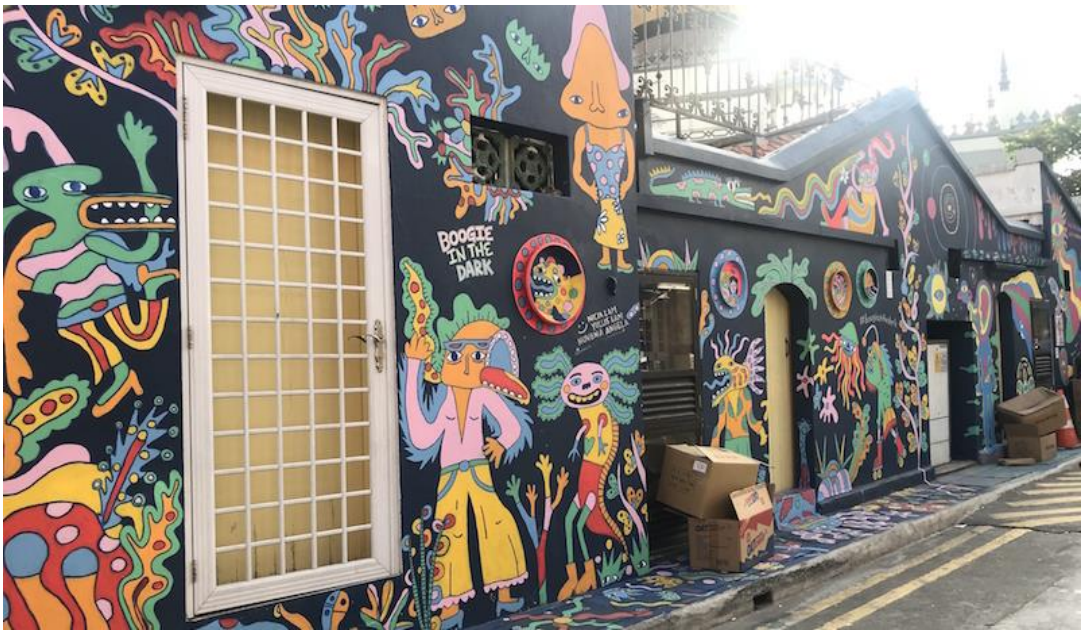
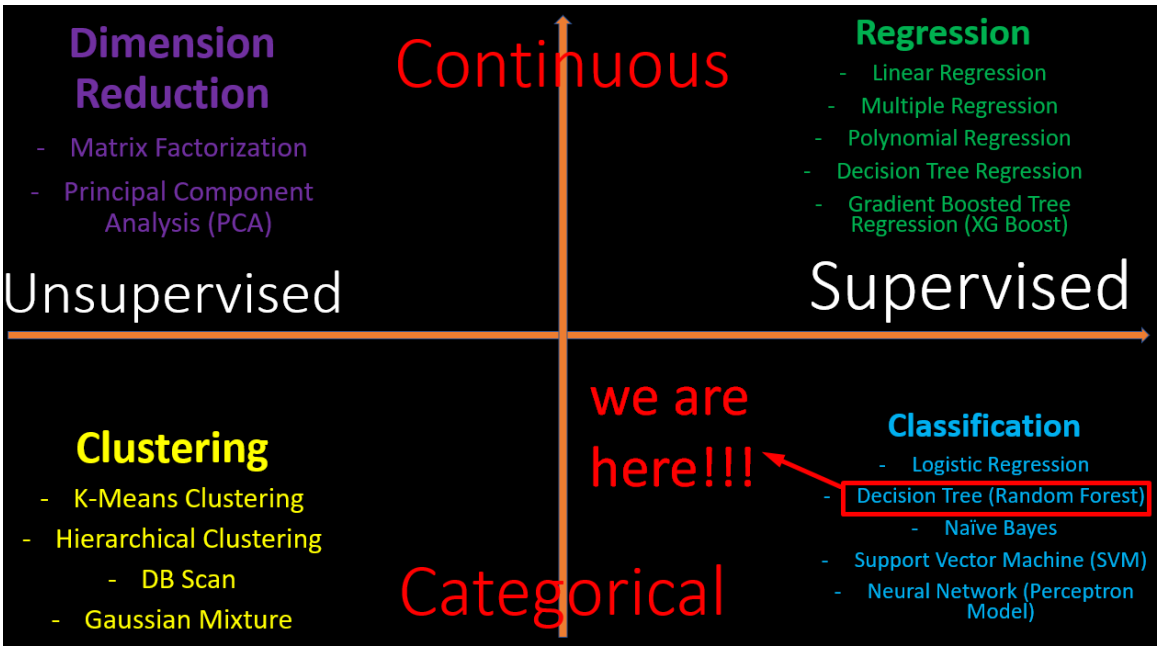
DECISION TREE (CLASSIFICATION)

