

# [Weka]

# 連續型資料轉離散型資料

國立臺中科技大學資訊工程系  
張家瑋 博士

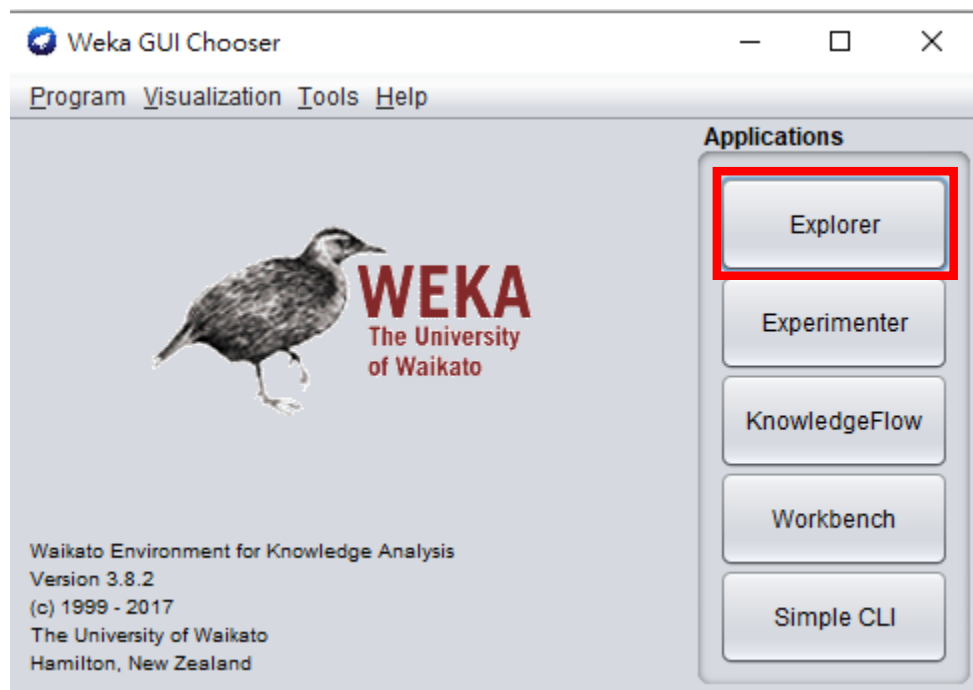
# Step 1

	A	B	C	D
1	Age	Marriage	income	buy_computer
2	24	single	High	no
3	28	single	mediate	no
4	35	single	low	yes
5	32	Married	mediate	no
6	40	Married	low	no
7	42	Married	low	no
8	38	Married	mediate	no
9	29	single	High	no
10	22	Married	low	no
11	33	Married	mediate	no
12	25	Married	High	yes
13	50	Married	mediate	no
14	35	single	mediate	yes
15	45	Married	low	no
16	37	single	mediate	yes
17	18	single	low	no

