

DR. ALVIN'S PUBLICATIONS

# DECISION TREE (CLASSIFICATION)

---

USING WEKA  
DR. ALVIN ANG



---

1 | PAGE

COPYRIGHTED BY DR ALVIN ANG  
WWW.ALVINANG.SG

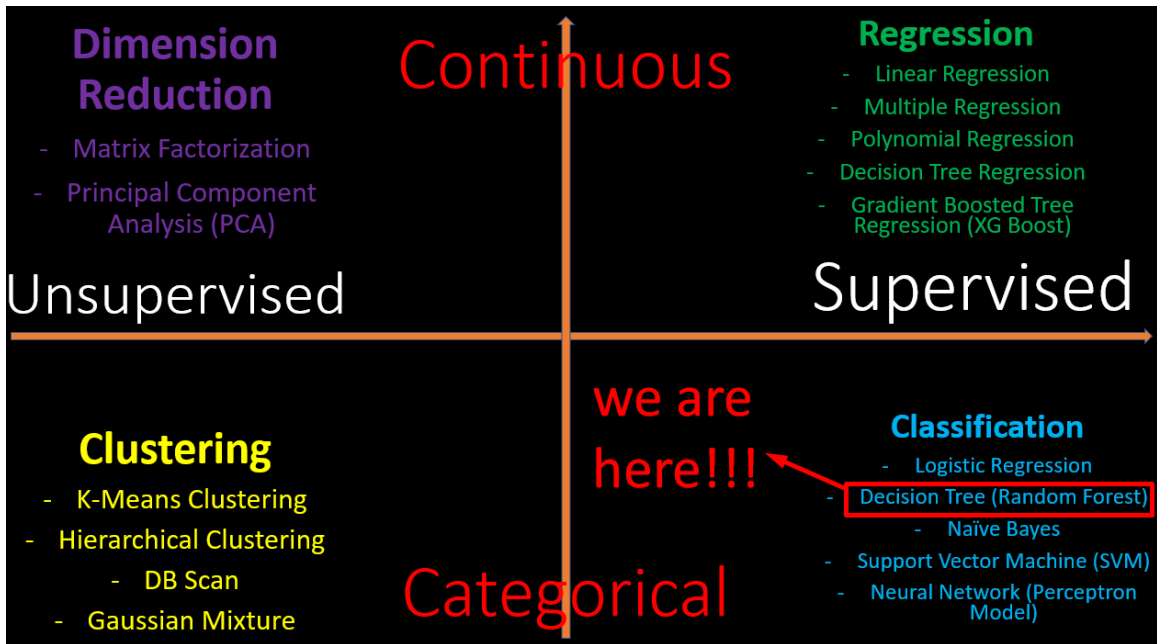
# CONTENTS

<b>I. Decision Tree (Classification) – Concept</b> .....	<b>3</b>
<b>A. Step 1: Calculate Entropy for Target (Insurance Premium)</b> .....	<b>6</b>
1. What is Entropy & Information Gain? .....	7
2. Back to our Example.... Calculate Entropy (Target) .....	9
<b>B. Step 2: Calculate Entropy of Target   Feature</b> .....	<b>10</b>
1. Entropy (Target, Smoker) .....	10
2. Entropy (Target, Age Group) .....	11
3. Entropy (Target, Medical Condition) .....	11
4. Entropy (Target, Salary Level) .....	11
<b>C. Step 3: Calculating Information Gain (IG)</b> .....	<b>12</b>
1. IG (Smoker) .....	12
2. IG (Age Group) .....	12
3. IG (Medical Condition) .....	12
4. IG (Salary Level) .....	13
<b>D. Step 4: Drawing the Decision Tree (Root Node)</b> .....	<b>13</b>
<b>E. Step 5: Which Node Next?</b> .....	<b>14</b>
1. Let’s take a look at Teenagers .....	15
2. Let’s take a look at Young .....	15
3. Finally .....	16
<b>II. WEKA for Decision Tree (Classification)</b> .....	<b>17</b>
<b>A. Step 1: Install Weka</b> .....	<b>17</b>
<b>B. Step 2: Bring in the Dataset</b> .....	<b>18</b>
<b>C. Step 3: Visualize the Dataset</b> .....	<b>19</b>
<b>D. Step 4: Choose Decision Tree Classifier</b> .....	<b>20</b>
<b>E. Step 5: Visualize the Tree</b> .....	<b>21</b>
<b>F. Step 6: Save the Trained Model</b> .....	<b>23</b>
<b>G. Step 7: Load the Saved Trained Model</b> .....	<b>24</b>
<b>H. Step 8: Using the Tree to Make a Prediction</b> .....	<b>26</b>
<b>III. Conclusion</b> .....	<b>31</b>
<b>About Dr. Alvin Ang</b> .....	<b>32</b>

---

I. DECISION TREE (CLASSIFICATION) – CONCEPT

---



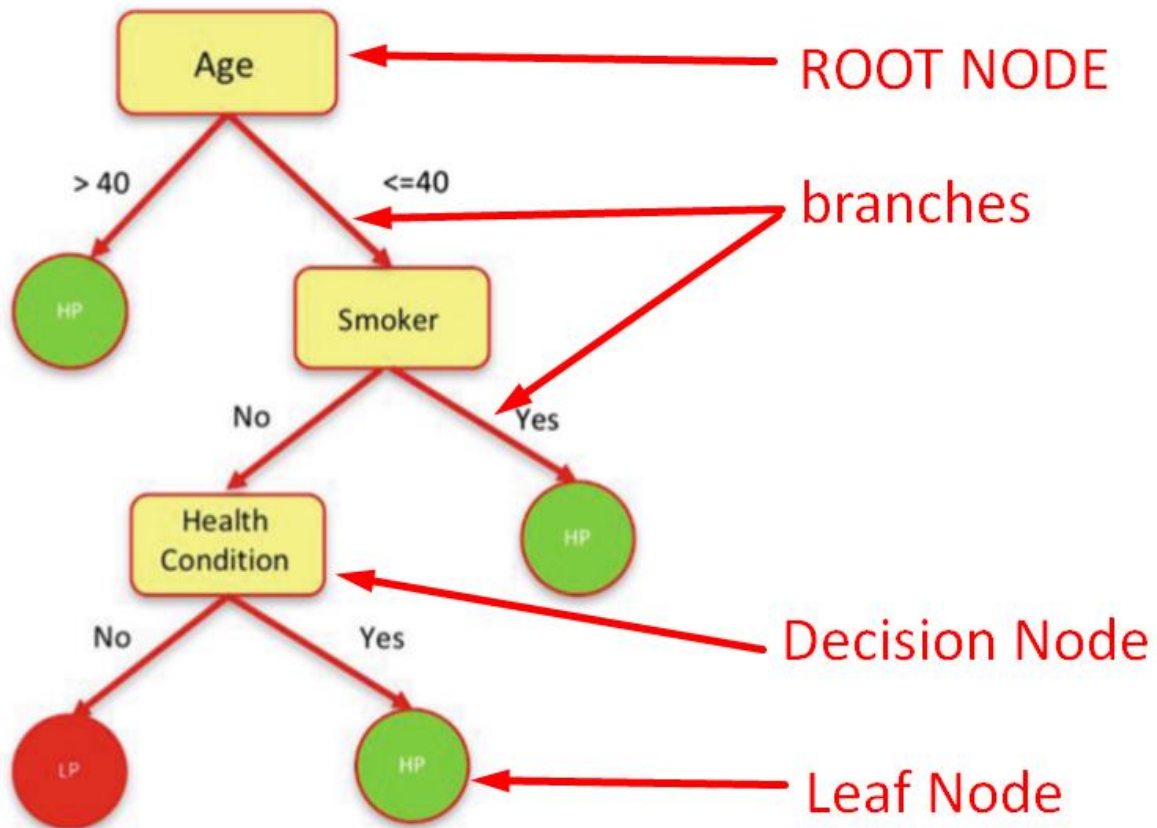
Let's say we are given the dataset below...

<https://www.alvinang.sg/s/Insurance-Premium-Datascsv.csv>

Predictors

Target

<b>Age Group</b>	<b>Smoker</b>	<b>Medical Condition</b>	<b>Salary Level</b>	<b>Insurance Premium</b>
Old	Yes	Yes	High	High
Teenager	Yes	Yes	Medium	High
Young	Yes	Yes	Medium	Low
Old	No	Yes	High	High
Young	Yes	Yes	High	Low
Teenager	No	Yes	Low	High
Teenager	No	No	Low	Low
Old	No	No	Low	High
Teenager	No	Yes	Medium	High
Young	No	Yes	Low	High
Young	Yes	No	High	Low
Teenager	Yes	No	Medium	Low
Young	No	No	Medium	High
Old	Yes	No	Medium	High



Question: How do we place the Root Node / Decision Nodes in the Decision Tree?

That is, how do we rank them in terms of importance?

How do we know that the Root Node (Age) is the most important Feature, followed by Smoker then Health Condition?

Features (in terms of importance)

1. Age
2. Smoker?
3. Medical Condition

Note: Salary is NOT an important feature! It is left out!

A. STEP 1: CALCULATE ENTROPY FOR TARGET (INSURANCE PREMIUM)

$$E = - \sum_{i=1}^N p_i \log_2 p_i$$

(equation for Entropy of Target... don't really go bother about it...is not that important...)