USING WEKA

DR. ALVIN ANG

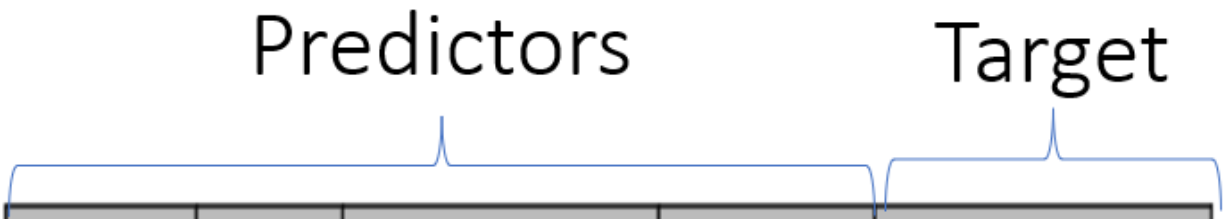


CONTENTS

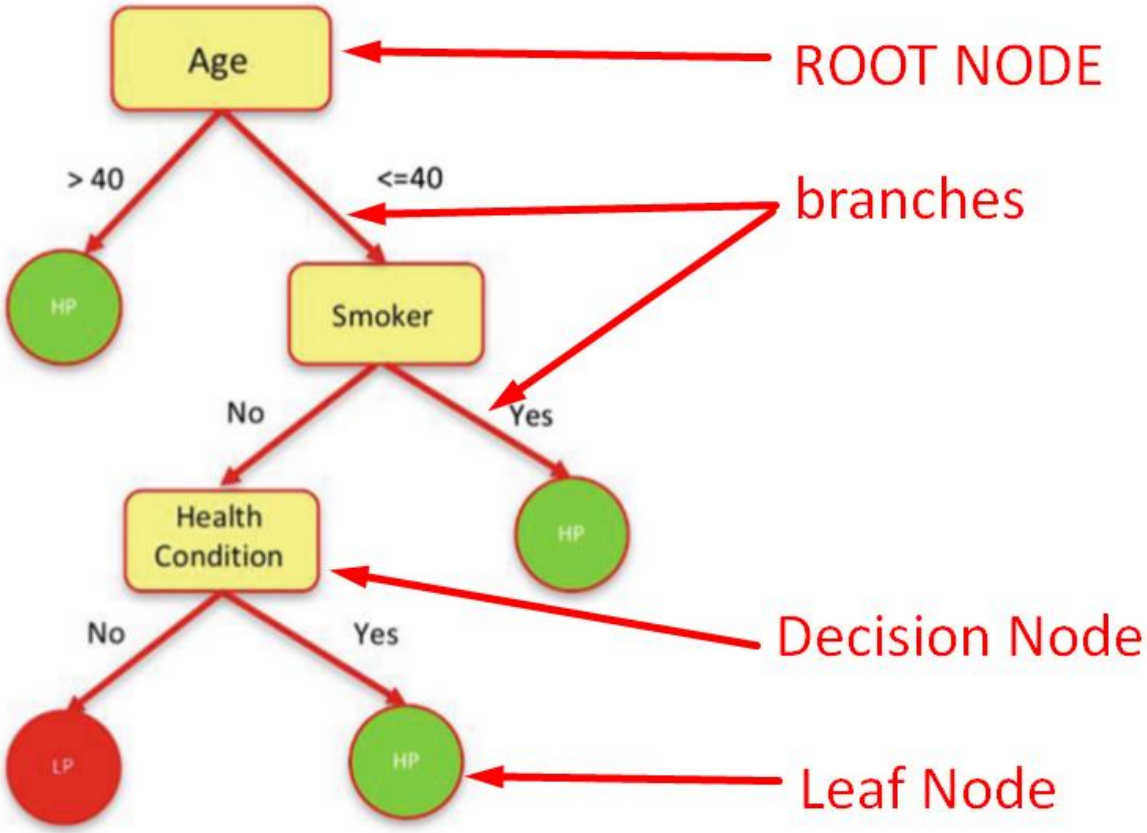


Let's say we are given the dataset below…

https://www.alvinang.sg/s/Insurance-Premium-Datacsv.csv



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



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

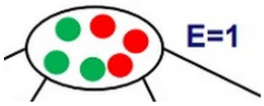
- **Pure means cannot split anymore.**
- **Pure means Entropy = 0.**
- **Pure means Homogeneous.**
- **Pure means Leaf Node (Terminal Node).**

