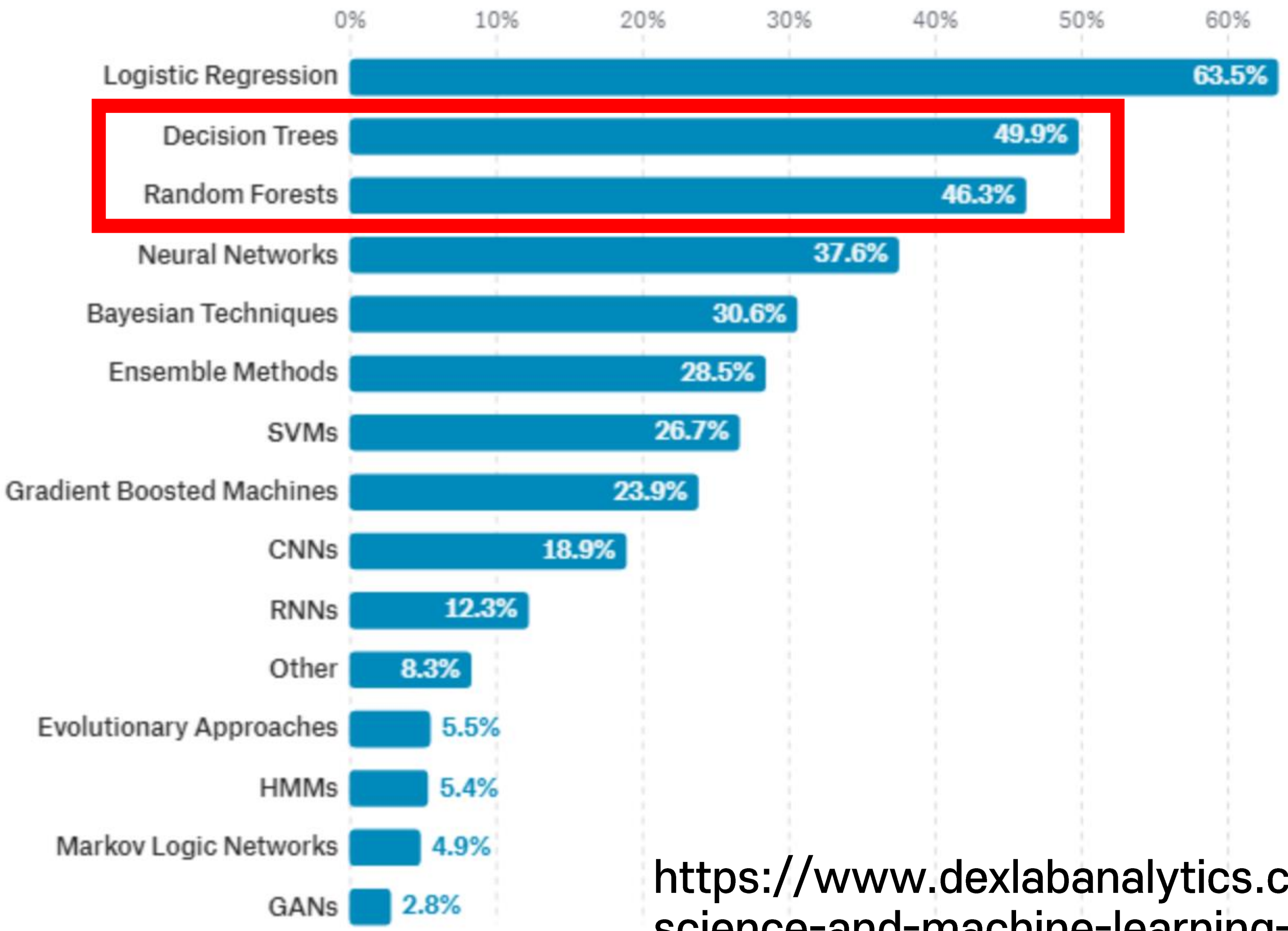
Machine learning

Decision Tree
Random Forest

엔트로피?

Python code

데이터과학자들이 많이 사용하는 머신러닝 기법



https://www.dexlabanalytics.com/blog/data-science-and-machine-learning-in-what-state-they-

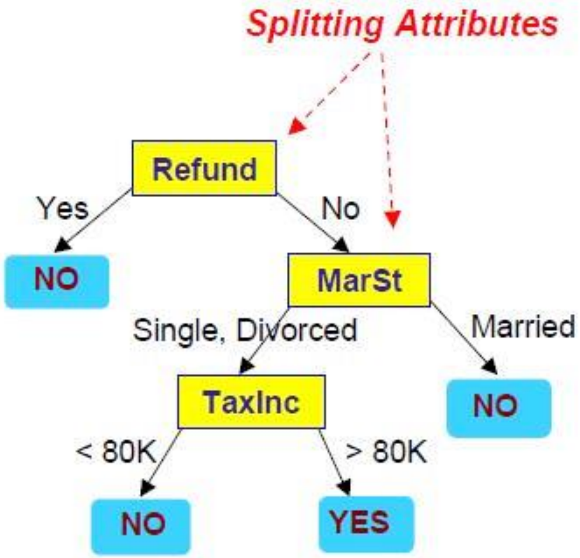
Decision Tree Classifier

- Feature에서부터 Label을 가장 잘 구분하는 선택지 힌트 구성

= Feature라는 뿌리에서 Label 나뭇잎까지 Tree 구성.

