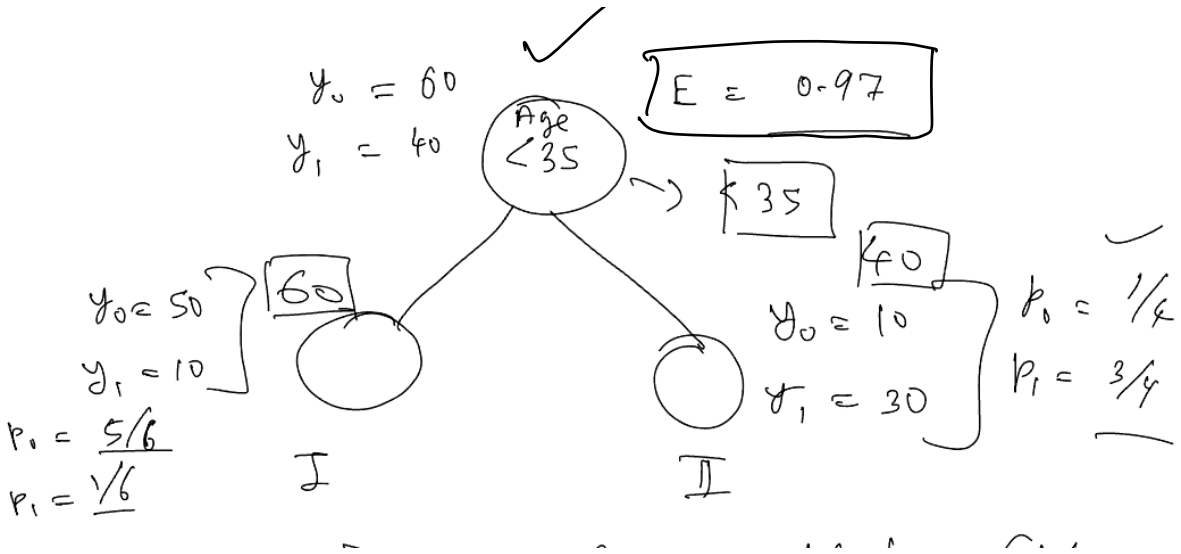


$$WE = 0.874$$

$$IG = 0.97 - 0.874 = \boxed{0.096}$$

↓

$$0.2566 > 0.096$$



$$WE = 0.7144$$

$$IG = 0.97 - 0.7144 = 0.2566$$

$$f_1 = ['A', 'B', 'C', 'D']$$

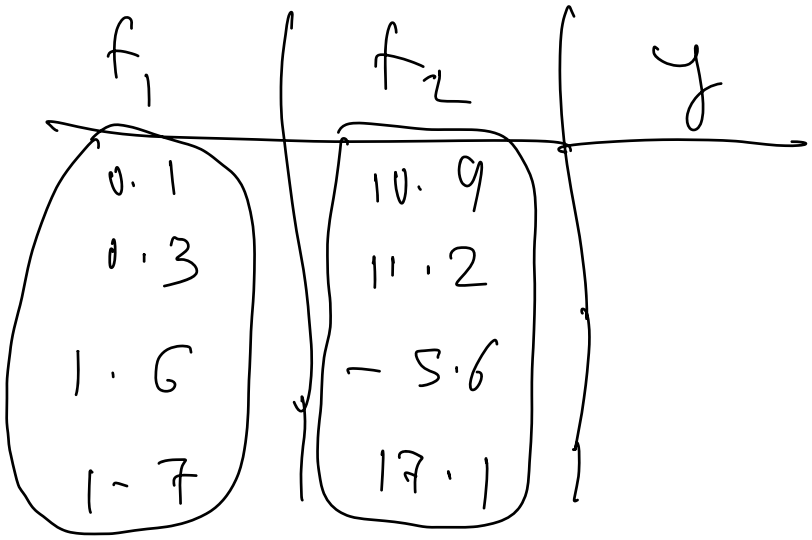
$$f_2 = ['E', 'F', 'G']$$

$f_1 = 'A'$: Yes / No

