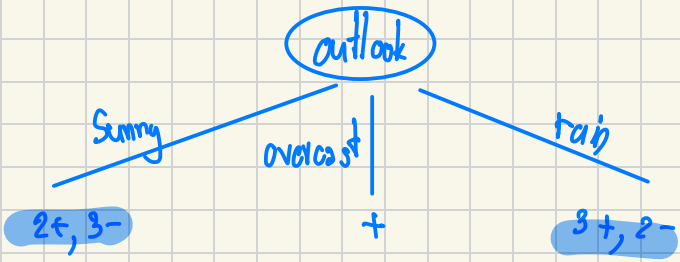
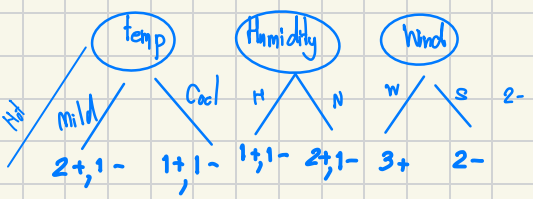
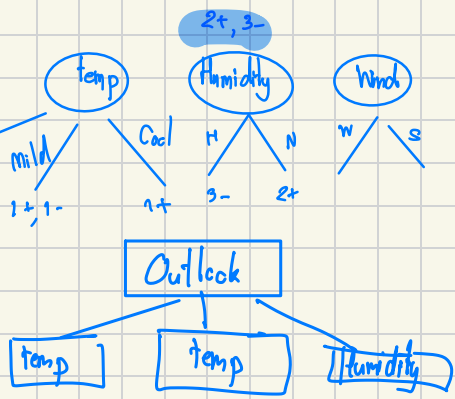
# Decision Tree

| Day | Outlook | Temperature | Humidity | Wind | PlayTennis |
|---|---|---|---|---|---|
| D1 | Sunny | Hot | High | Weak | No |
| D2 | Sunny | Hot | High | Strong | No |
| D3 | Overcast | Hot | High | Weak | Yes |
| D4 | Rain | Mild | High | Weak | Yes |
| D5 | Rain | Cool | Normal | Weak | Yes |
| D6 | Rain | Cool | Normal | Strong | No |
| D7 | Overcast | Cool | Normal | Strong | Yes |
| D8 | Sunny | Mild | High | Weak | No |
| D9 | Sunny | Cool | Normal | Weak | Yes |
| D10 | Rain | Mild | Normal | Weak | Yes |
| D11 | Sunny | Mild | Normal | Strong | Yes |
| D12 | Overcast | Mild | High | Strong | Yes |
| D13 | Overcast | Hot | Normal | Weak | Yes |
| D14 | Rain | Mild | High | Strong | No |