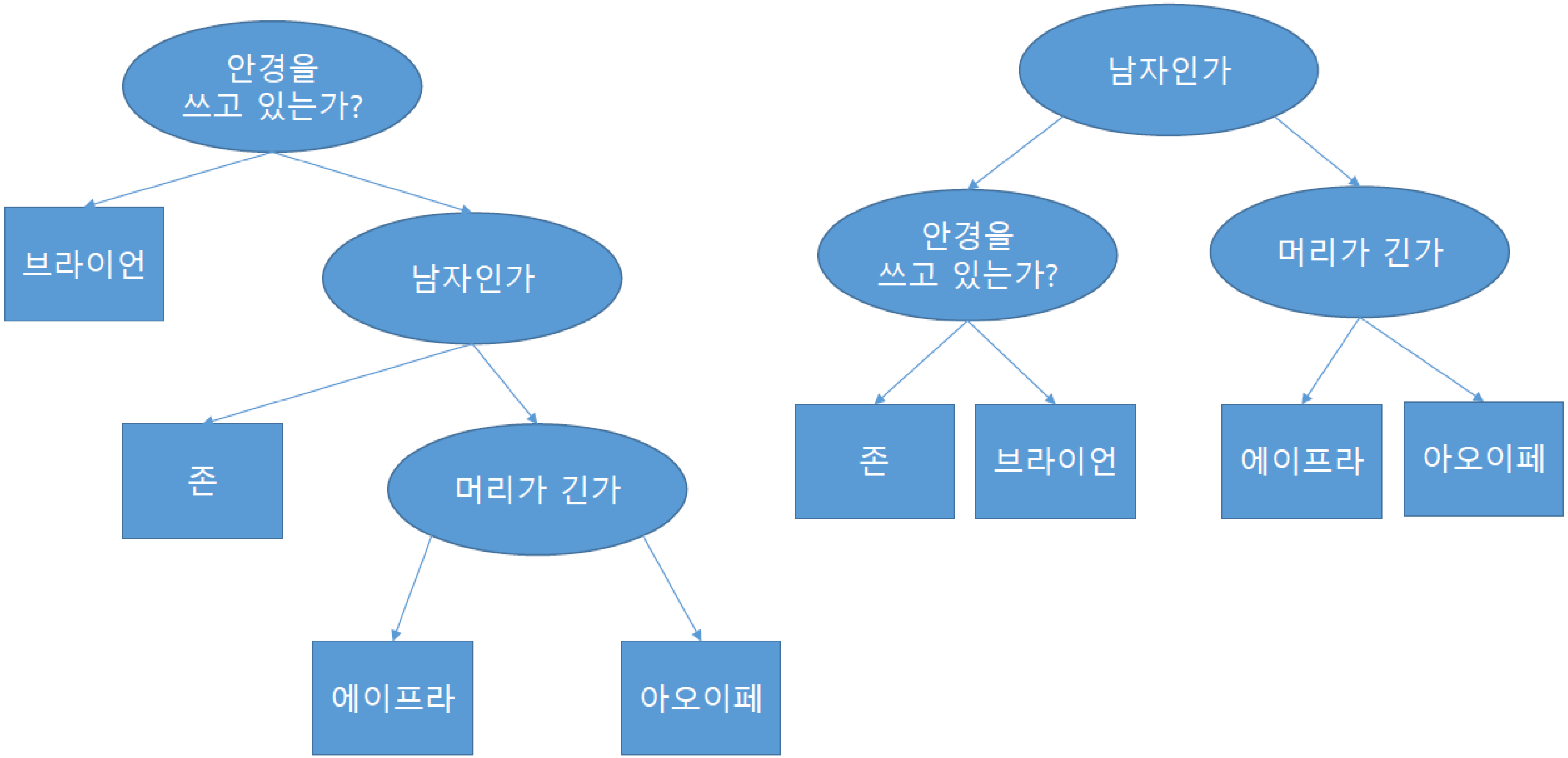
	categorical	categorical	continuous	class
Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Data



Model: Decision Tree

Guess Who



안경을
쓰고 있는가?
브라이언
남자인가
존
머리가 긴가
에이프라
아오이페
남자인가
안경을
쓰고 있는가?
머리가 긴가
존
브라이언
에이프라
아오이페

Decision Tree 만들기

- 어떤 질문이 가장 많은 해답을 줄 것인가?
  - 어떤 질문이 답의 모호성을 줄여줄 것인가?

- 데이터를 이용하여 splitting point 주요 힌트를 설정

Entropy

= 엉망

Entropy

- Entropy
  = 엉망 (무질서, 어원: 안쪽 변화) 정도를 표현

- ‘Entropy가 커진다’는 의미는
  = 에너지가 분산 = 일이 안됨

  예) 폰은 사용하다보면 느려지기만 하는 경험.
  배터리 수명 (에너지 저장 능력) 줄어듦.

Entropy

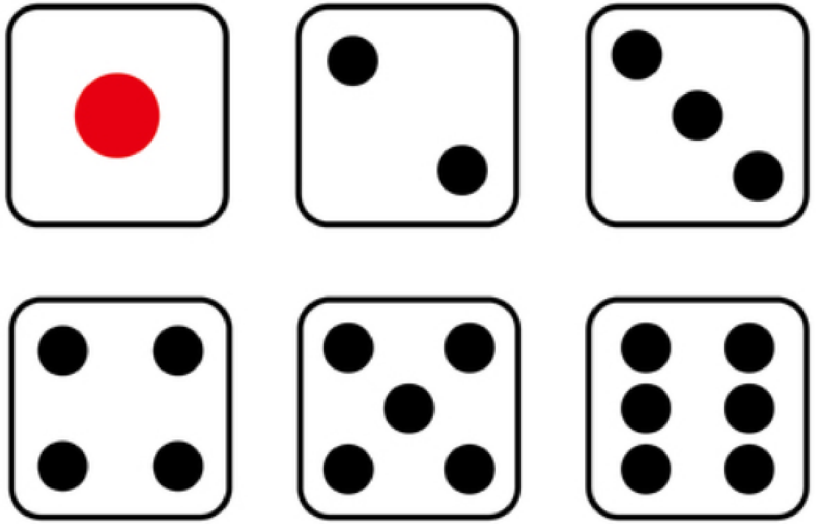
- Entropy
  = 엉망 (무질서, 어원: 안쪽 변화) 정도 표현

- ‘Entropy가 커진다’는 의미는
  = 더 불확실 해진다.
  = 더 무질서 정보의 양 → ‘정보의 양?’
  (경우의 수) 많아진다.

(컨텐츠가 여기서? 단, s가 없음)

Information content
정보 량

$$I(X) = log_2\left(\frac{1}{P(X)}\right)$$



[1] 동전 던져 앞면이 나오는 사건
[2] 주사위 눈이 1이 나오는 사건

두 사건의 정보량을 비교해봅시다.