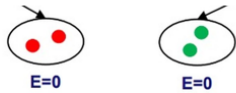
Note:

For more information, visit <https://towardsdatascience.com/entropy-how-decision-trees-make-decisions-2946b9c18c8>

# Entropy Measures Impurity or Disorder Ness!!!

## PURE

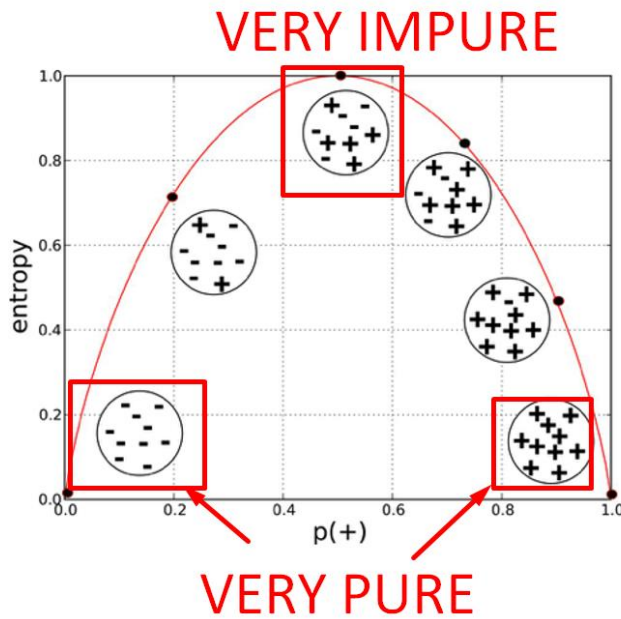
- Pure means cannot split anymore.
- Pure means Entropy = 0.
- Pure means Homogeneous.
- Pure means Leaf Node (Terminal Node).



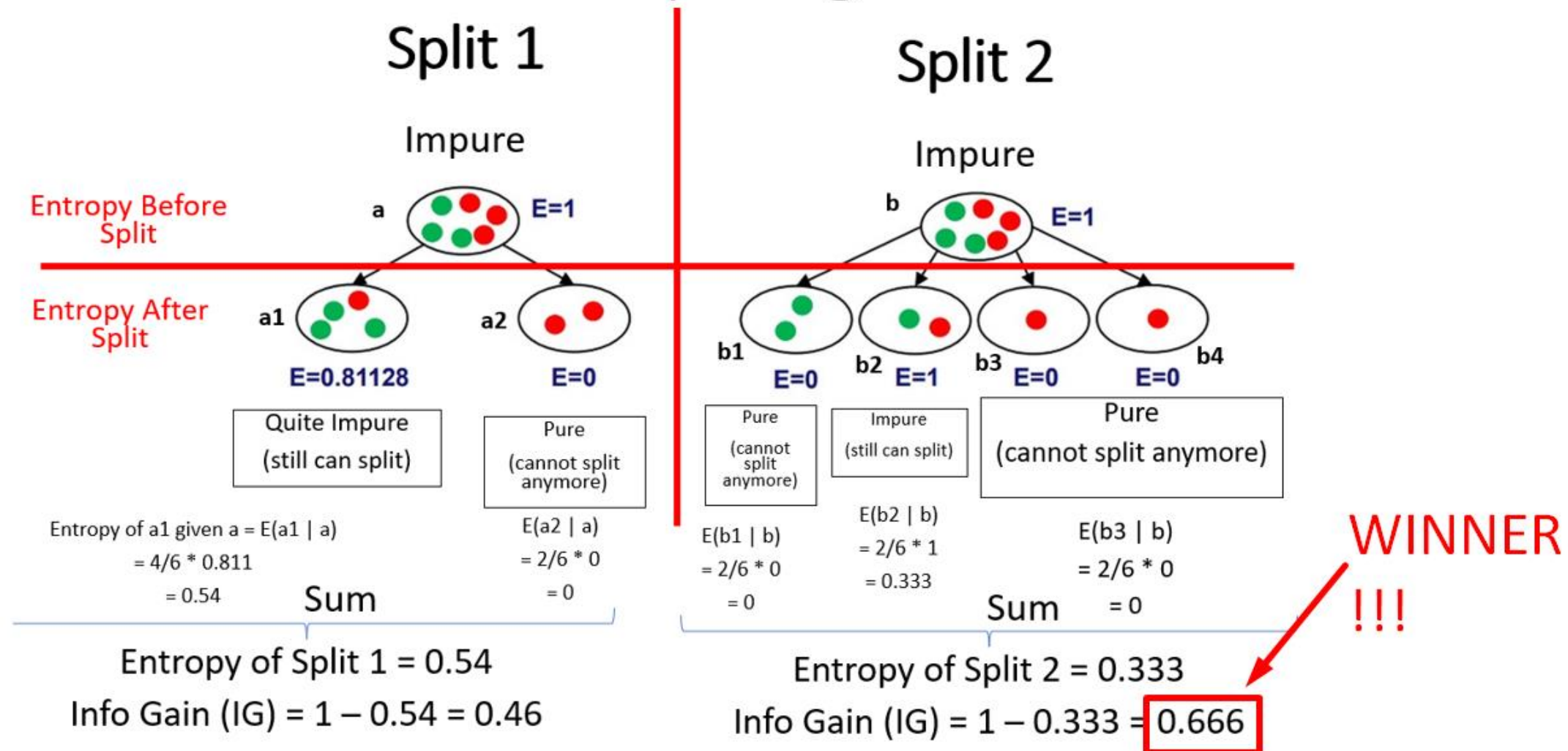
PURE

## IMPURE

- Impure means still can split.
- Impure means Entropy > 0.
- Impure means NOT Homogeneous,
- Impure means Decision Node.



# Comparing



- Say we have 2 ways to split... should we choose Split 1 or 2? Which is better?
  - Answer: Depends on the Information Gain (IG).
  - Split 2 has a higher IG than Split 1, thus we choose Split 2.
- Thus, you can see that the purpose of getting the Entropy is to help us get the IG.
- Because the IG is the change in Entropy (of the Original Node) BEFORE and AFTER splitting.
  - In other words, IG measures how much Homogeneity remains after splitting!
    - More Homogeneity (lower IG) means you can't split much later on.
    - Lower Homogeneity (higher IG) means you can split more later on!
  - Thus, the Feature with the HIGHEST IG will be the TOP Decision Node
- Because it means that there's a lot of impurity which can be split further down the tree!



2. BACK TO OUR EXAMPLE.... CALCULATE ENTROPY (TARGET)

Predictors

Target

Age Group	Smoker	Medical Condition	Salary Level	Insurance Premium
Old	Yes	Yes	High	High
Teenager	Yes	Yes	Medium	High
Young	Yes	Yes	Medium	Low
Old	No	Yes	High	High
Young	Yes	Yes	High	Low
Teenager	No	Yes	Low	High
Teenager	No	No	Low	Low
Old	No	No	Low	High
Teenager	No	Yes	Medium	High
Young	No	Yes	Low	High
Young	Yes	No	High	Low
Teenager	Yes	No	Medium	Low
Young	No	No	Medium	High
Old	Yes	No	Medium	High

Insurance Premium	
High	Low
9 cases	5 cases
Probability 9 / 14 = 0.64	5 / 14 = 0.36

$$\begin{aligned}
 & \text{Entropy}(\text{Insurance Premium}) \\
 &= \text{Entropy}(9, 5) \\
 &= -(0.64 \log_2 0.64) - (0.36 \log_2 0.36) \\
 &= 0.94
 \end{aligned}$$

**B. STEP 2: CALCULATE ENTROPY OF TARGET | FEATURE**

1. ENTROPY (TARGET, SMOKER)

		Insurance Premium (Target)		
		High	Low	
Smoker	Yes (7)	3 cases	4 cases	Probability (Yes) = 7/14 = 0.5
(Feature)	No (7)	6 cases	1 case	Probability (No) = 7/14 = 0.5

Probability (High | Yes) = 3/7      Probability (Low | Yes) = 4/7  
 Probability (High | No) = 6/7      Probability (Low | No) = 1/7

$$Entropy_{yes} = -\frac{3}{7} \left[ \log_2 \left( \frac{3}{7} \right) \right] - \frac{4}{7} \left[ \log_2 \left( \frac{4}{7} \right) \right] = 0.99$$

$$Entropy_{no} = -\frac{6}{7} \left[ \log_2 \left( \frac{6}{7} \right) \right] - \frac{1}{7} \left[ \log_2 \left( \frac{1}{7} \right) \right] = 0.59$$

$$\begin{aligned}
 Entropy_{(Target, Smoker)} &= [P_{yes} * Entropy_{yes}] + [P_{no} * Entropy_{no}] \\
 &= (0.5 * 0.99) + (0.5 * 0.59) \\
 &= 0.79
 \end{aligned}$$

2. ENTROPY (TARGET, AGE GROUP)

We repeat the steps above and calculate the ENTROPY for the rest of the other features...

$$\text{Entropy}_{(Target, Age Group)} = 0.69$$

3. ENTROPY (TARGET, MEDICAL CONDITION)

$$\text{Entropy}_{(Target, Medical Condition)} = 0.89$$

4. ENTROPY (TARGET, SALARY LEVEL)

$$\text{Entropy}_{(Target, Salary Level)} = 0.91$$

C. STEP 3: CALCULATING INFORMATION GAIN (IG)

$$\text{Information Gain} = \text{Entropy}_{\text{Target}} - \text{Entropy}_{\text{Target, Features}}$$

1. IG (SMOKER)

$$\begin{aligned} IG_{\text{Smoker}} &= \text{Entropy}_{\text{Target}} - \text{Entropy}_{\text{Target, Smoker}} \\ &= 0.94 - 0.79 \\ &= 0.15 \end{aligned}$$

2. IG (AGE GROUP)

$$\begin{aligned} IG_{\text{Age Group}} &= \text{Entropy}_{\text{Target}} - \text{Entropy}_{\text{Target, Age Group}} \\ &= 0.94 - 0.69 \\ &= 0.25 \end{aligned}$$

