# C4.5 and CART

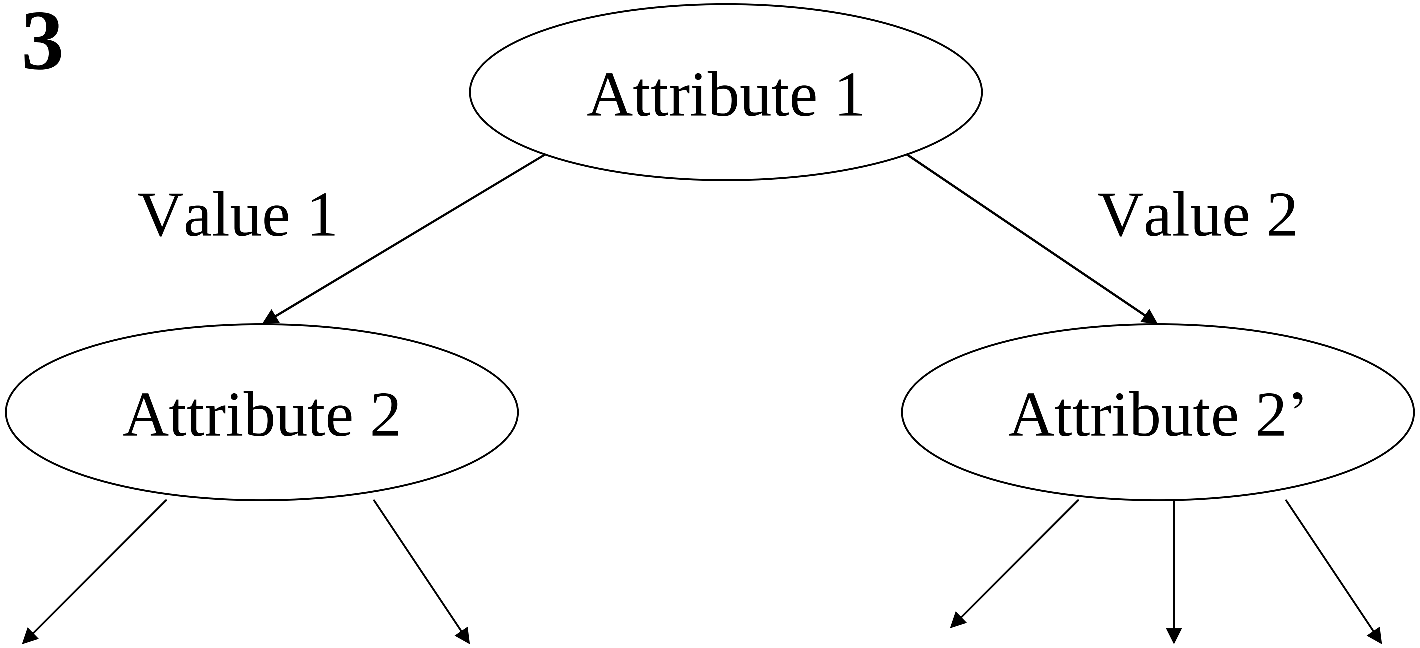
# ID3

- Creates tree using information theory concepts and tries to reduce expected number of comparison..
- ID3 chooses split attribute with the highest information gain:

- For Attribute *A*, relative to a collection of data

$$Gain(D, A) \equiv Entropy(D) - \sum_{v \in Values(A)} \frac{|Dv|}{|D|} Entropy(Dv)$$

# ID 3




- **Which Attribute is Best?**
  - Select the attribute that is most useful for classifying examples.
  - Quantitative Measure
    - *Information Gain*
    - For Attribute *A*, relative to a collection of data

$$Gain(D, A) \equiv Entropy(D) - \sum_{v \in Values(A)} \frac{|Dv|}{|D|} Entropy(Dv)$$

    - Expected Reduction of Entropy

| Outlook | Temperature | Humidity | Wind | PlayTennis |
|---|---|---|---|---|
| Sunny | Hot | High | Weak | No |
| Sunny | Hot | High | Strong | No |
| Overcast | Hot | High | Weak | Yes |
| Rain | Mild | High | Weak | Yes |
| Rain | Cool | Normal | Weak | Yes |
| Rain | Cool | Normal | Strong | No |
| Overcast | Cool | Normal | Strong | Yes |
| Sunny | Mild | High | Weak | No |
| Sunny | Cool | Normal | Weak | Yes |
| Rain | Mild | Normal | Weak | Yes |
| Sunny | Mild | Normal | Strong | Yes |
| Overcast | Mild | High | Strong | Yes |
| Overcast | Hot | Normal | Weak | Yes |
| Rain | Mild | High | Strong | No |

# Entropy of D

$$Entropy(D) = Entropy([9+,5-])$$

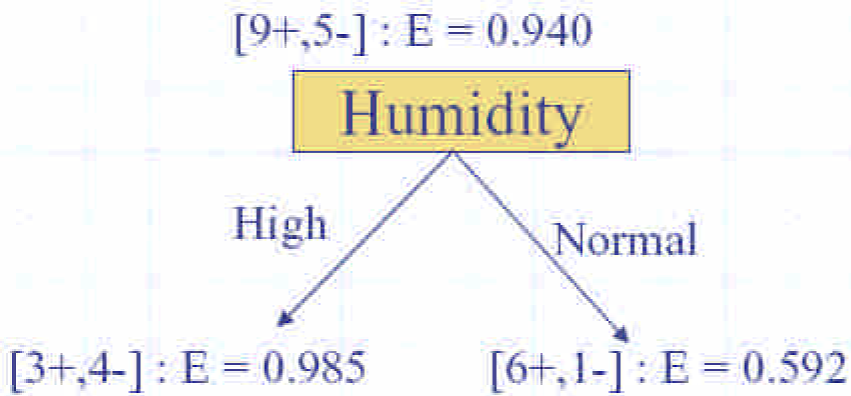
$$= -\frac{9}{14}\log\left(\frac{9}{14}\right) - \frac{5}{14}\log\left(\frac{5}{14}\right)$$

$$= 0.940$$

| Outlook | Temperature | Humidity | Wind | PlayTennis |
|---|---|---|---|---|
| Sunny | Hot | High | Weak | No |
| Sunny | Hot | High | Strong | No |
| Overcast | Hot | High | Weak | Yes |
| Rain | Mild | High | Weak | Yes |
| Rain | Cool | Normal | Weak | Yes |
| Rain | Cool | Normal | Strong | No |
| Overcast | Cool | Normal | Strong | Yes |
| Sunny | Mild | High | Weak | No |
| Sunny | Cool | Normal | Weak | Yes |
| Rain | Mild | Normal | Weak | Yes |
| Sunny | Mild | Normal | Strong | Yes |
| Overcast | Mild | High | Strong | Yes |
| Overcast | Hot | Normal | Weak | Yes |
| Rain | Mild | High | Strong | No |

# Attribute Humidity

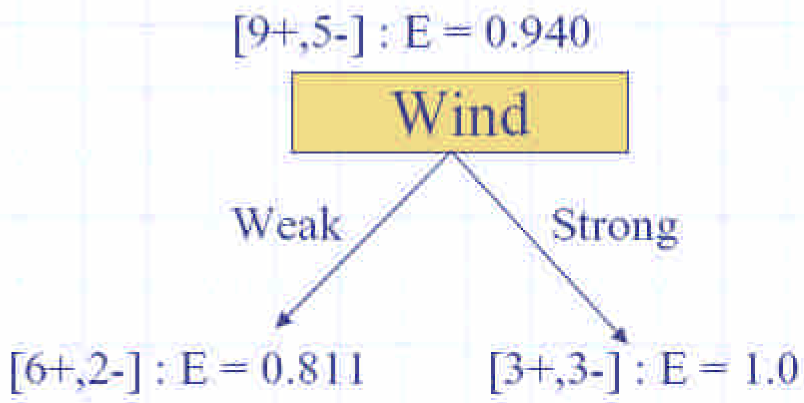
- Attribute *Humidity*
  - $D_{high}$ = [3+,4-]
  - $D_{normal}$=[6+,1-]




$$
\begin{aligned}
Gain(D, Wind) &= Entropy(D) - \sum_{v \in \{high, normal\}} \frac{|Dv|}{|D|} Entropy(Dv) \\
&= Entropy(D) - \frac{7}{14} Entropy(D_{high}) - \frac{7}{14} Entropy(D_{normal}) \\
&= 0.940 - \frac{7}{14} 0.985 - \frac{7}{14} 0.592 \\
&= 0.151
\end{aligned}
$$