PURE

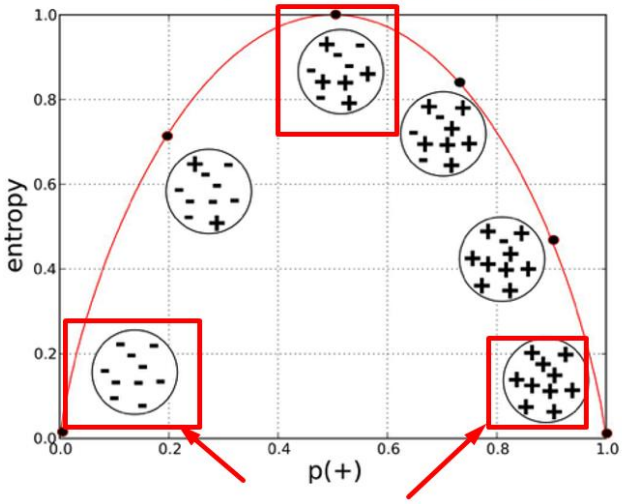
IMPURE

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

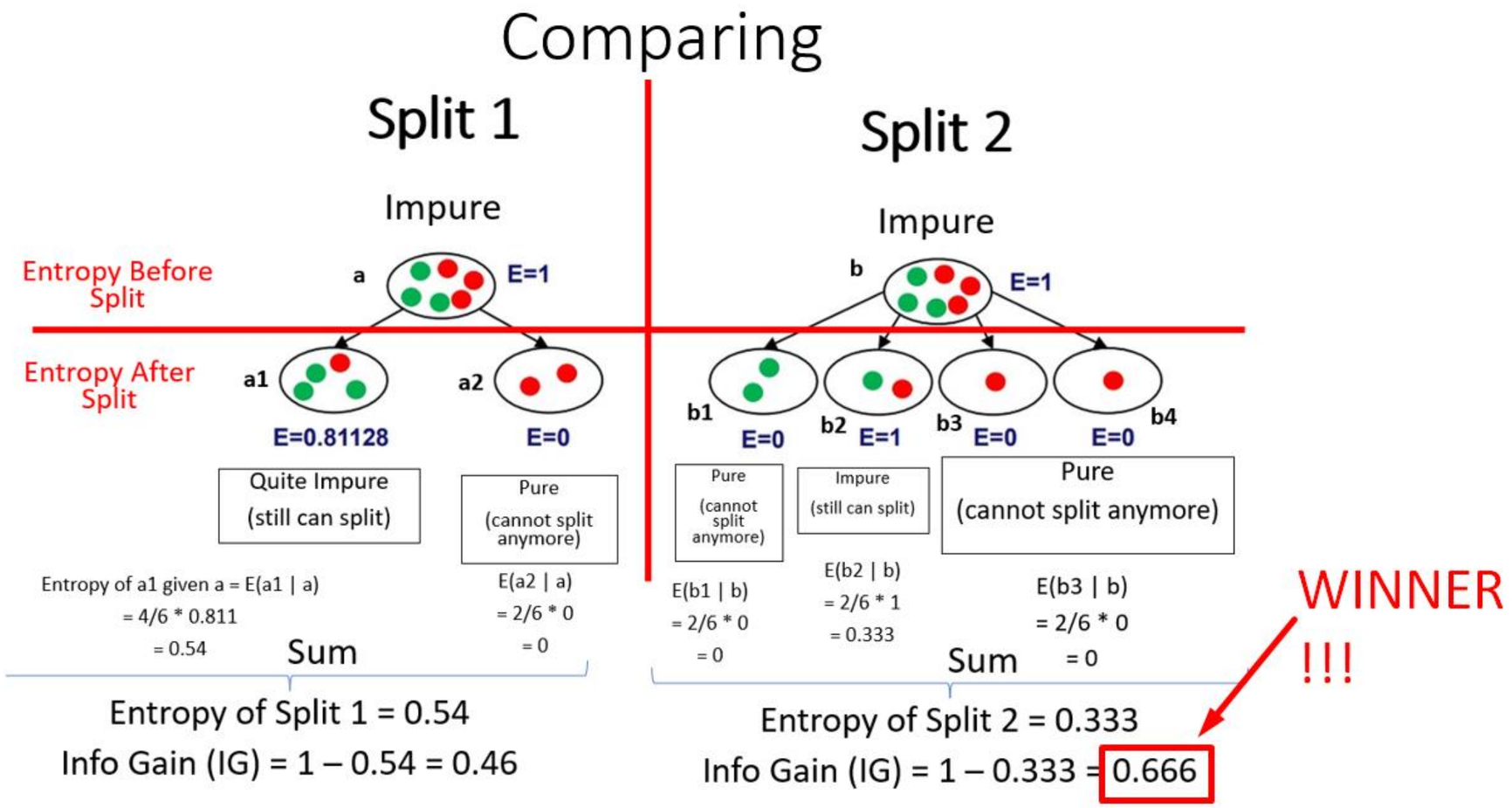
IMPURE



VERY IMPURE



VERY PURE



- 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 Homogenity remains after splitting!
    - More Homogenity (lower IG) means you can't split much later on.
    - Lower Homogenity (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

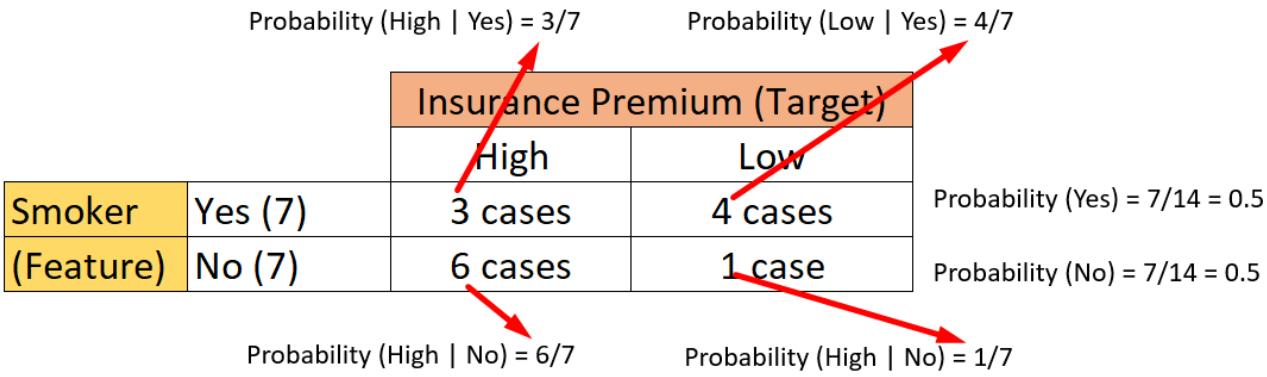
Age Group	Smoker	Medical Condition	Salary Level	Insurance Premium
Old	Yes	Yes	High	High
Teenager	Yes	Yes	Medium	High
Young	Yes	Yes	Medium	Low
Old	No	Yes	High	High
Young	Yes	Yes	High	Low
Teenager	No	Yes	Low	High
Teenager	No	No	Low	Low
Old	No	No	Low	High
Teenager	No	Yes	Medium	High
Young	No	Yes	Low	High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