Information content
정보 량

$$I(X) = log_2\left(\frac{1}{P(X)}\right)$$

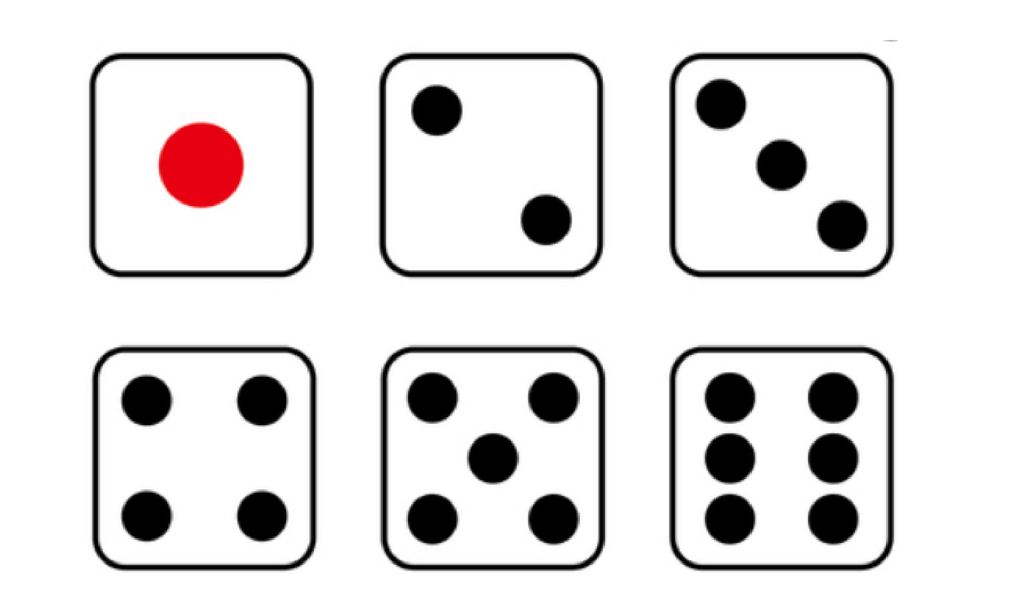
[1] 동전(2면체) 던져 앞면이 나오는 사건

$$I(X) = log_2\left(\frac{1}{1/2}\right) = 1$$

[2] 주사위(6면체) 눈이 1이 나오는 사건

$$I(X) = log_2\left(\frac{1}{1/6}\right) = 2.5849$$

우리가 봐도 경우의 수 2개와 6개는 다를 정보량이 다름

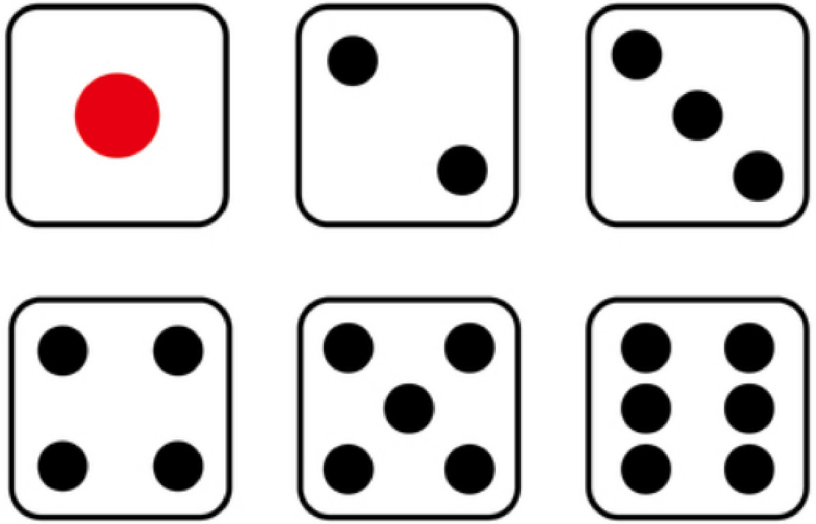


Information content
정보 량

$$I(X) = log_2\left(\frac{1}{\boldsymbol{P(X)}}\right)$$

확률 적은 사건이 일어나면 정보가 많다

= 기사거리가 많다 = 새로 학습할 양이 많다
= 드문일이라 놀라움이라는 감정 변화가 많다
= 드문일이라 불확실성이 높다
여러 사건들의 정보량 평균값에 이름을 붙이자!



Information Entropy, H(x)

$H(X) = E[I(X)]$ = 정보량의 기대값

$$= E\left[-log\left(P(X)\right)\right]$$

$$= -\sum P(x_i)\log(P(x_i))$$

Q. "얼마나 정보가 많길래?" 라는 질문에

*A. "정보량 * 나타날 확률을 곱해서 다 덧셈"*

Information Entropy, H(x)

$$H(X) = -\sum P(x_i)\log(P(x_i))$$

*"정보량 * 나타날 확률을 곱해서 다 덧셈"*

정보량: Log의 마법

(어느 한쪽 확률이 1에 가까우면 0으로 급격히 감소)

= 너무 당연하면 엔트로피가 작은 상태

= 확률 비등비등해야 엔트로피가 큰 상태

Information Entropy, H(x)

$$H(X) = -\sum P(x_i)\log(P(x_i))$$

브라질 vs 아르헨티나 축구, 승리 확률 (0.5, 0.5)라면,

$H1 =$ 0.5 $*$ -np.log(0.5) + 0.5 $*$ -np.log(0.5)

≒ 0.69

국가대표 vs 조기축구팀, 승리 확률 (0.99, 0.01),

$H2 =$ 0.99 $*$ -np.log(0.99) + 0.01 $*$ -np.log(0.01)

≒ 0.06

목표! : 엉망(엔트로피) 감소하는 것

$$Ent\,(D) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

전체 데이터 D의 엔트로피

$$Ent_{\,A}(D) = -\sum_{j=1}^{v} \frac{|D_j|}{D} * \; Ent\,(D_j)$$

속성 A로 분류시 엔트로피

$$Gain(A) = Ent\,(D) - Ent_{\,A}(D)$$

A 속성의 정보 소득

마케팅 미션: 어떤 사람이 컴퓨터를 살까?

C : class

class_buys_computer
no
no
yes
yes
yes
no
yes
no
yes
yes
yes
yes
yes
no

Case, 개인

$$E = -\sum_{i}^{C} p_i \log_2 p_i$$

$$h(D) = -\frac{9}{14}\log_2\frac{9}{14} + -\frac{5}{14}\log_2\frac{5}{14}$$

Yes No

$$h(9,5) = 0.940$$

```
x = np.array([9/14,5/14])
y = np.log2(x)

- sum(x * y)
```

0.9402859586706311

마케팅 미션: 어떤 사람이 컴퓨터를 살까?

Case, 개인

	age	income	student	credit_rating	class_buys_computer
0	youth	high	no	fair	no
1	youth	high	no	excellent	no
2	middle_aged	high	no	fair	yes
3	senior	medium	no	fair	yes
4	senior	low	yes	fair	yes
5	senior	low	yes	excellent	no
6	middle_aged	low	yes	excellent	yes
7	youth	medium	no	fair	no
8	youth	low	yes	fair	yes
9	senior	medium	yes	fair	yes
10	youth	medium	yes	excellent	yes
11	middle_aged	medium	no	excellent	yes
12	middle_aged	high	yes	fair	yes
13	senior	medium	no	excellent	no

v: 해당 속성 기준으로 나눠진 그룹 수

$$Ent_{A}(D) = -\sum_{j=1}^{v} \frac{|D_j|}{D} * Ent(D_j)$$

속성 A로 분류시 엔트로피

age 연령대로 구분해보면 될까요?

Yes　No

$$Ent_{age}(D) = \frac{5}{14} \times \left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right)$$

Youth

$$+ \frac{4}{14} \times \left(-\frac{4}{4}\log_2\frac{4}{4}\right)$$

Middle

Senior

$$+ \frac{5}{14} \times \left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right)$$

class

마케팅 미션: 어떤 사람이 컴퓨터를 살까?

```
entropy_allage = sum(group_age * entropy_group_age)
print('entropy_allage: ', entropy_allage)
```

```
entropy_allage:  0.6935361388961918
```

```
information_gain_of_age = entropy_parent - entropy_allage
print('information_gain_of_age: ', information_gain_of_age)
```

```
information_gain_of_age:  0.2467498197744391
```

목표! : 엉망(엔트로피) 감소하는 것

$$Ent\,(D) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

전체 데이터 D의 엔트로피

$$Ent_{\,A}(D) = -\sum_{j=1}^{v} \frac{|D_j|}{D} * \; Ent\,(D_j)$$

속성 A로 분류시 엔트로피

$$Gain(A) = Ent\,(D) - Ent_{\,A}(D)$$

A 속성의 정보 소득

Information 이득 Gain

- 한 속성을 기준으로 구분 후 '감소되는 entropy'
  (불확실성 감소 = 확실 정보 획득!)

$$Gain(A) = Ent(D) - Ent_A(D)$$

A 속성의 정보 소득

마케팅 미션: 어떤 사람이 컴퓨터를 살까?

```
entropy_allage = sum(group_age * entropy_group_age)
print('entropy_allage: ', entropy_allage)
```

```
entropy_allage:  0.6935361388961918
```

```
information_gain_of_age = entropy_parent - entropy_allage
print('information_gain_of_age: ', information_gain_of_age)
```

```
information_gain_of_age:  0.2467498197744391
```

축구 경기 진행할까요? 안할까요?

www.theweatheroutlook.com ▾ 이 페이지 번역하기

TheWeatherOutlook - latest UK weather forecasts

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	***No***
Sunny	Hot	High	True	***No***
Overcast	Hot	High	False	***Yes***
Rainy	Mild	High	False	***Yes***
Rainy	Cool	Normal	False	***Yes***
Rainy	Cool	Normal	True	***No***
Overcast	Cool	Normal	True	***Yes***
Sunny	Mild	High	False	***No***
Sunny	Cool	Normal	False	***Yes***
Rainy	Mild	Normal	False	***Yes***
Sunny	Mild	Normal	True	***Yes***
Overcast	Mild	High	True	***Yes***
Overcast	Hot	Normal	False	***Yes***
Rainy	Mild	High	True	***No***

https://www3.nd.edu/~rjohns15/cse40647.sp14/www/home.php

축구 경기 진행할까요? 안할까요?