# Attribute Wind

- Attribute *Wind*
  - $D = [9+,5-]$
  - $D_{weak} = [6+,2-]$
  - $D_{strong} = [3+,3-]$



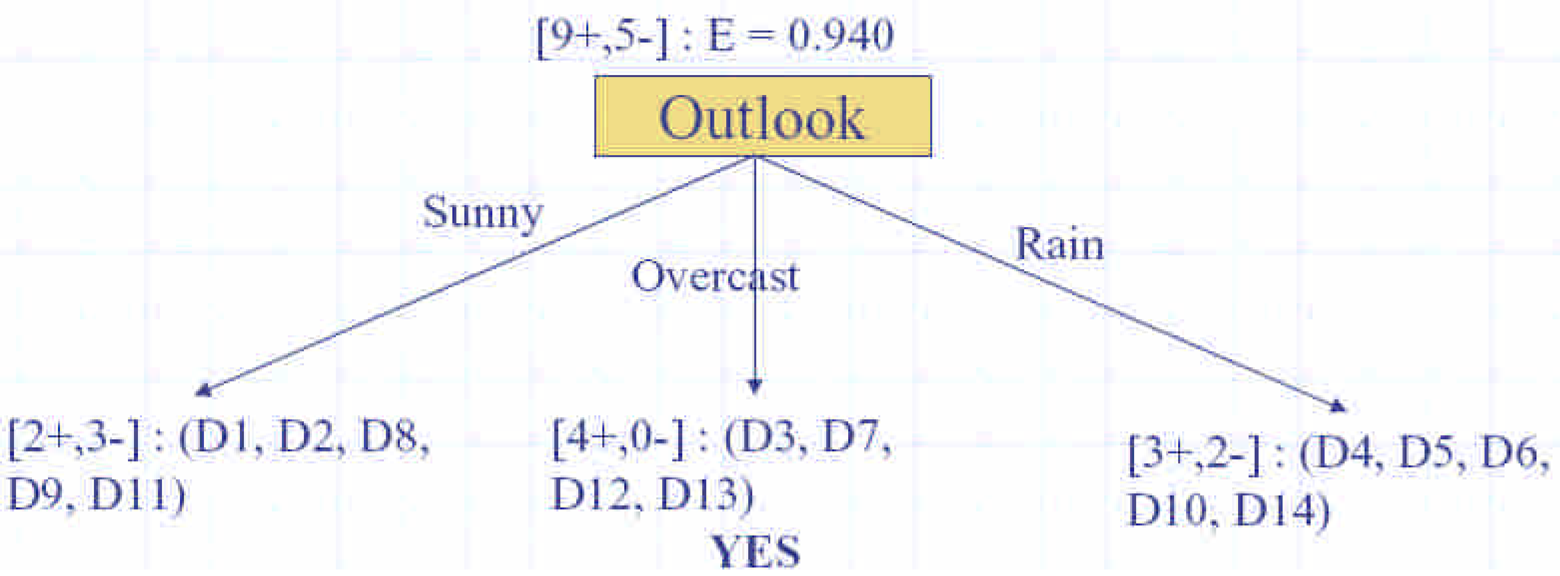

$$\begin{aligned}
Gain(D, Wind) &= Entropy(D) - \sum_{v \in \{weak, strong\}} \frac{|Dv|}{|D|} Entropy(Dv) \\
&= Entropy(D) - \frac{8}{14} Entropy(D_{weak}) - \frac{6}{14} Entropy(D_{strong}) \\
&= 0.940 - \frac{8}{14} 0.811 - \frac{6}{14} 1.00 \\
&= 0.048
\end{aligned}$$

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Yes:+ ; No:- | | | | | | | | |
| 2 | **Root** | p | n | p+n | -p/(p+n)*log2(p/(p+n)) | -n/(p+n)*log2(n/(p+n)) | sum of Products | Probability | Product |
| 3 | D | 9 | 5 | 14 | 0.410 | 0.531 | 0.940 | 1 | 0.940 |
| 4 | | | | | | | | | |
| 5 | **Outlook** | p | n | p+n | -p/(p+n)*log2(p/(p+n)) | -n/(p+n)*log2(n/(p+n)) | sum of Products | Probability | Product |
| 6 | D_Sunny | 2 | 3 | 5 | 0.529 | 0.442 | 0.971 | 0.36 | 0.347 |
| 7 | D_Overcast | 4 | 0 | 4 | 0.000 | 0.000 | 0.000 | 0.29 | 0.000 |
| 8 | D_Rain | 3 | 2 | 5 | 0.442 | 0.529 | 0.971 | 0.36 | 0.347 |
| 9 | SUM | | | | | | | | 0.694 |
| 10 | GAIN | | | | | | Max. information gain | | **0.247** |
| 11 | | | | | | | | | |
| 12 | **Temp.** | p | n | p+n | -p/(p+n)*log2(p/(p+n)) | -n/(p+n)*log2(n/(p+n)) | sum of Products | Probability | Product |
| 13 | D_Hot | 2 | 2 | 4 | 0.500 | 0.500 | 1.000 | 0.29 | 0.286 |
| 14 | D_Mild | 4 | 2 | 6 | 0.390 | 0.528 | 0.918 | 0.43 | 0.394 |
| 15 | D_Cool | 3 | 1 | 4 | 0.311 | 0.500 | 0.811 | 0.29 | 0.232 |
| 16 | SUM | | | | | | | | 0.911 |
| 17 | GAIN | | | | | | | | 0.029 |
| 18 | | | | | | | | | |
| 19 | **Humidity** | p | n | p+n | -p/(p+n)*log2(p/(p+n)) | -n/(p+n)*log2(n/(p+n)) | sum of Products | Probability | Product |
| 20 | D_High | 3 | 4 | 7 | 0.524 | 0.461 | 0.985 | 0.50 | 0.493 |
| 21 | D_Normal | 6 | 1 | 7 | 0.191 | 0.401 | 0.592 | 0.50 | 0.296 |
| 22 | SUM | | | | | | | | 0.788 |
| 23 | GAIN | | | | | | | | 0.152 |

