

DECISION TREE



IKM Lab

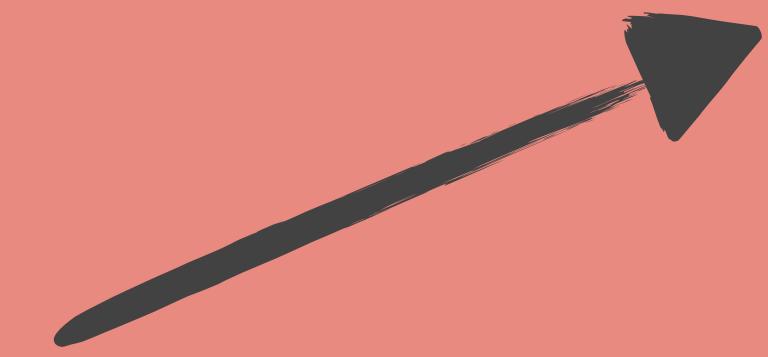
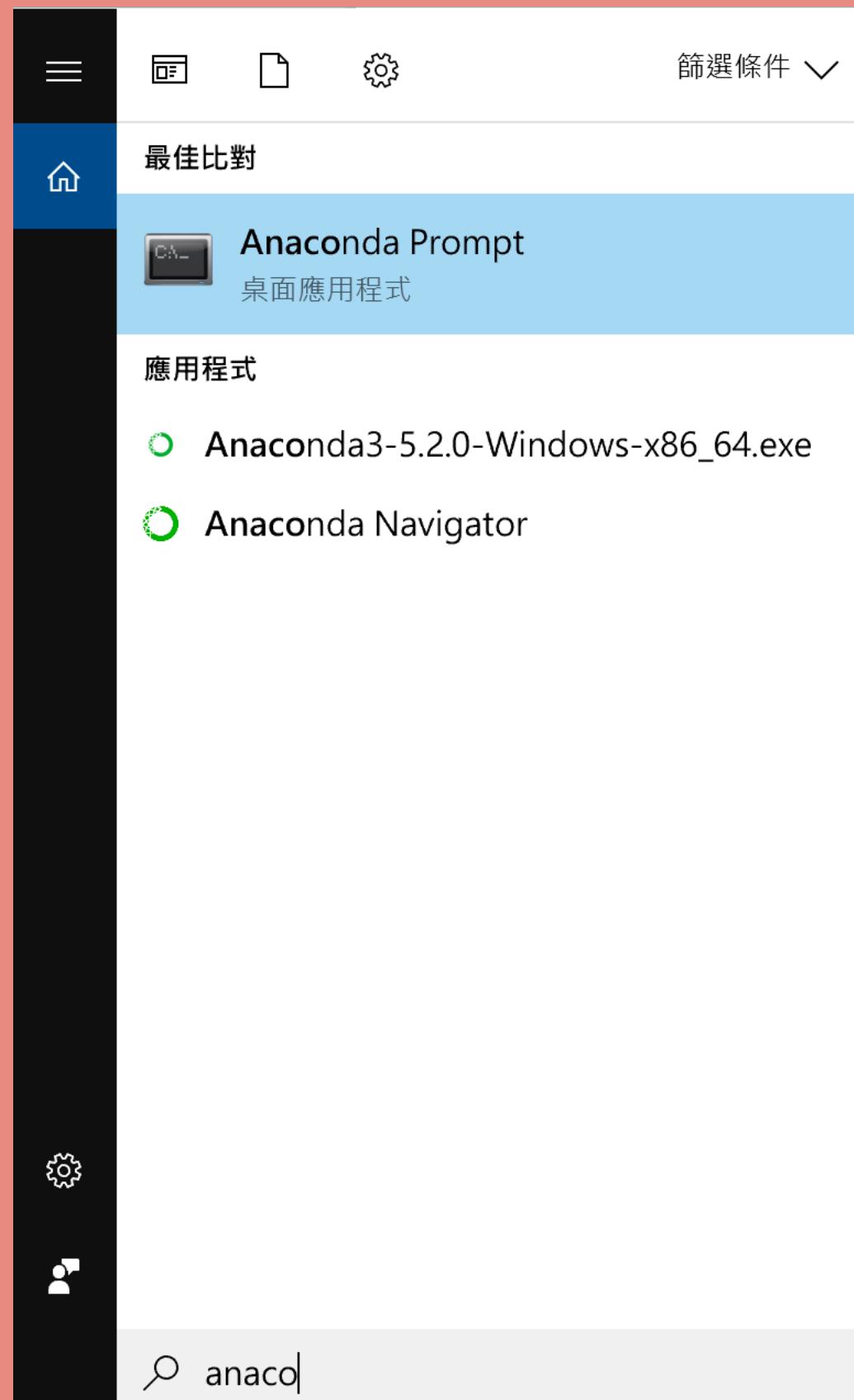