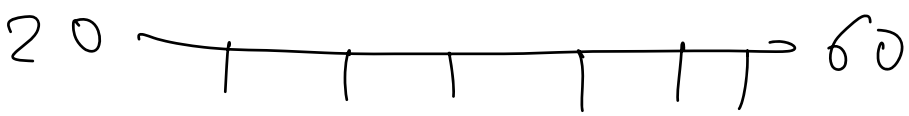
$f_1 = 'B'$: Yes / No

$f_2 = 'E'$: Yes / No

$4 + 3 = 7$



↳ unique values of every feature



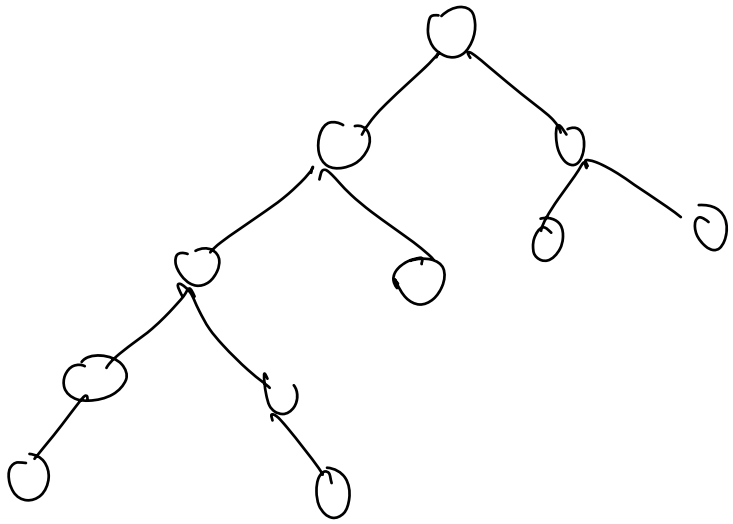
$\frac{40}{10} = 4$

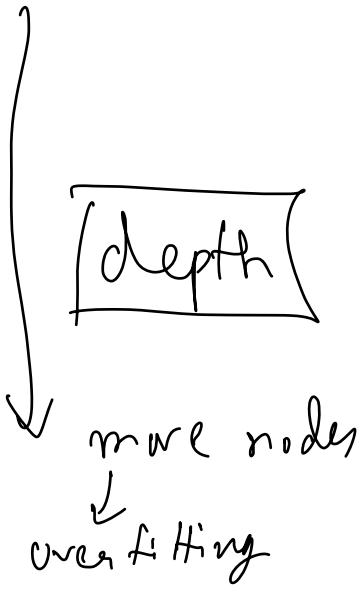
20, 24, 28, 32, 36, ..., 60

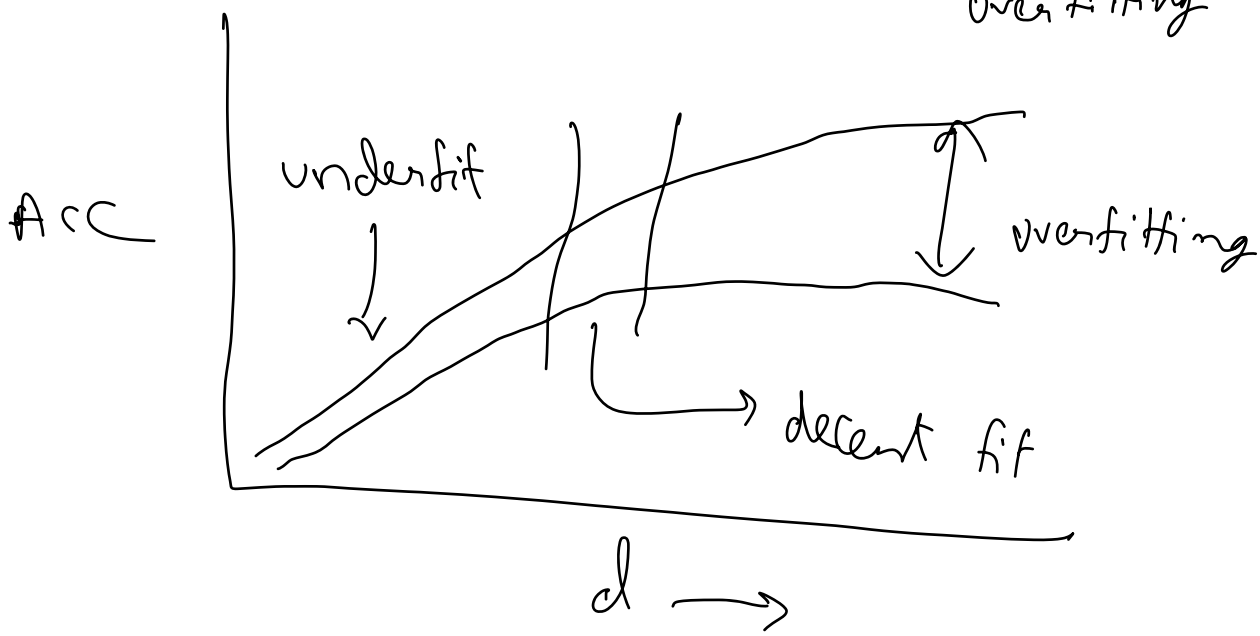
Underfit vs overfit