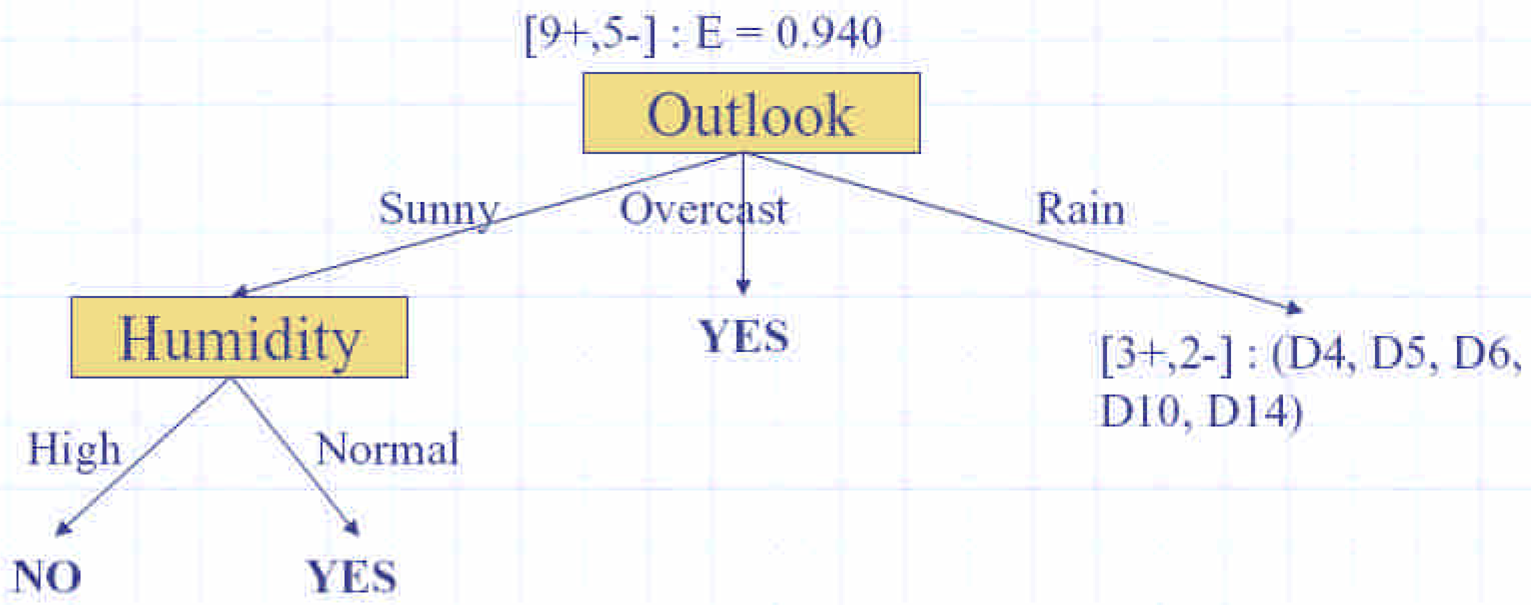
Sheet1 / Sheet2 / Sheet3

# Best Attribution Chosen

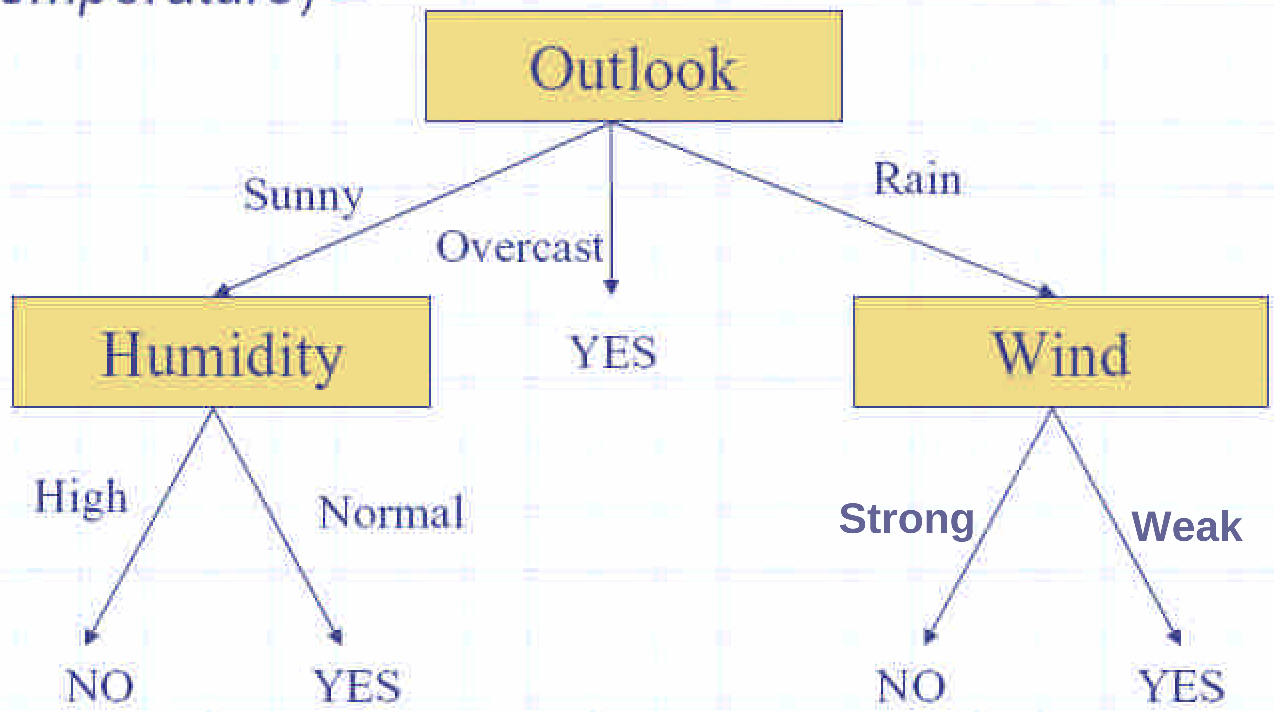
- Best Attribute?
  - ***Gain(D, Outlook) = 0.246***
  - *Gain(D, Humidity)* = 0.151
  - *Gain(D, Wind)* = 0.048
  - *Gain(D, Temperature)* = 0.029



- Best Attribute?
  - ***Gain*(*D*, *Humidity*) = 0.971**
  - *Gain*(*D*, *Wind*) = 0.020
  - *Gain*(*D*, *Temperature*) = 0.571



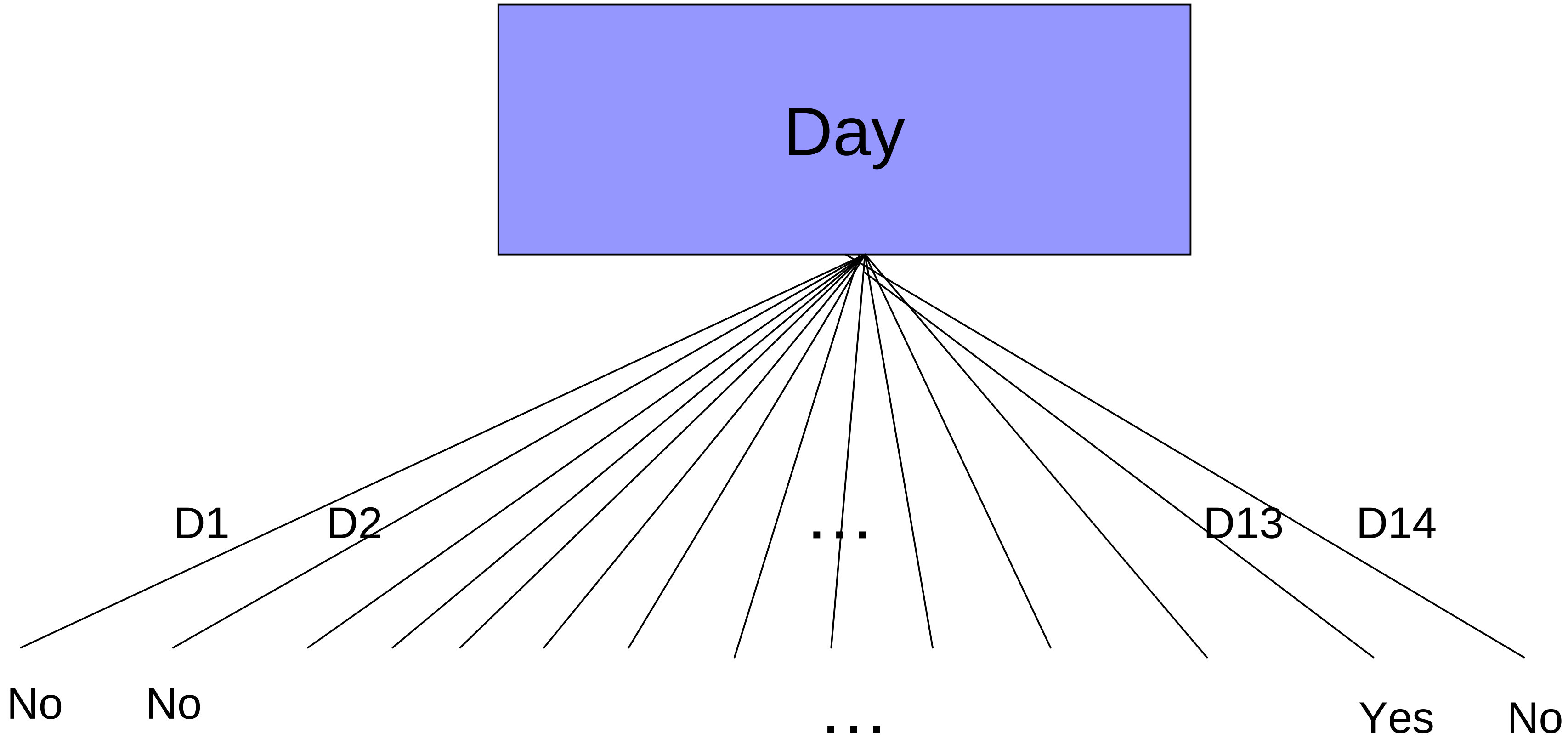
- Best Attribute?
  - *Gain(D, Humidity)* = 0.020
  - ***Gain(D, Wind)* = 0.971**
  - *Gain(D, Temperature)* = 0.020