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

$$\mathrm{H}(Y) = -\sum_{i=1}^{K} p_k \log_2 p_k$$

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

$$= 0.94$$

https://www3.nd.edu/~rjohns15/cse40647.sp14/www/home.php

축구 경기 진행할까요? 안할까요?

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	*No*
Sunny	Hot	High	True	*No*
Overcast	Hot	High	False	*Yes*
Rainy	Mild	High	False	*Yes*
Rainy	Cool	Normal	False	*Yes*
Rainy	Cool	Normal	True	*No*
Overcast	Cool	Normal	True	*Yes*
Sunny	Mild	High	False	*No*
Sunny	Cool	Normal	False	*Yes*
Rainy	Mild	Normal	False	*Yes*
Sunny	Mild	Normal	True	*Yes*
Overcast	Mild	High	True	*Yes*
Overcast	Hot	Normal	False	*Yes*
Rainy	Mild	High	True	*No*

$$InfoGain(\text{Humidity}) =$$
$$H(Y) - \frac{m_L}{m}H_L - \frac{m_R}{m}H_R$$
$$= 0.94 - \frac{7}{14}H_L - \frac{7}{14}H_R$$

https://www3.nd.edu/~rjohns15/cse40647.sp14/www/home.php

축구 경기 진행할까요? 안할까요?

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	***No***
Sunny	Hot	High	True	***No***
Overcast	Hot	High	False	***Yes***
Rainy	Mild	High	False	***Yes***
Rainy	Cool	**Normal**	False	***Yes***
Rainy	Cool	**Normal**	True	***No***
Overcast	Cool	**Normal**	True	***Yes***
Sunny	Mild	High	False	***No***
Sunny	Cool	**Normal**	False	***Yes***
Rainy	Mild	**Normal**	False	***Yes***
Sunny	Mild	**Normal**	True	***Yes***
Overcast	Mild	High	True	***Yes***
Overcast	Hot	**Normal**	False	***Yes***
Rainy	Mild	High	True	***No***

$$InfoGain(\text{Humidity}) =$$

$$H(Y) - \frac{m_L}{m} H_L - \frac{m_R}{m} H_R$$

$$0.94 - \frac{7}{14} H_L - \frac{7}{14} H_R$$

$$H_L = -\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7}$$

https://www3.nd.edu/~rjohns15/cse40647.sp14/www/home.php

축구 경기 진행할까요? 안할까요?

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	**High**	False	*No*
Sunny	Hot	**High**	True	*No*
Overcast	Hot	**High**	False	*Yes*
Rainy	Mild	**High**	False	*Yes*
Rainy	Cool	Normal	False	*Yes*
Rainy	Cool	Normal	True	*No*
Overcast	Cool	Normal	True	*Yes*
Sunny	Mild	**High**	False	*No*
Sunny	Cool	Normal	False	*Yes*
Rainy	Mild	Normal	False	*Yes*
Sunny	Mild	Normal	True	*Yes*
Overcast	Mild	**High**	True	*Yes*
Overcast	Hot	Normal	False	*Yes*
Rainy	Mild	**High**	True	*No*

$$InfoGain(\text{Humidity}) =$$
$$H(Y) - \frac{m_L}{m}H_L - \frac{m_R}{m}H_R$$
$$0.94 - \frac{7}{14}H_L - \frac{7}{14}H_R$$

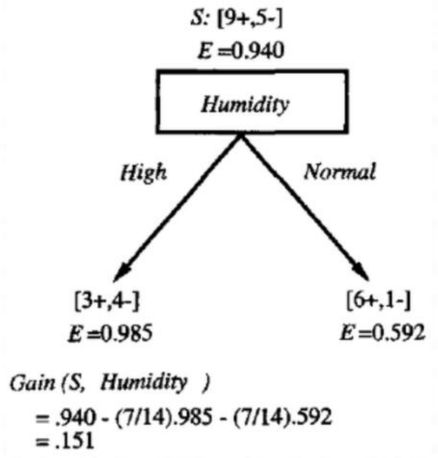
$$H_L = -\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7}$$
$$= 0.592$$
$$H_R = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7}$$
$$= 0.985$$

https://www3.nd.edu/~rjohns15/cse40647.sp14/www/home.php

축구 경기 진행할까요? 안할까요?

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	***No***
Sunny	Hot	High	True	***No***
Overcast	Hot	High	False	***Yes***
Rainy	Mild	High	False	***Yes***
Rainy	Cool	Normal	False	***Yes***
Rainy	Cool	Normal	True	***No***
Overcast	Cool	Normal	True	***Yes***
Sunny	Mild	High	False	***No***
Sunny	Cool	Normal	False	***Yes***
Rainy	Mild	Normal	False	***Yes***
Sunny	Mild	Normal	True	***Yes***
Overcast	Mild	High	True	***Yes***
Overcast	Hot	Normal	False	***Yes***
Rainy	Mild	High	True	***No***

$$InfoGain(\text{Humidity}) =$$
$$H(Y) - \frac{m_L}{m} H_L - \frac{m_R}{m} H_R$$
$$0.94 - \frac{7}{14} 0.592 - \frac{7}{14} 0.985$$
$$= 0.94 - 0.296 - 0.4925$$
$$= 0.1515$$



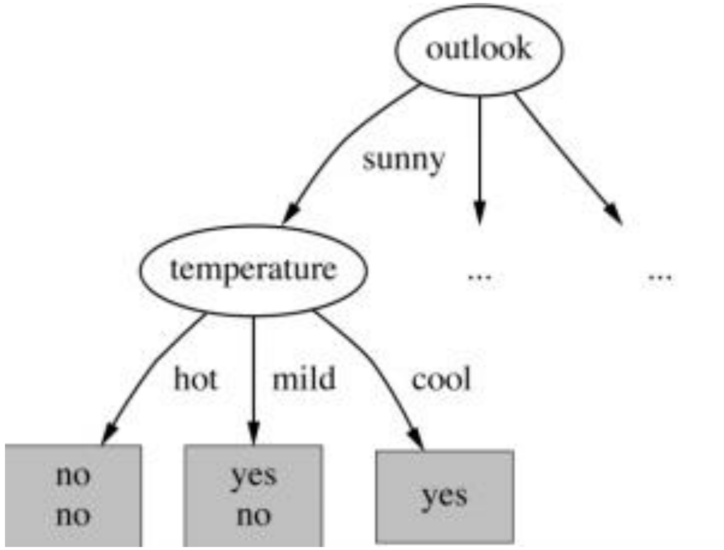
Information Gain: 축구 사례

- Information gain for each feature:
  - Outlook = 0.247
  - Temperature = 0.029
  - Humidity = 0.152
  - Windy = 0.048
- Initial split is on outlook, because it is the feature with the highest information gain.