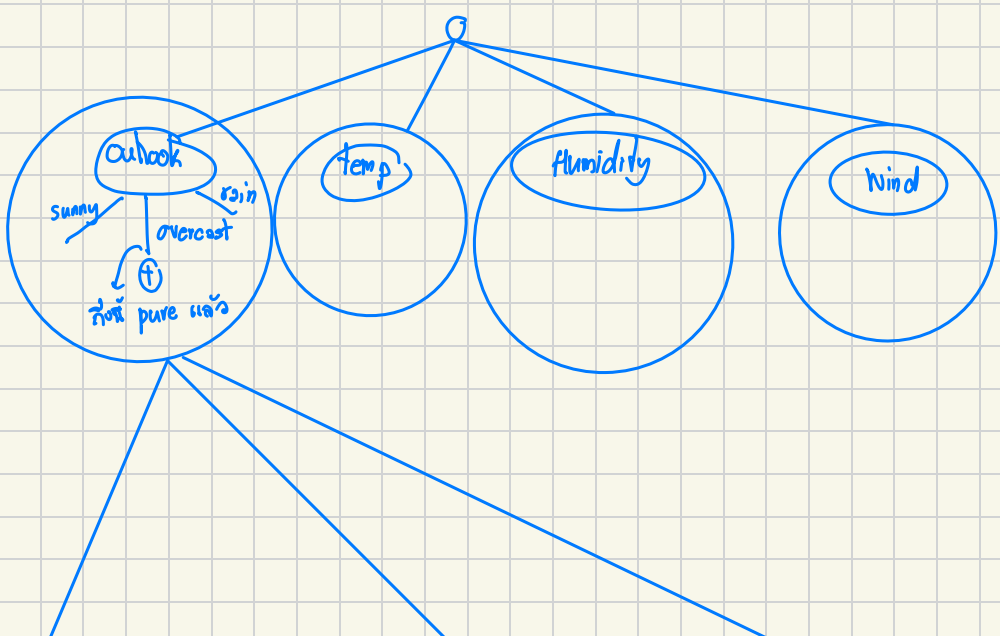

3+, 2-

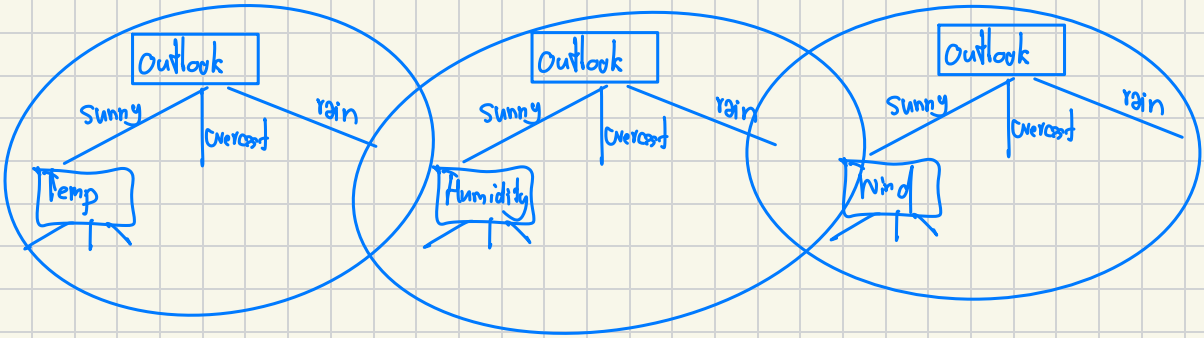


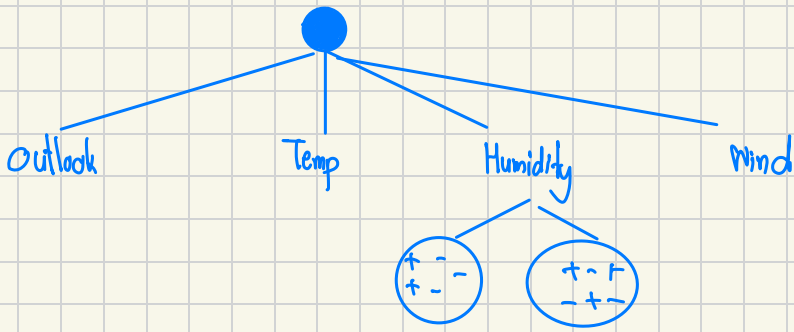

Prefer smaller trees

AI : BFS





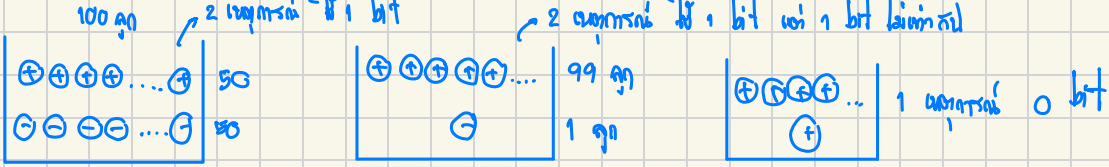

Greedy




| Day | Outlook | Temperature | Humidity | Wind | PlayTennis |
|---|---|---|---|---|---|
| D1 | Sunny | Hot | High | Weak | No − |
| D2 | Sunny | Hot | High | Strong | No − |
| D3 | Overcast | Hot | High | Weak | Yes + |
| D4 | Rain | Mild | High | Weak | Yes + |
| D5 | Rain | Cool | Normal | Weak | Yes + |
| D6 | Rain | Cool | Normal | Strong | No − |
| D7 | Overcast | Cool | Normal | Strong | Yes + |
| D8 | Sunny | Mild | High | Weak | No − |
| D9 | Sunny | Cool | Normal | Weak | Yes + |
| D10 | Rain | Mild | Normal | Weak | Yes + |
| D11 | Sunny | Mild | Normal | Strong | Yes + |
| D12 | Overcast | Mild | High | Strong | Yes + |
| D13 | Overcast | Hot | Normal | Weak | Yes + |
| D14 | Rain | Mild | High | Strong | No − |