| 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 |

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 2 | **Root** | p | n | p+n | -p/(p+n)*log2(p/(p+n)) | -n/(p+n)*log2(n/(p+n)) | sum of Products | Probability | Product |
| 3 | D | 9 | 5 | 14 | 0.410 | 0.531 | 0.940 | 1 | 0.940 |
| 4 | | | | | | | | | |
| 5 | **Day** | p | n | p+n | -p/(p+n)*log2(p/(p+n)) | -n/(p+n)*log2(n/(p+n)) | sum of Products | Probability | Product |
| 6 | **D1** | 0 | 1 | 1 | 0.000 | 0.000 | 0.000 | 0.07 | 0.000 |
| 7 | **D2** | 0 | 1 | 1 | 0.000 | 0.000 | 0.000 | 0.07 | 0.000 |
| 8 | **D3** | 1 | 0 | 1 | 0.000 | 0.000 | 0.000 | 0.07 | 0.000 |
| 9 | **D4** | 1 | 0 | 1 | 0.000 | 0.000 | 0.000 | 0.07 | 0.000 |
| 10 | **D5** | 1 | 0 | 1 | 0.000 | 0.000 | 0.000 | 0.07 | 0.000 |
| 11 | **...** | | | | | | | | |
| 12 | **D14** | 0 | 1 | 1 | 0.000 | 0.000 | 0.000 | 0.07 | 0.000 |
| 13 | SUM | | | | | | | | 0.000 |
| 14 | GAIN | | | | | | Max. information gain | | **0.940** |
| 15 | **Outlook** | p | n | p+n | -p/(p+n)*log2(p/(p+n)) | -n/(p+n)*log2(n/(p+n)) | sum of Products | Probability | Product |
| 16 | D_Sunny | 2 | 3 | 5 | 0.529 | 0.442 | 0.971 | 0.36 | 0.347 |
| 17 | D_Overcast | 4 | 0 | 4 | 0.000 | 0.000 | 0.000 | 0.29 | 0.000 |
| 18 | D_Rain | 3 | 2 | 5 | 0.442 | 0.529 | 0.971 | 0.36 | 0.347 |
| 19 | SUM | | | | | | | ? | 0.694 |
| 20 | GAIN | | | | | | | | **0.247** |
| 21 | | | | | | | | | |
| 22 | **Temp.** | p | n | p+n | -p/(p+n)*log2(p/(p+n)) | -n/(p+n)*log2(n/(p+n)) | sum of Products | Probability | Product |
| 23 | D_Hot | 2 | 2 | 4 | 0.500 | 0.500 | 1.000 | 0.29 | 0.286 |
| 24 | D_Mild | 4 | 2 | 6 | 0.390 | 0.528 | 0.918 | 0.43 | 0.394 |