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

B. STEP 2: CALCULATE ENTROPY OF TARGET | FEATURE

1. ENTROPY (TARGET, SMOKER)



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

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

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

2. ENTROPY (TARGET, AGE GROUP)

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

$$Entropy_{(Target\ ,\ Age\ Group)} = 0.69$$

3. ENTROPY (TARGET, MEDICAL CONDITION)

$$Entropy_{(Target\ ,\ Medical\ Condition)} = 0.89$$

4. ENTROPY (TARGET, SALARY LEVEL)

$$Entropy_{(Target\ ,\ Salary\ Level)} = 0.91$$

C. STEP 3: CALCULATING INFORMATION GAIN (IG)

$$Information\,Gain = Entropy_{Target} - Entropy_{Target,\,Features}$$

1. IG (SMOKER)

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

2. IG (AGE GROUP)

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

3. IG (MEDICAL CONDITION)

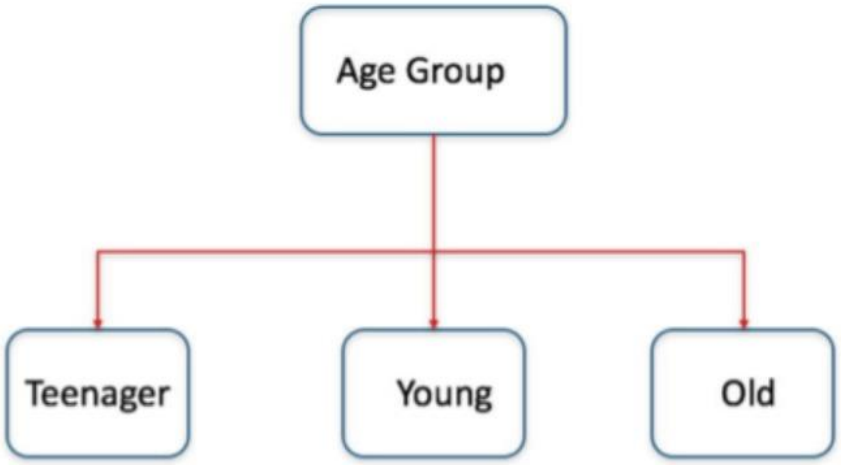
$$\begin{aligned} IG_{Medical\,Condition} &= Entropy_{Target} - Entropy_{Target,\,Medical\,Condition} \\ &= 0.94 - 0.89 \\ &= 0.05 \end{aligned}$$

4. IG (SALARY LEVEL)

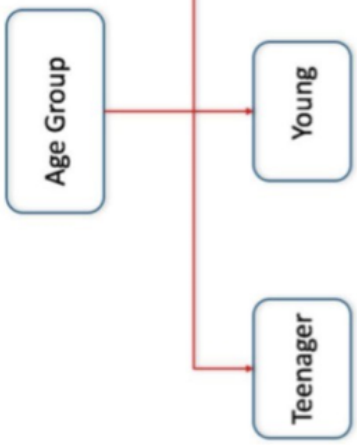
$$IG_{Salary\ Level} = Entropy_{Target} - Entropy_{Target,\ Salary\ Level}$$
$$= 0.94 - 0.91$$
$$= 0.03$$

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

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



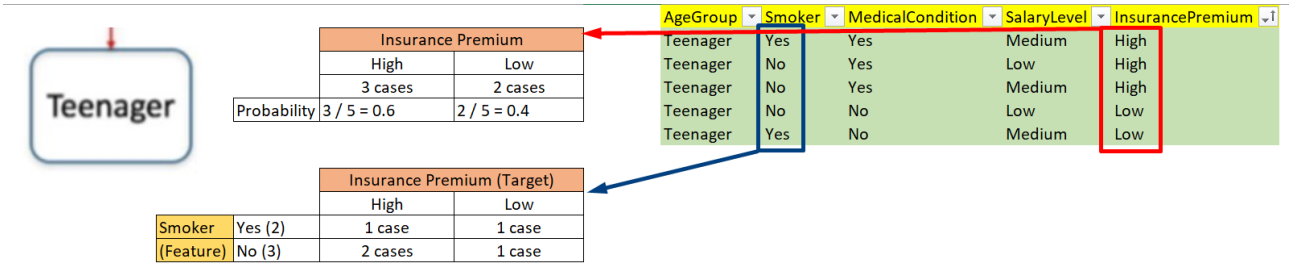
E. STEP 5: WHICH NODE NEXT?



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

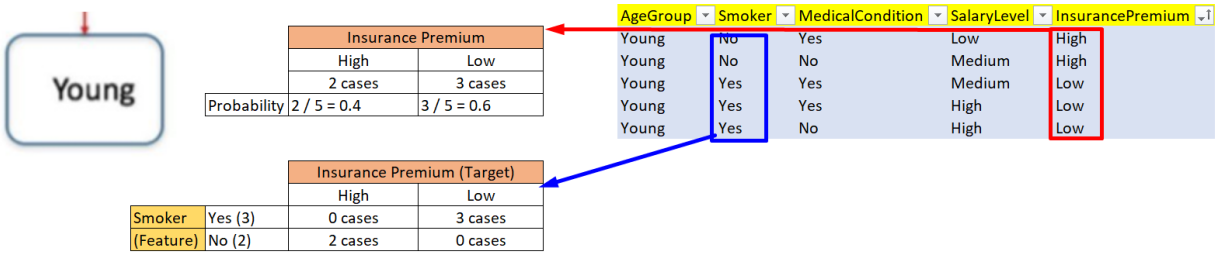
All High which means OLD is a Terminal Node (Leaf Node)