9+, 5−

Entropy ยืมศัพท์มาจาก thermodynamic คือค่าความสับสนวุ่นวาย (Information theory)


$$\text{Entropy}(S) = \sum_{v} -P_v \log_2 P_v$$

Entropy ({+,+,+,+,−,−,−,−}) = $-p_{\oplus}\log_2 p_{\oplus} - p_{\ominus}\log_2 p_{\ominus}$

$= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}$

$= -\frac{1}{2}\log_2 2^{-1} - \frac{1}{2}\log_2 2^{-1}$

$= \frac{1}{2} + \frac{1}{2} = 1$ # บอกว่าใช้ 1 bit แทนข้อมูลนี้ได้

Entropy ({+,+,+,+,+,+,+}) = $-p_{\oplus}\log_2 p_{\oplus} - p_{\ominus}\log_2 p_{\ominus}$

$= -1\log_2 1 - 0\log_2 0$ (0, 0)

$= 0$ bit

Entropy ({+,+,+,+,+,−}) = $-0.99\log_2 0.99 - 0.01\log_2 0.01$ (≈0, ≈0)

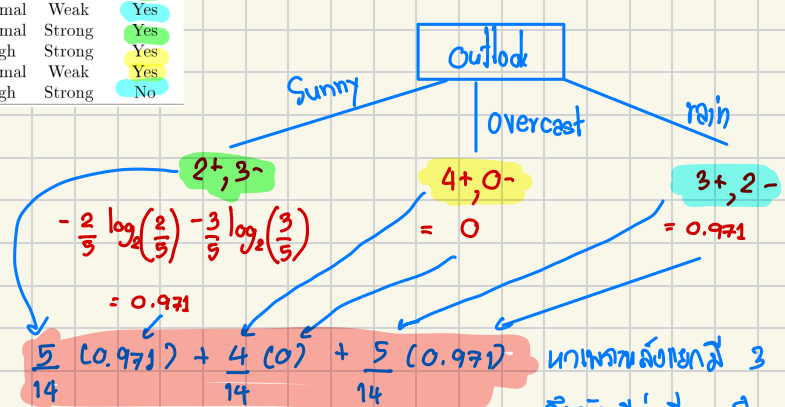
$\approx 0.052\ldots$

| Day | Outlook | Temperature | Humidity | Wind | PlayTennis |
|---|---|---|---|---|---|
| D1 | Sunny | Hot | High | Weak | No |
| D2 | Sunny | Hot | High | Strong | No |
| D3 | Overcast | Hot | High | Weak | Yes |
| D4 | Rain | Mild | High | Weak | Yes |
| D5 | Rain | Cool | Normal | Weak | Yes |
| D6 | Rain | Cool | Normal | Strong | No |
| D7 | Overcast | Cool | Normal | Strong | Yes |
| D8 | Sunny | Mild | High | Weak | No |
| D9 | Sunny | Cool | Normal | Weak | Yes |
| D10 | Rain | Mild | Normal | Weak | Yes |
| D11 | Sunny | Mild | Normal | Strong | Yes |
| D12 | Overcast | Mild | High | Strong | Yes |
| D13 | Overcast | Hot | Normal | Weak | Yes |
| D14 | Rain | Mild | High | Strong | No |