OUTLINE

- Package installation
- What is Decision Tree ?
- How can we use it ?
- How can we know which attribute is important ?
- My decision tree 😍



- Search for “Anaconda Prompt”



- Pip install pydotplus

```
(base) C:\Users\yuying>pip install pydotplus
```

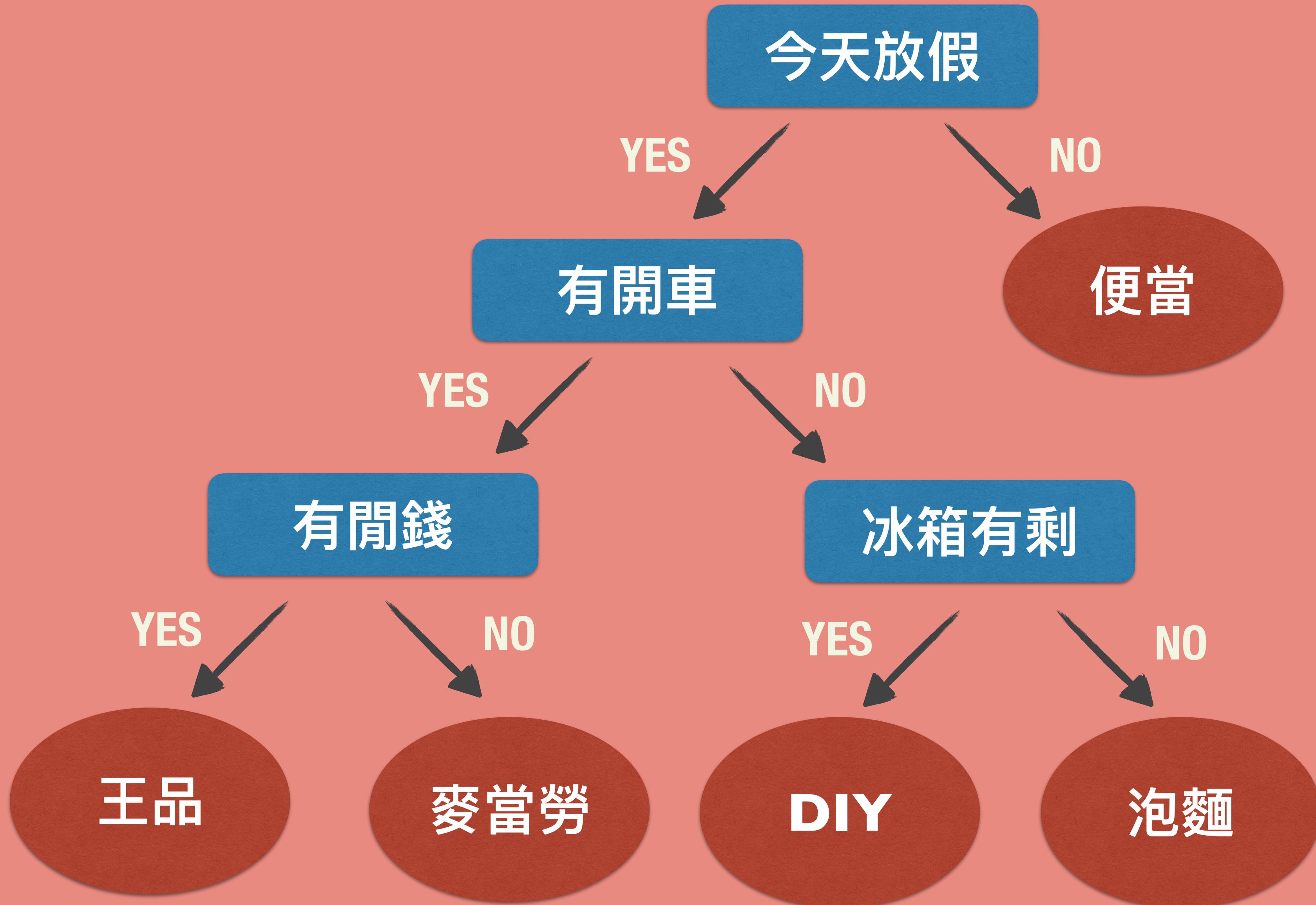


- Just wait a moment :)

```
(base) C:\Users\yuying>pip install pydotplus
Collecting pydotplus
  Downloading https://files.pythonhosted.org/packages/60/bf/62567830b700d9f6930e9ab6831d6ba256f7b0b730acb37278b0ccdfaf/pydotplus-2.0.2.tar.gz (278kB)
    100% |██████████| 286kB 546kB/s
Requirement already satisfied: pyparsing>=2.0.1 in c:\programdata\anaconda3\lib\site-packages (from pydotplus) (2.2.0)
Building wheels for collected packages: pydotplus
  Running setup.py bdist_wheel for pydotplus ... done
  Stored in directory: C:\Users\yuying\AppData\Local\pip\Cache\wheels\35\7b\ab\66fb7b2ac1f6df87475b09dc48e707b6e0de80a6d8444e3628
Successfully built pydotplus
distributed 1.21.8 requires msgpack, which is not installed.
Installing collected packages: pydotplus
Successfully installed pydotplus-2.0.2
```

WHAT IS DECISION TREE?