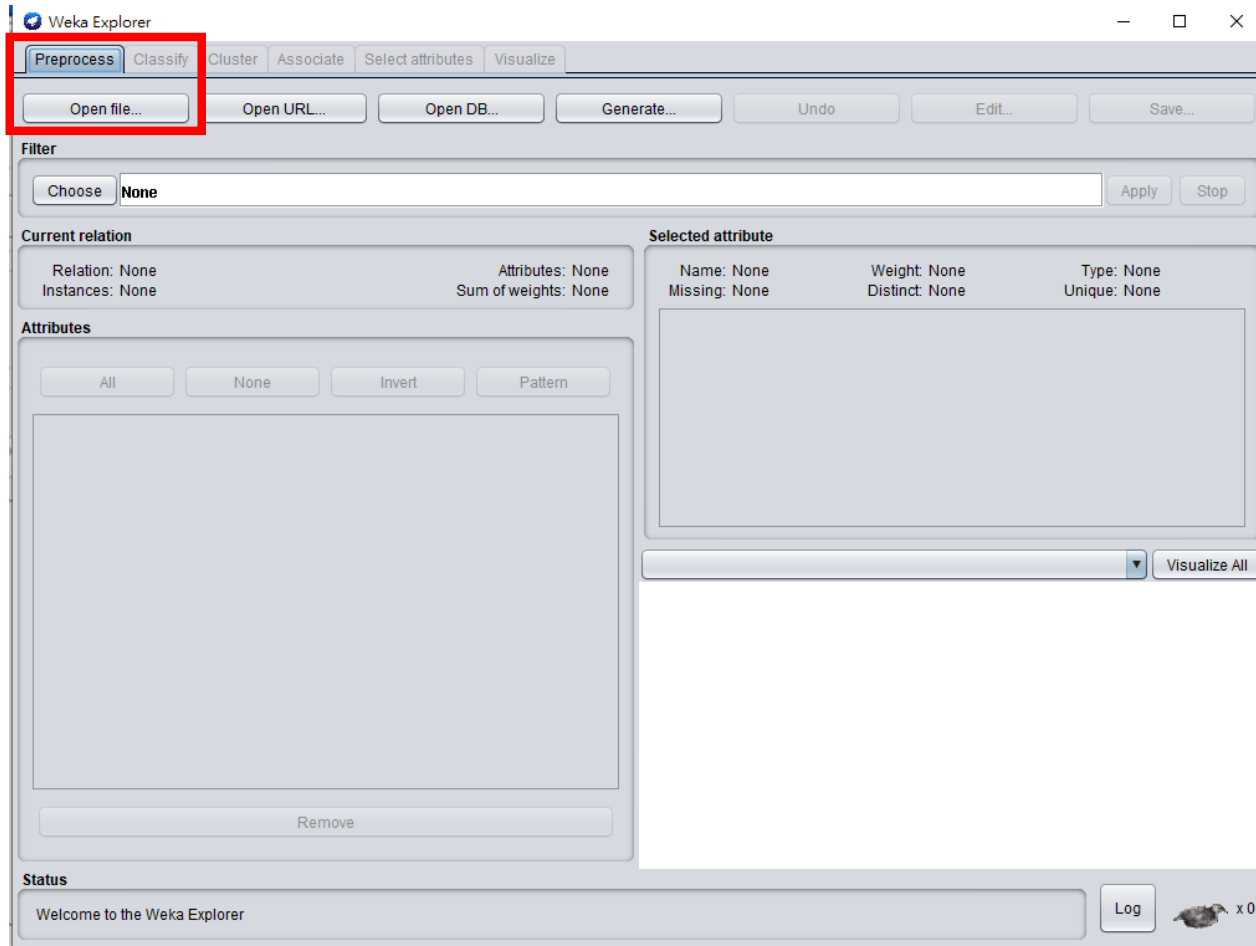
- 為了使用特定分類演算法，如決策樹系列方法時，較好的輸入的資料型態是離散型資料。
- 年齡是連續數值，如何轉成離散型資料呢？

# Step 2

- 準備好 CSV 或 ARFF 格式的檔案。
- Weka 啟動後，點選 Explorer。



# Step 3



- 選擇 Preprocess 。
- 點擊 Open files ，選取欲處理的 csv, arff 資料集 。

# Step 4

- Choose 前處理方法

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

Filter

Choose None Apply Stop

Current relation

Relation: test\_class\_1  
Instances: 16  
Attributes: 4  
Sum of weights: 16

Attributes

All | None | Invert | Pattern

No.	Name
1	<input checked="" type="checkbox"/> Age
2	<input type="checkbox"/> Marriage
3	<input type="checkbox"/> income
4	<input type="checkbox"/> buy_computer

Remove

Selected attribute

Name: Age  
Missing: 0 (0%)  
Distinct: 15  
Type: Numeric  
Unique: 14 (88%)

