Decision Tree (ID3 Algorithm) Classification Example

1 Toy Dataset: Early Diagnosis of Diabetes

AgeGroup	BMI	BP	Glucose	Diabetes
Young	Normal	Normal	Normal	No
Young	Overweight	Normal	High	Yes
Young	Obese	High	High	Yes
Middle	Normal	Normal	Normal	No
Middle	Overweight	High	Normal	No
Middle	Obese	High	High	Yes
Middle	Overweight	Normal	High	Yes
Old	Normal	High	High	Yes
Old	Overweight	Normal	Normal	No
Old	Obese	High	Normal	~~Yes~~ NO
Young	Obese	Normal	High	Yes
Middle	Normal	High	Normal	No
Old	Overweight	High	High	Yes
Old	Obese	High	High	Yes

Table 1: Toy dataset for early diagnosis of Diabetes

Question: If a person is young, overweight, with normal blood pressure and glucose level, does he have Diabetes?

X ,

$y \in [100, 200]$

0~100 : N

100~200 : H

Decision Tree (ID3)

→ ID3 (Iterative Dichotomiser 3) is an early decision tree algorithm, which is used ~~[illegible]~~ for classification problem.

Machine Learning

- Supervised
  - Classification
    - → Binary
    - → Multi-class
  - Regression
- Unsupervised
- Reinforcement Learning

Supervised, Unsupervised → Semi-supervised

Main Idea of Decision Tree (ID3)

→ Build a decision tree top down

→ At each node, choose the feature with maximum Information Gain (IG).
(Reducing uncertainty/entropy the most)

Key Concepts

→ Entropy is the measure of uncertainty.

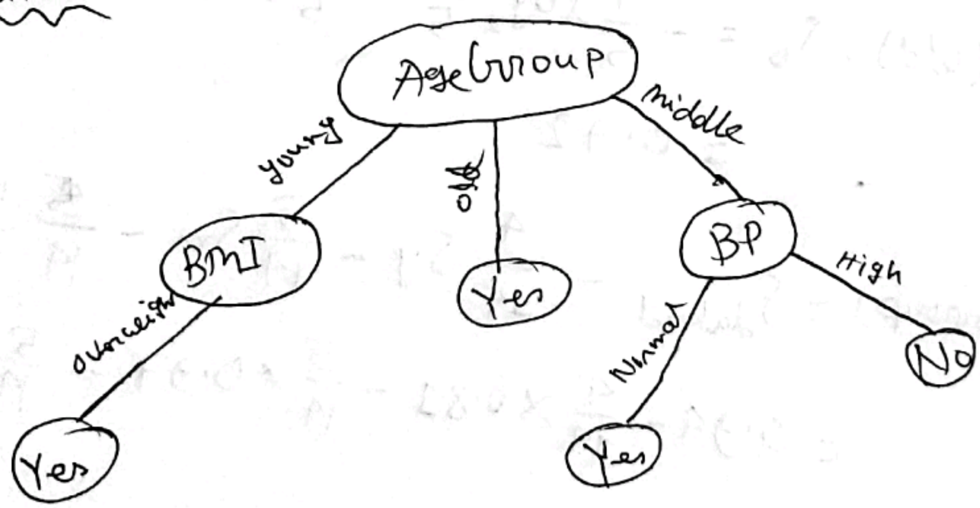
$$H(S) = -\sum_{i=1}^{c} p_i \log_2(p_i) \quad ; \quad H(S) \in [0, 1]$$

if $H(S) = 0$; pure set

if $H(S) = 1$; maximum uncertainty

→ Information Gain (IG) measures reduction in entropy after splitting a dataset on a feature. The feature ~~is~~ with the highest IG is chosen as a split.

Intuition



Classify a person who is ~~with~~ young, overweight, high BP, normal glucose.