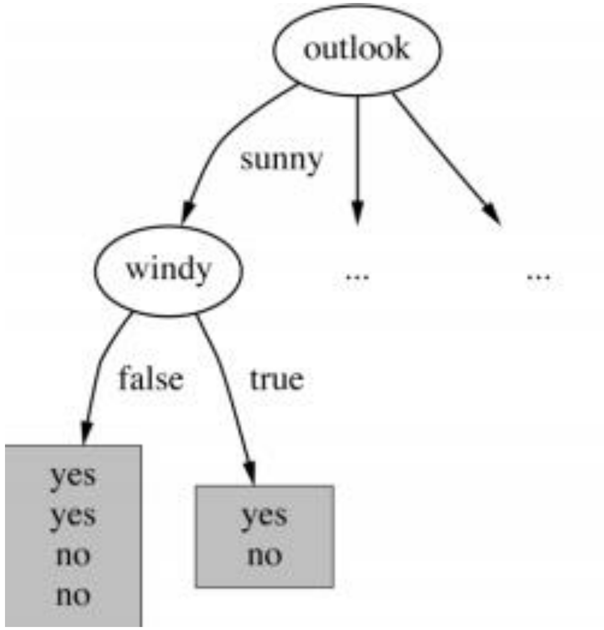
https://www3.nd.edu/~rjohns15/cse40647.sp14/www/home.php

Information Gain: 축구 사례

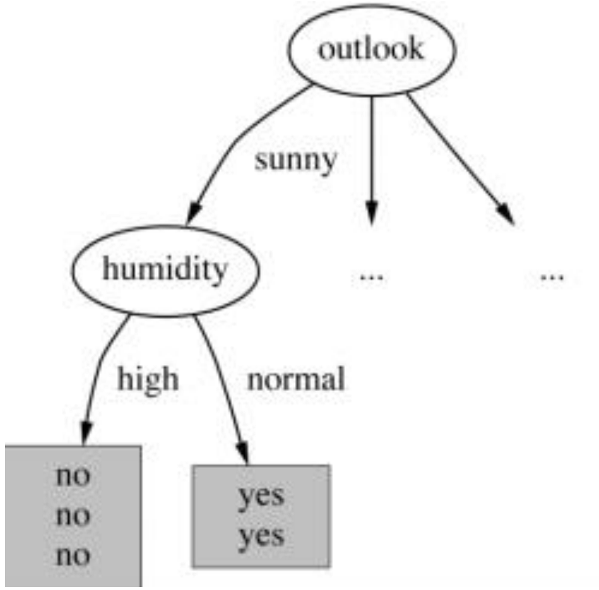
- Now we search for the best split at the next level:



Temperature = 0.571



Windy = 0.020

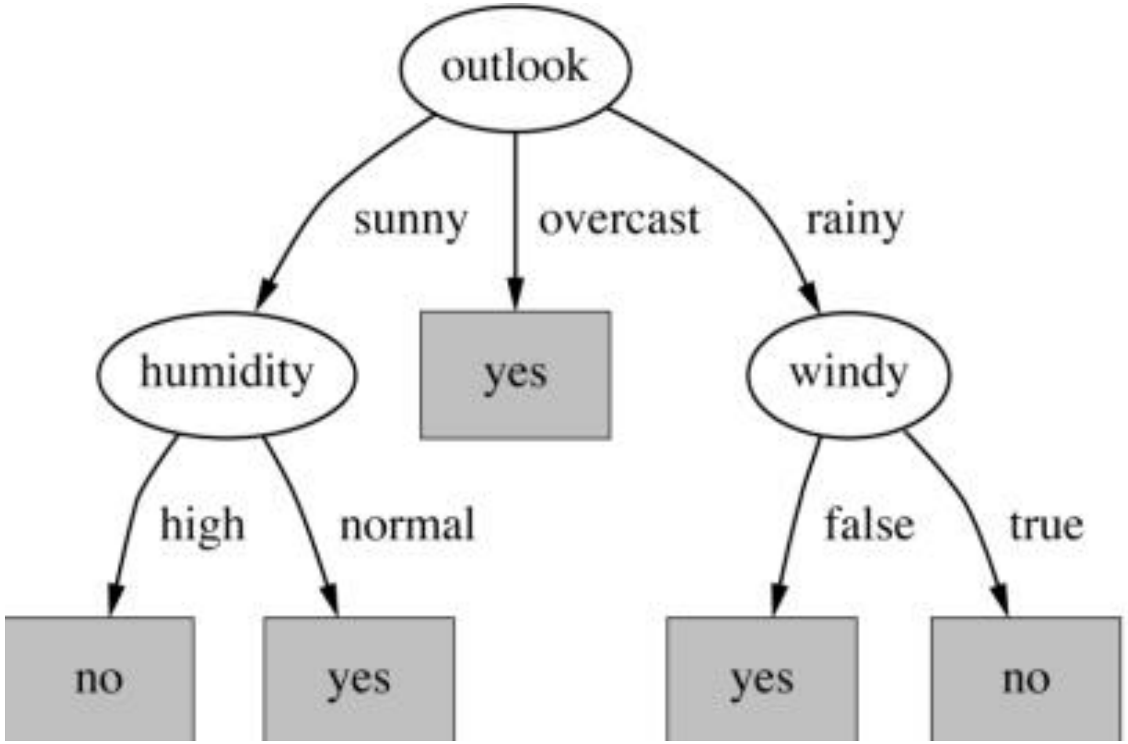


Humidity = 0.971

https://www3.nd.edu/~rjohns15/cse40647.sp14/www/home.php

Information Gain: 축구 사례

- The final decision tree:



Note that not all leaves need to be pure;
Sometimes similar (even identical) instances have different classes.
Splitting stops when data cannot be split any further.

https://www3.nd.edu/~rjohns15/cse40647.sp14/www/home.php

sklearn.tree.DecisionTreeClassifier

class `sklearn.tree.DecisionTreeClassifier`(*, *criterion='gini', splitter='best', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=None, random_state=None, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, class_weight=None, presort='deprecated', ccp_alpha=0.0*) ¶

Parameters:	**criterion : *{"gini", "entropy"}, default="gini"*** The function to measure the quality of a split. Supported criteria are "gini" for the Gini impurity and "entropy" for the information gain.