Statistic	Value
Minimum	18
Maximum	50
Mean	33.312
StdDev	8.731

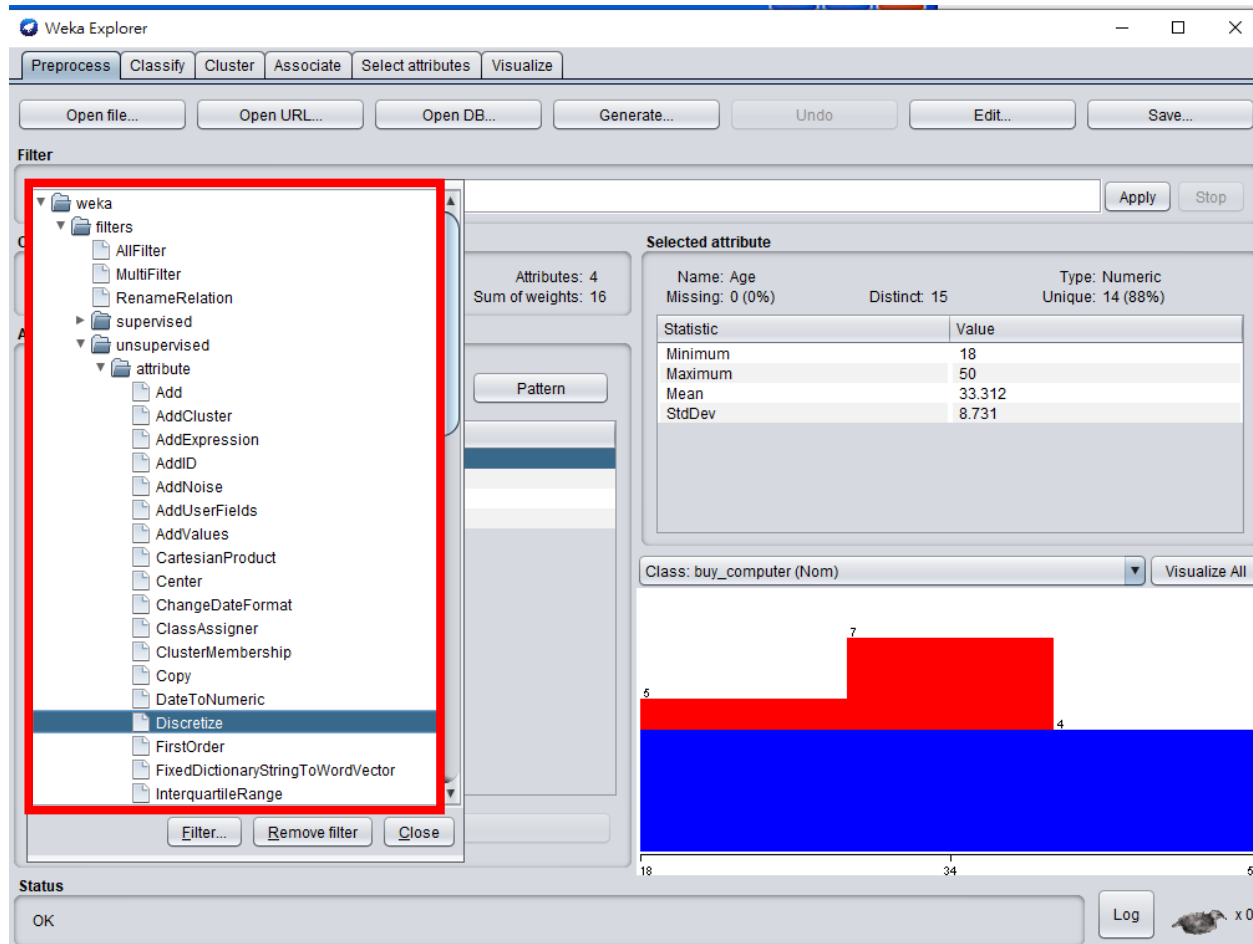
Class: buy\_computer (Nom) Visualize All

Bar chart showing the distribution of the 'Age' attribute. The x-axis represents Age values (18, 34, 50). The y-axis represents frequency. The chart shows a red bar for Age 18 (frequency 7) and a blue bar for Age 34 (frequency 5). The total frequency is 12.