Sheet1 / Sheet2 / Sheet3


Day
D1
D2
. . .
D13
D14
No
No
. . .
Yes
No

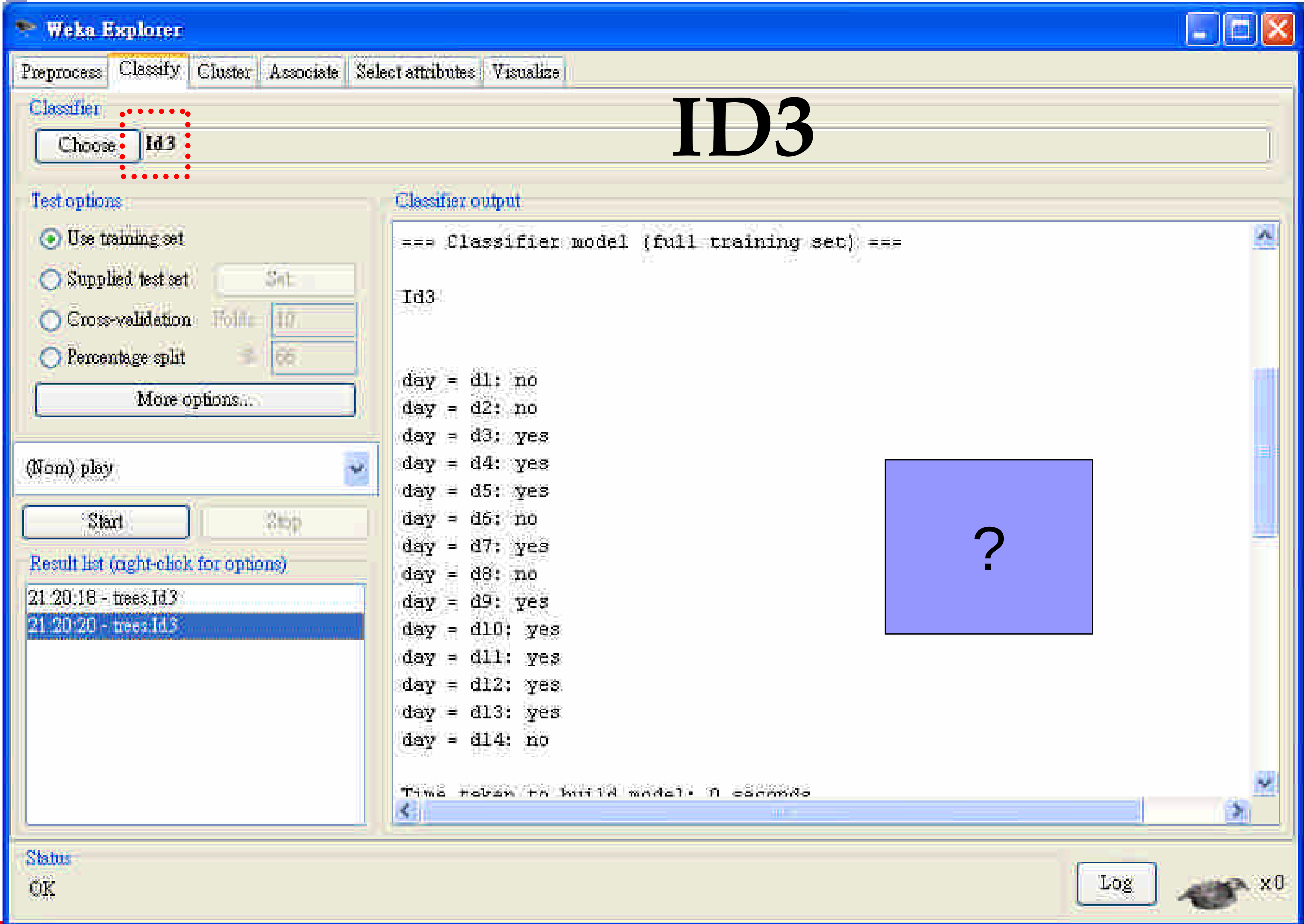
# C4.5

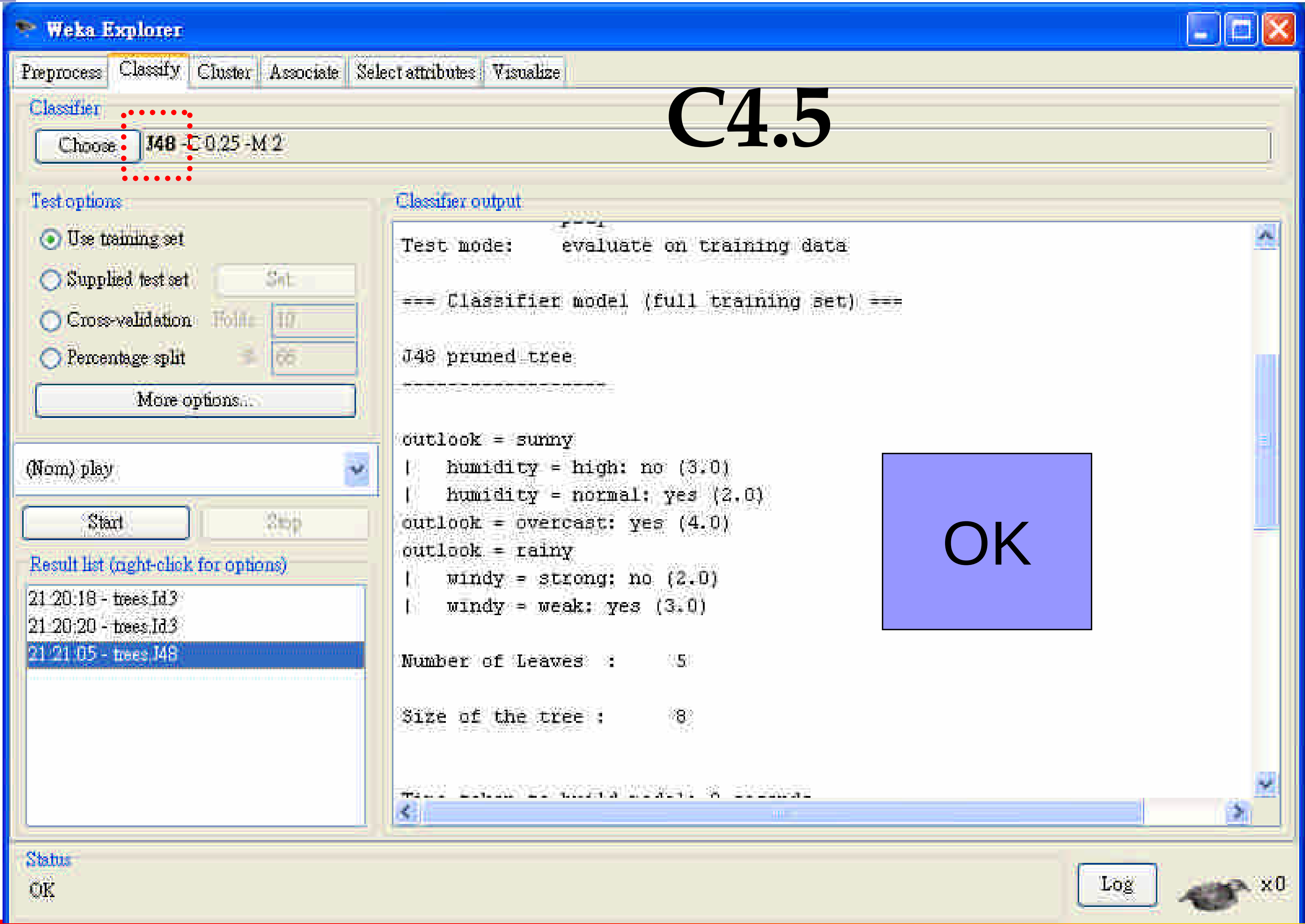
- ID3 favors attributes with large number of divisions
- Improved version of ID3:
  - Missing Data
  - Continuous Data
  - Pruning
  - Rules
  - GainRatio:

$$GainRatio(S, A) \equiv \frac{Gain(S, A)}{SplitInformation(S, A)}$$

$$SplitInformation(S, A) \equiv - \sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

where $S_i$ is subset of $S$ for which $A$ has value $v_i$

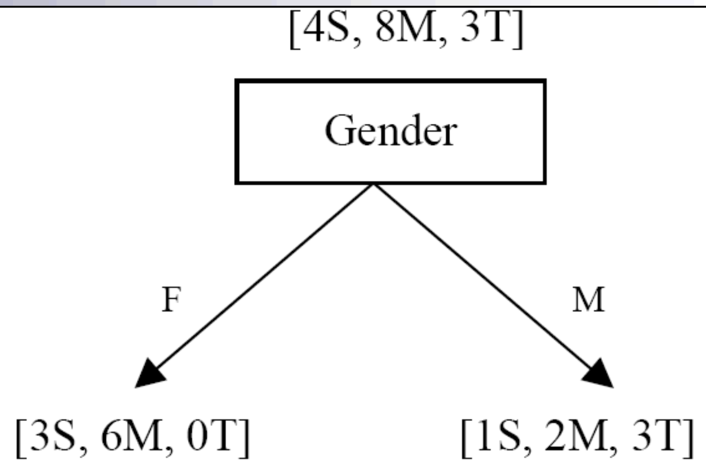

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
Id3
ID3
Test options
Use training set
Supplied test set
Set
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) play
Start
Stop
Result list (right-click for options)
21:20:18 - trees.Id3
21:20:20 - trees.Id3
Classifier output
=== Classifier model (full training set) ===
Id3
day = d1: no
day = d2: no
day = d3: yes
day = d4: yes
day = d5: yes
day = d6: no
day = d7: yes
day = d8: no
day = d9: yes
day = d10: yes
day = d11: yes
day = d12: yes
day = d13: yes
day = d14: no
?
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) play
Start
Stop
Result list (right-click for options)
21:20:18 - trees.Id3
21:20:20 - trees.Id3
21:21:05 - trees.J48
Classifier output
Test mode:    evaluate on training data
=== Classifier model (full training set) ===
J48 pruned tree
------------------
outlook = sunny
|   humidity = high: no (3.0)
|   humidity = normal: yes (2.0)
outlook = overcast: yes (4.0)
outlook = rainy
|   windy = strong: no (2.0)
|   windy = weak: yes (3.0)
Number of Leaves  :     5
Size of the tree :      8
Status
OK
Log
x 0

# C4.5

OK

# CART




$$\Phi(\text{Gender})=2*\frac{6}{15}\times\frac{9}{15}\times\left(\frac{2}{15}+\frac{4}{15}+\frac{3}{15}\right)$$
$$=0.224$$

- Create Binary Tree
- Uses entropy
- Formula to choose split point, s, for node t:

$$\Phi(s/t) = 2P_L P_R \sum_{j=1}^{m} | P(C_j \mid t_L) - P(C_j \mid t_R) |$$

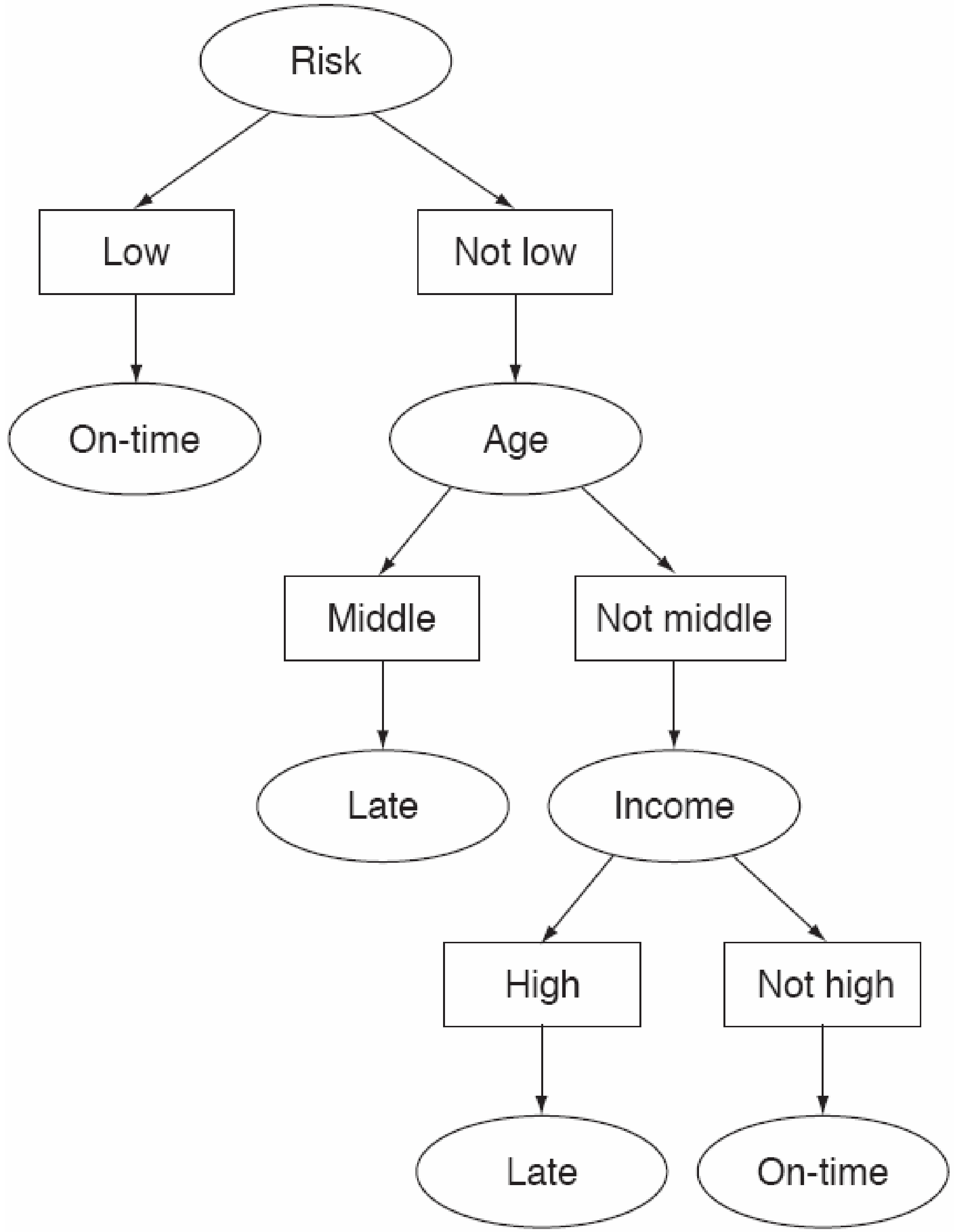
➔Maximum

- $P_L$,$P_R$ probability that a tuple in the training set will be on the left or right side of the tree.

