



INTRODUCTION TO DATA ANALYTICS

Classification Analysis

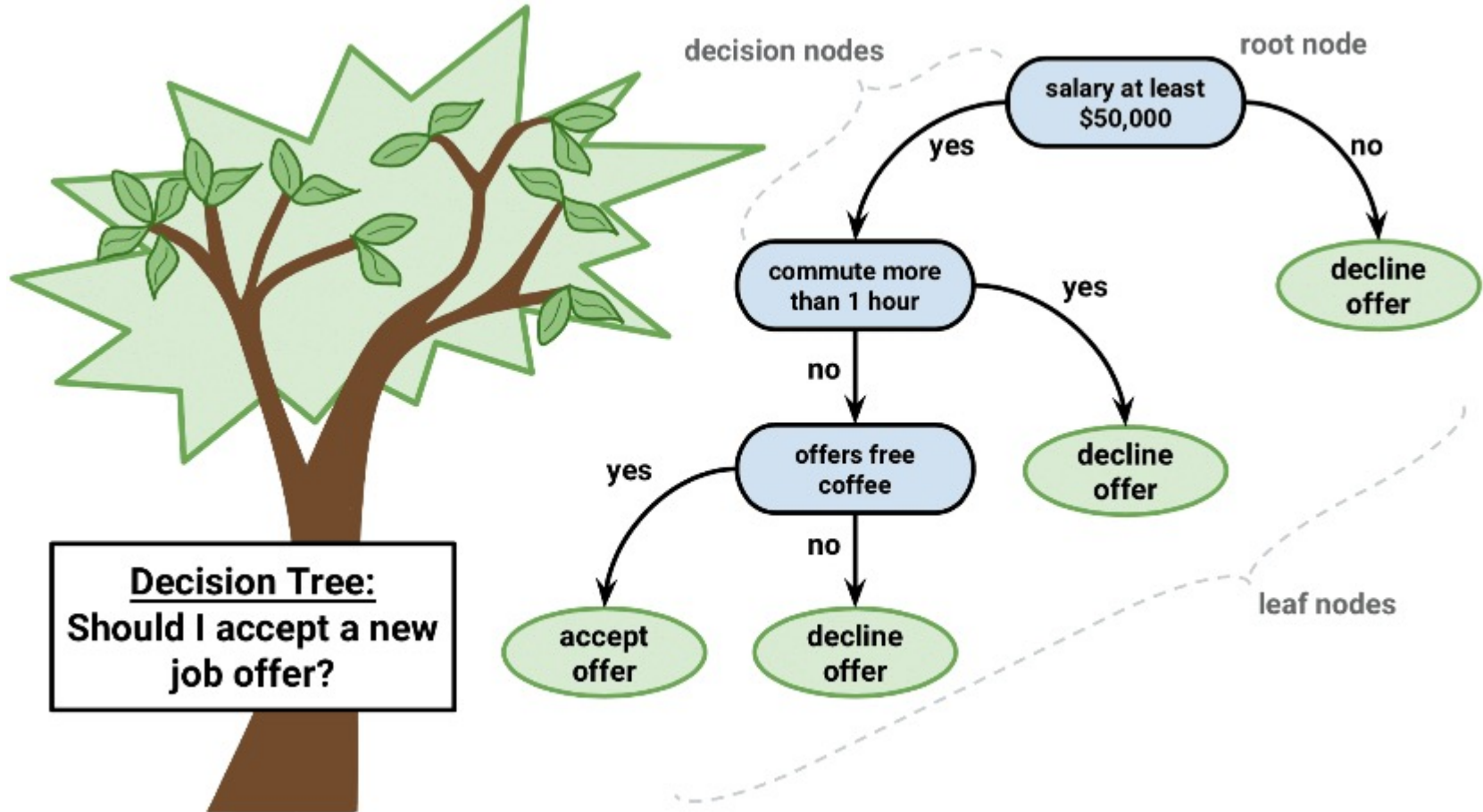
Asst. Prof. Dr. Rathachai Chawuthai

Department of Computer Engineering

Faculty of Engineering

King Mongkut's Institute of Technology Ladkrabang

Decision Tree



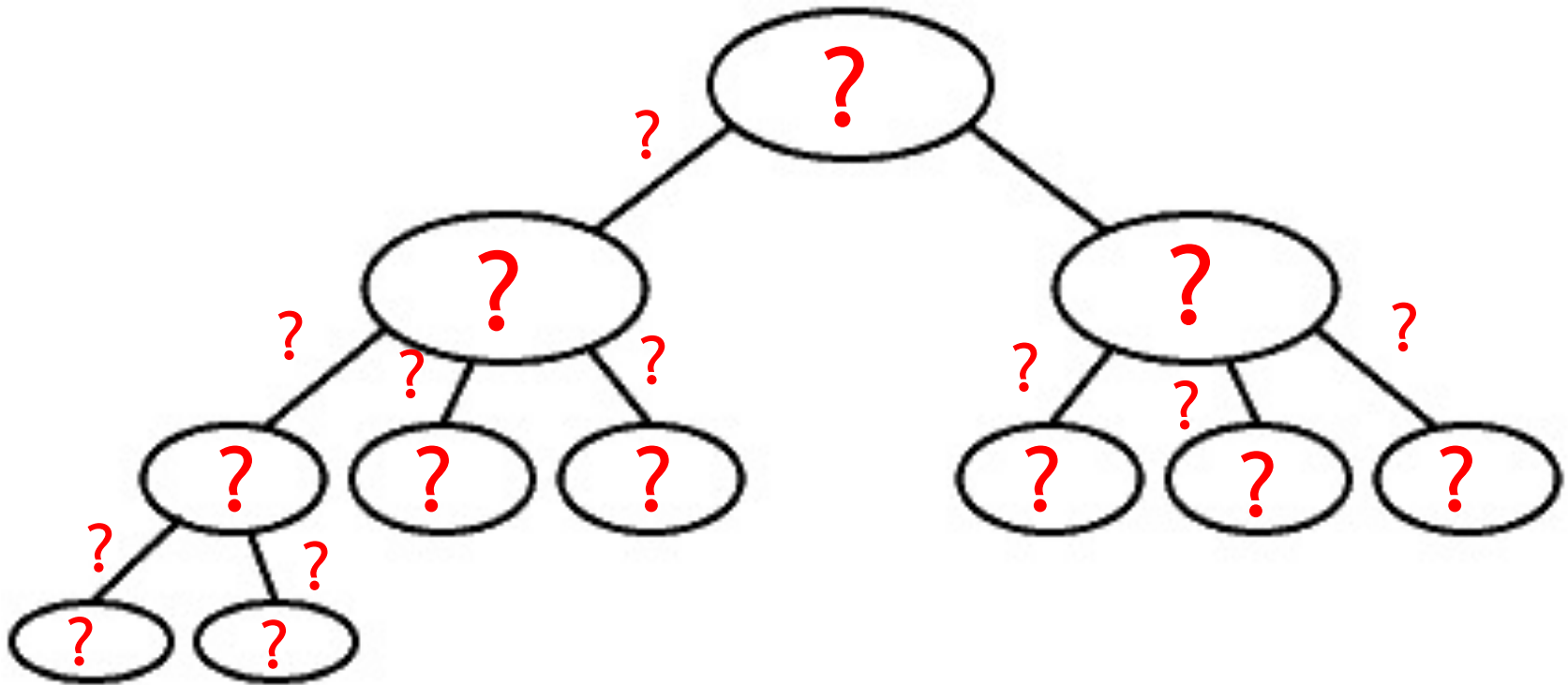
Decision Tree

- A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.
- Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning.
- A decision tree consists of three types of nodes:
 - Decision nodes – typically represented by squares
 - Chance nodes – typically represented by circles
 - End nodes – typically represented by triangles

Case

CASE	TEMPERATURE	HEADACHE	NAUSEA	Flu
1	high	YES	no	YES
2	very_high	YES	YES	YES
3	normal	no	no	no
4	high	YES	YES	YES
5	high	no	YES	no
6	normal	YES	no	no
7	normal	no	YES	no