엉망 = 불순도

- Impure (Not pure)   vs   순수 pure

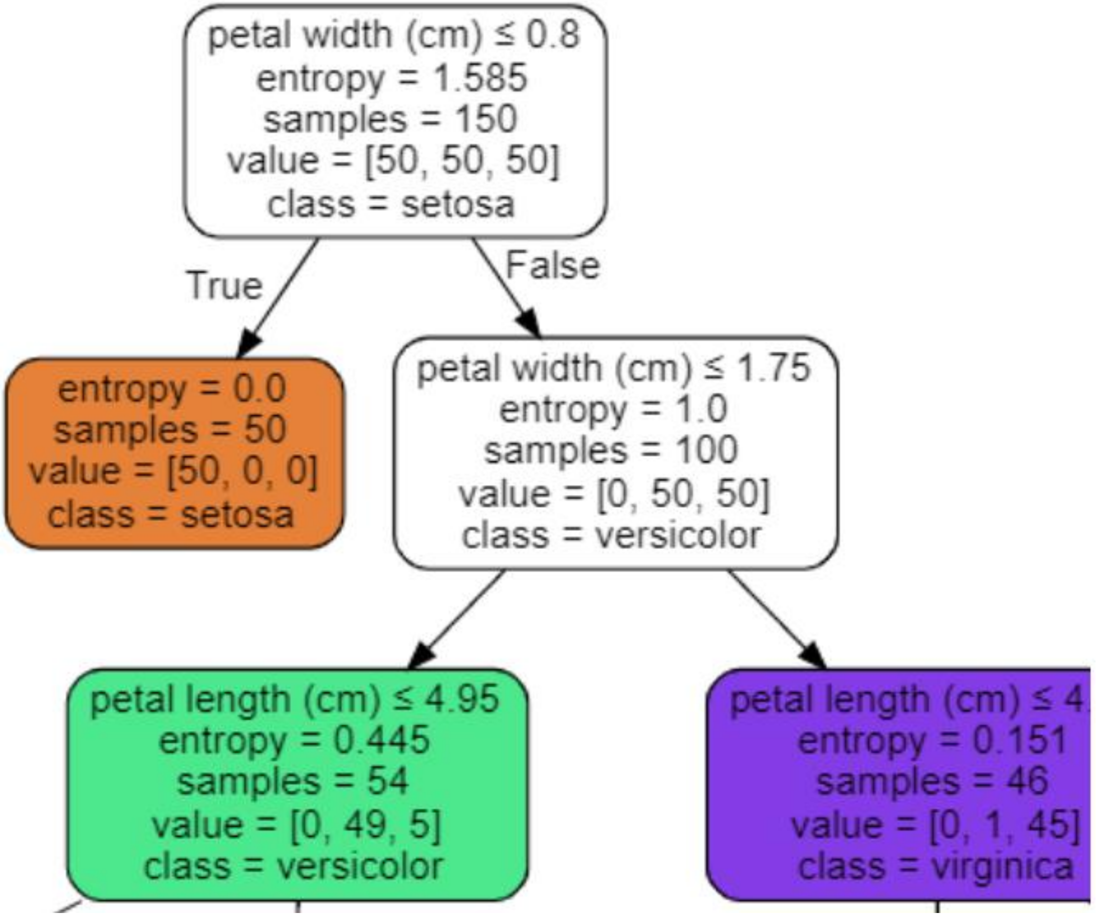
= Label   섞임   vs   모두 같음

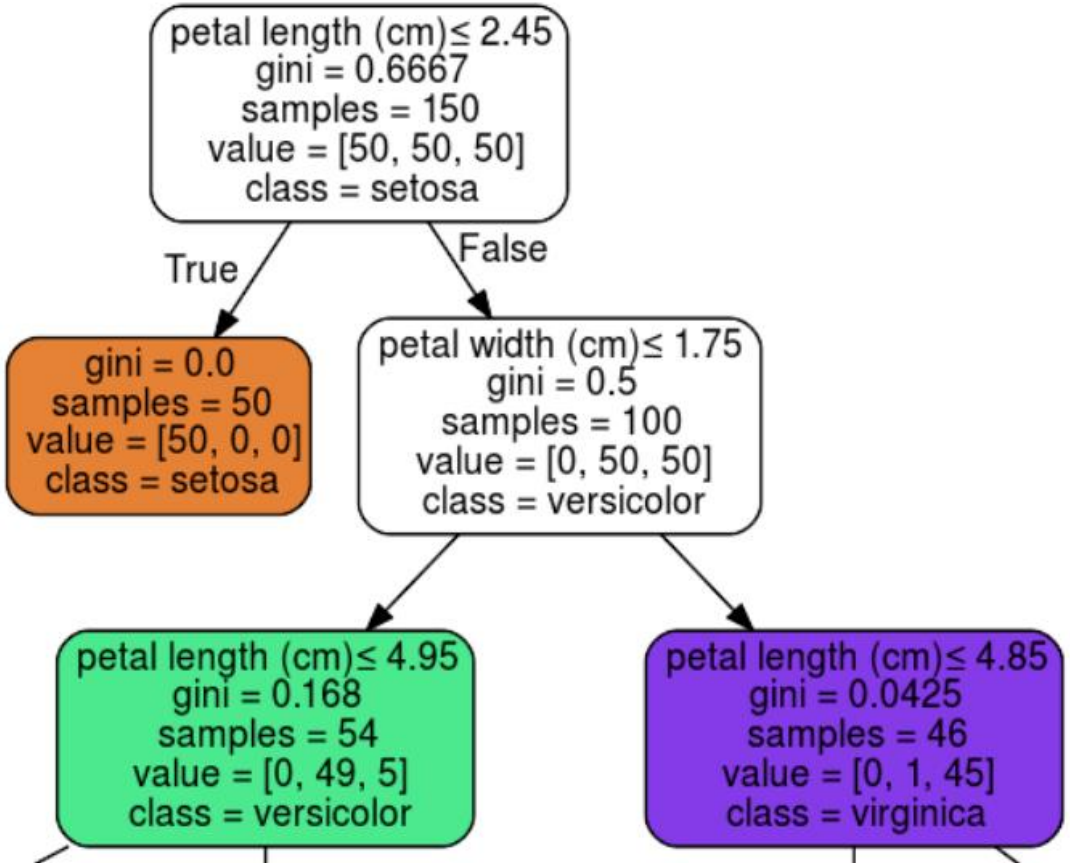
= Impurity 지표로 판단

'entropy' or 'gini'