# 案例

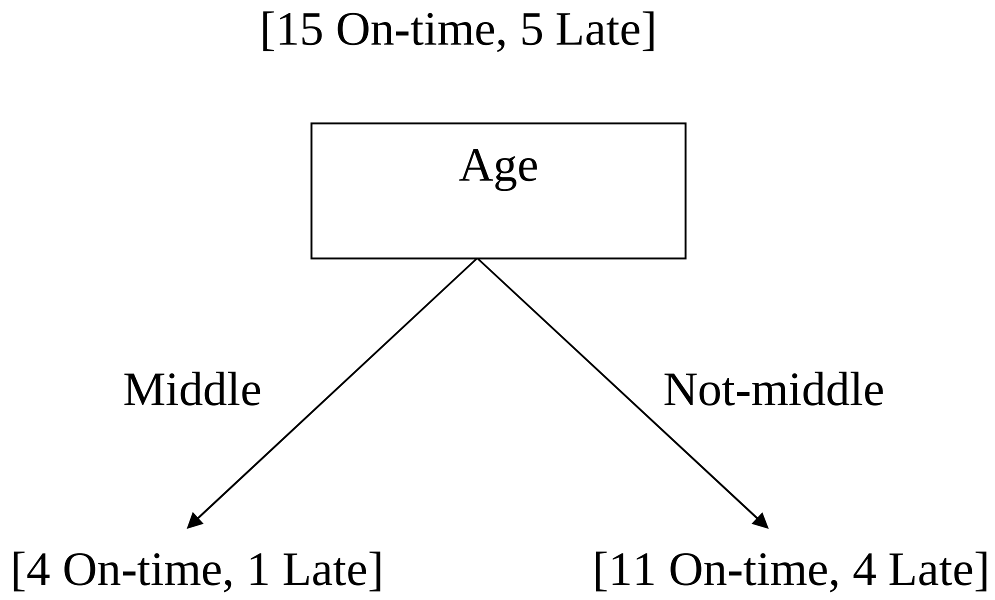
| 年齡 | 收入 | 資產 | 負債 | 貸款金額 | 風險 | 信用 | 貸款繳交情況 |
|---|---|---|---|---|---|---|---|
| 20（年輕） | 17,152（低） | 11,090 | 20,455 | 400 | 高 | 綠 | 準時 |
| 23（年輕） | 25,862（低） | 24,756 | 30,083 | 2,300 | 高 | 綠 | 準時 |
| 28（年輕） | 26,169（低） | 47,355 | 49,341 | 3,100 | 高 | 黃 | 遲繳 |
| 23（年輕） | 21,117（低） | 21,242 | 30,278 | 300 | 高 | 紅 | 拖欠 |
| 22（年輕） | 7,127（低） | 23,903 | 17,231 | 900 | 低 | 黃 | 準時 |
| 26（年輕） | 42,083（平均） | 35,726 | 41,421 | 300 | 高 | 紅 | 遲繳 |
| 24（年輕） | 55,557（平均） | 27,040 | 48,191 | 1,500 | 高 | 綠 | 準時 |
| 27（年輕） | 34,843（平均） | 0 | 21,031 | 2,100 | 高 | 紅 | 準時 |
| 29（年輕） | 74,295（平均） | 88,827 | 100,599 | 100 | 高 | 黃 | 準時 |
| 23（年輕） | 38,887（平均） | 6,260 | 33,635 | 9,400 | 低 | 綠 | 準時 |
| 28（年輕） | 31,758（平均） | 58,492 | 49,268 | 1,000 | 低 | 綠 | 準時 |
| 25（年輕） | 80,180（高） | 31,696 | 69,529 | 1,000 | 高 | 綠 | 遲繳 |
| 33（中年） | 40,921（平均） | 91,111 | 90,076 | 2,900 | 平均 | 黃 | 遲繳 |
| 36（中年） | 63,124（平均） | 164,631 | 144,697 | 300 | 低 | 綠 | 準時 |
| 39（中年） | 59,006（平均） | 195,759 | 161,750 | 600 | 低 | 綠 | 準時 |
| 39（中年） | 125,713（高） | 382,180 | 315,396 | 5,200 | 低 | 黃 | 準時 |
| 55（中年） | 80,149（高） | 511,937 | 21,923 | 1,000 | 低 | 綠 | 準時 |
| 62（老年） | 101,291（高） | 783,164 | 23,052 | 1,800 | 低 | 綠 | 準時 |
| 71（老年） | 81,723（高） | 776,344 | 20,277 | 900 | 低 | 綠 | 準時 |
| 63（老年） | 99,522（高） | 783,491 | 24,643 | 200 | 低 | 綠 | 準時 |

| Age | Income | Risk | Result |
|---|---|---|---|
| not-midlle | not-high | not-low | On-time |
| not-midlle | not-high | not-low | On-time |
| not-midlle | not-high | not-low | Late |
| not-midlle | not-high | not-low | Late |
| not-midlle | not-high | low | On-time |
| not-midlle | not-high | not-low | Late |
| not-midlle | not-high | not-low | On-time |
| not-midlle | not-high | not-low | On-time |
| not-midlle | not-high | not-low | On-time |
| not-midlle | not-high | low | On-time |
| not-midlle | not-high | low | On-time |
| not-midlle | high | not-low | Late |
| midlle | not-high | not-low | Late |
| midlle | not-high | low | On-time |
| midlle | not-high | low | On-time |
| midlle | high | low | On-time |
| midlle | high | low | On-time |
| Not-midlle | high | low | On-time |
| not-midlle | high | low | On-time |
| not-midlle | high | low | On-time |


Risk
Low
Not low
On-time
Age
Middle
Not middle
Late
Income
High
Not high
Late
On-time

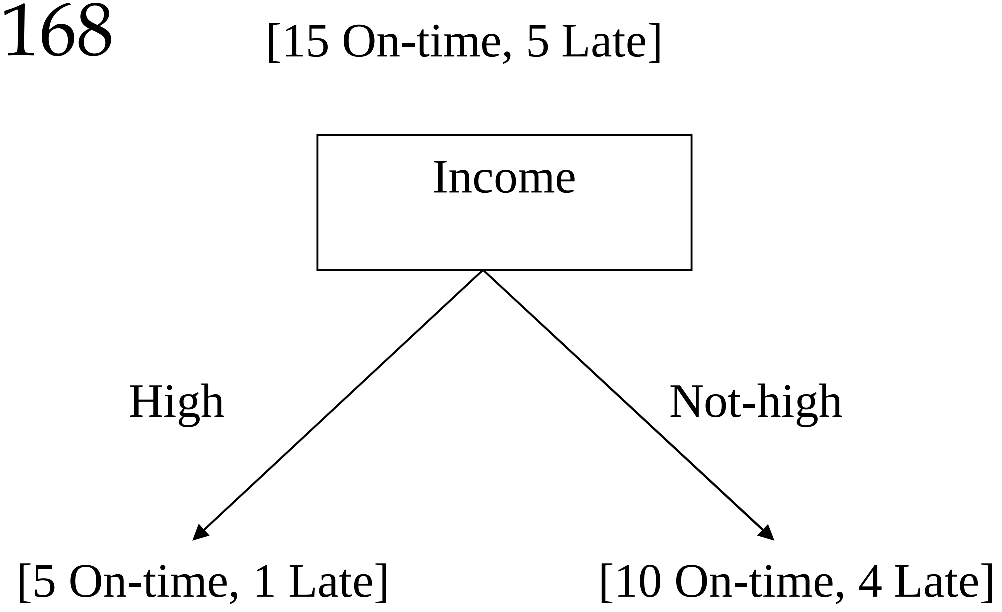
# CART Analysis

- At the start, there are three choices for split point:
  - $\Phi(\text{Age})=2(5/20)(15/20)(7/20 + 3/20)=0.1875$