$S_0 = \{+,+,+,+,+,+,+,+,+,-,-,-,-,-\}$

Entropy (S) : $-\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right)$

$= 0.940$



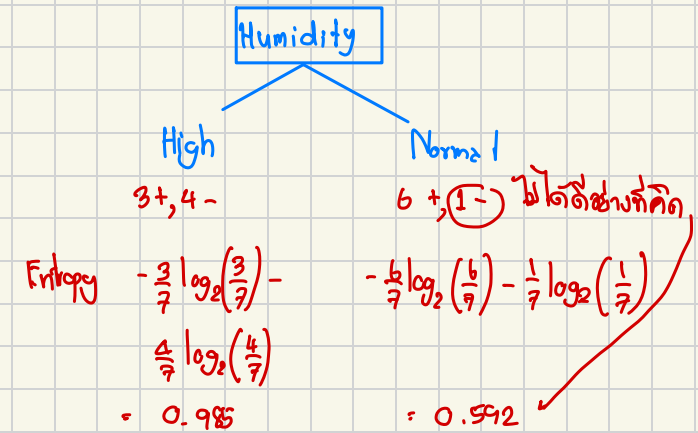

หาเพราะมีลูกหลาน 3 กิ่ง ซึ่งต้องมีค่าที่จะคำนวณต่อแทน

ก่อนแยก outlook − หลังแยก outlook

$Gain(S_o, Outlook) = 0.940 - \text{Weighted Average}$

$= 0.940 - 0.694 = 0.246$

| Day | Outlook | Temperature | Humidity | Wind | PlayTennis |
|---|---|---|---|---|---|
| D1 | Sunny | Hot | High | Weak | No |
| D2 | Sunny | Hot | High | Strong | No |
| D3 | Overcast | Hot | High | Weak | Yes |
| D4 | Rain | Mild | High | Weak | Yes |
| D5 | Rain | Cool | Normal | Weak | Yes |
| D6 | Rain | Cool | Normal | Strong | No |
| D7 | Overcast | Cool | Normal | Strong | Yes |
| D8 | Sunny | Mild | High | Weak | No |
| D9 | Sunny | Cool | Normal | Weak | Yes |
| D10 | Rain | Mild | Normal | Weak | Yes |
| D11 | Sunny | Mild | Normal | Strong | Yes |
| D12 | Overcast | Mild | High | Strong | Yes |
| D13 | Overcast | Hot | Normal | Weak | Yes |
| D14 | Rain | Mild | High | Strong | No |

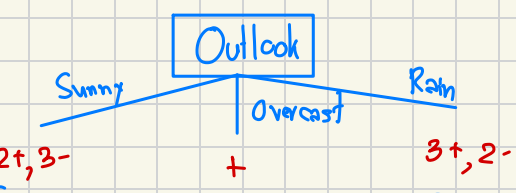



$\text{Weighted Average} = \frac{7}{14}(0.985) + \frac{7}{14}(0.592)$

$= 0.789$

$Gain(S_o, Humidity) = 0.940 - 0.789 = 0.151$

pure คือไม่มี +, − ปะปนกัน




$S_1 = \{2+, 3-\}$

$Gain(S_1, Temp)$

$Gain(S_1, Humidity)$

$Gain(S_1, Wind)$

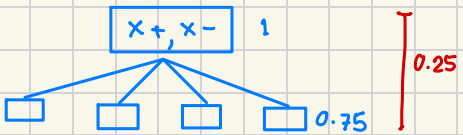
$S_2 = \{3+, 2-\}$

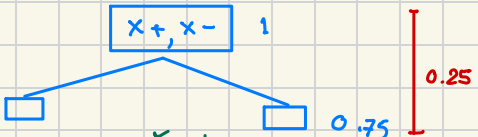
$Gain(S_2, Temp)$

$Gain(S_2, Humidity)$

$Gain(S_2, Wind)$

4/2/2024 Gain → Information Gain หมายถึง represent เหตุการณ์ที่เกิดขึ้น




ต้นนี้ดีกว่า ✓

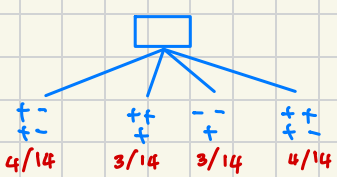
| Day | Outlook | Temperature | Humidity | Wind | PlayTennis |
|---|---|---|---|---|---|
| D1 | Sunny | Hot | High | Weak | No |
| D2 | Sunny | Hot | High | Strong | No |
| D3 | Overcast | Hot | High | Weak | Yes |
| D4 | Rain | Mild | High | Weak | Yes |
| D5 | Rain | Cool | Normal | Weak | Yes |
| D6 | Rain | Cool | Normal | Strong | No |
| D7 | Overcast | Cool | Normal | Strong | Yes |
| D8 | Sunny | Mild | High | Weak | No |
| D9 | Sunny | Cool | Normal | Weak | Yes |
| D10 | Rain | Mild | Normal | Weak | Yes |
| D11 | Sunny | Mild | Normal | Strong | Yes |
| D12 | Overcast | Mild | High | Strong | Yes |
| D13 | Overcast | Hot | Normal | Weak | Yes |
| D14 | Rain | Mild | High | Strong | No |


## GainRatio

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)}$$

$$SplitInfo(D) = -\sum_{j=1}^{n} \frac{|D_j|}{|D|} \log_2 \frac{|D_j|}{|D|}$$

(Entropy ในการแยก)