1. LET'S TAKE A LOOK AT TEENAGERS…



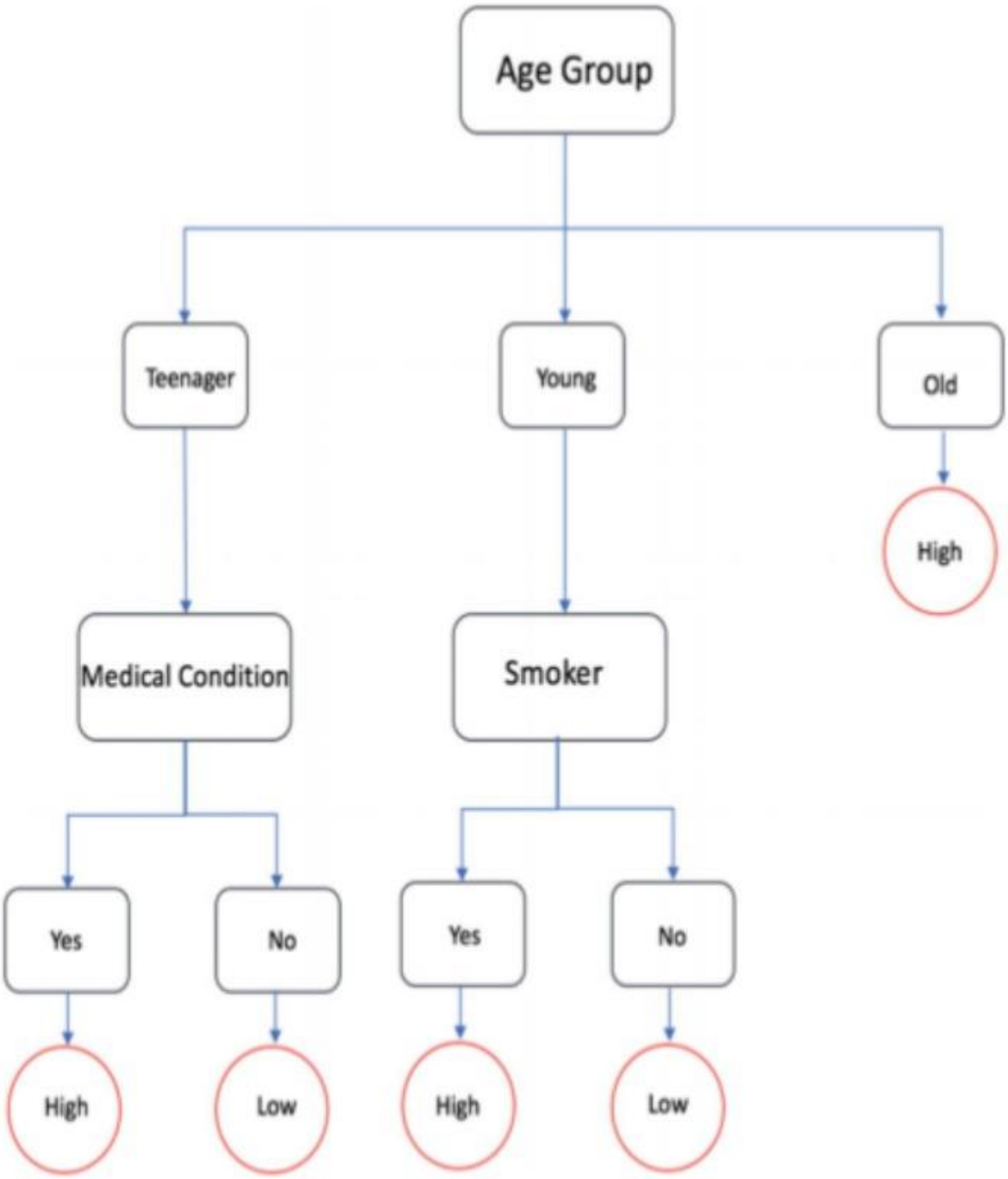
- We can now go through the entire Step 1 to 3 for Entropy / IG calculation for the Teenager Node.
- Objective to find out which is the next important feature that comes after Teenager?
    - Is it Smoker?
    - Medical Condition?
    - Salary Level?

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…..

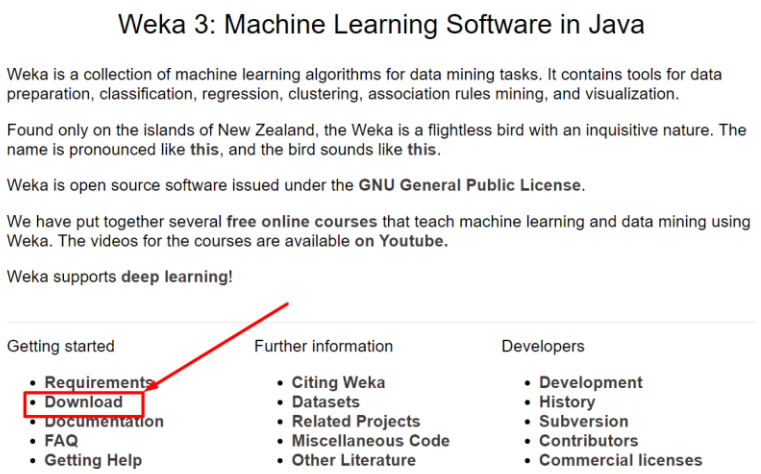


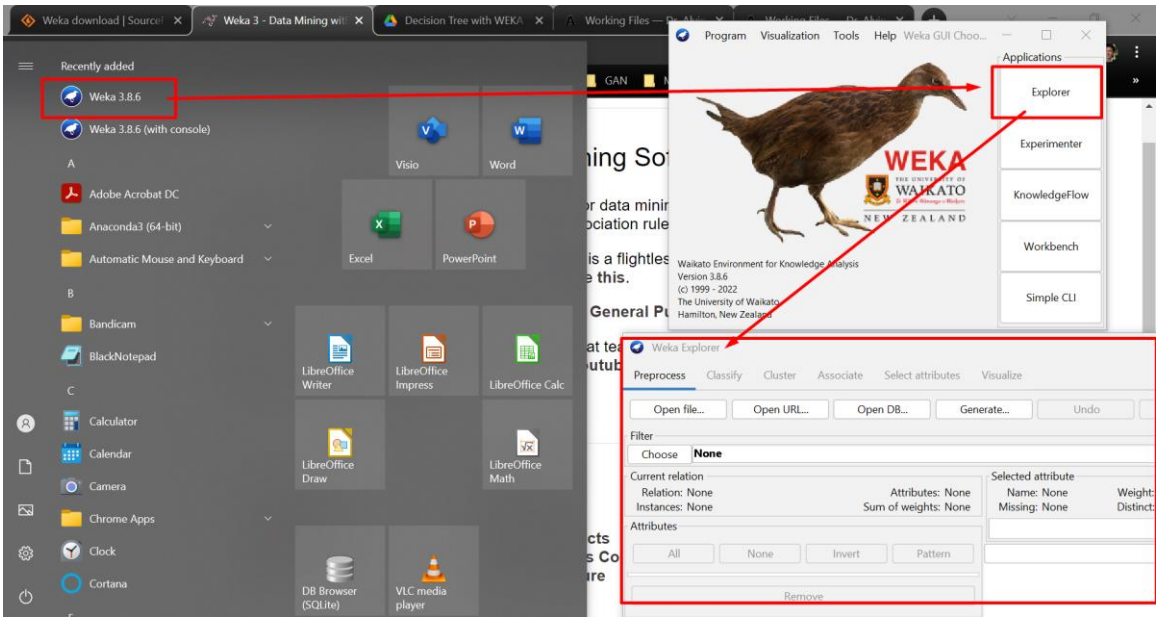
II. WEKA FOR DECISION TREE (CLASSIFICATION)