Decision Tree, 꽃잎 examples

속성: 너비, 불순도 기준 : entropy(=IG) vs 속성: 길이, 불순도 기준 : gini

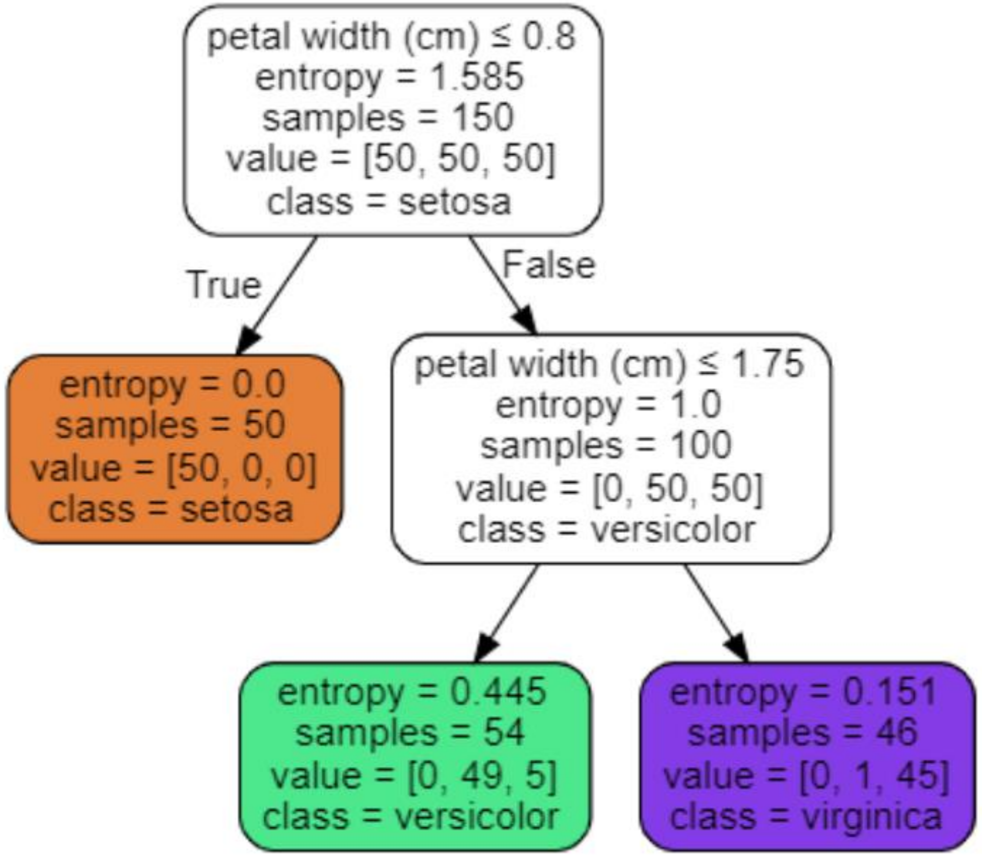




가지치기(프루닝 pruning)

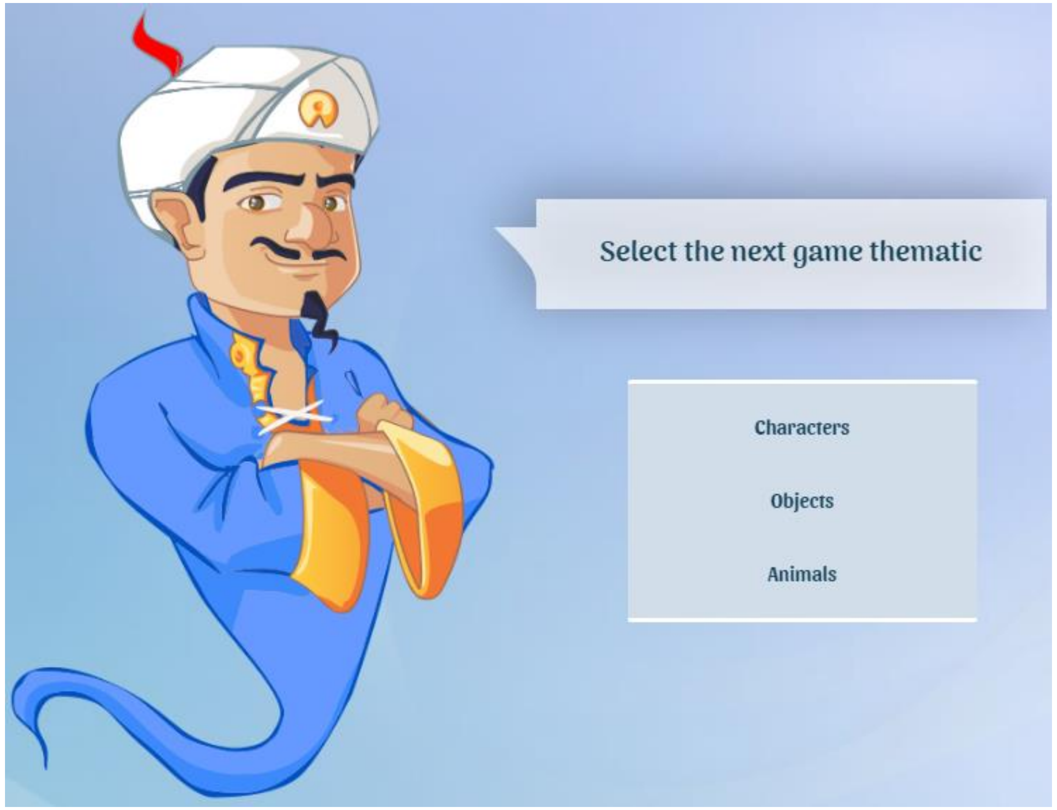
```
clf3 = tree.DecisionTreeClassifier(criterion='entropy', max_depth=2)
clf3.fit(iris.data, iris.target)
```

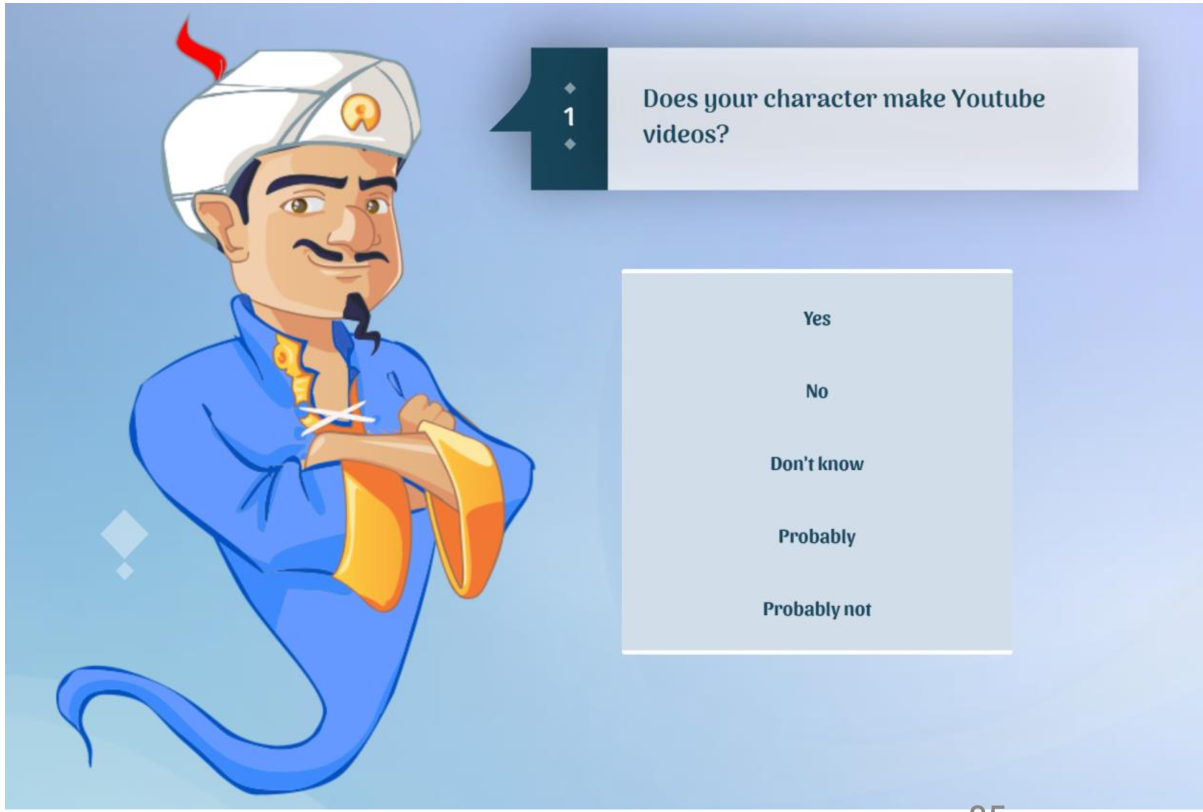
속성: 너비, 불순도 기준: entropy(=IG)