3. IG (MEDICAL CONDITION)

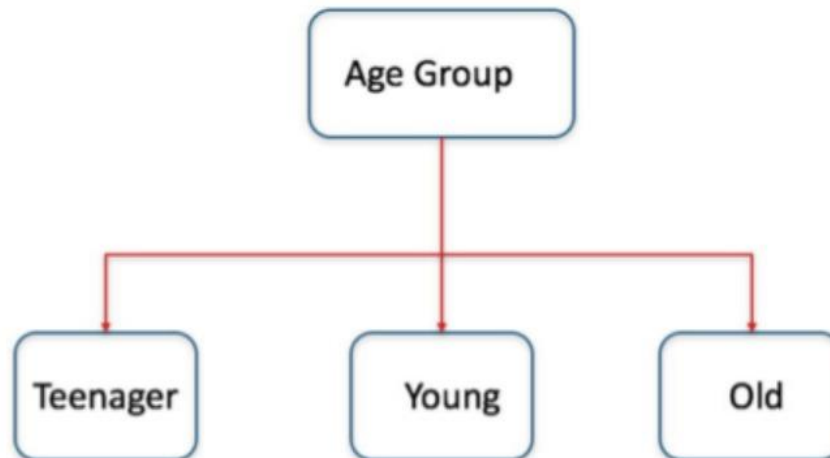
$$\begin{aligned} IG_{\text{Medical Condition}} &= \text{Entropy}_{\text{Target}} - \text{Entropy}_{\text{Target, Medical Condition}} \\ &= 0.94 - 0.89 \\ &= 0.05 \end{aligned}$$

4. IG (SALARY LEVEL)

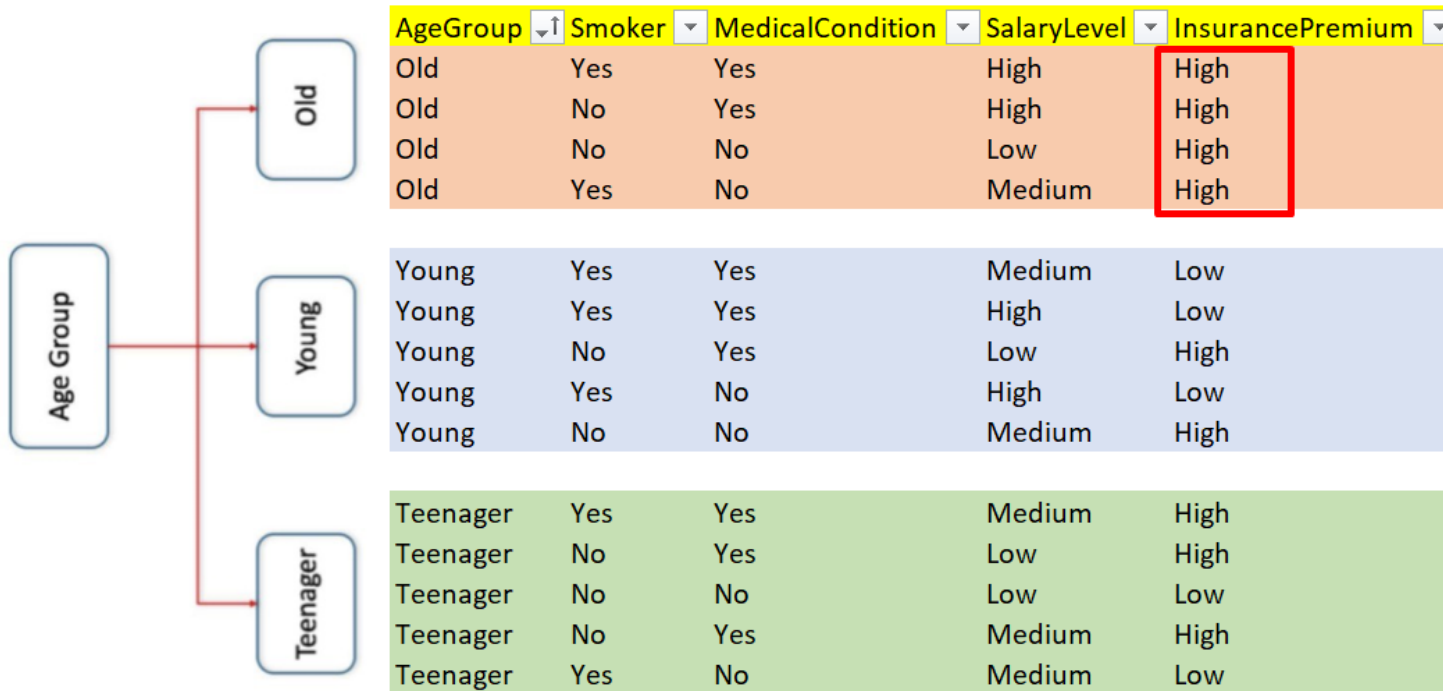
$$\begin{aligned}IG_{Salary\ Level} &= Entropy_{Target} - Entropy_{Target, Salary\ Level} \\ &= 0.94 - 0.91 \\ &= 0.03\end{aligned}$$

- Highest IG = Age Group = 0.25
- Thus, ROOT NODE = AGE GROUP

D. STEP 4: DRAWING THE DECISION TREE (ROOT NODE)

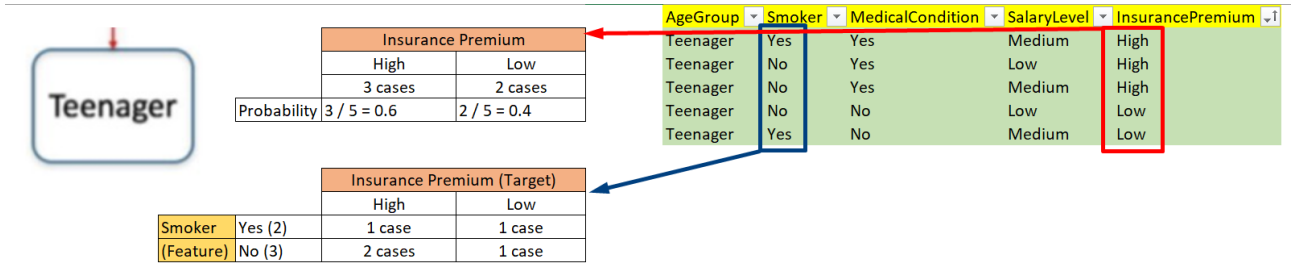


E. STEP 5: WHICH NODE NEXT?



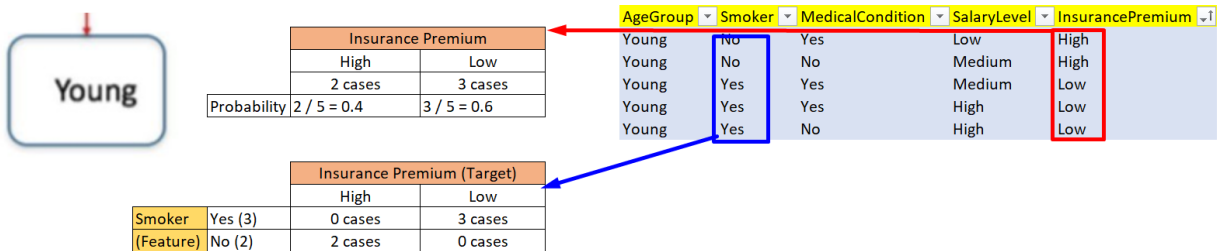
All High  
which  
means  
OLD is a  
Terminal  
Node  
(Leaf Node)

1. LET'S TAKE A LOOK AT TEENAGERS...



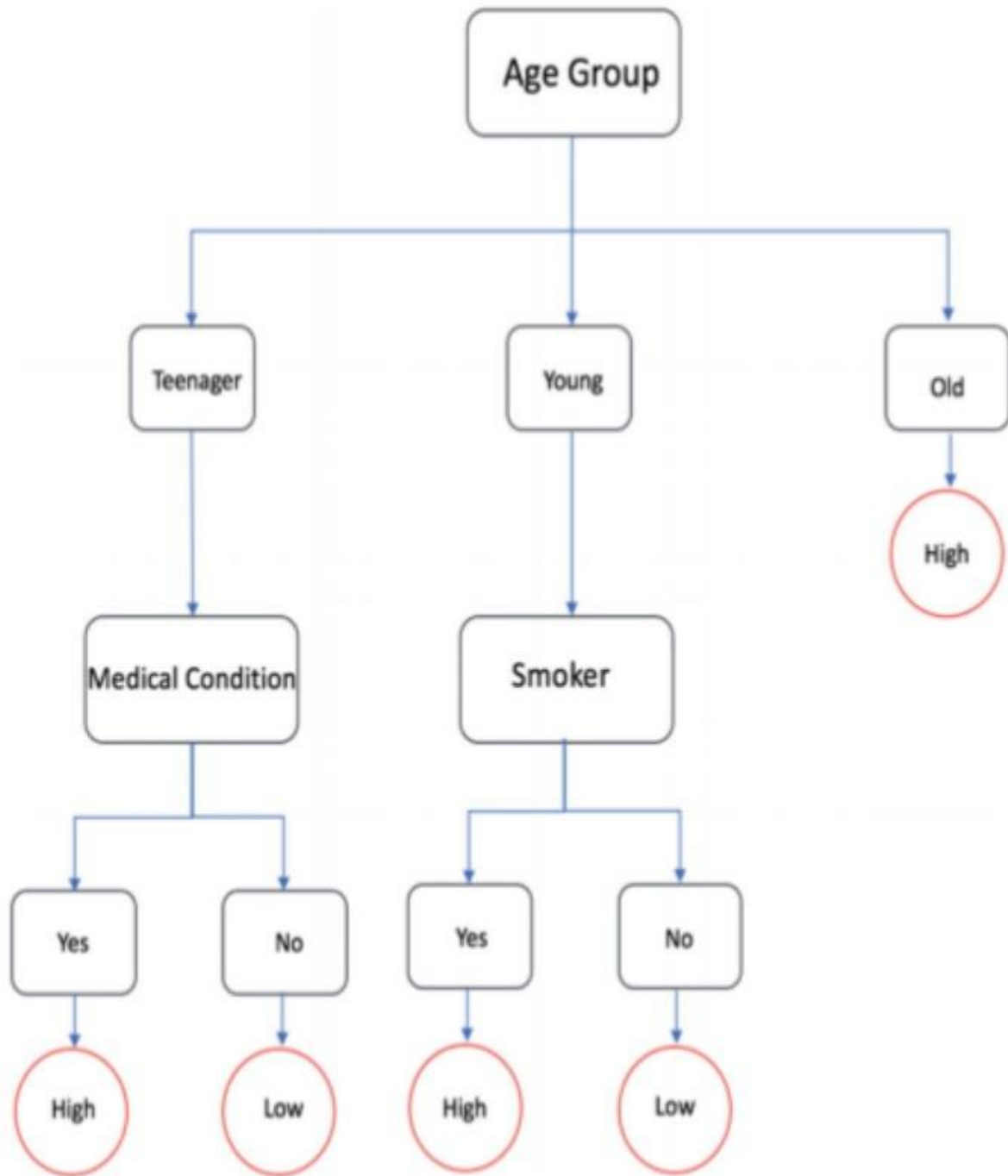
- We can now go through the entire Step 1 to 3 for Entropy / IG calculation for the Teenager Node.
- Objective to find out which is the next important feature that comes after Teenager?
  - Is it Smoker?
  - Medical Condition?
  - Salary Level?