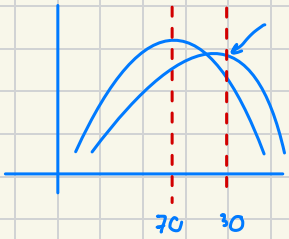

$$= -\frac{4}{14}\log_2\left(\frac{4}{14}\right) - \frac{3}{14}\log_2\left(\frac{3}{14}\right) - \frac{3}{14}\log_2\left(\frac{3}{14}\right) - \frac{4}{14}\log_2\left(\frac{4}{14}\right)$$

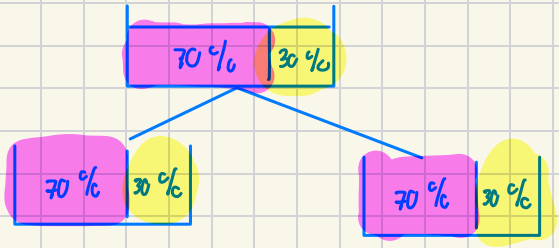
2 กลุ่ม กลุ่มละเท่าๆกัน : $-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1$

4 กลุ่ม กลุ่มละเท่าๆกัน $= -\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)$

$= -1\log_2\left(\frac{1}{4}\right)$

$= -1\log_2 2^{-2} = 2$

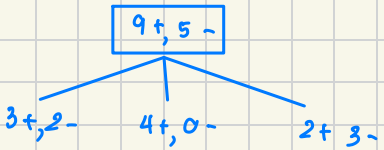


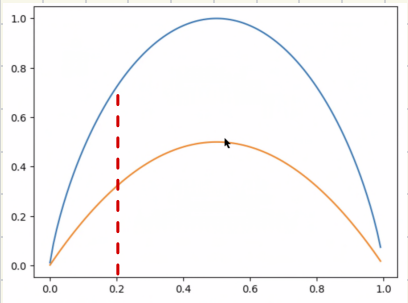

ในโลก ML มอง dataset เป็น distributed data
และข้อมูลจริงไม่ได้เข้ามาแบบ 50 : 50

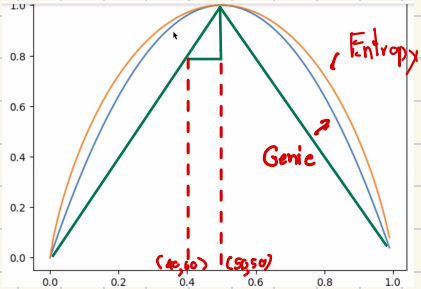
## Gini Impurity

$$Gini(D) = 1 - \sum_{v} p_v^2$$

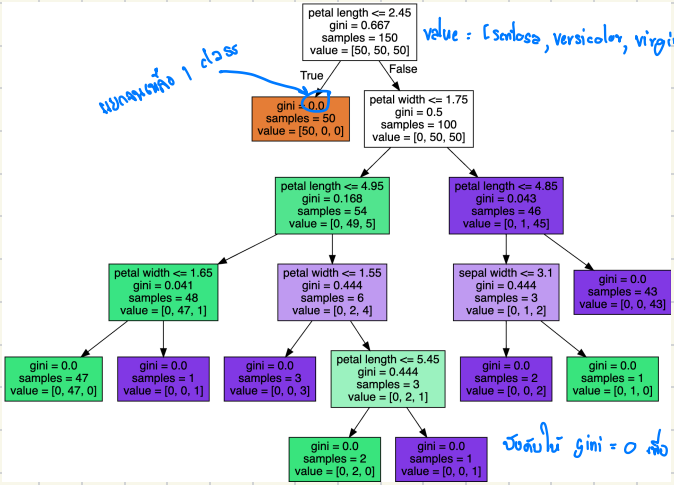


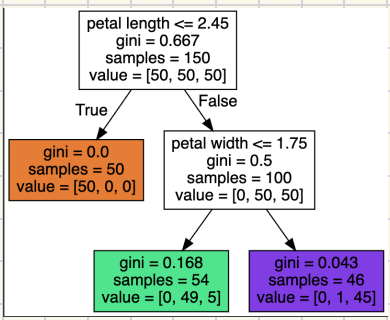

$$= \frac{5}{14}(Gini(3+,2-)) + \frac{4}{14}(Gini(4+,0-)) + \frac{5}{14}(Gini(2+,3-))$$