Solution:

Entropy (~~entire~~ entire dataset), $S_{dataset}$

$$= -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$$

$$= 0.94$$

Calculating IG (Age Group):

$$\text{Entropy (Young)}, S_y = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}$$

$$= 0.81$$

$$\text{Entropy (Middle)}, S_m = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$$

$$= 0.97$$

$$\text{Entropy (Old)}, S_o = -\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5}$$

$$= 0.72$$

$$\text{IG (Age Group)} = S_{dataset} - \frac{4}{14}S_y - \frac{5}{14}S_m - \frac{5}{14}S_o$$

$$= 0.94 - \frac{4}{14}\times 0.81 - \frac{5}{14}\times 0.97 - \frac{5}{14}\times 0.72$$

$$= 0.10$$

Calculating IG (BMI):

Entropy (Normal), $E_N = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}$

$= 0.81$

Entropy (Overweight), $E_O = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$

$= 0.97$

Entropy (obese), $E_{Ob} = -\frac{5}{5}\log_2\frac{5}{5} - \frac{0}{5}\log_2\frac{0}{5}$

$= 0$

$IG(BMI) = S_{dataset} - \frac{4}{14}\times E_N - \frac{5}{14}\times E_O - \frac{5}{14}\times E_b$

$= 0.94 - \frac{4}{14}\times 0.81 - \frac{5}{14}\times 0.97 - 0$

$= 0.36$

Calculating IG (BP):

$$\text{Entropy (Normal)}, E_N = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}$$

$$= 1$$

$$\text{Entropy (High)}, E_H = -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8}$$

$$= 0.81$$

$$IG(BP) = 0.94 - \frac{6}{14}\times 1 - \frac{8}{14}\times 0.81 = 0.04$$

Calculating IG (Glucose):

$$\text{Entropy (Normal)}, E_N = -\frac{1}{6}\log_2\frac{1}{6} - \frac{5}{6}\log_2\frac{5}{6}$$

$$= 0.65$$

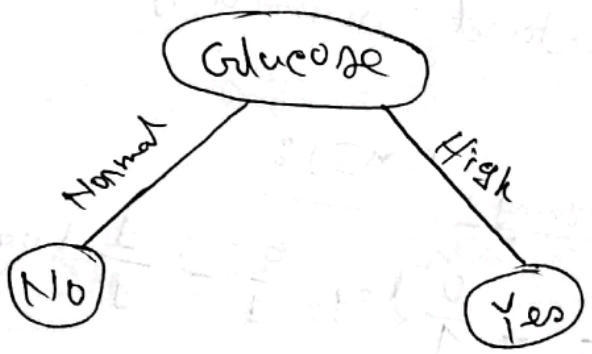
$$\text{Entropy (High)}, E_H = -\frac{8}{8}\log_2\frac{8}{8} - \frac{0}{8}\log_2\frac{0}{8}$$

$$= 0$$

$$IG(\text{Glucose}) = 0.94 - \frac{6}{14}\times 0.65 - 0$$

$$= 0.66$$

Features	IG
AgeGroup	0.10
BMI	0.36
BP	0.04
Glucose	0.66 → max



Note

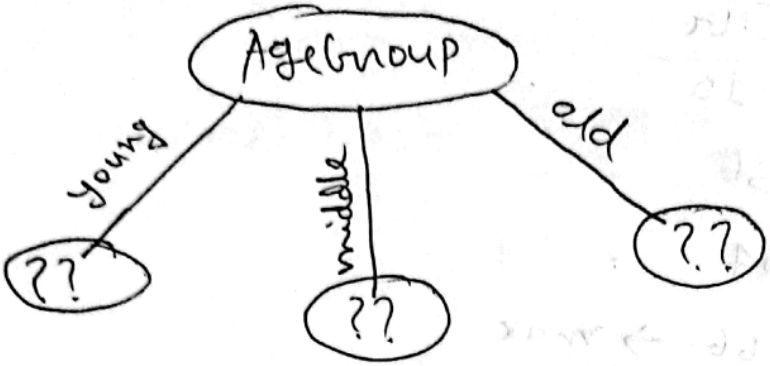
correct format of writing IG:

$IG(S, AgeGroup) = 0.10$

$IG(S, BMI) = 0.36$

$IG(S, BP) = 0.04$

$IG(S, Glucose) = 0.66$



Create a new dataset subsetting "Young"