2. LET'S TAKE A LOOK AT YOUNG...



- We can now go through the entire Step 1 to 3 for Entropy / IG calculation for the Teenager Node.
- Objective to find out which is the next important feature that comes after Teenager?
  - Is it Smoker?
  - Medical Condition?
  - Salary Level?

3. FINALLY.....





---

## II. WEKA FOR DECISION TREE (CLASSIFICATION)

---

<https://www.cs.waikato.ac.nz/ml/weka/>

<https://www.alvinang.sg/s/Insurance-Premium-Datascv.csv>

### A. STEP 1: INSTALL WEKA

#### Weka 3: Machine Learning Software in Java

Weka is a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization.

Found only on the islands of New Zealand, the Weka is a flightless bird with an inquisitive nature. The name is pronounced like this, and the bird sounds like this.

Weka is open source software issued under the **GNU General Public License**.

We have put together several **free online courses** that teach machine learning and data mining using Weka. The videos for the courses are available on **Youtube**.

Weka supports **deep learning!**

#### Getting started

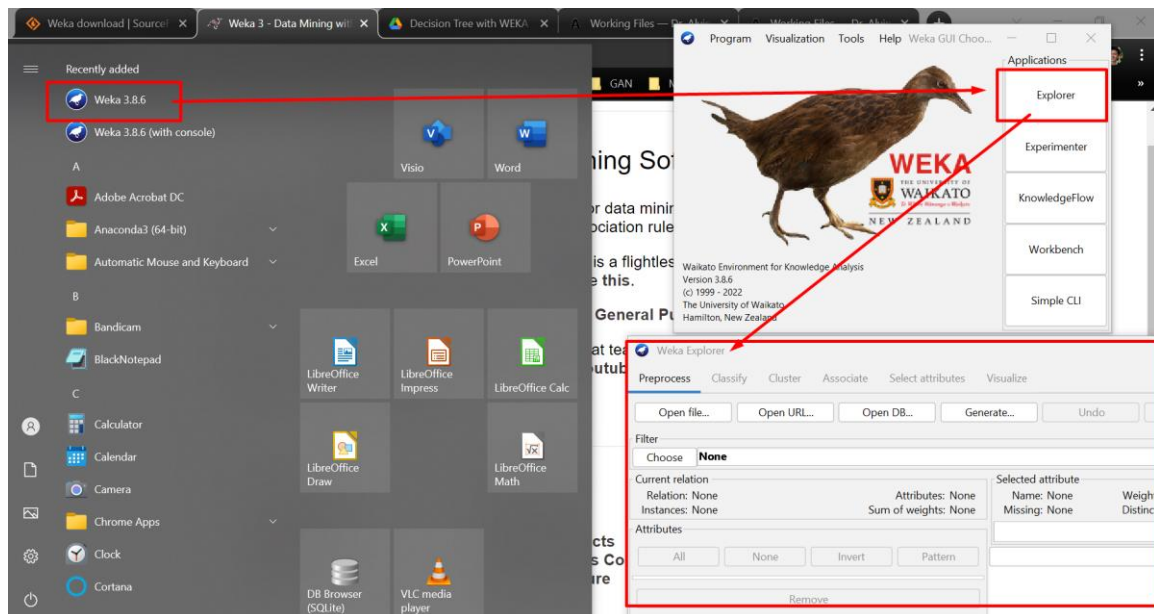
- Requirements
- **Download**
- Documentation
- FAQ
- Getting Help

#### Further information

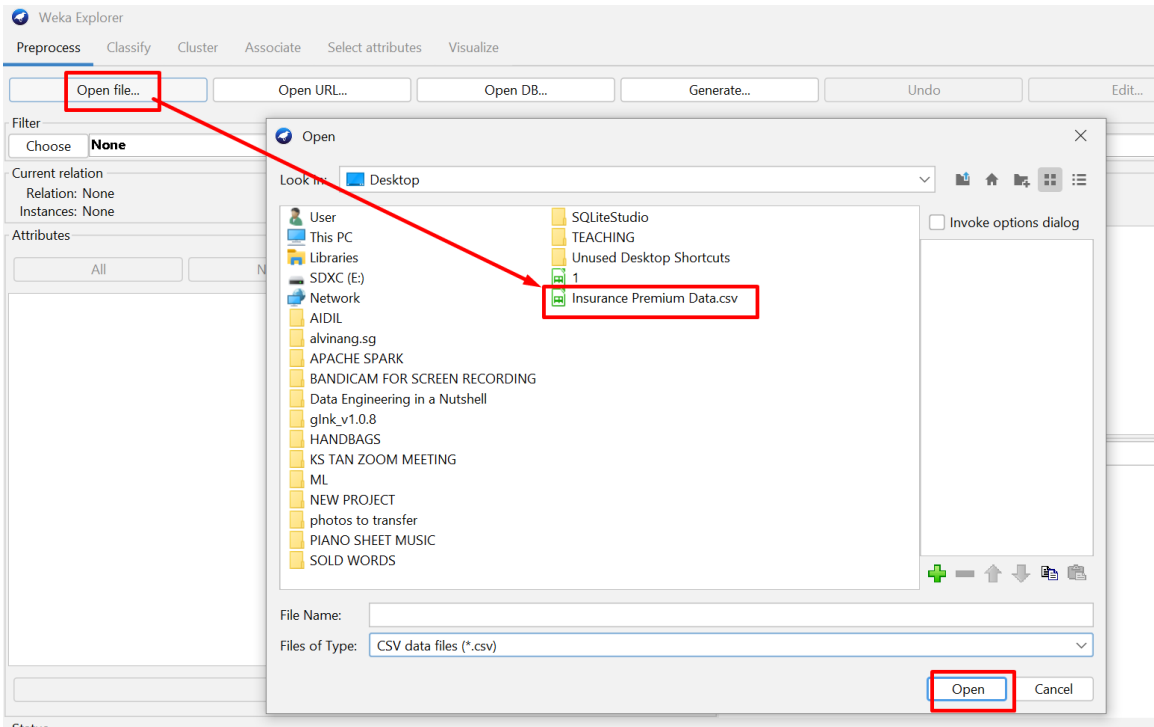
- Citing Weka
- Datasets
- Related Projects
- Miscellaneous Code
- Other Literature

#### Developers

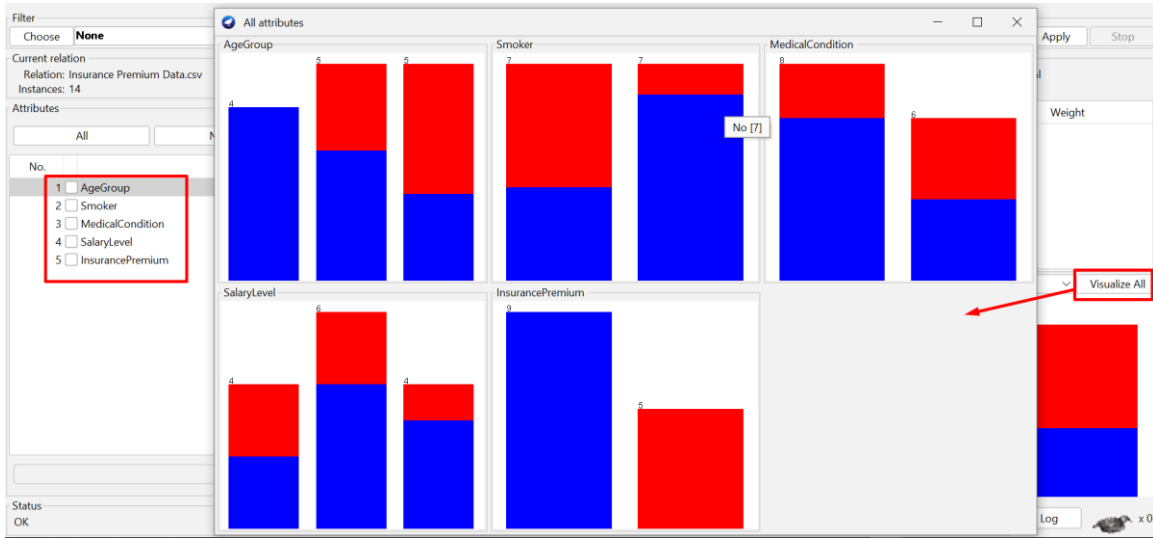
- Development
- History
- Subversion
- Contributors
- Commercial licenses