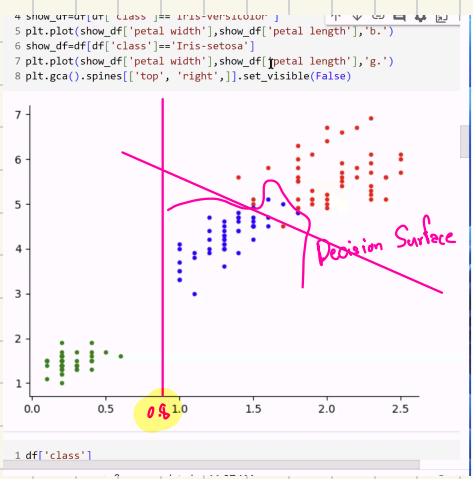

ข้อดีของ Genie คือ ทำให้ได้สัดส่วนที่ sensitive กว่า entropy
ทำให้การเปลี่ยนแปลงสะท้อนถึงสัดส่วนดีกว่า

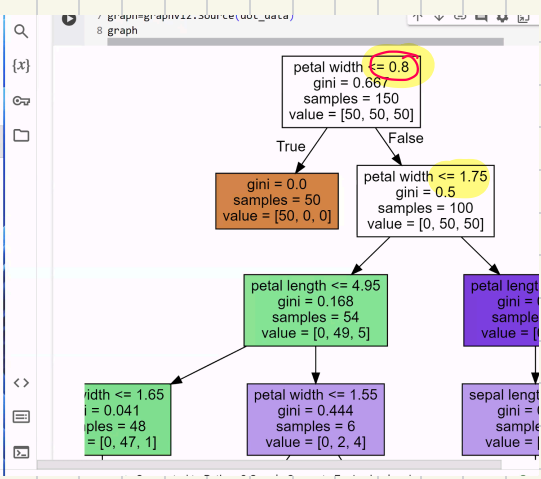


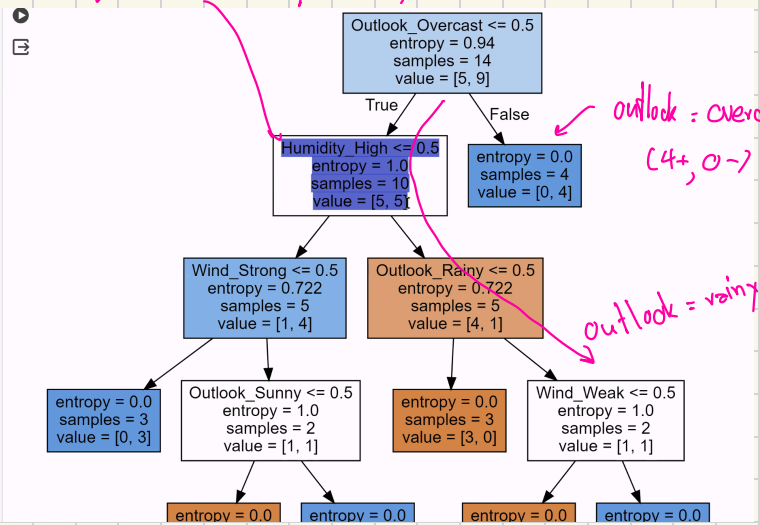




min_sample_split = 4 คือ sample < 4 ไม่แตก

min_impurity_decrease = 0.05 คือถ้าเล็ก > 0.05 ถึง split ต่อ





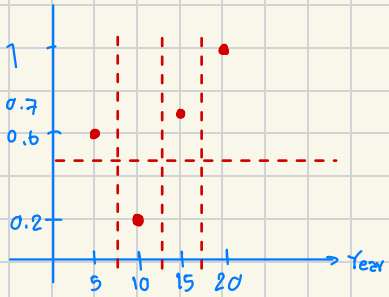

ก่อน outlook sunny กับ rainy

outlook : overcast (4+, 0-)

outlook = rainy

sklearn แยกต้นไม้โดยใช้ค่า continuous ทำให้มีแค่ 3 กิ่งไม่ได้

# 11/2/2024 Random Forest & XGBoost




$R = [0.1, -0.3, 0.1, 0.2]$

$\text{Similarity Score} = \frac{(\sum R)^2}{\#R + \lambda}$ $\lambda = 0$

ก็จะแทนให้ output R เริ่มต้น = 0.5

Year < 75

| 0.1 | -0.3, 0.1, 0.2 |
|---|---|
| $\frac{(0.1)^2}{1+0}$ | $\frac{(-0.3+0.1+0.2)^2}{3+0}$ |
| = 0.01 | = 0 |

