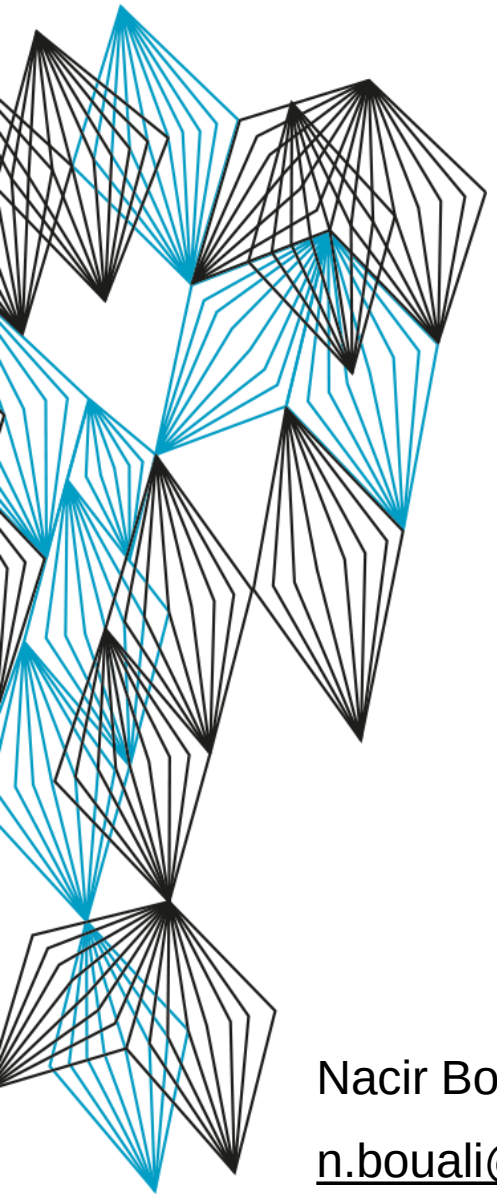

ARTIFICIAL INTELLIGENCE & CYBER SECURITY

MACHINE LEARNING

INFORMATION GAIN

Nacir Bouali

n.bouali@utwente.nl

UNIVERSITY OF TWENTE.

ALGORITHMIC METHOD

- We will work with information gain.
- Information gain is based on the entropy of a dataset.

$$Gain(S, A) = Entropy(S) - \sum_{v \in A} \frac{|Sv|}{|S|} Entropy(Sv)$$

INFORMATION GAIN - FEVER

Ex.	Fever	Headache	Fatigue	Cough	TasteSmell Loss	Covid
1	Yes	Mild	No	Wet	Taste	Yes
2	No	Strong	No	Wet	Both	Yes
3	Yes	Mild	No	Absent	Taste	No
4	No	Mild	No	Wet	Smell	Yes
5	No	Mild	Yes	Wet	Taste	Yes
6	Yes	Lite	No	Absent	Smell	Yes
7	No	Lite	No	Dry	Smell	No
8	No	Strong	No	Absent	Taste	Yes
9	No	Absent	Yes	Wet	Both	No
10	No	Mild	No	Wet	Smell	Yes
11	No	Absent	Yes	Absent	Taste	No
12	No	Lite	No	Dry	Both	No
13	Yes	Strong	No	Wet	Both	Yes
14	No	Absent	Yes	Wet	Both	No
15	No	Mild	No	Absent	Taste	Yes
16	Yes	Lite	No	Dry	Smell	No
17	No	Absent	No	Absent	Taste	Yes
18	No	Absent	No	Wet	Smell	Yes
19	No	Lite	No	Dry	Both	No
20	No	Lite	No	Dry	Smell	No

We calculated this before

$$G(S,A) = E(S) - \sum_{v \in A} \frac{|Sv|}{|S|} E\,(Sv)$$

$$G(S,Fever) = 0.99 - \sum_{v \in \{yes,no\}} \frac{|Sv|}{|S|} E\,(Sv)$$

Attribute/Feature Fever takes two values: yes/no.
For the value yes, we have:

$$\left[\frac{|Syes|}{|S|} E\,(S_{yes})\right.$$

INFORMATION GAIN - FEVER

Ex.	Fever	Headache	Fatigue	Cough	TasteSmell Loss	Covid
1	Yes	Mild	No	Wet	Taste	Yes
3	Yes	Mild	No	Absent	Taste	No
6	Yes	Lite	No	Absent	Smell	Yes
13	Yes	Strong	No	Wet	Both	Yes
16	Yes	Lite	No	Dry	Smell	No

We calculated this before

$$G(S,A) = E(S) - \sum_{v \in A} \frac{|Sv|}{|S|} E\,(Sv)$$

$$G(S,Fever) = 0.99 - \sum_{v \in \{yes,no\}} \frac{|Sv|}{|S|} E\,(Sv)$$

$$\frac{|Syes|}{|S|} E\,(S_{yes}) = \frac{5}{20} * \left(-\frac{3}{5} log_2\left(\frac{3}{5}\right) - \frac{2}{5} log_2\left(\frac{2}{5}\right)\right)$$

$E(S|Fever = yes)$

Attribute/Feature Fever takes two values: yes/no.
For the value yes, we have:

$$\frac{|Syes|}{|S|} E\,(S_{yes}) = \frac{5}{20} * \left(-\frac{3}{5} log_2\left(\frac{3}{5}\right) - \frac{2}{5} log_2\left(\frac{2}{5}\right)\right) = 0.25 * 0.97 =$$