Tree



Case

CASE	TEMPERATURE	HEADACHE	NAUSEA	Flu
1	high	YES	no	YES
2	very_high	YES	YES	YES
3	normal	no	no	no
4	high	YES	YES	YES
5	high	no	YES	no
6	normal	YES	no	no
7	normal	no	YES	no

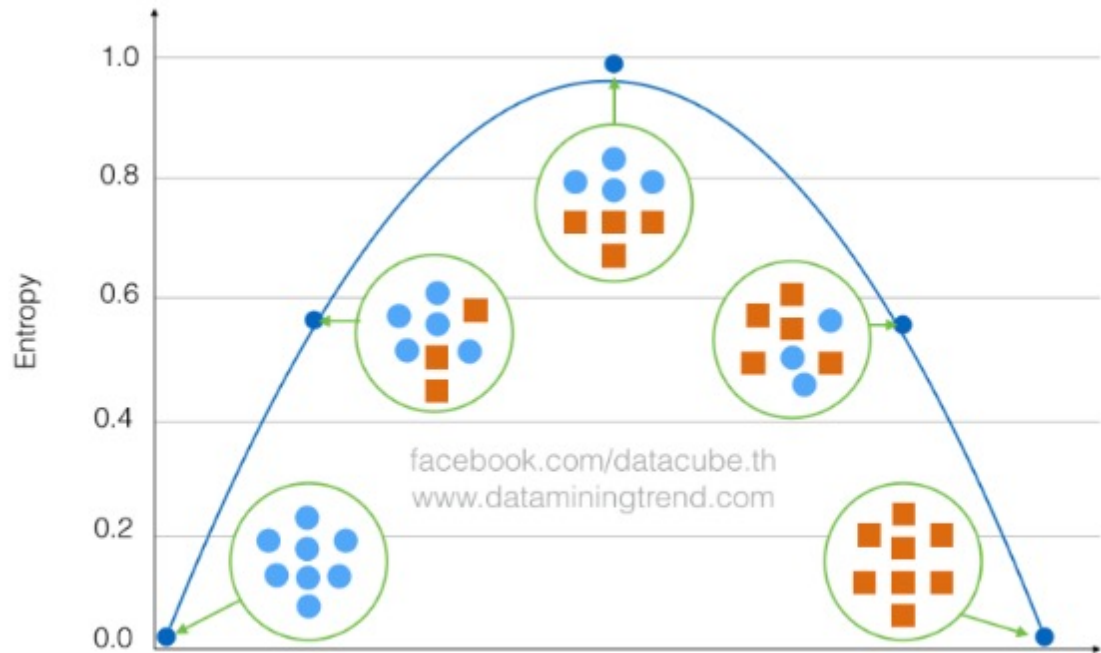
Case

CASE	TEMPERATURE	HEADACHE	NAUSEA	Flu
1	high	YES	no	YES
2	very_high	YES	YES	YES
3	normal	no	no	no
4	high	YES	YES	YES
5	high	no	YES	no
6	normal	YES	no	no
7	normal	no	YES	no

Information Gain

- Same things --> Low
- Different things --> High

$$Entropy = \sum_{i=1}^C -p_i * \log_2(p_i)$$



Decision Tree: Construction

Finding entropy H of every attribute

1) Headache

CASE	TEMPERATURE	HEADACHE	NAUSEA	Flu
1	high	YES	no	YES
2	very_high	YES	YES	YES
3	normal	no	no	no
4	high	YES	YES	YES
5	high	no	YES	no
6	normal	YES	no	no
7	normal	no	YES	no

Headache = YES

Headache = no

$H(\text{Flu} \mid \text{Headache}) =$

-

-

Decision Tree: Construction

Finding entropy H of
every attribute

1) Headache

CASE	TEMPERATURE	HEADACHE	NAUSEA	Flu
1	high	YES	no	YES
2	very_high	YES	YES	YES
3	normal	no	no	no
4	high	YES	YES	YES
5	high	no	YES	no
6	normal	YES	no	no
7	normal	no	YES	no