More nodes ⟶ Overfitting

less nodes ⟶



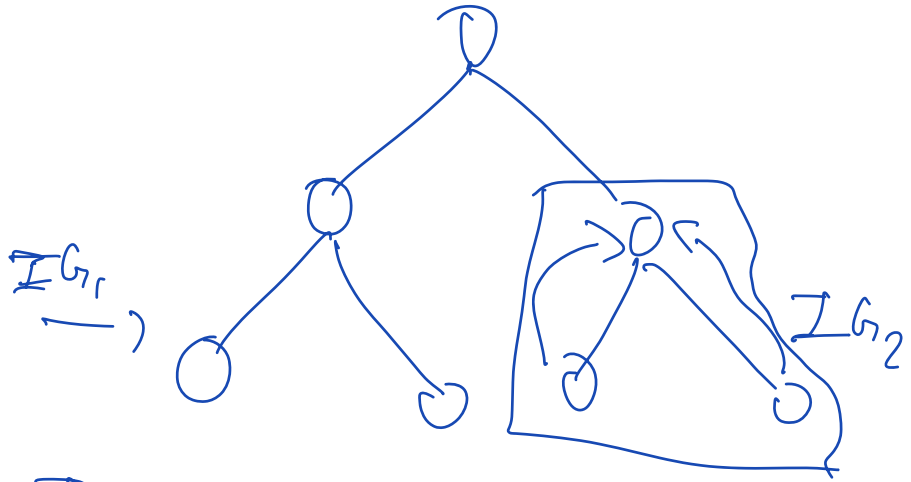




$$-\left[\, p_0 \log_2(p_0) + p_1 \log(p_1) \,\right]$$

$$-\left[\, p \log p + (1-p) \log(1-p) \,\right]$$

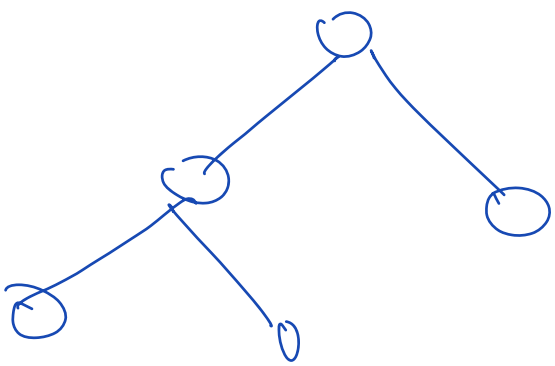
$$-\left[\, w_1\, p \log p + w_2 (1-p) \log(1-p) \,\right]$$