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

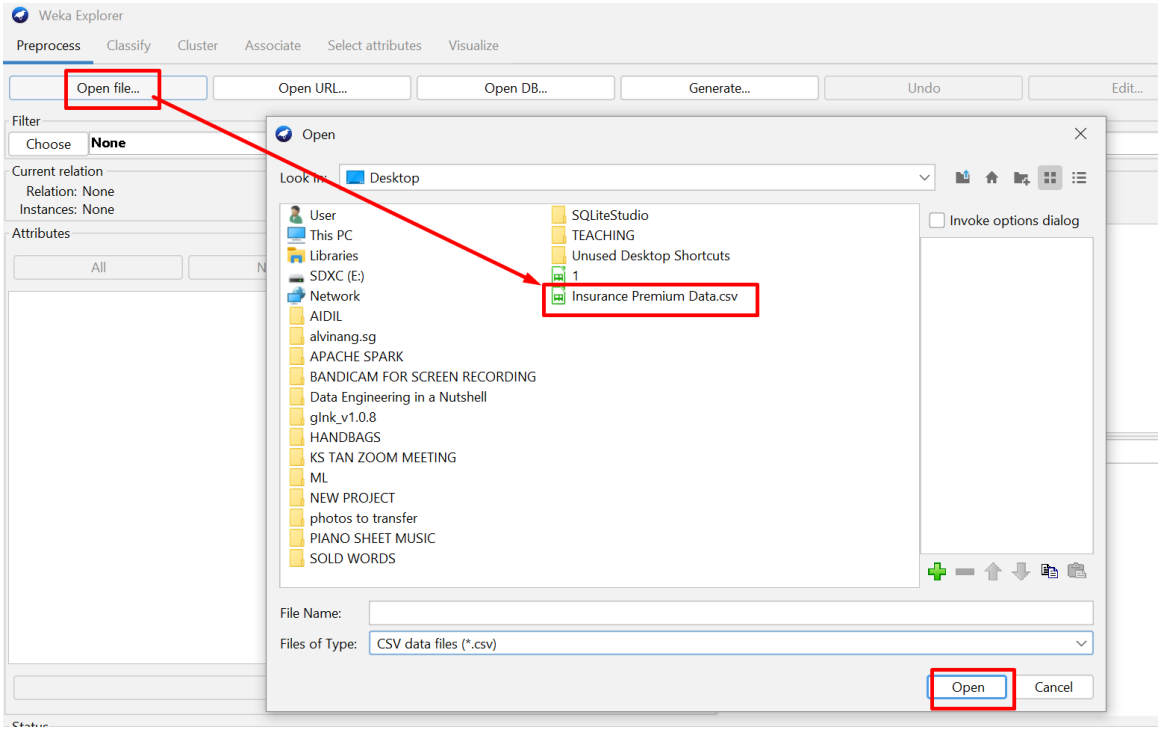
https://www.alvinang.sg/s/Insurance-Premium-Datacsv.csv