0.24

INFORMATION GAIN - FEVER

Ex.	Fever	Headache	Fatigue	Cough	TasteSmell Loss	Covid
2	No	Strong	No	Wet	Both	Yes
4	No	Mild	No	Wet	Smell	Yes
5	No	Mild	Yes	Wet	Taste	Yes
7	No	Lite	No	Dry	Smell	No
8	No	Strong	No	Absent	Taste	Yes
9	No	Absent	Yes	Wet	Both	No
10	No	Mild	No	Wet	Smell	Yes
11	No	Absent	Yes	Absent	Taste	No
12	No	Lite	No	Dry	Both	No
14	No	Absent	Yes	Wet	Both	No
15	No	Mild	No	Absent	Taste	Yes
17	No	Absent	No	Absent	Taste	Yes
18	No	Absent	No	Wet	Smell	Yes
19	No	Lite	No	Dry	Both	No
20	No	Lite	No	Dry	Smell	No

We calculated this before

$$G(S,A) = E(S) - \sum_{v \in A} \frac{|Sv|}{|S|} E\,(Sv)$$

$$G(S,Fever) = 0.99 - \sum_{v \in Fever} \frac{|Sv|}{|S|} E\,(Sv)$$

Attribute/Feature Fever takes two values: yes/no.
For the value no, we have:

$$\frac{|Sno|}{|S|} E\,(S_{no}) = \frac{15}{20} * \left(-\frac{8}{15} log_2\left(\frac{8}{15}\right) - \frac{7}{15} log_2\left(\frac{7}{15}\right)\right)$$

$E(S|Fever = no)$

$$\frac{|Sno|}{|S|} E\,(S_{no}) = 0.99 * 0.75 = 0.74$$

INFORMATION GAIN - FEVER

Ex.	Fever	Headache	Fatigue	Cough	TasteSmell Loss	Covid
2	No	Strong	No	Wet	Both	Yes
4	No	Mild	No	Wet	Smell	Yes
5	No	Mild	Yes	Wet	Taste	Yes
7	No	Lite	No	Dry	Smell	No
8	No	Strong	No	Absent	Taste	Yes
9	No	Absent	Yes	Wet	Both	No
10	No	Mild	No	Wet	Smell	Yes
11	No	Absent	Yes	Absent	Taste	No
12	No	Lite	No	Dry	Both	No
14	No	Absent	Yes	Wet	Both	No
15	No	Mild	No	Absent	Taste	Yes
17	No	Absent	No	Absent	Taste	Yes
18	No	Absent	No	Wet	Smell	Yes
19	No	Lite	No	Dry	Both	No
20	No	Lite	No	Dry	Smell	No

We calculated this before

$$G(S,A) = E(S) - \sum_{v \in A} \frac{|Sv|}{|S|} E\,(Sv)$$

$$G(S,Fever) = 0.99 - \sum_{v \in Fever} \frac{|Sv|}{|S|} E\,(Sv)$$

$$G(S,Fever) = 0.99 - (0.24 + 0.74) = 0.01$$

The information gain from the feature fever is 0.01

INFORMATION GAIN - COUGH

Ex.	Fever	Headache	Fatigue	Cough	TasteSmell Loss	Covid
1	Yes	Mild	No	Wet	Taste	Yes
2	No	Strong	No	Wet	Both	Yes
3	Yes	Mild	No	Absent	Taste	No
4	No	Mild	No	Wet	Smell	Yes
5	No	Mild	Yes	Wet	Taste	Yes
6	Yes	Lite	No	Absent	Smell	Yes
7	No	Lite	No	Dry	Smell	No
8	No	Strong	No	Absent	Taste	Yes
9	No	Absent	Yes	Wet	Both	No
10	No	Mild	No	Wet	Smell	Yes
11	No	Absent	Yes	Absent	Taste	No
12	No	Lite	No	Dry	Both	No
13	Yes	Strong	No	Wet	Both	Yes
14	No	Absent	Yes	Wet	Both	No
15	No	Mild	No	Absent	Taste	Yes
16	Yes	Lite	No	Dry	Smell	No
17	No	Absent	No	Absent	Taste	Yes
18	No	Absent	No	Wet	Smell	Yes
19	No	Lite	No	Dry	Both	No
20	No	Lite	No	Dry	Smell	No

$$G(S,A) = E(S) - \sum_{v \in A} \frac{|Sv|}{|S|} E\,(Sv)$$

$$G(S,Cough) = 0.99 - \sum_{v \in \{wet,absent,dry\}} \frac{|Sv|}{|S|} E\,(Sv)$$

$$G(S,Cough) = 0.99 - \left(\frac{S_{wet}}{20} E(S|wet) + \frac{S_{absent}}{20} E(S|absent) + \frac{S_{dry}}{20} E(S|dry)\right)$$

$$G(S,Cough) = 0.99 - \left(\frac{9}{20} * 0.76 + \frac{6}{20} * 0.91 + \frac{5}{20} * 0\right) =$$

$$0.99 - 0.342 - 0.273 - 0 = 0.375$$

Feature cough provides an information gain of 0.375

ALGORITHMIC METHOD

- We worked with information gain.
- Information gain is based on the entropy of a dataset.
- We should selected the attribute that maximizes the information gain.