## B. STEP 2: BRING IN THE DATASET



### C. STEP 3: VISUALIZE THE DATASET

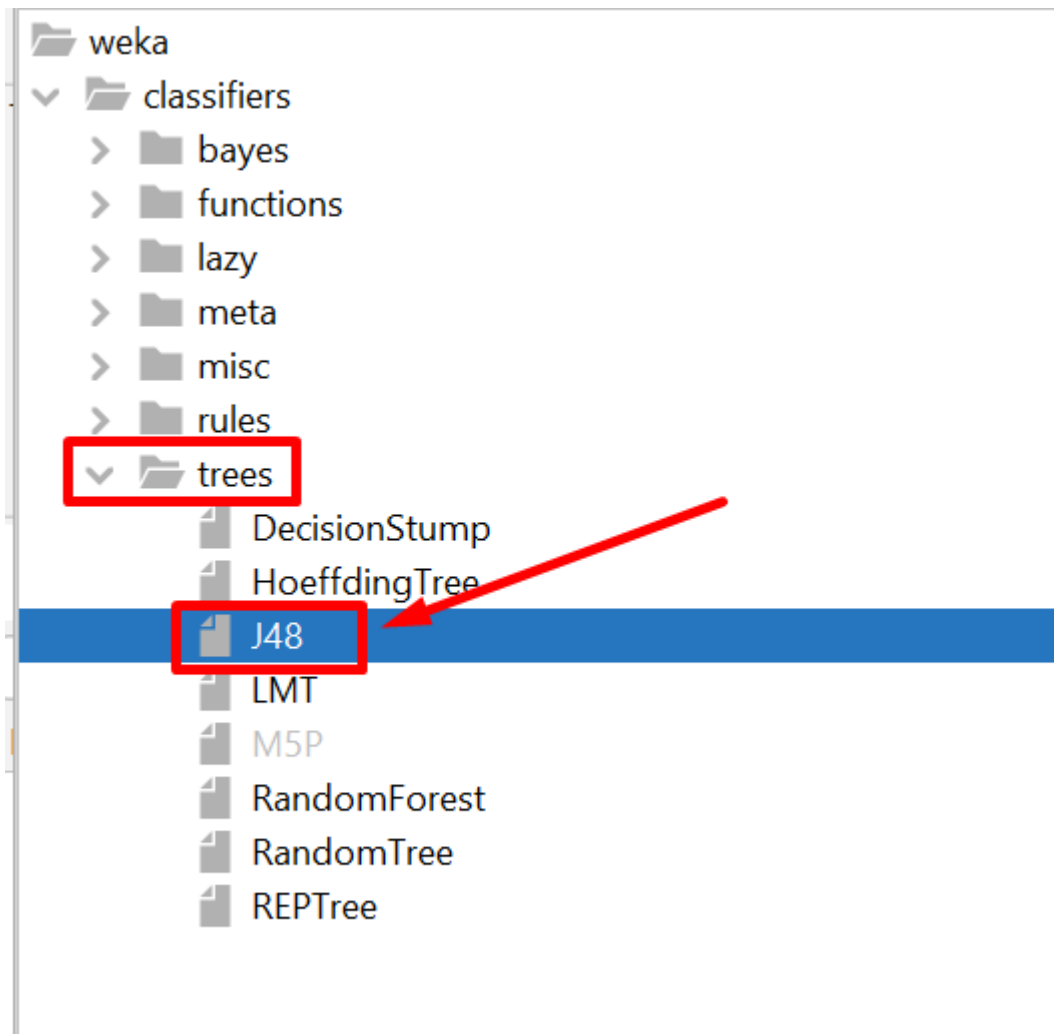
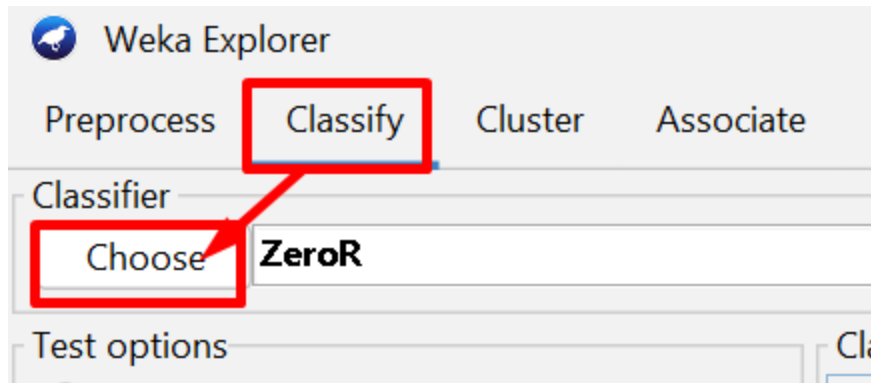


Color code:

- Red = Low Insurance Premium Cases
- Blue = High Insurance Premium Cases
- E.g. you see the AgeGroup has 3 bars....
  - 4 OLD = All Blue (coz 4 Low Premium)
  - 5 Young = 3 Blue 2 Red
  - 5 Teenagers = 3 Red 2 Blue

Age Group	AgeGroup	Smoker	MedicalCondition	SalaryLevel	InsurancePremium
Old	Old	Yes	Yes	High	High
	Old	No	Yes	High	High
	Old	No	No	Low	High
	Old	Yes	No	Medium	High
Young	Young	Yes	Yes	Medium	Low
	Young	Yes	Yes	High	Low
	Young	No	Yes	Low	High
	Young	Yes	No	High	Low
	Young	No	No	Medium	High
Teenager	Teenager	Yes	Yes	Medium	High
	Teenager	No	Yes	Low	High
	Teenager	No	No	Low	Low
	Teenager	No	Yes	Medium	High
	Teenager	Yes	No	Medium	Low
	Teenager	Yes	No	Medium	Low