$\text{Similarity Score} = \frac{(0.1-0.3+0.1+0.2)^2}{4+0}$

$= \frac{0.01}{4} = 0.0025$

$\text{Gain} = S_{left} + S_{right} - S_{root}$

$= 0.01 + 0 - 0.0025$

$= 0.0075$

Year < 12.5

| 0.1, -0.3 | 0.1, 0.2 |
|---|---|
| $\frac{(-0.2)^2}{2} = 0.02$ | $\frac{(0.1+0.2)^2}{2+0} = 0.045$ |

$\text{Gain} = 0.02 + 0.045 - 0.0025$

$= 0.065 - 0.0025$

$= 0.0625$

Year < 17.5

| 0.1, -0.3, 0.1 | 0.2 |
|---|---|
| $\frac{(-0.1)^2}{3} = 0.0033$ | $\frac{(0.2)^2}{1} = 0.04$ |

$\text{Gain} = 0.0033 + 0.04 - 0.0025$

$= 0.040$

Year < 12.5

Gain > 8

0.1, -0.3 | 0.02 | 0.1, 0.2 | 0.09/2 = 0.045

0.1 | -0.3 | 0.1 | 0.2

= 0.01 | 0.09 | 0.01 | 0.04

Gain = 0.01 + 0.09 - 0.02 = 0.08

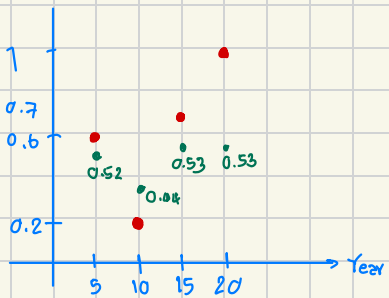
= 0.01 + 0.04 - 0.045 = 0.005

Year < 12.5

Year < 7.5 | 0.1 0.2

0.1 | -0.3

Output = $\frac{\Sigma R}{\#R + \lambda}$ = 0.15




new_output = 0.5 + learning rate × output

= 0.5 + 0.2 × output

0.5 + (0.2)(0.1) = 0.52

0.5 + (0.2)(-0.3) = 0.44

0.5 + (0.2)(0.15) = 0.53

R | 0.08 | -0.24 | 0.07 | 0.17

Year < 7.5

0.08 | -0.24, 0.07 0.17

Similarity score = $\frac{(0.08 - 0.24 + 0.07 + 0.17)^2}{4 + 0}$

= $\frac{(0.08)^2}{4}$

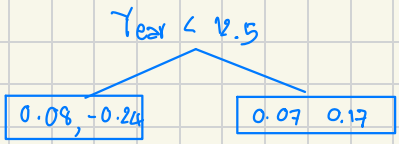
= 0.0016

Similarity = $\frac{(0.08)^2}{1}$ = $\frac{(-0.24 + 0.07 - 0.17)^2}{3}$

= 0.0064 = 0

Gain = 0.0064 + 0 - 0.0016

= 0.0048




$\frac{0.16^2}{2}$ = 0.0128

$\frac{0.24^2}{2}$ = 0.0288

Gain = 0.0128 + 0.0288 − 0.0016 = 0.0240

```
[27]  clf.fit(X,y)
```

```
XGBClassifier(base_score=None, booster=None, callbacks=None,
              colsample_bylevel=None, colsample_bynode=None,
              colsample_bytree=None, device=None, early_stopping_rounds=None,
              enable_categorical=False, eval_metric=None, feature_types=None,
              gamma=None, grow_policy=None, importance_type=None,
              interaction_constraints=None, learning_rate=None, max_bin=None,
              max_cat_threshold=None, max_cat_to_onehot=None,
              max_delta_step=None, max_depth=None, max_leaves=None,
              min_child_weight=None, missing=nan, monotone_constraints=None,
              multi_strategy=None, n_estimators=None, n_jobs=None,
              num_parallel_tree=None, random_state=None, ...)
```

```
[28]  clf.feature_importances_
```

ค่าความสำคัญของ attribute

```
array([0.20695841, 0.03330886, 0.07800627, 0.04108072, 0.02753872,
       0.53931534, 0.        , 0.02886179, 0.01932171, 0.02560823],
      dtype=float32)
```

