Last Class (Sept 6)

1) Recap
2) Entropy
3) Information Gain
4) How to split a node — categorical & Numerical
5) Issue with Entropy
6) Gini Impurity
7) Feature Scaling
8) Overfit vs Underfit

Today's class

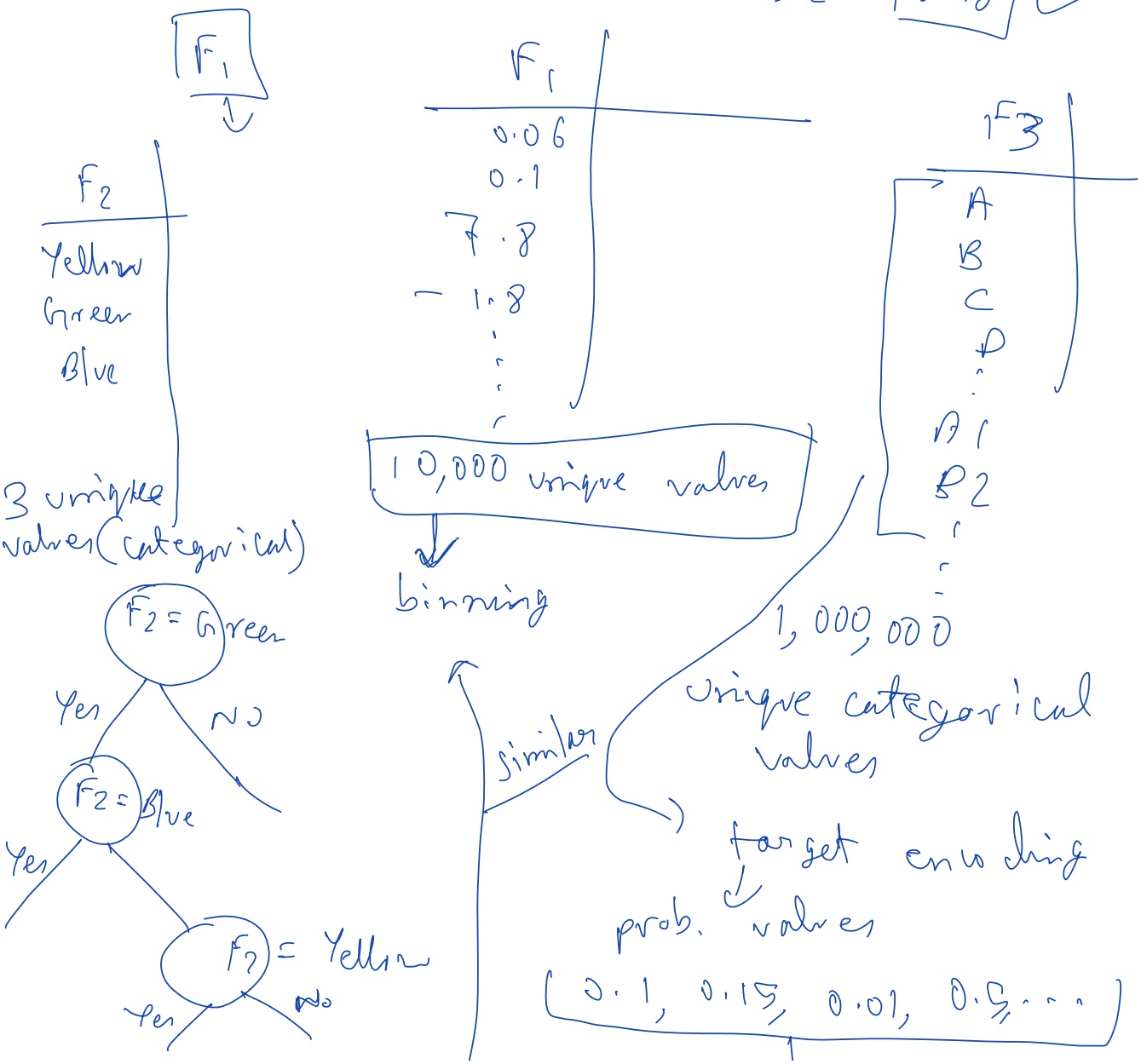
1) Quizzes + Recap
2) Recap of splitting based on numerical & categorical features (encoding them)
3) Recap of max_depth for overfit vs underfit
4) Hyperparameter tuning
5) Visualizing DT
6) High Dimensional Data
7) Data Imbalance
8) Feature Importance
9) Regression using DT
10) Overview of Bagging and Random Forests