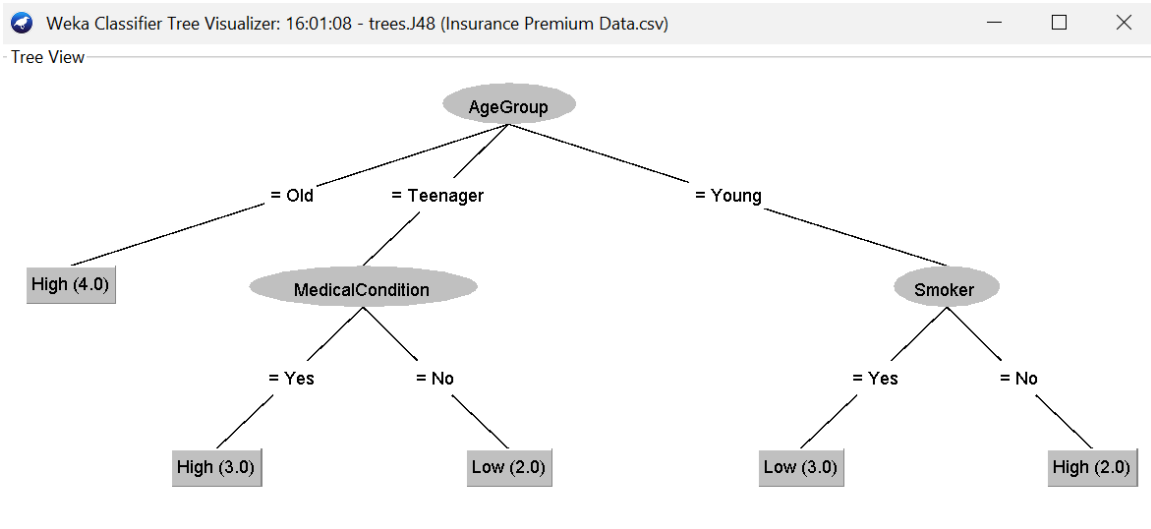
D. STEP 4: CHOOSE DECISION TREE CLASSIFIER

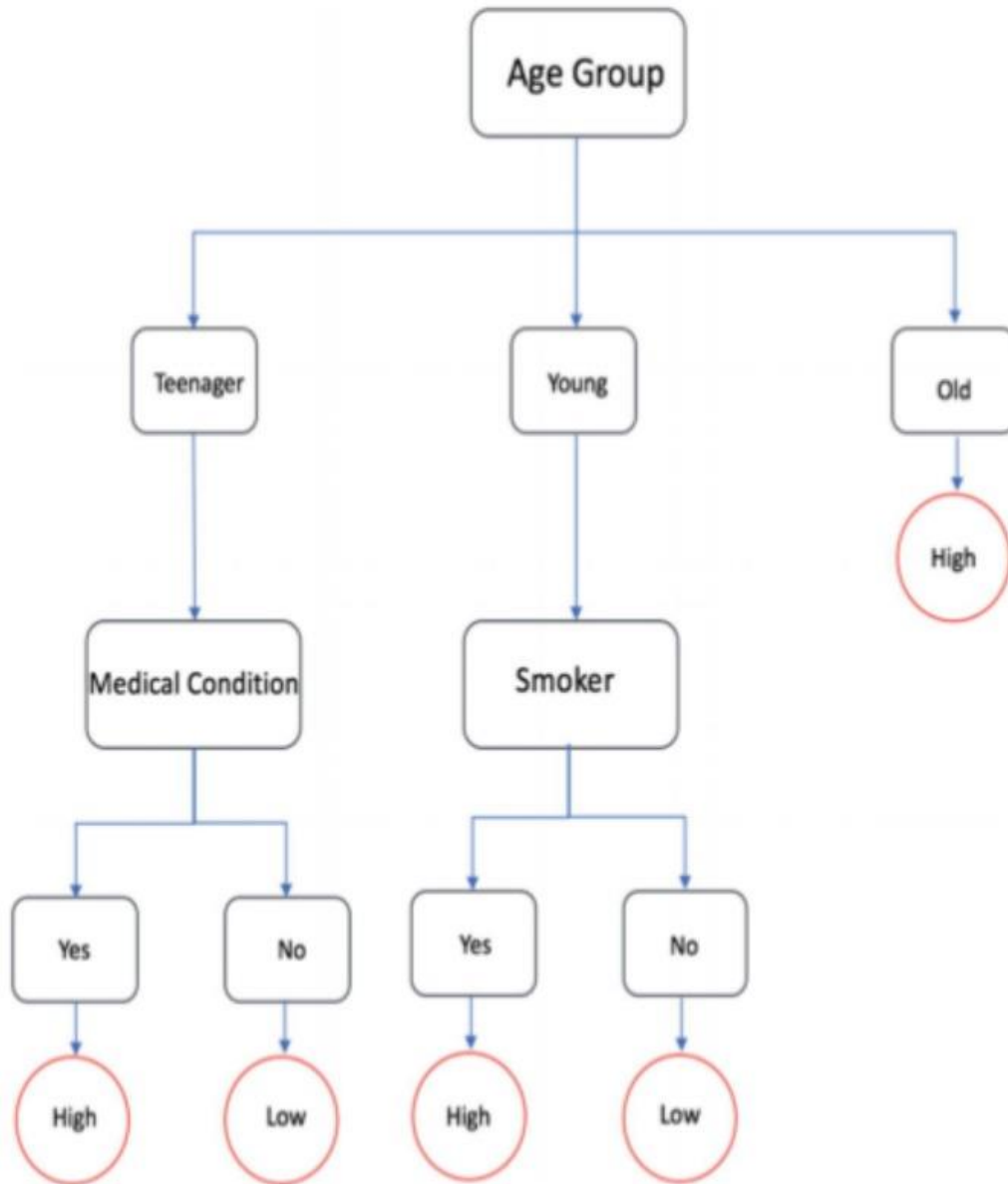


### E. STEP 5: VISUALIZE THE TREE

The screenshot shows the Weka Classifier window with the 'trees.J48' classifier selected. The 'Test options' section has 'Cross-validation' selected with 10 folds. The 'Classifier output' section shows 'Smoker = No: High (2.0)'. A context menu is open over the 'Result list' (right-click for options), with 'Visualize tree' highlighted. A red arrow points to the 'Start' button, and another red arrow points to the 'Visualize tree' option in the menu. The 'Result list' contains a table of performance metrics:

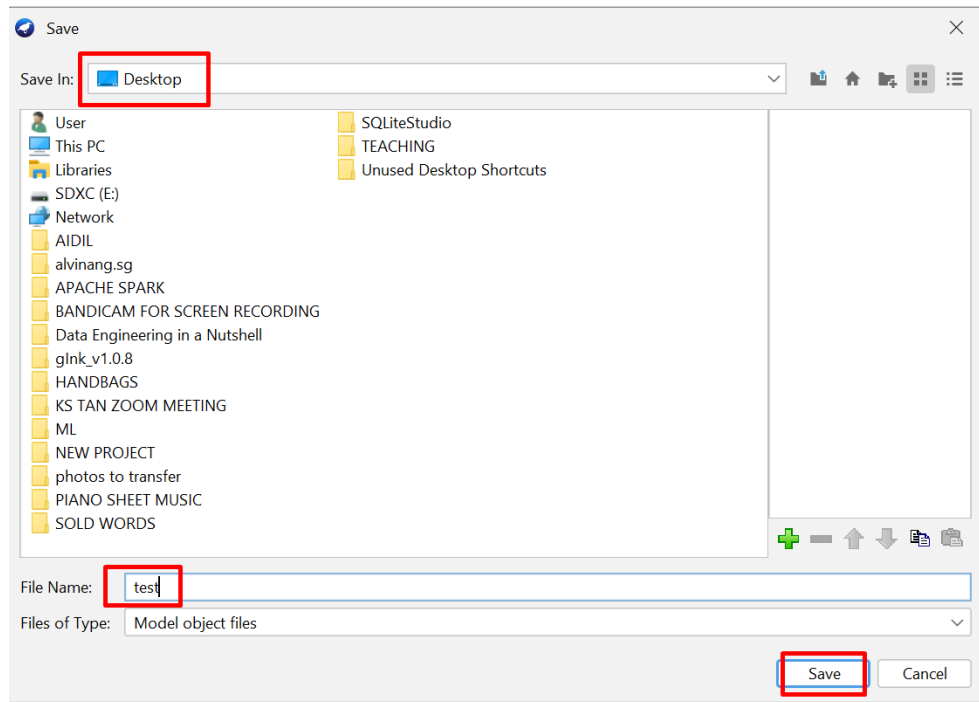
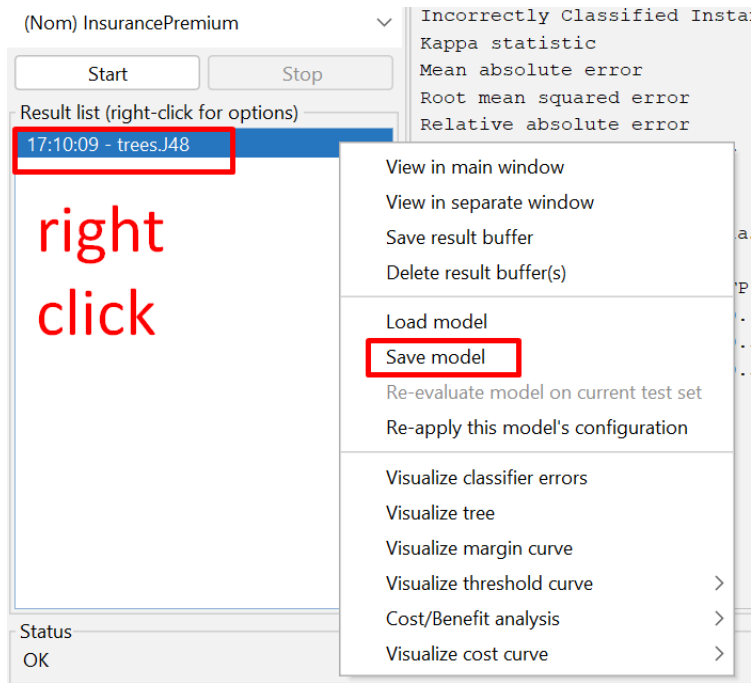
Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
00	0.571	0.444	0.500	-0.149	0.511	0.702	High
56	0.286	0.400	0.333	-0.149	0.511	0.359	Low
84	0.469	0.429	0.440	-0.149	0.511	0.580	



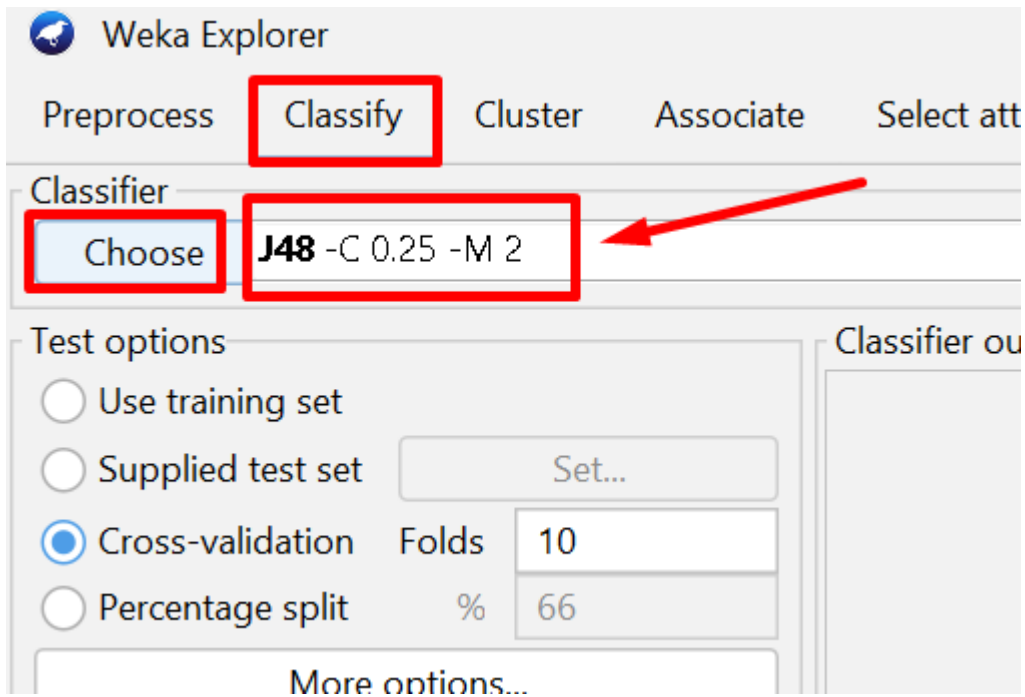
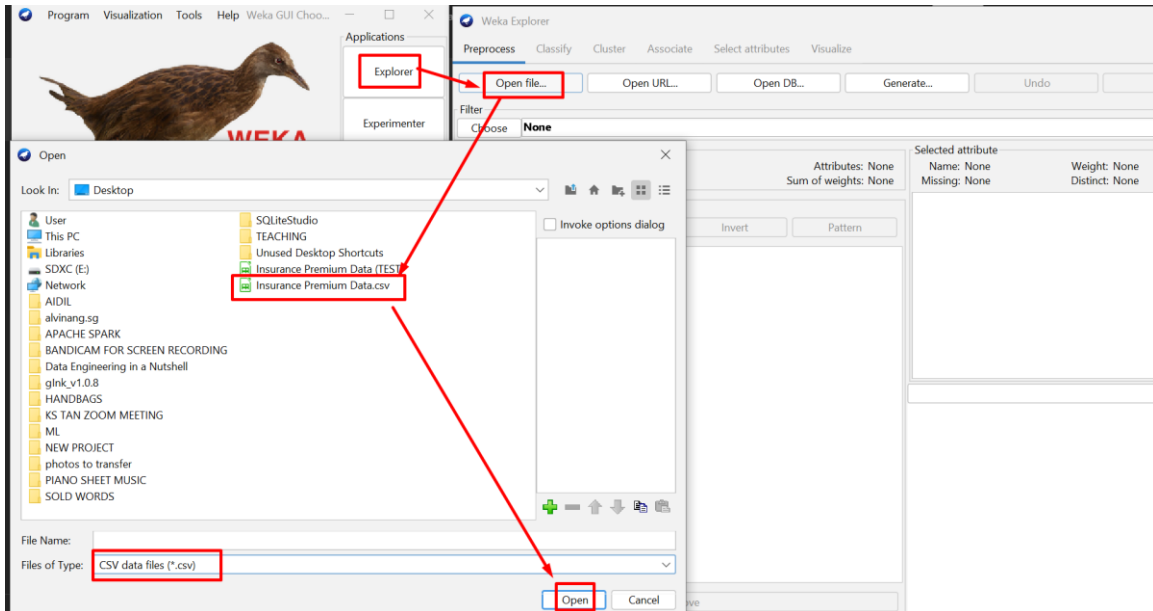