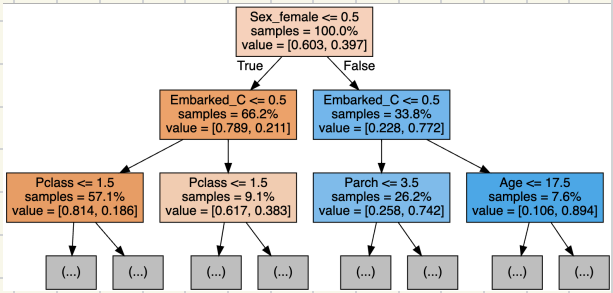
```
[29]  clf.feature_names_in_
```

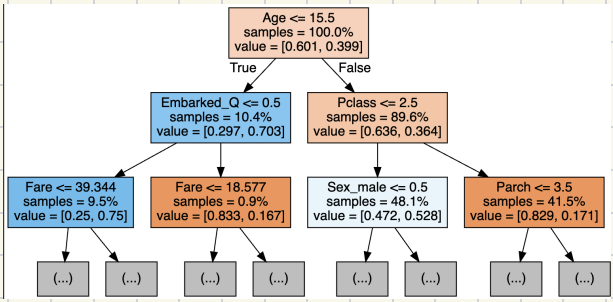
```
array(['Pclass', 'Age', 'SibSp', 'Parch', 'Fare', 'Sex_female',
       'Sex_male', 'Embarked_C', 'Embarked_Q', 'Embarked_S'], dtype='<U10')
```

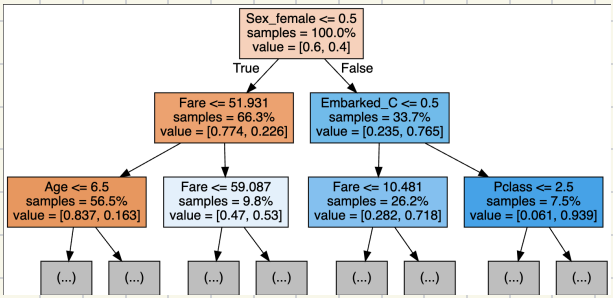
```
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
X_train,X_test,Y_train,Y_test=train_test_split(X,y,test_size=0.2)

rf=RandomForestClassifier()
xgb=XGBClassifier()
rf.fit(X_train,Y_train)
xgb.fit(X_train,Y_train)
y_pred=rf.predict(X_test)
print(f'RandomForest {accuracy_score(y_pred,Y_test)}')
y_pred=xgb.predict(X_test)
print(f'XGBoost {accuracy_score(y_pred,Y_test)}')
```

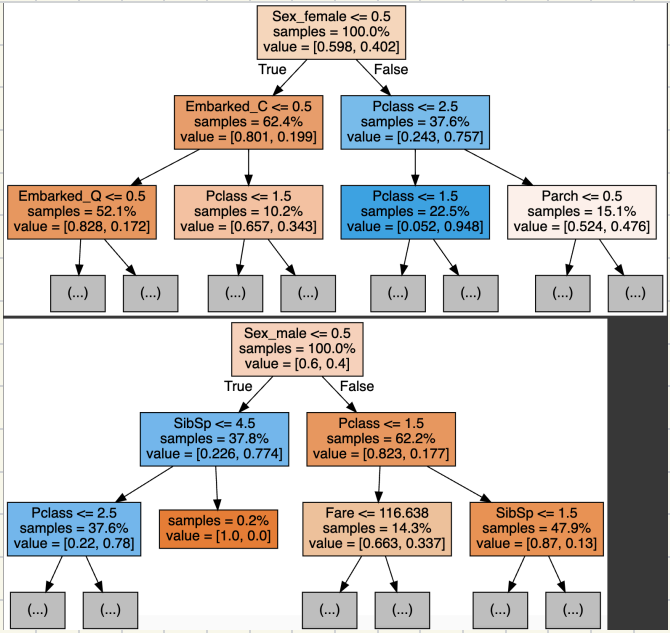
```
RandomForest 0.7482517482517482
XGBoost 0.7692307692307693
```

Random Forest display






https://colab.research.google.com/drive/1WWx_aR1LWnVA_wuFgVKgLuwr1Px_I5Ic?usp=sharing