Status

OK Log x 0

# Step 5



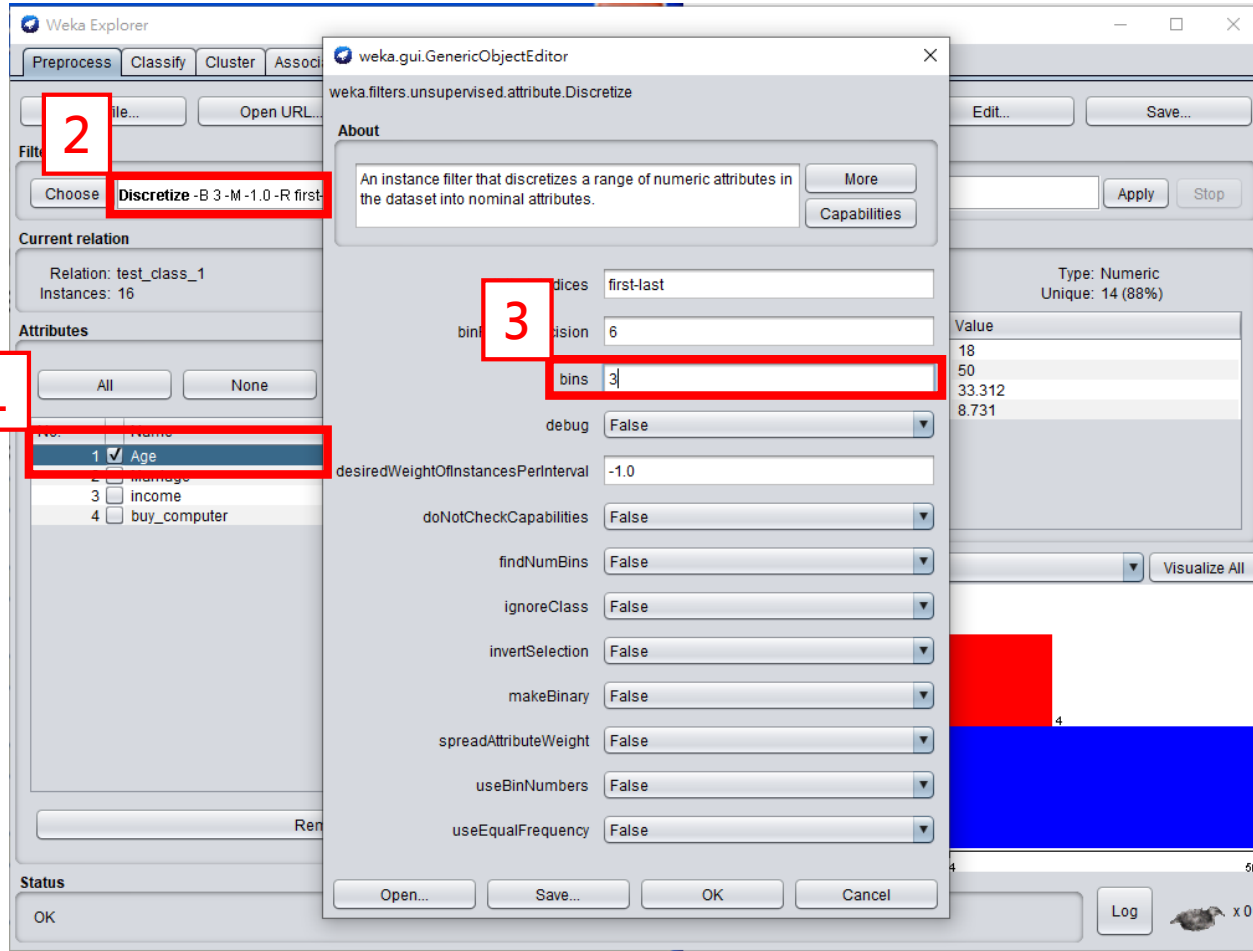
- 請依照下列階層找到 Discretize 方法
- weka
  - unsupervised
    - attribute
      - Discretize

# Step 6

1. 選取 Age

2. 點擊 Discretize 處

3. 將 bins 設定成 3



# Step 7

The screenshot shows the Weka Explorer window with the 'Preprocess' tab selected. The 'Filter' dropdown menu is open, and 'Discretize -B 3 -M -1.0 -R first-last-precision 6' is highlighted in a red box. The 'Current relation' is 'test\_class\_1' with 16 instances. The 'Attributes' list on the left shows 'Age' selected with a checkmark. The 'Selected attribute' panel on the right displays statistics for 'Age': Minimum 18, Maximum 50, Mean 33.312, and StdDev 8.731. Below this, a histogram for the 'buy\_computer' class is shown, with the x-axis representing age values from 18 to 50. The histogram has two bars: a blue bar from 18 to 34 and a red bar from 34 to 50. The red bar is labeled with '7' above it and '4' to its right. The blue bar is labeled with '5' above it. The 'Status' bar at the bottom shows 'OK' and a 'Log' button.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter

Choose **Discretize -B 3 -M -1.0 -R first-last-precision 6** Apply Stop

Current relation

Relation: test\_class\_1  
Instances: 16  
Attributes: 4  
Sum of weights: 16

Attributes

All None Invert Pattern

No.	Name
1	<input checked="" type="checkbox"/> Age
2	<input type="checkbox"/> Marriage
3	<input type="checkbox"/> income
4	<input type="checkbox"/> buy_computer

Remove

Selected attribute

Name: Age  
Missing: 0 (0%)  
Distinct: 15  
Type: Numeric  
Unique: 14 (88%)

Statistic	Value
Minimum	18
Maximum	50
Mean	33.312
StdDev	8.731

Class: buy\_computer (Nom) Visualize All

5 7 4

18 34 50

Status

OK Log x 0

1. 確認

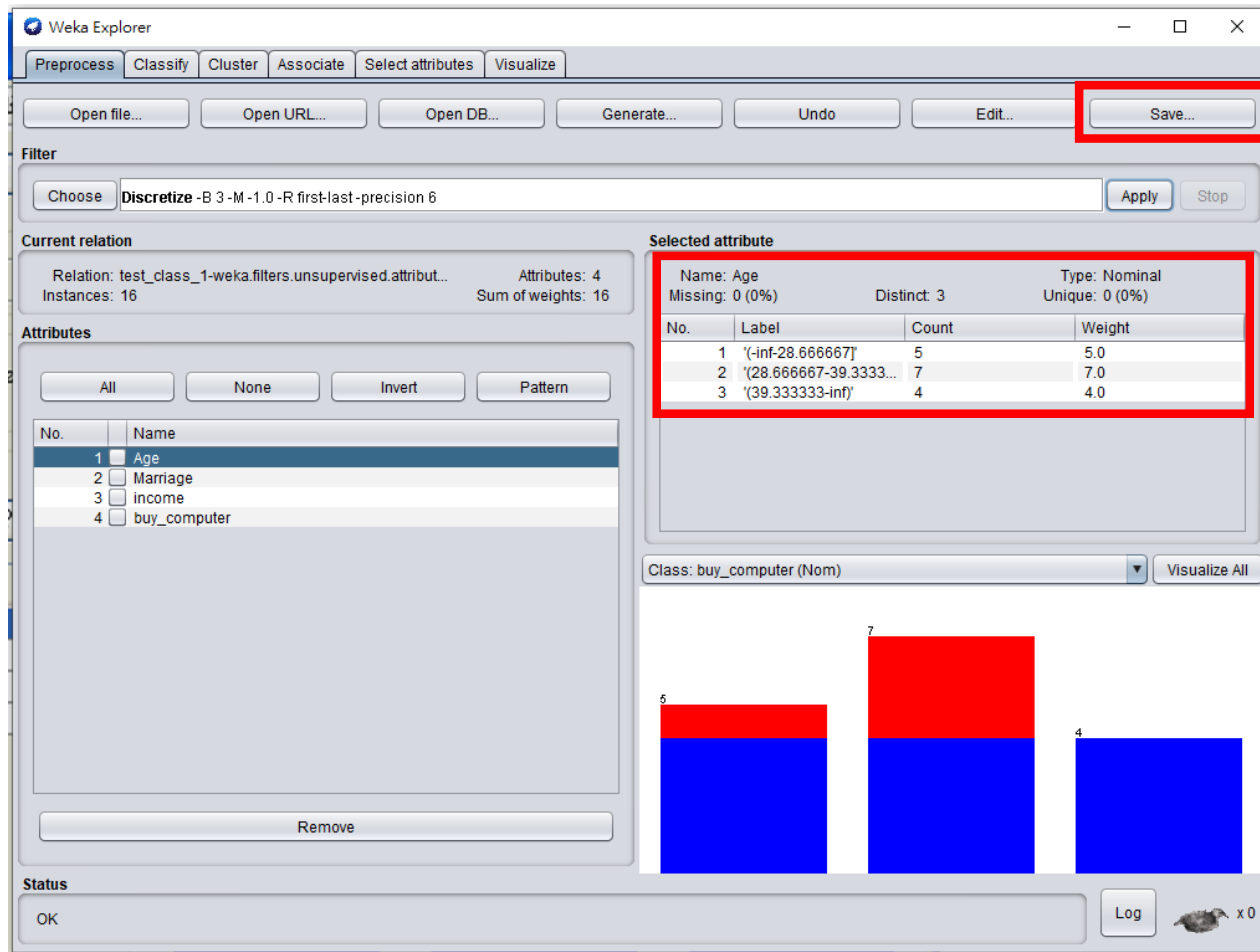
Discretize -B 3 -M -1.0 -R first-last-precision 6