Headache = YES

$H(\text{Flu} \mid \text{Headache}) =$

-

$\frac{4}{7}$

Headache = no

$\frac{3}{7}$

Decision Tree: Construction

Finding entropy H of every attribute

1) Headache

CASE	TEMPERATURE	HEADACHE	NAUSEA	Flu
1	high	YES	no	YES
2	very_high	YES	YES	YES
3	normal	no	no	no
4	high	YES	YES	YES
5	high	no	YES	no
6	normal	YES	no	no
7	normal	no	YES	no

Headache = YES

$H(\text{Flu} \mid \text{Headache}) =$

$$- \frac{4}{7} \left[\boxed{} + \boxed{} \right]$$

Flu=YES

Flu=no

Headache = no

$$- \frac{3}{7} \left[\boxed{} + \boxed{} \right]$$

Flu=YES

Flu=no

Decision Tree: Construction

Finding entropy H of every attribute

1) Headache

CASE	TEMPERATURE	HEADACHE	NAUSEA	Flu
1	high	YES	no	YES
2	very_high	YES	YES	YES
3	normal	no	no	no
4	high	YES	YES	YES
5	high	no	YES	no
6	normal	YES	no	no
7	normal	no	YES	no

Headache = YES

Headache = no

$H(\text{Flu} \mid \text{Headache}) =$

$$- \frac{4}{7} \left[\frac{3}{4} \log_2 \left(\frac{3}{4} \right) + \frac{1}{4} \log_2 \left(\frac{1}{4} \right) \right] - \frac{3}{7} \left[\quad + \quad \right]$$

↑
↑

Flu=YES
Flu=no

Decision Tree: Construction

Finding entropy H of every attribute

1) Headache

CASE	TEMPERATURE	HEADACHE	NAUSEA	Flu
1	high	YES	no	YES
2	very_high	YES	YES	YES
3	normal	no	no	no
4	high	YES	YES	YES
5	high	no	YES	no
6	normal	YES	no	no
7	normal	no	YES	no

Headache = YES

$H(\text{Flu} \mid \text{Headache}) =$

$$- \frac{4}{7} \left[\frac{3}{4} \log_2 \left(\frac{3}{4} \right) + \frac{1}{4} \log_2 \left(\frac{1}{4} \right) \right]$$

Headache = no

$$- \frac{3}{7} \left[\frac{0}{3} \log_2 \left(\frac{0}{3} \right) + \frac{3}{3} \log_2 \left(\frac{3}{3} \right) \right]$$

Flu=YES

Flu=no

Decision Tree: Construction

Finding entropy H of
every attribute

1) Headache

CASE	TEMPERATURE	HEADACHE	NAUSEA	Flu
1	high	YES	no	YES
2	very_high	YES	YES	YES
3	normal	no	no	no
4	high	YES	YES	YES
5	high	no	YES	no
6	normal	YES	no	no
7	normal	no	YES	no

$$\begin{aligned} H(\text{Flu} \mid \text{Headache}) = & -\frac{4}{7} \left[\frac{3}{4} \log_2 \left(\frac{3}{4} \right) + \frac{1}{4} \log_2 \left(\frac{1}{4} \right) \right] - \frac{3}{7} \left[\frac{0}{3} \log_2 \left(\frac{0}{4} \right) + \frac{3}{3} \log_2 \left(\frac{3}{3} \right) \right] \\ & = 0.464 \end{aligned}$$

Note: In the original image, red arrows point from the terms $\frac{0}{3} \log_2 \left(\frac{0}{4} \right)$ and $\frac{3}{3} \log_2 \left(\frac{3}{3} \right)$ to a red '0' above the second bracket, indicating that these terms are zero.