gini

- 선택을 모아서 지니가 대상을 추측하는 게임. a.k.a. 스무고개

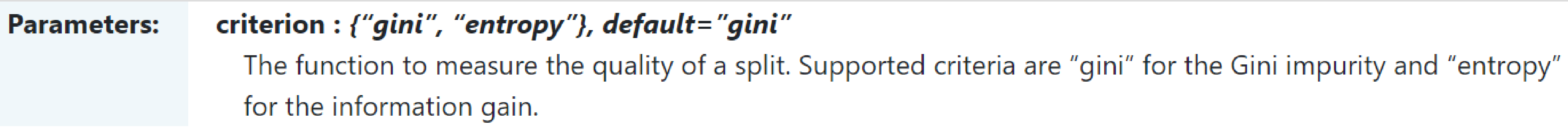




https://en.akinator.com/game

gini

- **Measurement of inequality 같지않음 지표**
- **by Corrado Gini (Italian statistician)**



gini

- x_1 속성 → y 두 label로 나누고 싶을 때

x_1	1	2	3	4	5	6	7	8
y	0	0	0	1	1	1	1	1

모인 샘플 들끼리 비슷함 = 순수함

If we split at $x_1 < 3.5$, we get an optimal split.
If we split at $x_1 < 4.5$, we make a mistake (misclassification).

Idea: *A better split should make the samples "pure" (homogeneous).*

Gini Index

The Gini index is defined as:

$$\text{Gini} = 1 - \sum_{i=1}^{K} {p_k}^2$$

where p_k denotes the proportion of instances belonging to class k $(K = 1, \ldots, k)$.

마케팅 미션: 어떤 사람이 컴퓨터를 살까?

Case, 개인

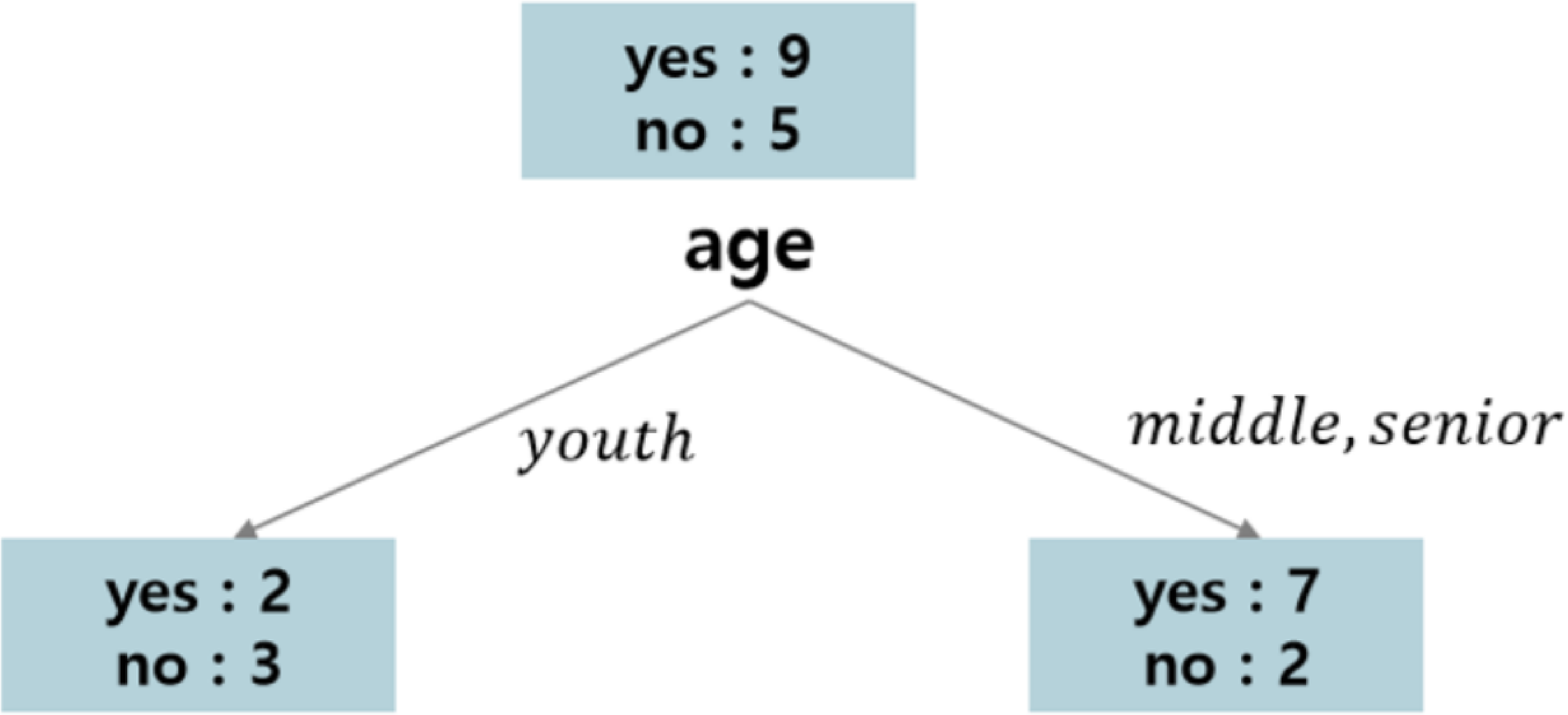
	age	income	student	credit_rating	class_buys_computer
0	youth	high	no	fair	no
1	youth	high	no	excellent	no
2	middle_aged	high	no	fair	yes
3	senior	medium	no	fair	yes
4	senior	low	yes	fair	yes
5	senior	low	yes	excellent	no
6	middle_aged	low	yes	excellent	yes
7	youth	medium	no	fair	no
8	youth	low	yes	fair	yes
9	senior	medium	yes	fair	yes
10	youth	medium	yes	excellent	yes
11	middle_aged	medium	no	excellent	yes
12	middle_aged	high	yes	fair	yes
13	senior	medium	no	excellent	no

age 연령대로 구분해보면 될까요?

	Yes	No
Youth	2	3
Middle + Senior	7	2

Sklearn에서 제공하는

특정 함수는 **Binary Splitting**만 허용



$$\text{Gini} = 1 - \sum_{i=1}^{K} {p_k}^2$$

age = youth

(D$_{\text{Group A}}$/D) * Gini$_{\text{Group A}}$ + (D$_{\text{Group ~A}}$/D) * Gini$_{\text{Group ~A}}$

$$G(age = youth) = \frac{5}{14}\left(1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2\right) + \frac{9}{14}\left(1 - \left(\frac{7}{9}\right)^2 - \left(\frac{2}{9}\right)^2\right) = 0.394$$

$$G(age = middle) = \frac{4}{14}\left(1 - \left(\frac{4}{4}\right)^2\right) + \frac{10}{14}\left(1 - \left(\frac{5}{10}\right)^2 - \left(\frac{5}{10}\right)^2\right) = \mathbf{0.357}$$

$$G(age = senior) = \frac{5}{14}\left(1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2\right) + \frac{9}{14}\left(1 - \left(\frac{6}{9}\right)^2 - \left(\frac{3}{9}\right)^2\right) = 0.457$$