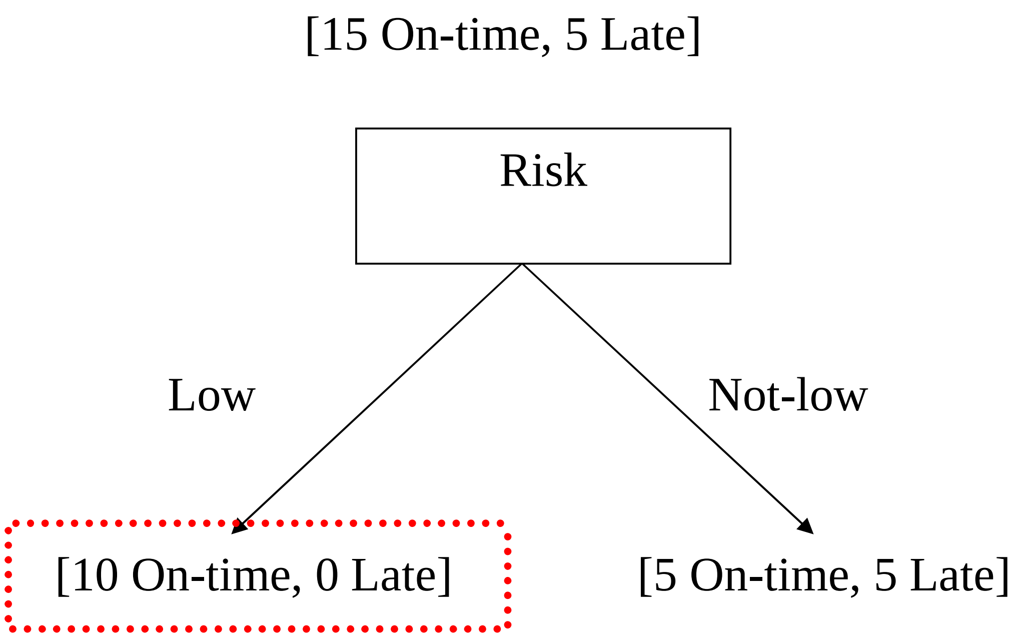
# CART Analysis

- At the start, there are three choices for split point:
  - $\Phi(\text{Income}) = 2(6/20)(14/20)(5/20 + 3/20) = 0.168$



# CART Analysis

- At the start, there are three choices for split point:
  - $\Phi(\text{Risk})=2(10/20)(10/20)(5/20 + 5/20)=0.25$



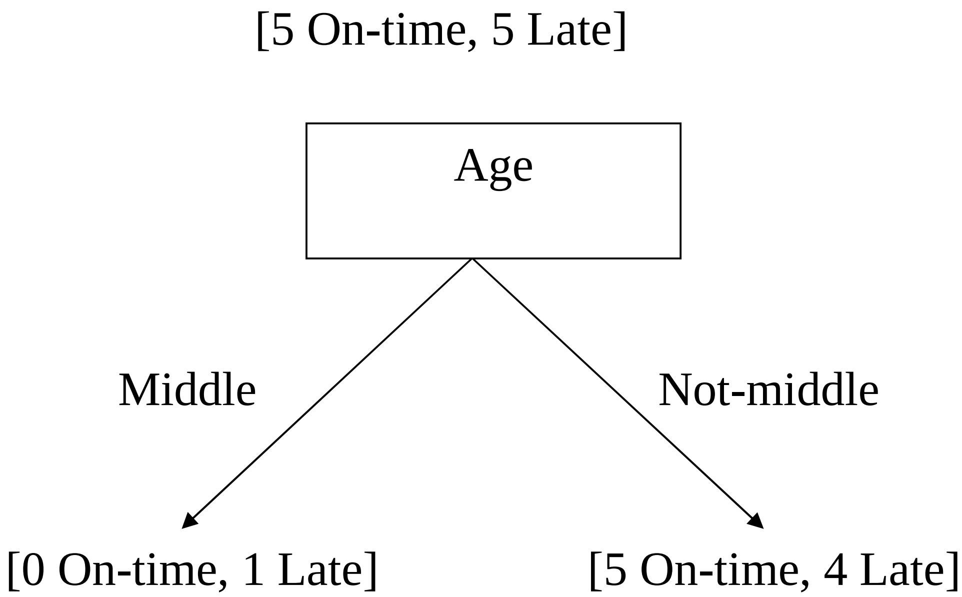

Maximum

# Step 2:

| Age | Income | Risk | Result |
|---|---|---|---|
| not-midlle | not-high | not-low | On-time |
| not-midlle | not-high | not-low | On-time |
| not-midlle | not-high | not-low | Late |
| not-midlle | not-high | not-low | Late |
| not-midlle | not-high | not-low | Late |
| not-midlle | not-high | not-low | On-time |
| not-midlle | not-high | not-low | On-time |
| not-midlle | not-high | not-low | On-time |
| not-midlle | high | not-low | Late |
| midlle | not-high | not-low | Late |

# CART Analysis

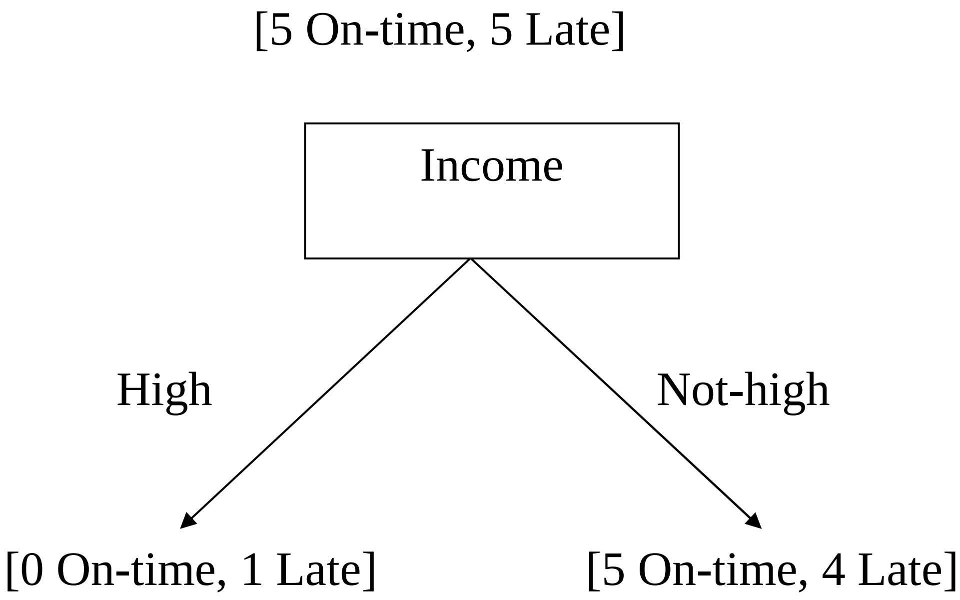
- At the start, there are three choices for split point:
  - $\Phi(\text{Age})=2(1/10)(9/10)(5/10 + 3/10)=0.144$