A. STEP 1: INSTALL WEKA

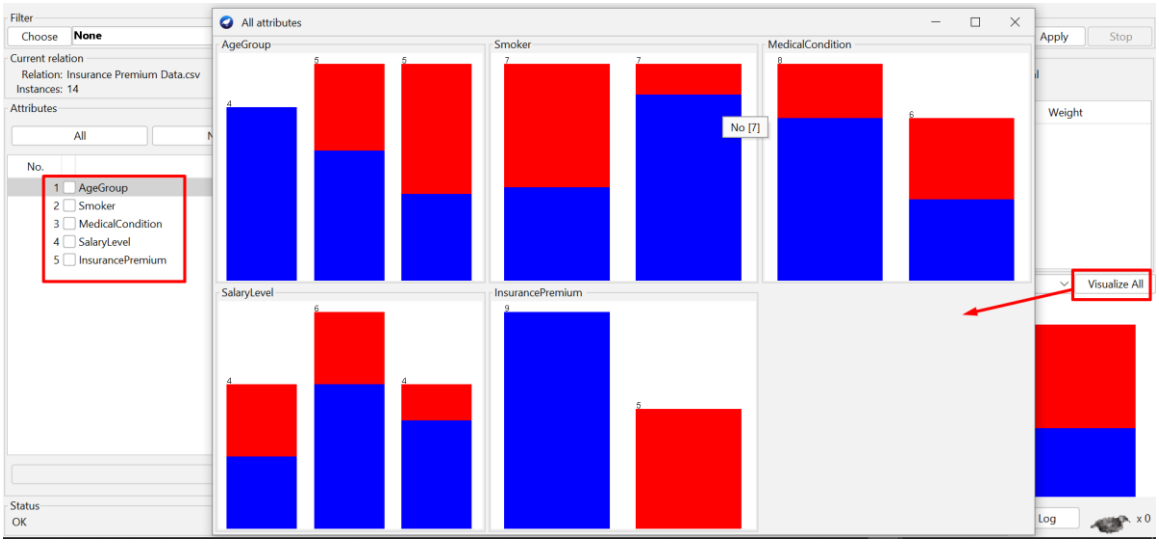




B. STEP 2: BRING IN THE DATASET

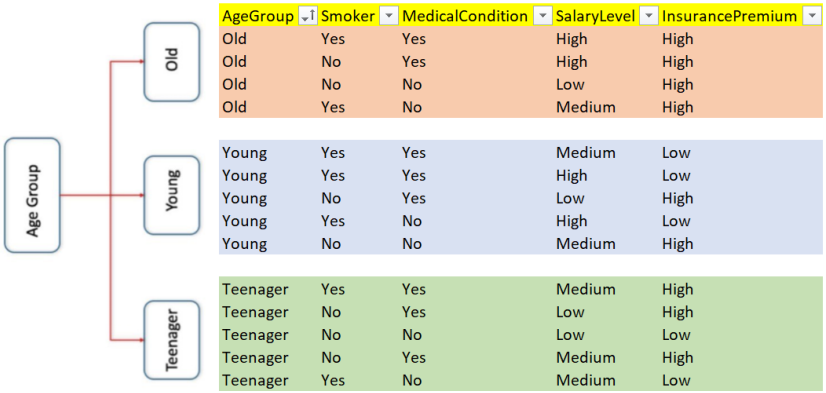


C. STEP 3: VISUALIZE THE DATASET



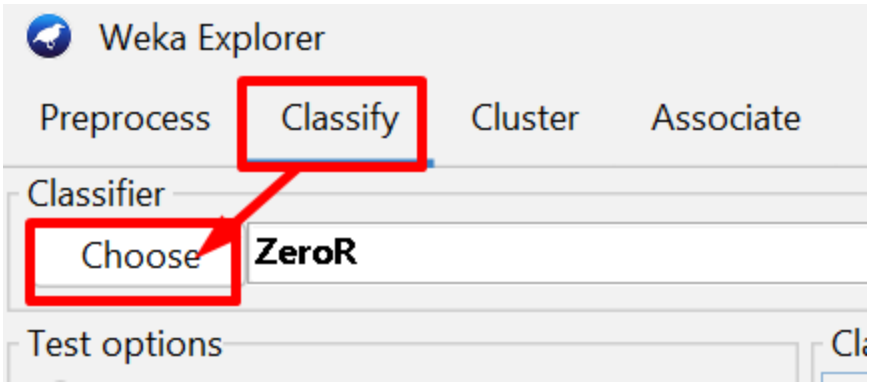
Color code:

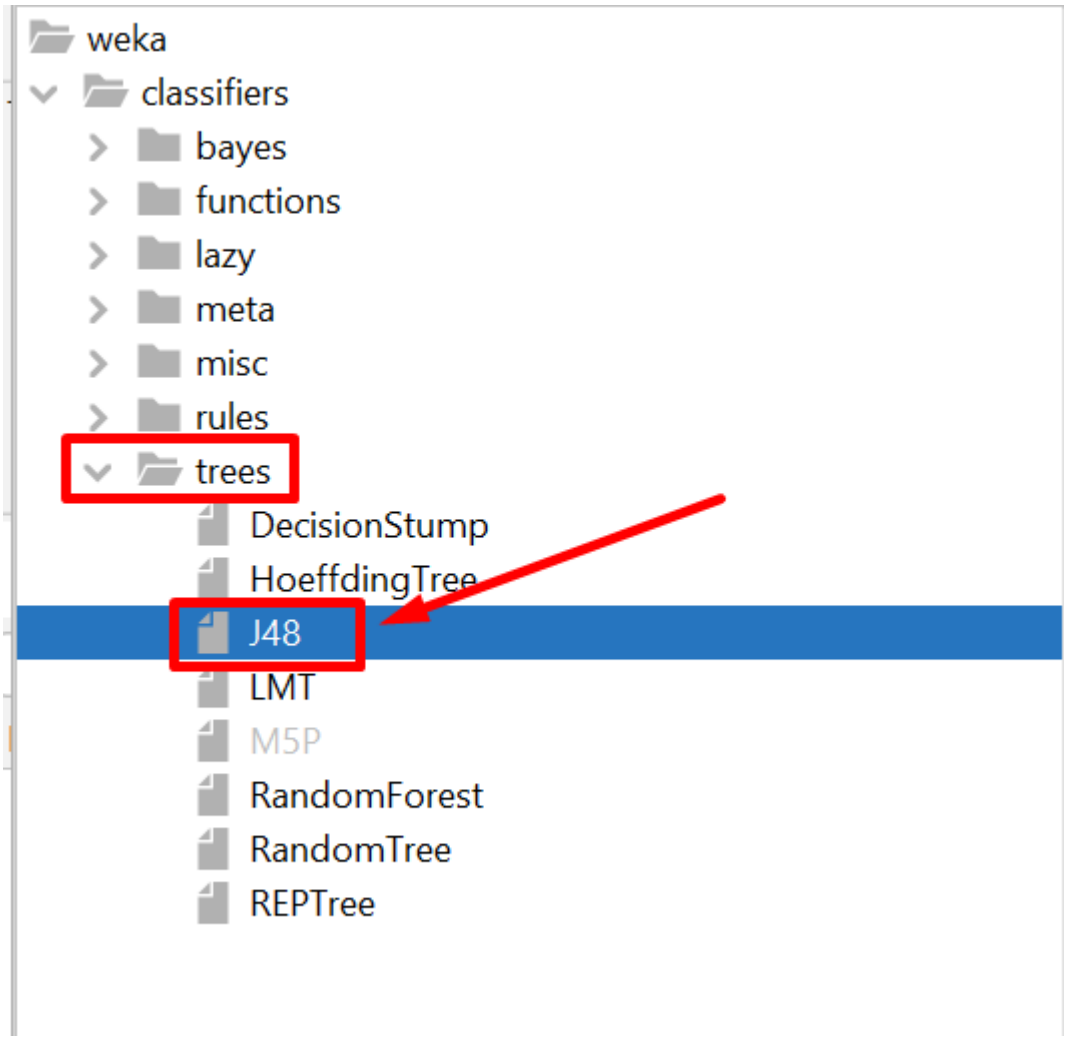
- Red = Low Insurance Premium Cases
- Blue = High Insurance Premium Cases
- E.g. you see the AgeGroup has 3 bars….
    - 4 OLD = All Blue (coz 4 Low Premium)
    - 5 Young = 3 Blue 2 Red
    - 5 Teenagers = 3 Red 2 Blue