- 藍色方形：決策點
- 紅色橢圓：結果

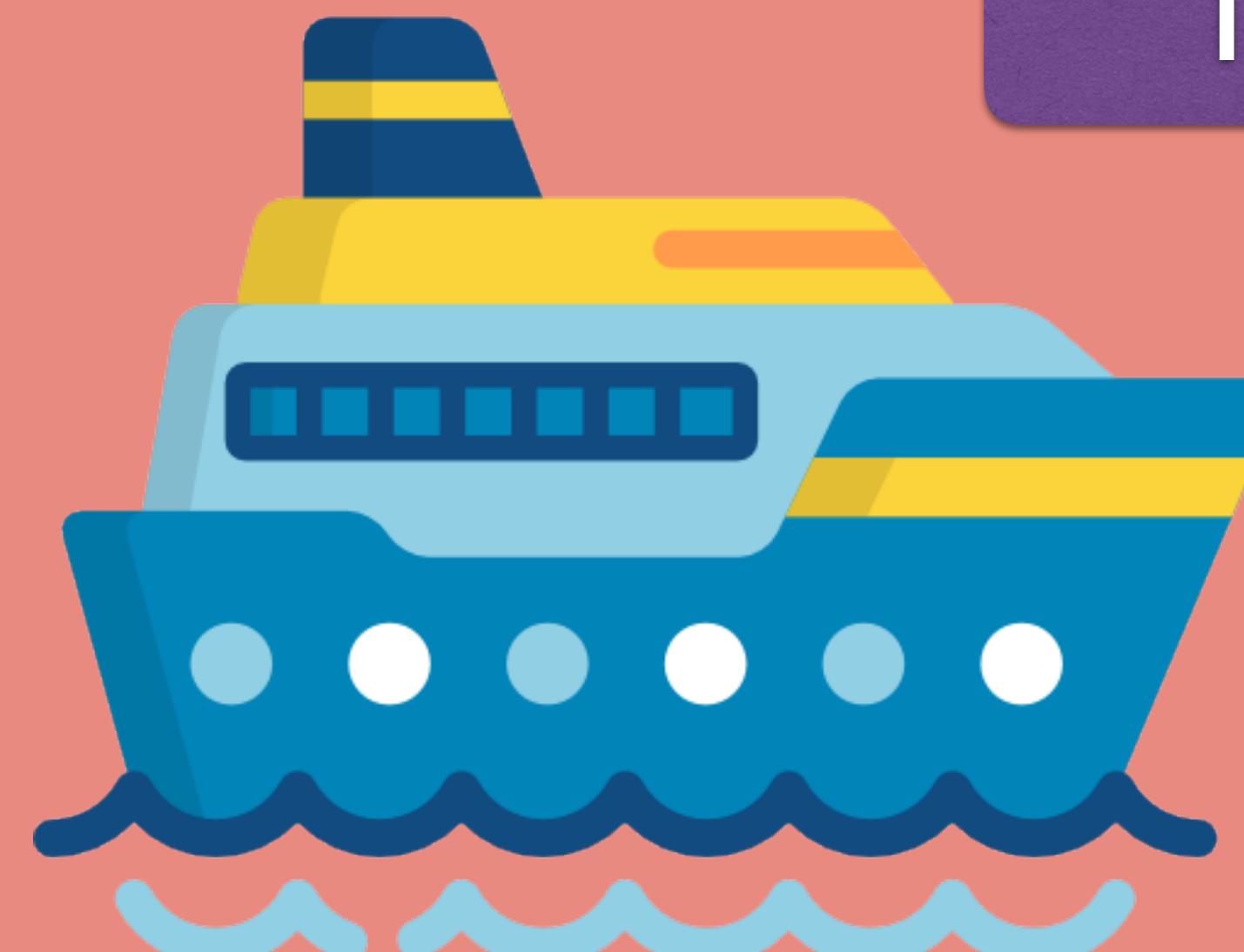


HOW CAN WE USE DECISION TREE?



Passenger 1

35-year-old
Ticket Class = 2nd
Parents.Wife



Titanic



Passenger 2

7-month-old
Ticket Class = 3rd
Parents



Passenger 3

62-year-old
Ticket Class = 1st
Husband. son

TRAINING DATA

- PClass : 艙等 (1~3)
- Sex : 性別
- Age : 年齡
- SibSp : 兄弟姐妹、丈夫（妻子）人數
- Parch : 父母、小孩人數
- Fare : 票價
- Result : 是否存活 (1:存活，0:死亡)

result	Pclass	Sex	Age	SibSp	Parch	Fare
0	3	0	22	1	0	7.25
1	1	1	38	1	0	71.2833
1	3	1	26	0	0	7.925
1	1	1	35	1	0	53.1
0	3	0	35	0	0	8.05
0	1	0	54	0	0	51.8625
0	3	0	2	3	1	21.075
1	3	1	27	0	2	11.1333
1	2	1	14	1	0	30.0708
1	3	1	4	1	1	16.7
1	1	1	58	0	0	26.55
0	3	0	20	0	0	8.05

TESTING DATA

- 現在已知以下資訊：
 - PClass : 艙等 (1~3)
 - Sex : 性別
 - Age : 年齡
 - SibSp : 兄弟姐妹、丈夫（妻子）人數
 - Parch : 父母、小孩人數
 - Fare : 票價
- 希望猜出該乘客是否存活

result	Pclass	Sex	Age	SibSp	Parch	Fare
1	2	1	40	1	1	39
0	1	0	31	1	0	52
0	2	0	70	0	0	10.5
1	2	0	31	0	0	13
0	3	0	18	0	0	7.775
0	3	0	24.5	0	0	8.05
1	3	1	18	0	0	9.8417
0	3	1	43	1	6	46.9
1	1	0	36	0	1	512.3292
1	1	0	27	0	0	76.7292
0	3	0	20	0	0	9.225