Decision Tree: Construction

Finding entropy H
of every attribute

CASE	TEMPERATURE	HEADACHE	NAUSEA	Flu
1	high	YES	no	YES
2	very_high	YES	YES	YES
3	normal	no	no	no
4	high	YES	YES	YES
5	high	no	YES	no
6	normal	YES	no	no
7	normal	no	YES	no

$$H(\text{Flu} \mid \text{Headache}) = 0.464$$

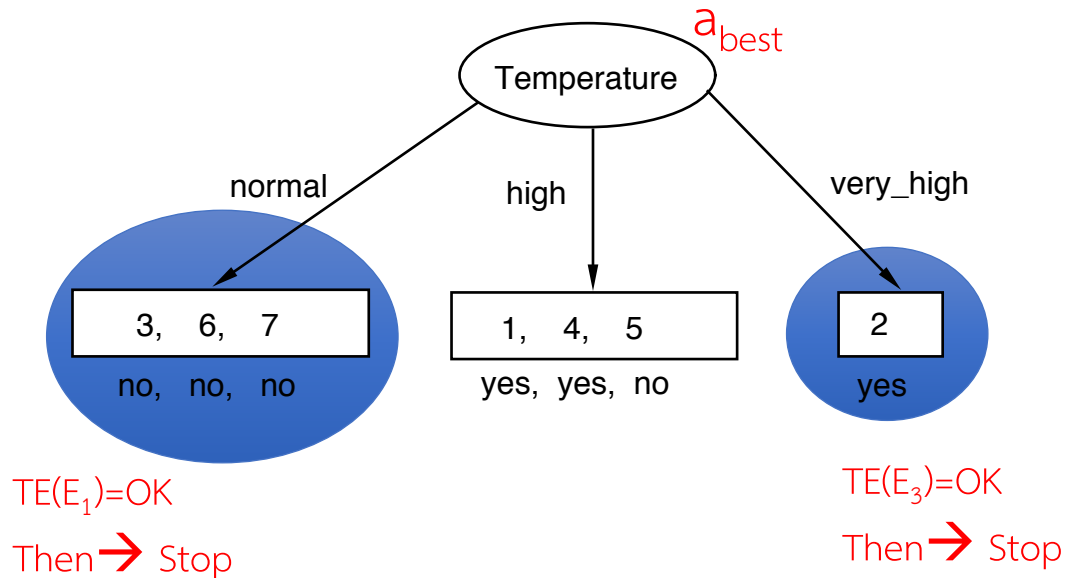
$$H(\text{Flu} \mid \text{Temperature}) = 0.394$$

$$H(\text{Flu} \mid \text{Nausea}) = 0.620$$

Lowest

Decision Tree: Construction

CASE	TEMPERATURE	HEADACHE	NAUSEA	Flu
1	high	YES	no	YES
2	very_high	YES	YES	YES
3	normal	no	no	no
4	high	YES	YES	YES
5	high	no	YES	no
6	normal	YES	no	no
7	normal	no	YES	no



$$H(\text{Flu} \mid \text{Headache}) = 0.464$$

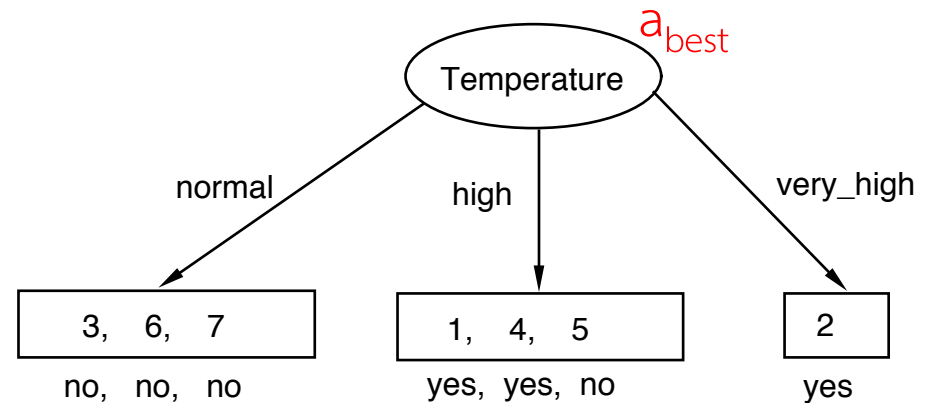
$$H(\text{Flu} \mid \text{Temperature}) = 0.394$$

$$H(\text{Flu} \mid \text{Nausea}) = 0.620$$

Lowest

Decision Tree: Construction

CASE	TEMPERATURE	HEADACHE	NAUSEA	Flu
1	high	YES	no	YES
2	very_high	YES	YES	YES
3	normal	no	no	no
4	high	YES	YES	YES
5	high	no	YES	no
6	normal	YES	no	no
7	normal	no	YES	no