2. 確認 Age 有勾選

3. 點擊 Apply



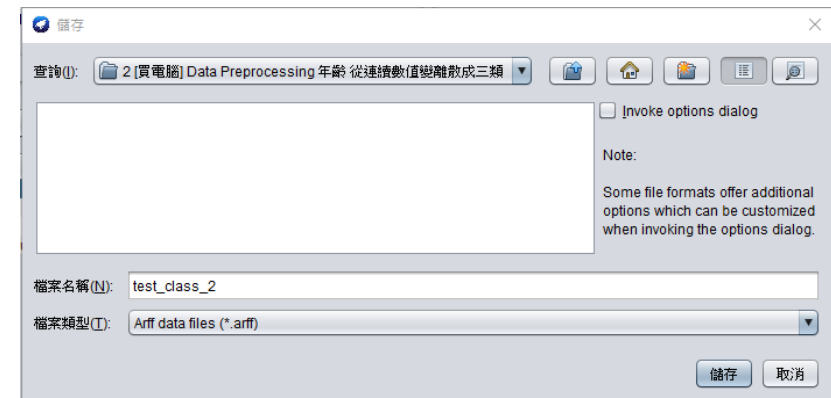
# Step 8



- 處理結果顯示出三個區間

1. -inf-28.666667
2. 28.66667-39.33333...
3. 39.333333-inf

- 點擊 Save 存檔



# Step 9

- 更換三個區間的表示法

1. 以  $\leq 29$  取代

`\'(-inf-28.666667]\'`

2. 以  $29 \sim \leq 39$  取代

`\'(28.666667-39.333333]\'`

3. 以  $> 39$  取代

`\'(39.333333-inf)\'`

```
1 @relation
2 test_class_1-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-Rfirst-last-precision6
3
4 @attribute Age {\'(-inf-28.666667]\',\'(28.666667-39.333333]\',\'(39.333333-inf)\'}
5 @attribute Marriage {single,Married}
6 @attribute income {High,mediate,low}
7 @attribute buy_computer {no,yes}
8
9 @data
10 \'(-inf-28.666667]\',single,High,no
11 \'(-inf-28.666667]\',single,mediate,no
12 \'(28.666667-39.333333]\',single,low,yes
13 \'(28.666667-39.333333]\',Married,mediate,no
14 \'(39.333333-inf)\',Married,low,no
15 \'(39.333333-inf)\',Married,low,no
16 \'(28.666667-39.333333]\',
17 \'(-inf-28.666667]\',Marr
18 \'(28.666667-39.333333]\',
19 \'(-inf-28.666667]\',Marr
20 \'(39.333333-inf)\',Marri
21 \'(28.666667-39.333333]\',
22 \'(39.333333-inf)\',Marri
23 \'(28.666667-39.333333]\',
24 \'(39.333333-inf)\',Marri
```

```
1 @relation
2 test_class_1-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-Rfirst-last-precision6
3
4 @attribute Age {<=29,29~<=39,>39}
5 @attribute Marriage {single,Married}
6 @attribute income {High,mediate,low}
7 @attribute buy_computer {no,yes}
8
9 @data
10 <=29,single,High,no
11 <=29,single,mediate,no
12 29~<=39,single,low,yes
13 29~<=39,Married,mediate,no
14 >39,Married,low,no
15 >39,Married,low,no
16 29~<=39,Married,mediate,no
17 29~<=39,single,High,no
18 <=29,Married,low,no
19 29~<=39,Married,mediate,no
20 <=29,Married,High,yes
21 >39,Married,mediate,no
22 29~<=39,single,mediate,yes
23 >39,Married,low,no
24 29~<=39,single,mediate,yes
25 <=29,single,low,no
```