FLOWCHART

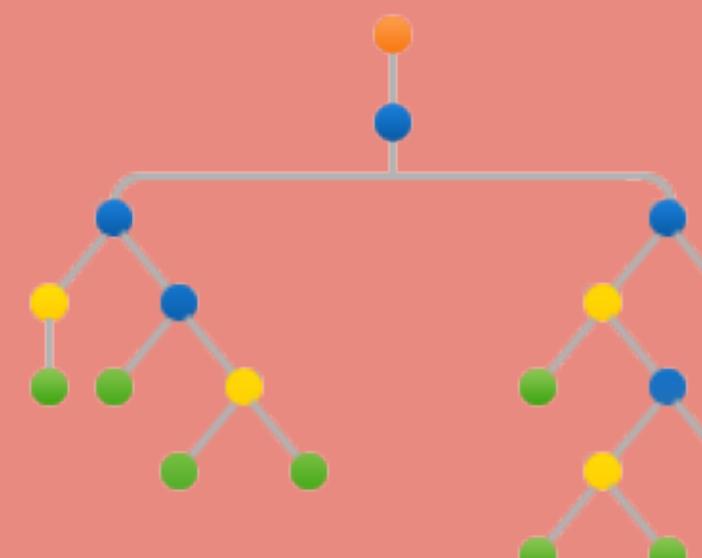
Training Data

Build
Decision Tree

Apply to
Testing Data

Prediction

result	Pclass	Sex	Age	SibSp	Parch	Fare
0	3	0	22	1	0	7.25
1	1	1	38	1	0	71.2833
1	3	1	26	0	0	7.925
1	1	1	35	1	0	53.1
0	3	0	35	0	0	8.05
0	1	0	54	0	0	51.8625
0	3	0	2	3	1	21.075
1	3	1	27	0	2	11.1333
1	2	1	14	1	0	30.0708
1	3	1	4	1	1	16.7
1	1	1	58	0	0	26.55
0	3	0	20	0	0	8.05



result	Pclass	Sex	Age	SibSp	Parch	Fare
1	2	1	40	1	1	39
0	1	0	31	1	0	52
0	2	0	70	0	0	10.5
1	2	0	31	0	0	13
0	3	0	18	0	0	7.775
0	3	0	24.5	0	0	8.05
1	3	1	18	0	0	9.8417
0	3	1	43	1	6	46.9
1	1	0	36	0	1	512.3292
1	1	0	27	0	0	76.7292
0	3	0	20	0	0	9.225



**HOW CAN WE KNOW
WHICH ATTRIBUTE IS
IMPORTANT?**

資料集	人數
存活	100
死亡	200

Attribute 1

年齡 > 60

YES

NO

人數	
存活	50
死亡	50

人數	
存活	50
死亡	150

Attribute 2

艙等 = 1st

YES

NO

人數	
存活	99
死亡	1

資料集	人數
存活	100
死亡	200

Attribute 1

年齡 > 60

YES

人數	
存活	50
死亡	50

NO

人數	
存活	50
死亡	150

Attribute 2

艙等 = 1st

YES

人數	
存活	99
死亡	1

NO

人數	
存活	1
死亡	199



GINI

$$1 - \sum_{i=1}^J p_i^2$$

- Gini impurity score is a measure of how the data would be incorrectly labeled.
- Gini = 0.0 -> best
 - when we use this attribute to classify the dataset, it can be clearly categorized.
- Gini = 0.5 -> worst
 - this attribute is not an important feature to make a decision.