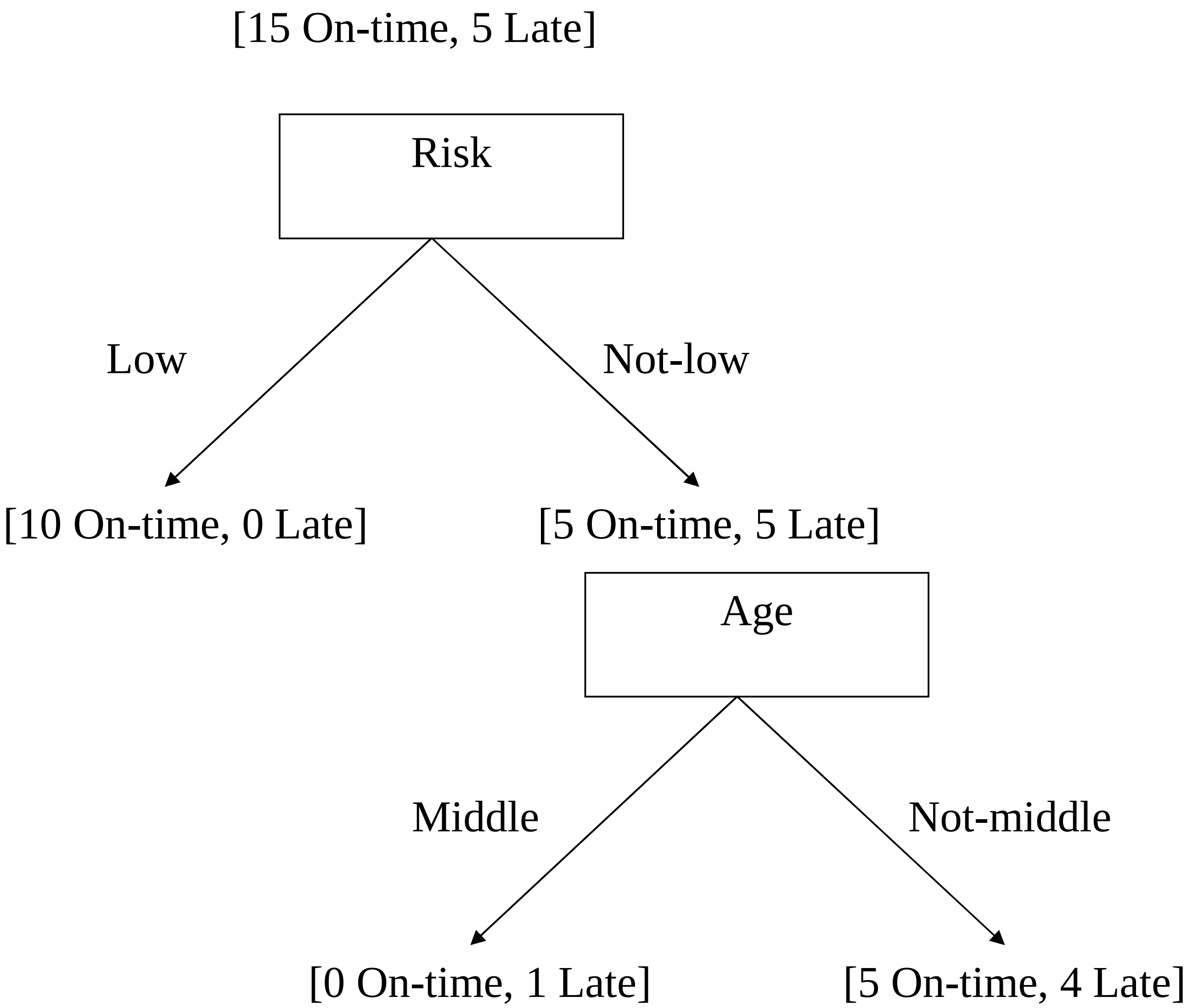



identical

# CART Analysis

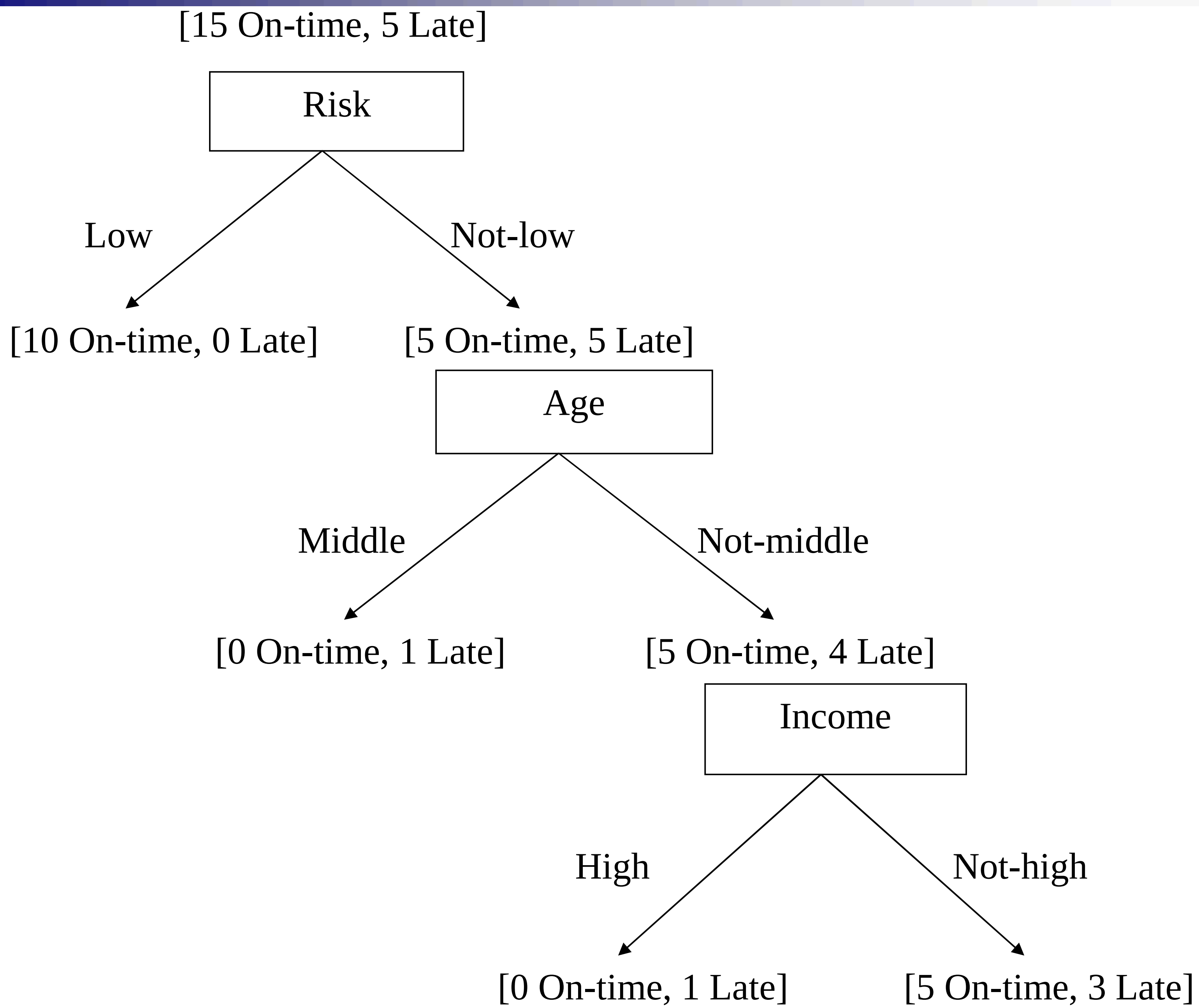
- At the start, there are three choices for split point:
  - $\Phi(\text{Income})=2(1/10)(9/10)(5/10 + 3/10)=0.144$




[15 On-time, 5 Late]
Risk
Low
Not-low
[10 On-time, 0 Late]
[5 On-time, 5 Late]
Age
Middle
Not-middle
[0 On-time, 1 Late]
[5 On-time, 4 Late]

# Step 3:

| Age | Income | Risk | Result |
|---|---|---|---|
| not-midlle | not-high | not-low | On-time |
| not-midlle | not-high | not-low | On-time |
| not-midlle | not-high | not-low | Late |
| not-midlle | not-high | not-low | Late |
| not-midlle | not-high | not-low | Late |
| not-midlle | not-high | not-low | On-time |
| not-midlle | not-high | not-low | On-time |
| not-midlle | not-high | not-low | On-time |
| not-midlle | high | not-low | Late |


[15 On-time, 5 Late]
Risk
Low
Not-low
[10 On-time, 0 Late]
[5 On-time, 5 Late]
Age
Middle
Not-middle
[0 On-time, 1 Late]
[5 On-time, 4 Late]
Income
High
Not-high
[0 On-time, 1 Late]
[5 On-time, 3 Late]