Calculating $IG(S_{young}, BMI)$:

$$\text{Entropy (Normal)} = -\frac{0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1}$$

$$= 0$$

$$\text{Entropy (Overweight)} = 0$$

$$\text{Entropy (obese)} = 0$$

$$IG(S_{young}, BMI) = \text{Entrop(young)} - 0$$

$$= -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{3}\log_2\frac{1}{3}$$

$$= 0.84$$

Calculating IG(Syoung, BP):

$$\text{Entropy (Normal)} = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}$$

$$= 0.91$$

$$\text{Entropy (High)} = 0$$

$$IG(S_{young}, BP) = 0.84 - \frac{3}{4} \times 0.91 = 0.15$$

Calculating IG(Syoung, Glucose):

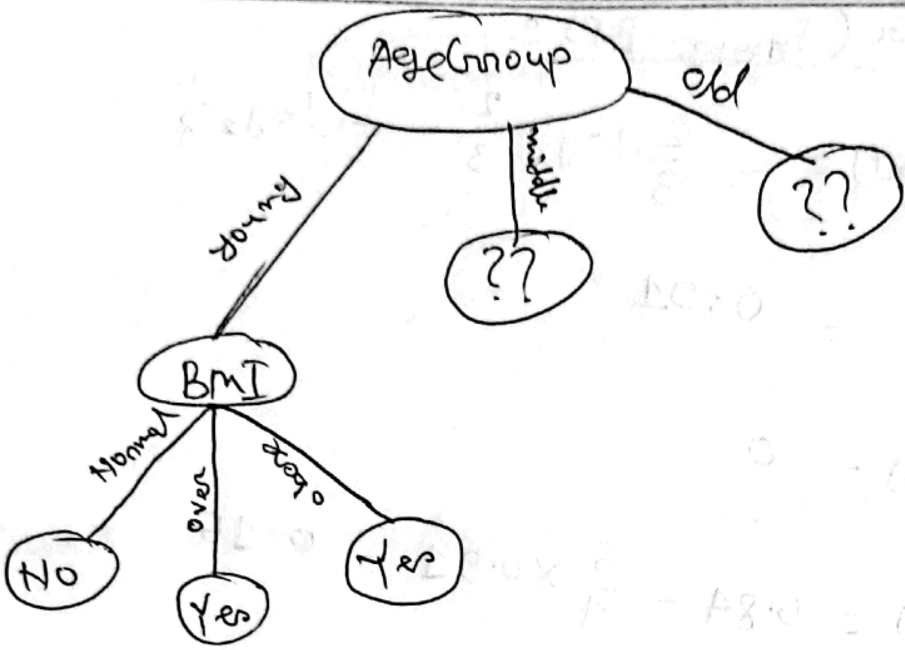
$$\text{Entropy (Normal)} = 0$$

$$\text{Entropy (High)} = 0$$

$$IG(S_{young}, \text{Glucose}) = 0.84$$

Features	IG	
BMI	0.84	→ max
BP	0.15	
Glucose	0.84	→ max

} pick 1 randomly



Create a new dataset subsetting 'Middle'

$$\text{Entropy}(S_{middle}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$$

$$= 0.97$$

Calculating $IG(S_{middle}, BMI)$:

$\text{Entropy}(\text{Normal}) = 0$

$$\text{Entropy}(\text{Overweight}) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}$$

$$= 1$$

$\text{Entropy}(\text{obese}) = 0$

$$IG(S_{middle}, BMI) = 0.97 - \frac{2}{5}\times 1 = 0.57$$

Calculating IG (Smiddle, BP):

$$\text{Entropy (Normal)} = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}$$