GINI

Training Data	人數
存活	100
死亡	200

$$1 - \frac{100}{300}^2 - \frac{200}{300}^2 = 0.44$$



年齡 > 60

YES

人數	
存活	50
死亡	50

$$1 - \frac{50}{100}^2 - \frac{50}{100}^2 = 0.5$$

NO

人數	
存活	50
死亡	150

$$1 - \frac{50}{200}^2 - \frac{150}{200}^2 = 0.375$$

艙等 = 1st

YES

人數	
存活	99
死亡	1

$$1 - \frac{99}{100}^2 - \frac{1}{100}^2 = 0.02$$

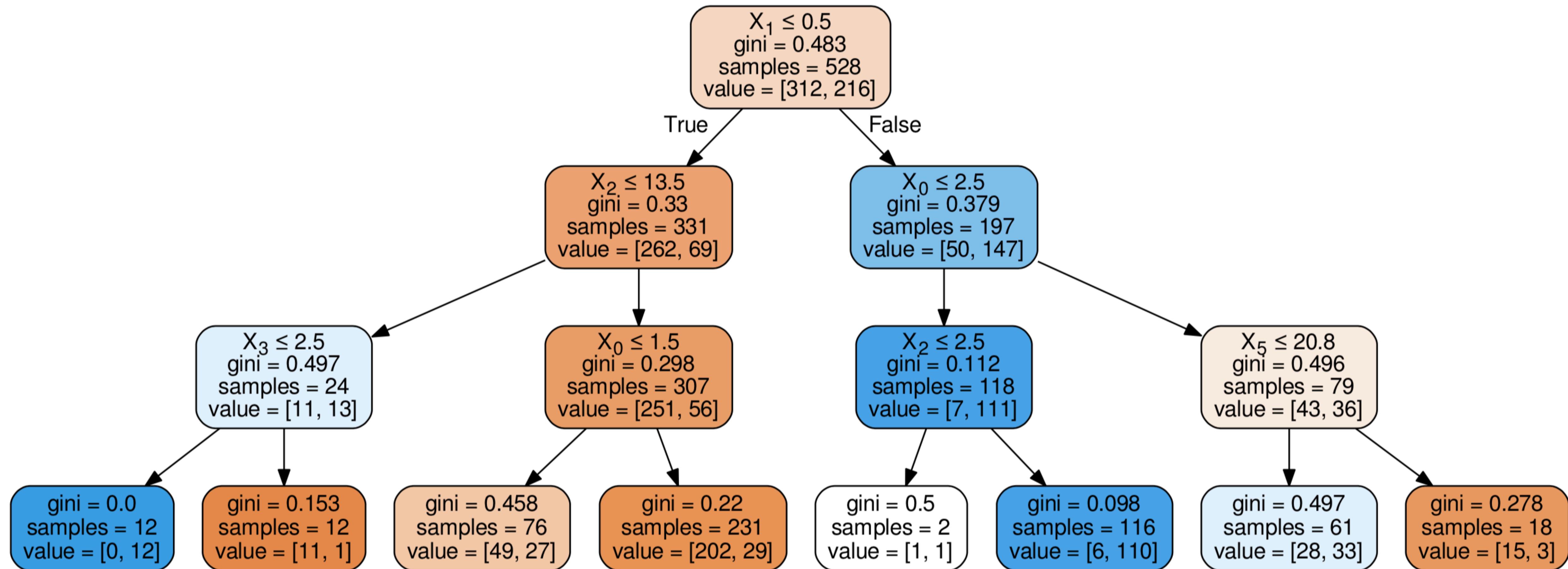
NO

人數	
存活	1
死亡	199

$$1 - \frac{1}{200}^2 - \frac{199}{200}^2 = 0.01$$

TREE.PDF

$X_0 : \text{PClass}$	$X_3 : \text{SibSp}$
$X_1 : \text{Sex}$	$X_4 : \text{Parch}$
$X_2 : \text{Age}$	$X_5 : \text{Fare}$



THANK YOU