It's the same Tree as we drew earlier!

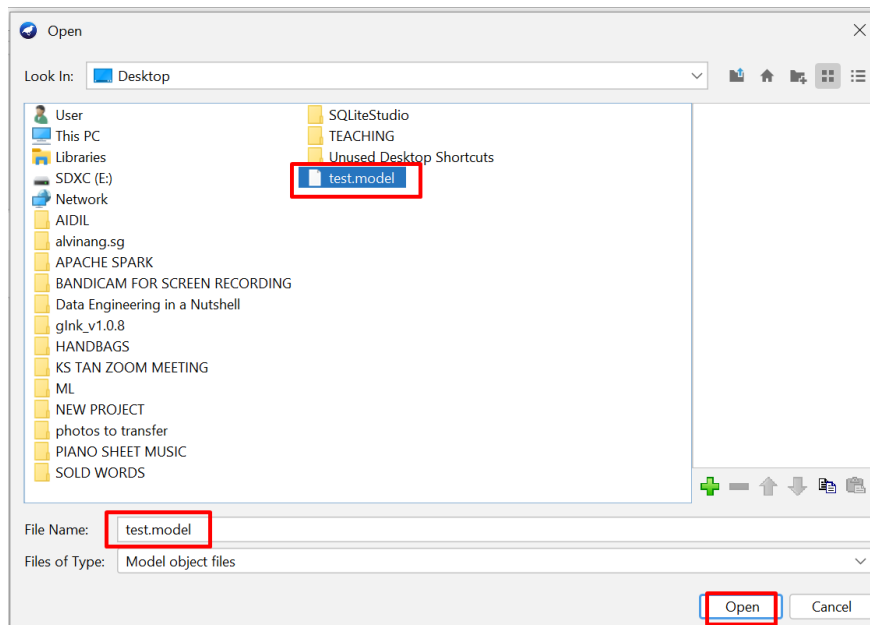
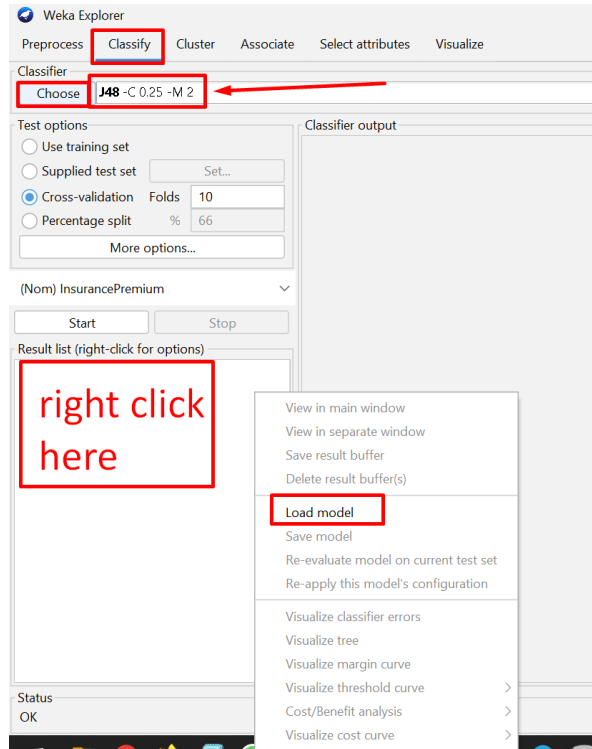
## F. STEP 6: SAVE THE TRAINED MODEL



### G. STEP 7: LOAD THE SAVED TRAINED MODEL







## H. STEP 8: USING THE TREE TO MAKE A PREDICTION

Go here and download the data again:

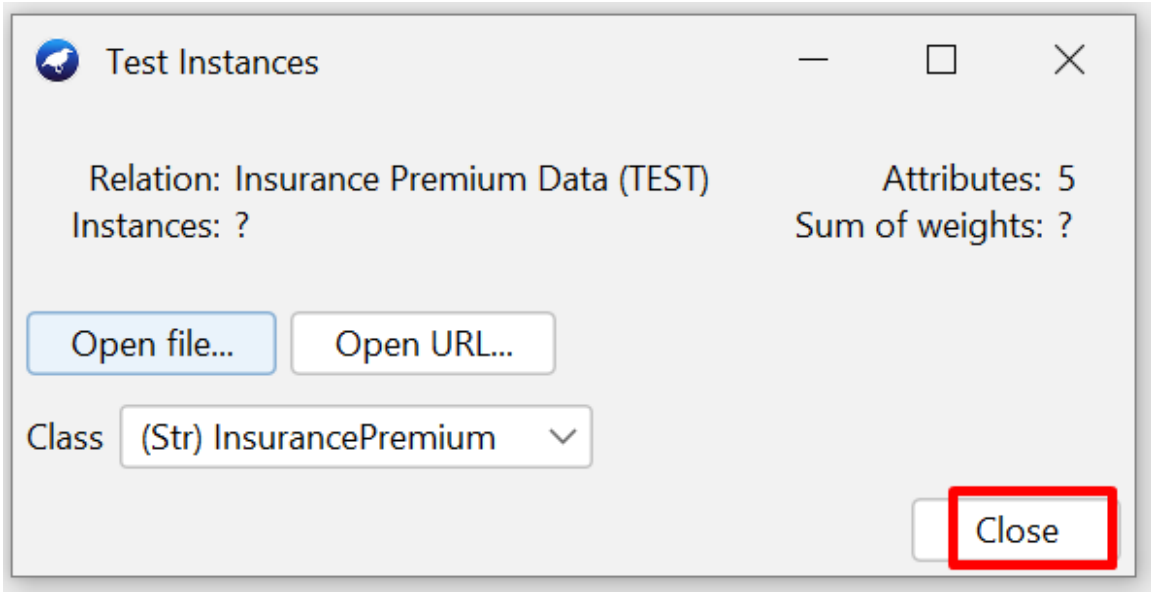
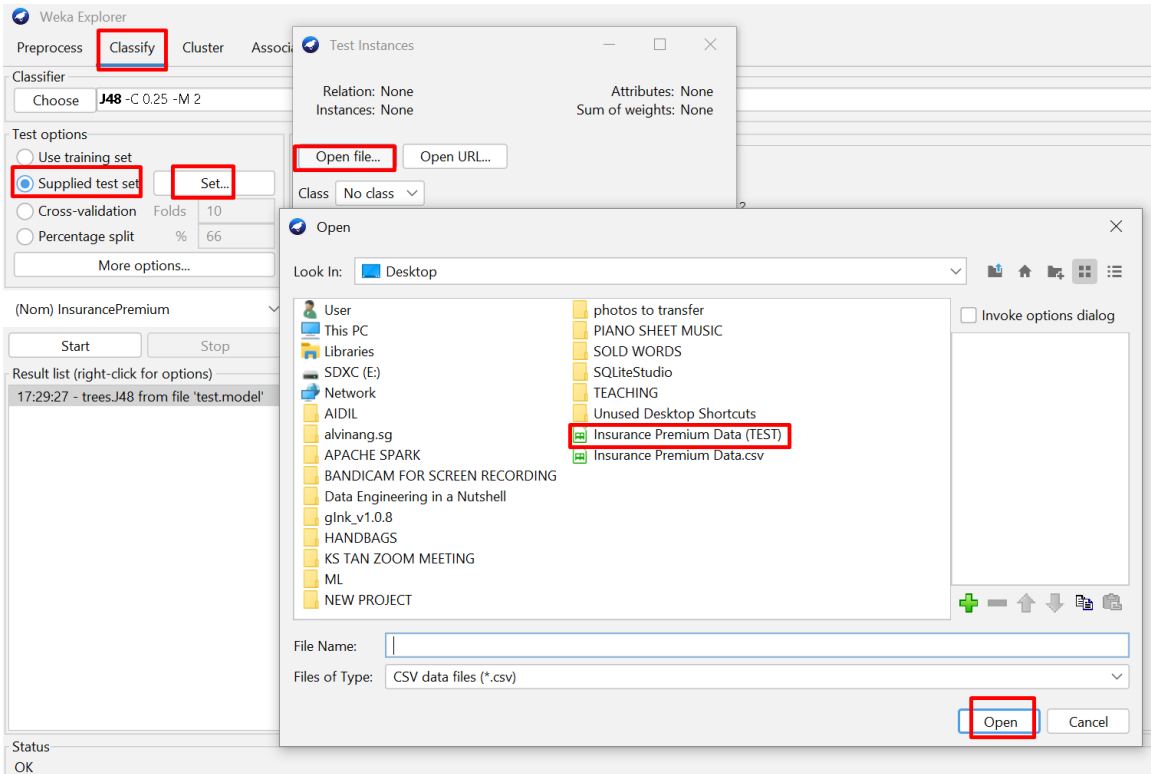
<https://www.alvinang.sg/s/Insurance-Premium-Data.csv>