$p_1 = \frac{0.4}{40/100}$ $p_2 = \frac{0.6}{60/100}$

$1 - [p_1^2 + p_2^2]$

$1 - [0.4^2 + 0.6^2] = 1 - [0.16 + 0.36]$

$= 1 - 0.52 = \boxed{0.48}$



1,000,000 prob values

↓

binning

Train Acc = 0.99
Test Acc = 0.75] Overfitting

→ DT with large depths → overfitting

1) DT with depth = 5

Train Acc — Test Acc

83% ← 2% → 81%

2) DT with depth = 20

Train Acc — Test Acc

92% ← 8% → 84%

Model 1 → more stable

Model 2 → less memory

Model 1 → less test error → less decisions being made

less time ← less learning happening

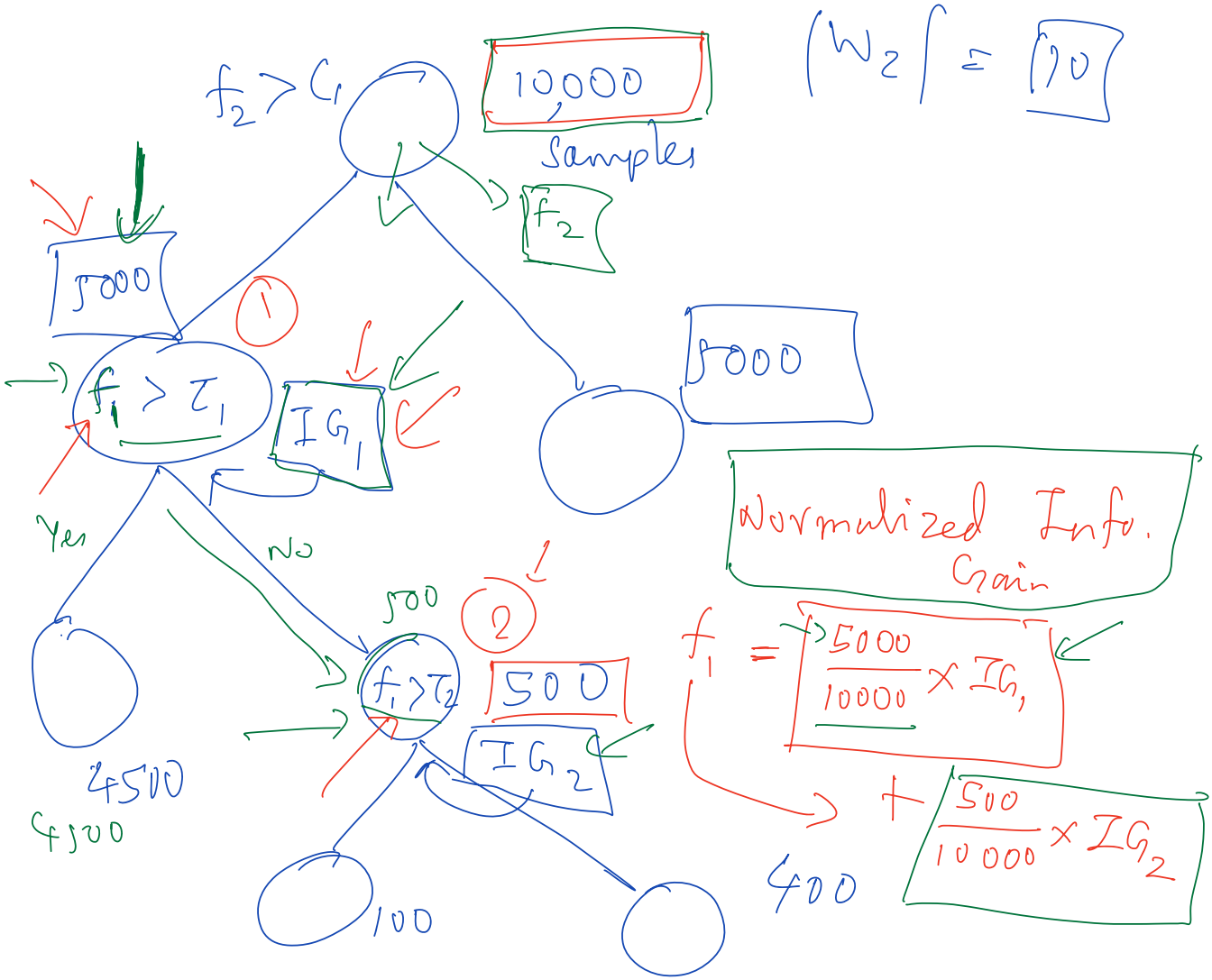
& resources needed