Finding entropy H of attributes from the remainings

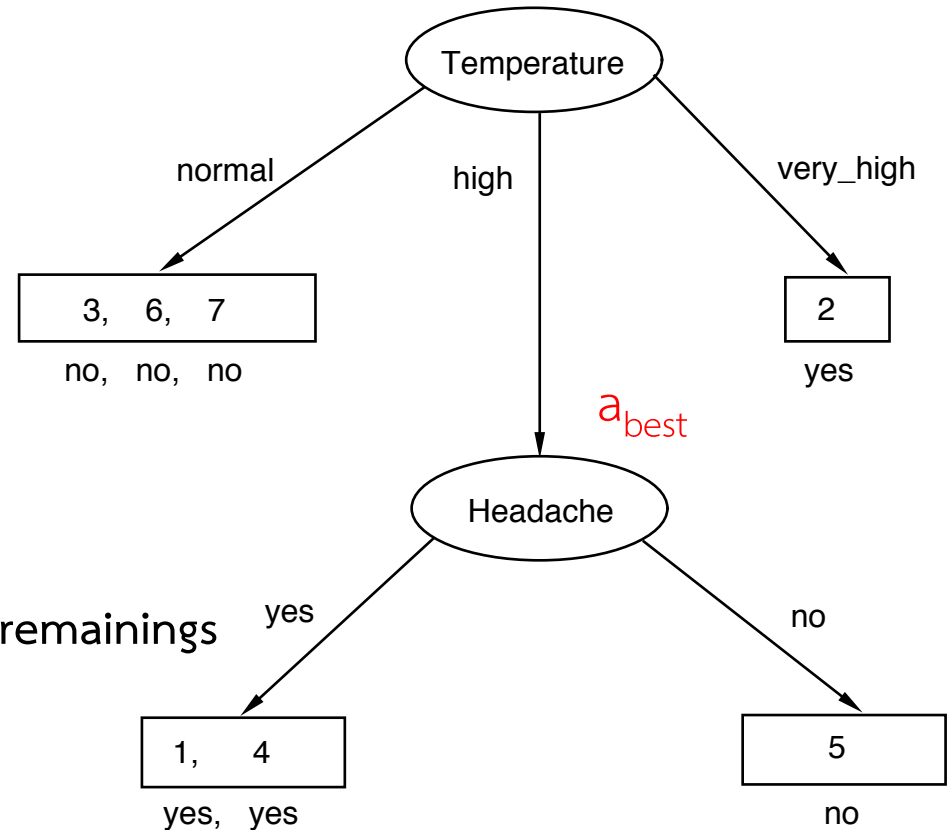
$$H(\text{Flu}_{\text{Temperature} = \text{high}} \mid \text{Headache}) = 0$$

Lowest

$$H(\text{Flu}_{\text{Temperature} = \text{high}} \mid \text{Nausea}) = 0.667$$

Decision Tree: Construction

CASE	TEMPERATURE	HEADACHE	NAUSEA	Flu
1	high	YES	no	YES
2	very_high	YES	YES	YES
3	normal	no	no	no
4	high	YES	YES	YES
5	high	no	YES	no
6	normal	YES	no	no
7	normal	no	YES	no



Finding entropy H of attributes from the remainings

$$H(\text{Flu}_{\text{Temperature} = \text{high}} \mid \text{Headache}) = 0$$

$$H(\text{Flu}_{\text{Temperature} = \text{high}} \mid \text{Nausea}) = 0.667$$