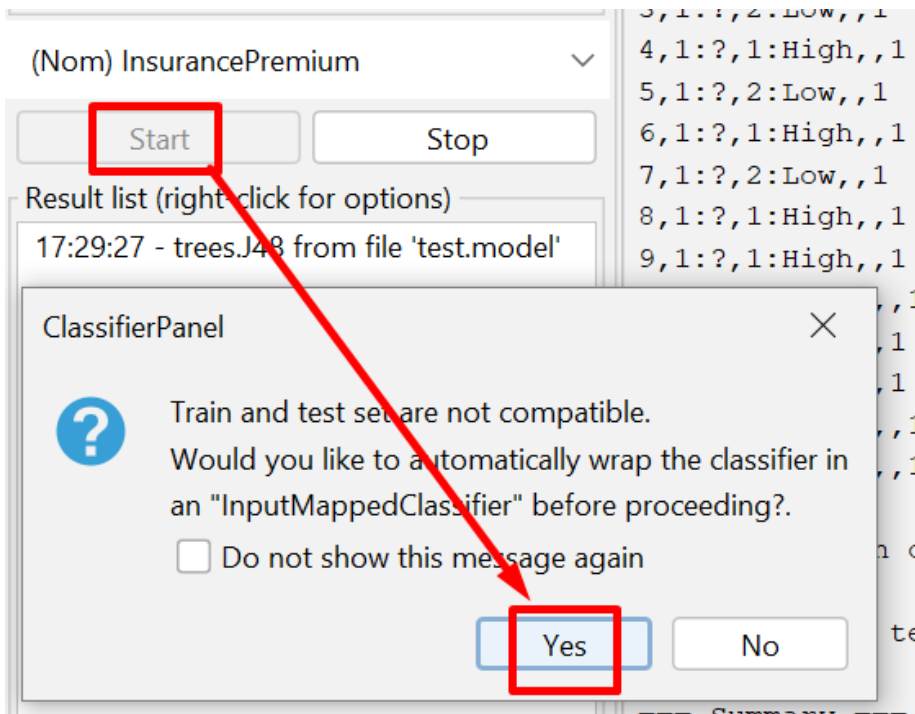
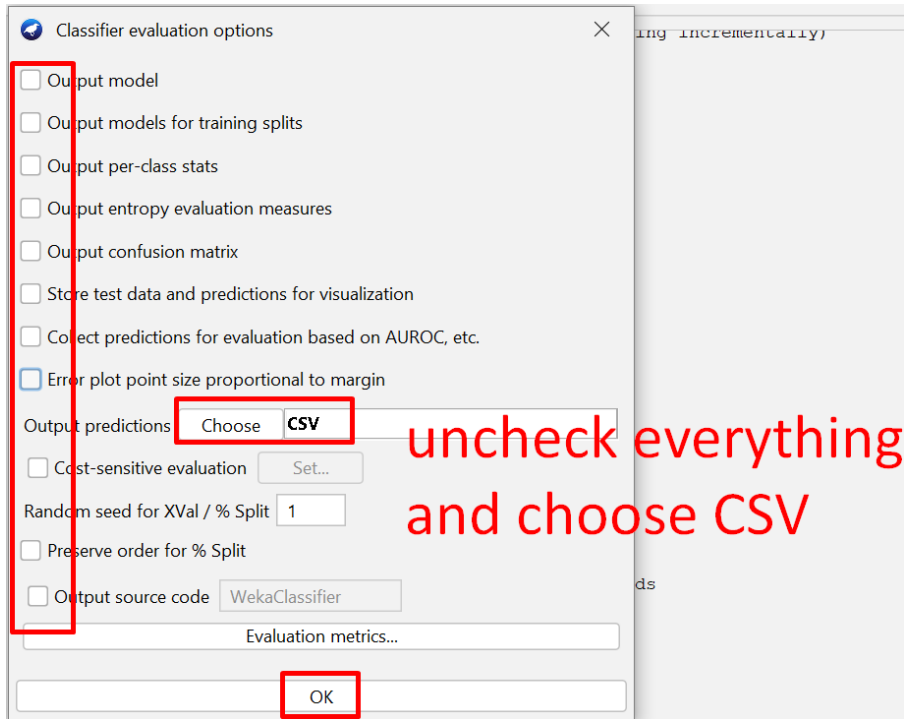
The screenshot shows a spreadsheet with the following data:

	A	B	C	D	E	F	G	H
1	AgeGroup	Smoker	MedicalCondition	SalaryLevel	InsurancePremium			
2	Old	Yes	Yes	High				
3	Teenager	Yes	Yes	Medium				
4	Young	Yes	Yes	Medium				
5	Old	No	Yes	High				
6	Young	Yes	Yes	High				
7	Teenager	No	Yes	Low				
8	Teenager	No	No	Low				
9	Old	No	No	Low				
10	Teenager	No	Yes	Medium				
11	Young	No	Yes	Low				
12	Young	Yes	No	High				
13	Teenager	Yes	No	Medium				
14	Young	No	No	Medium				
15	Old	Yes	No	Medium				

Annotations in the image:

- A red box highlights the 'InsurancePremium' column header.
- A red arrow points from the 'InsurancePremium' column header to the 'File' menu.
- A red box highlights the 'File' menu.
- A red arrow points from the 'File' menu to the 'Save As' option.
- A red box highlights the 'Save As' option.
- A red arrow points from the 'Save As' option to the file name 'Insurance Premium Data (TEST)' in the bottom bar.
- Red text on the right side of the spreadsheet says: "delete away the data in InsurancePremium column to simulate a new Dataset...." and "later, predictions will enter here".
- Red text at the bottom left says: "save it as a new file".





Test options

- Use training set
- Supplied test set Set...
- Cross-validation Folds: 10
- Percentage split %: 66
- More options...

(Nom) InsurancePremium

Start Stop

Result list (right-click for options)

- 17:29:27 - trees.148 from file 'test.model'
- 17:35:54 - misc.InputMappedClassifier

Classifier output

Test mode: user supplied test set: size unknown (reading incrementally)

=== Predictions on test set ===

```
inst#,actual,predicted,error,prediction
1,1:?,1:High,,1
2,1:?,1:High,,1
3,1:?,2:Low,,1
4,1:?,1:High,,1
5,1:?,2:Low,,1
6,1:?,1:High,,1
7,1:?,2:Low,,1
8,1:?,1:High,,1
9,1:?,1:High,,1
10,1:?,1:High,,1
11,1:?,2:Low,,1
12,1:?,2:Low,,1
13,1:?,1:High,,1
14,1:?,1:High,,1
```

select and CTRL + C

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.02 seconds

=== Summary ===

Total Number of Instances 0  
Ignored Class Unknown Instances 14

Text Import

Import

Character set: Unicode (UTF-16)

Language: Default - English (USA)

From row: 1

Separator Options

- Fixed width
- Separated by
- Tab  Comma  Semicolon  Space  Other
- Merge delimiters  Trim spaces String delimiter: "

Other Options

- Format quoted field as text  Detect special numbers
- Evaluate formulas  Skip empty cells

Fields

Column type:

	Standard	Standard	Standard	Standard	Standard
	inst#	actual	predicted	error	prediction
1	1	1:?	1:High		1
2	2	1:?	1:High		1
3	3	1:?	2:Low		1
4	4	1:?	1:High		1
5	5	1:?	2:Low		1
6	6	1:?	1:High		1
7	7	1:?	2:Low		1
8	8	1:?	1:High		1

in a new sheet, paste CTRL V

seperated by comma

OK Cancel

Insurance Premium Data (TEST) Sheet2

	A	B	C	D	E
1	inst#	actual	predicted	error	prediction
2		1:?	1:High		1
3		2:?	1:High		1
4		3:?	2:Low		1
5		4:?	1:High		1
6		5:?	2:Low		1
7		6:?	1:High		1
8		7:?	2:Low		1
9		8:?	1:High		1
10		9:?	1:High		1
11		10:?	1:High		1
12		11:?	2:Low		1
13		12:?	2:Low		1
14		13:?	1:High		1
15		14:?	1:High		1

copy this column

we see that the prediction is PERFECT!

paste here

---

### III. CONCLUSION

---

1. Age Group is the MOST IMPORTANT feature of the Dataset
2. Medical Condition and Smoker are next most important features.
3. Salary is NOT important at all for prediction!

---

## ABOUT DR. ALVIN ANG

---



Dr. Alvin Ang earned his Ph.D., Masters and Bachelor degrees from NTU, Singapore. He is a scientist, entrepreneur, as well as a personal/business advisor. More about him at [www.AlvinAng.sg](http://www.AlvinAng.sg).