$$= 1$$

$$\text{Entropy (High)} = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$$

$$= 0.92$$

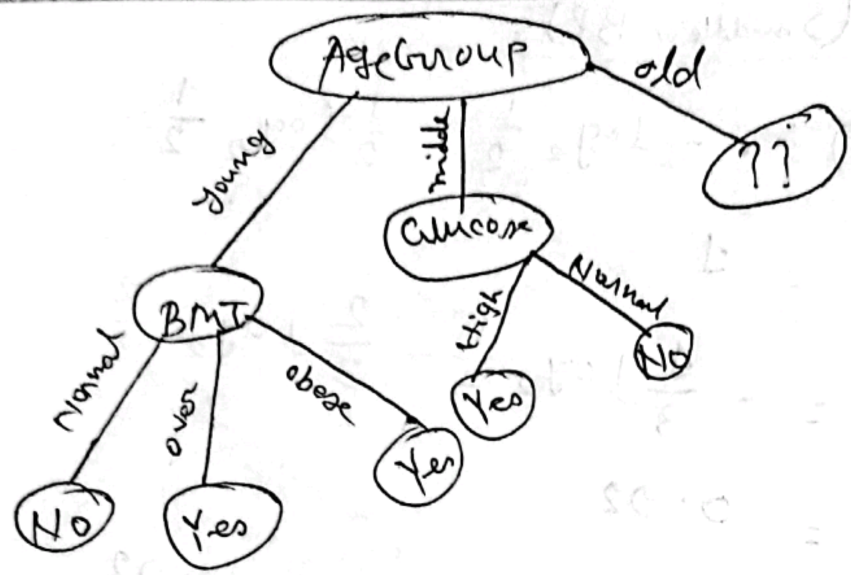
$$IG(\text{Smiddle, BP}) = 0.97 - \frac{2}{5}\times 1 - \frac{3}{5}\times 0.92$$

$$= 0.018$$

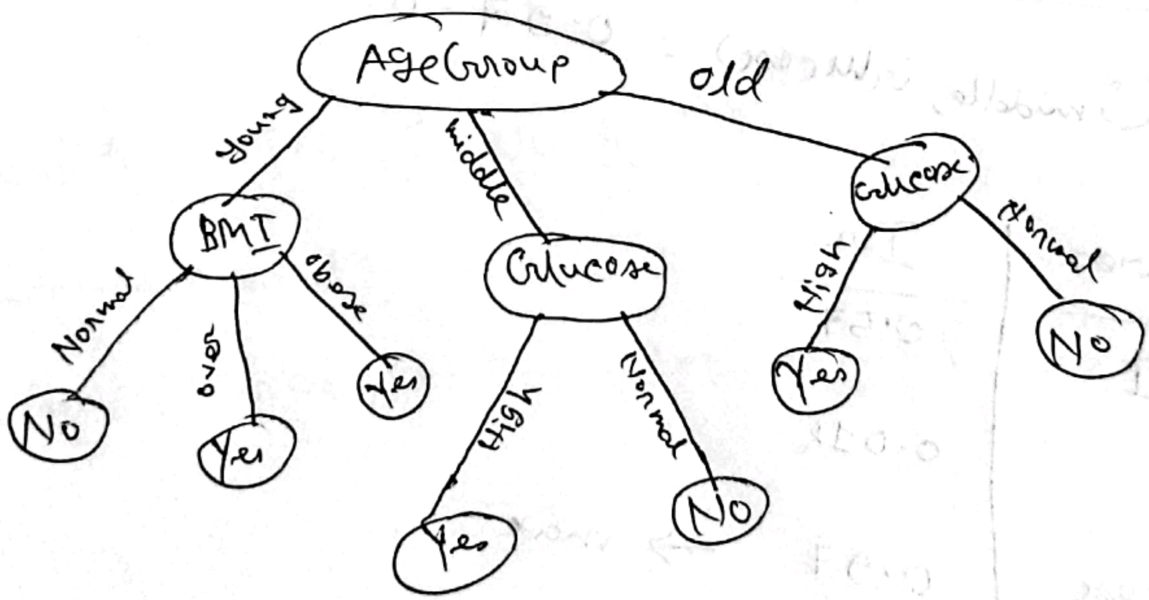
Calculating IG (Smiddle, Glucose):

$$IG(\text{Smiddle, Glucose}) = 0.97 - 0 = 0.97$$

Features	IG
BMI	0.57
BP	0.018
Glucose	0.97 → max



Similarly, calculate root node for old. You'll see that Glucose becomes the root node. So, our final tree :



Pros of Decision Tree (ID3)

→ Interpretable ML

→ Handles categorical data well

→ Works very fast

Cons

→ Not good for continuous data (binning required)

→ Can overfit training data