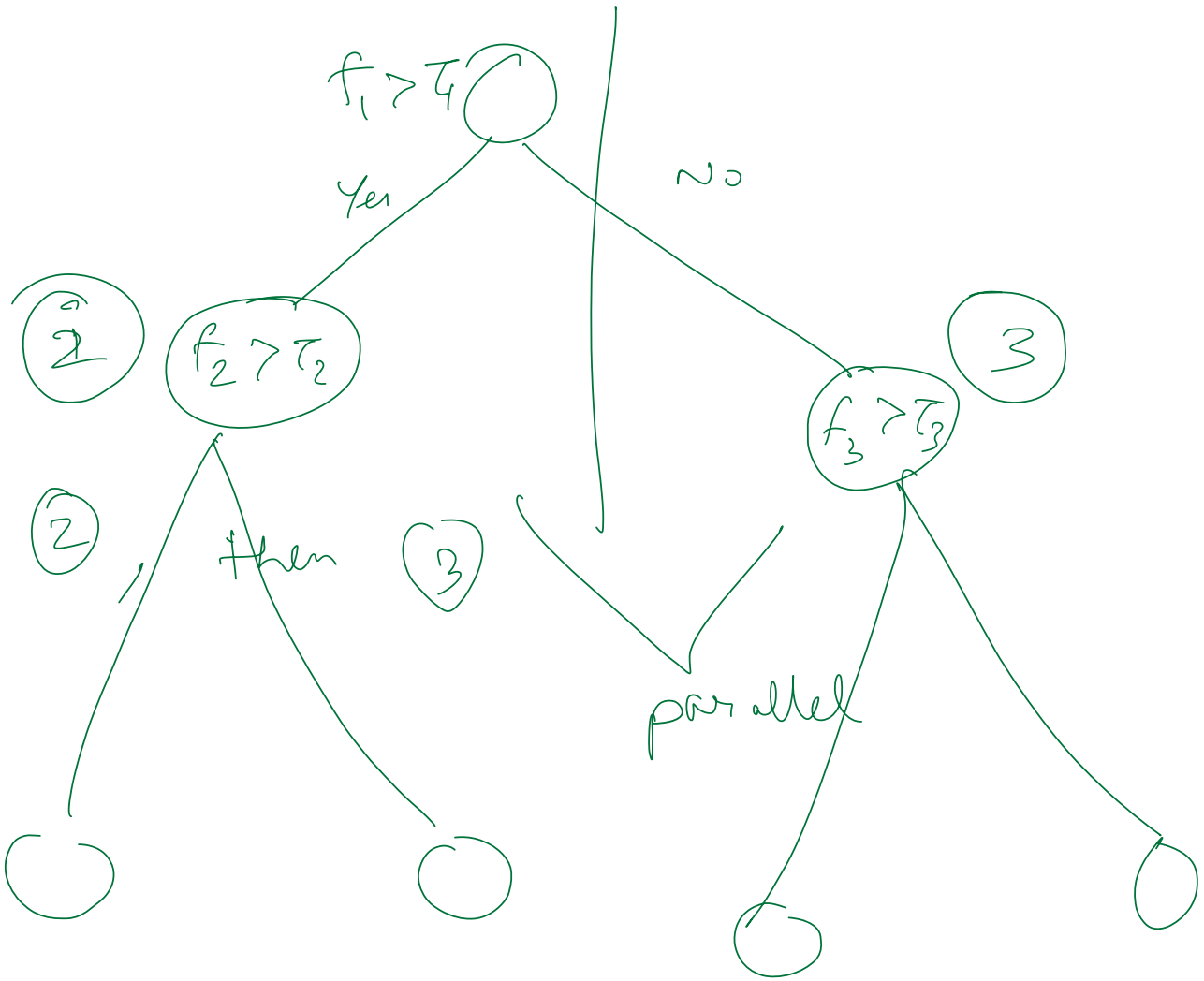
→ Hence model 1 is better in general

$y = 5x_1 \boxed{-10}x_2 + 1.5x_3$

$f_1, f_2, f_3, \ldots, f_{10}$ → x_2 $\quad W_2 = -10$

$|W_2| = \boxed{10}$



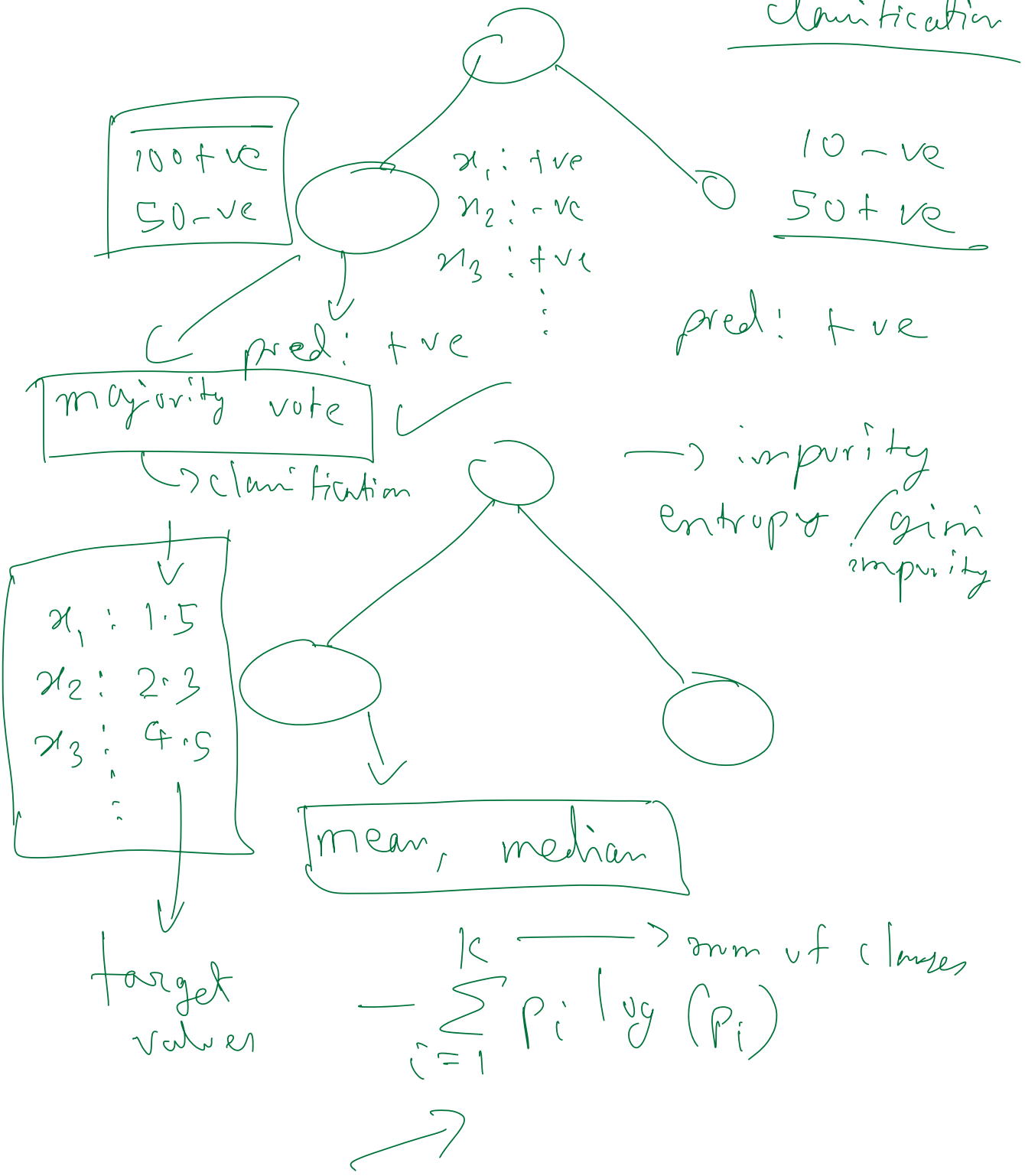


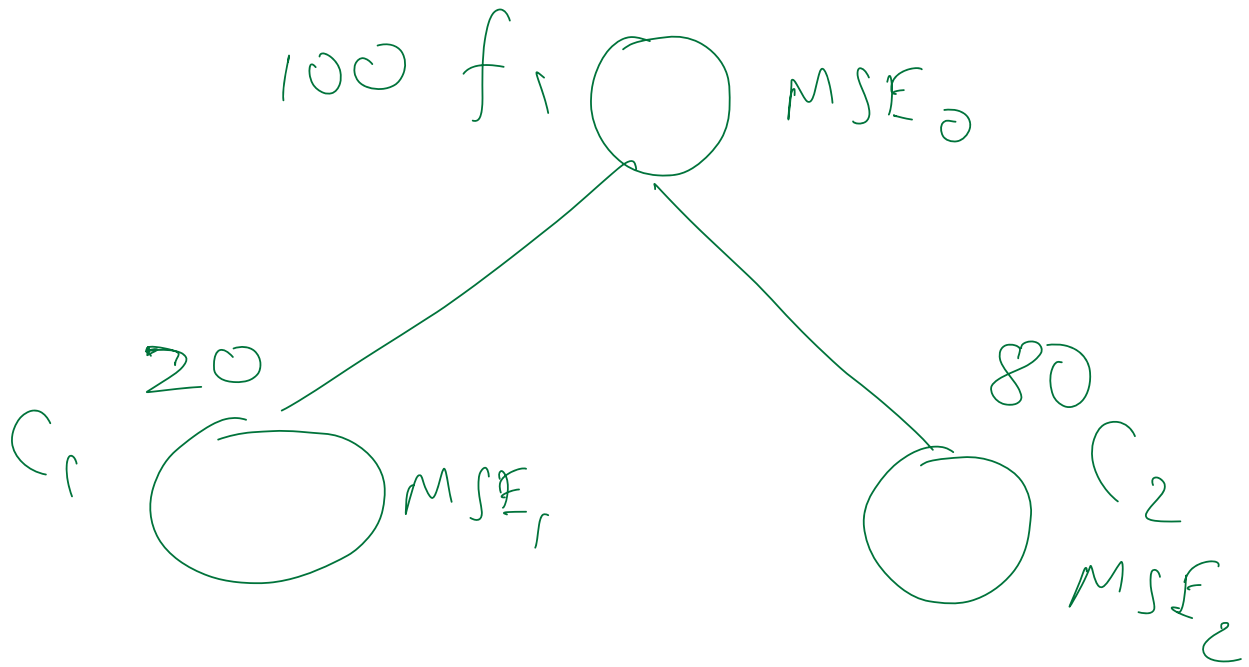
2 cores → $1/2$

4 cores → $1/4$

8 cores → $1/8$

Classification





$$WMSE = \frac{20}{100} \times MSE_1 + \frac{80}{100} \times MSE_2$$

$$IG = \boxed{MSE_0 - WMSE}$$

Generification

$$IG = E_p - W.E_C$$

feature = = category

Male = = Male Male / Female

Female = Male

Male

.

.

.

y [; ;]