IG_2 is very small

Pruning

removing of unnecessary branches



$y_+ = 50$
$y_- = 50$
100
P

Classification

$y_+ = 30$
$y_- = 20$
L

R
$y_+ = 20$
$y_- = 30$

+ve

-ve

Regression

$f_1 = [0.9, 1.8, -1.9, 17.0, 222.5]$

Mean

$$-[p_1 \log(p_1) + p_2 \log(p_2)]$$

↳ 2 classes

MSE

Feature imp

$f_1, f_2, f_3, \ldots, f_{10}$

$f_1 \rightarrow$ 5 times

$f_2 \rightarrow$ 2 times

$f_3 \rightarrow$ 1 time

$f_1 \propto 5$
$f_2 \propto 2$
$f_3 \propto 1$

$f_{4 \ldots 10} \propto 0$

Categorical features

→ Label encoding (0, 1) $|f| = 2$

→ One hot encoding. $|f| = 3$ to 6

→ Target encoding. $|f| > 6$

↳ probability

N unique values

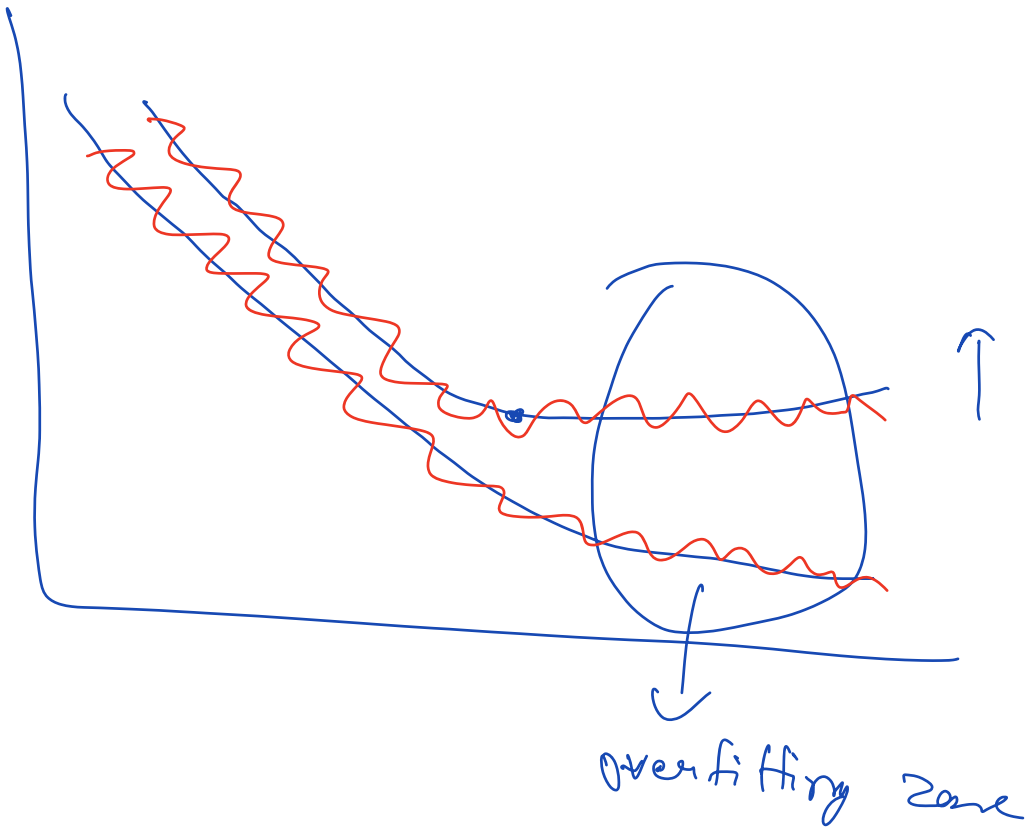
↳ $(N-1)$

$4 \rightarrow p(y=1 \mid f_i = 4)$

f_i	y
4	0
5	0
6	1

f_j
p_4
p_5

$p(y=1 \mid f_i = 5)$


overfitting zone