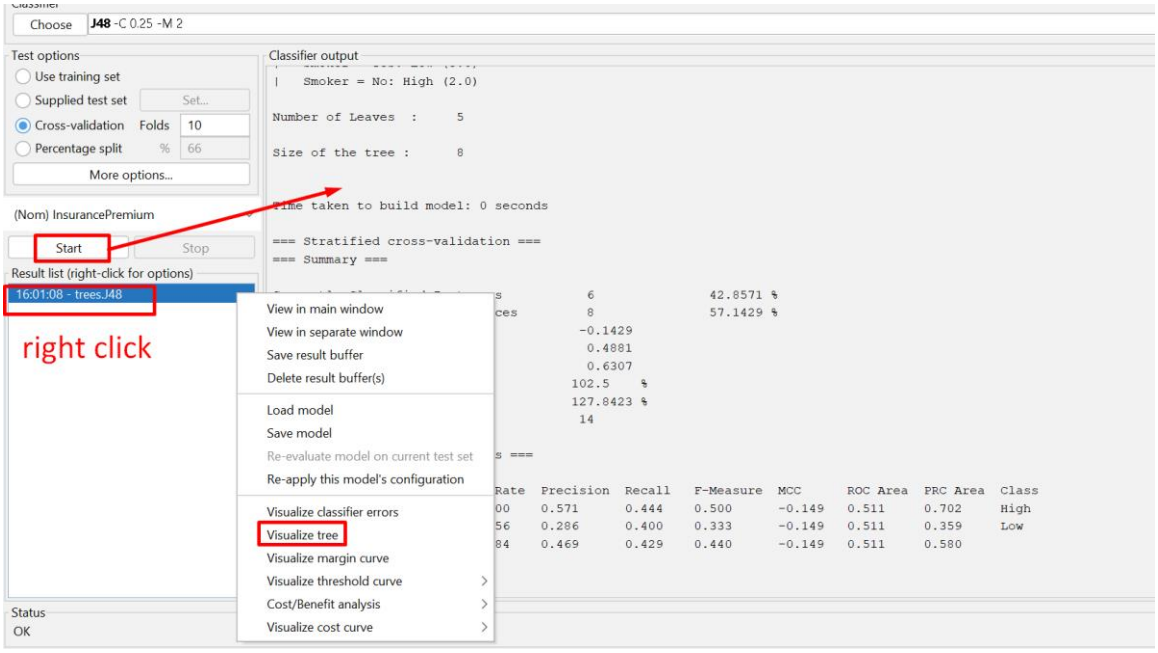
AgeGroup	Smoker	MedicalCondition	SalaryLevel	InsurancePremium
Old	Yes	Yes	High	High
Old	No	Yes	High	High
Old	No	No	Low	High
Old	Yes	No	Medium	High
Young	Yes	Yes	Medium	Low
Young	Yes	Yes	High	Low
Young	No	Yes	Low	High
Young	Yes	No	High	Low
Young	No	No	Medium	High
Teenager	Yes	Yes	Medium	High
Teenager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

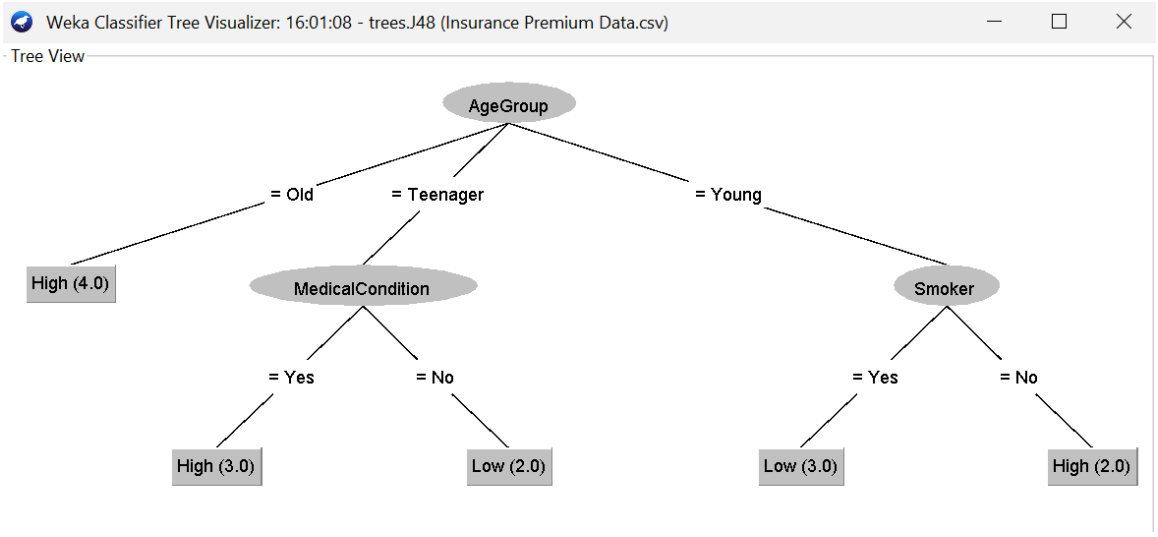
D. STEP 4: CHOOSE DECISION TREE CLASSIFIER

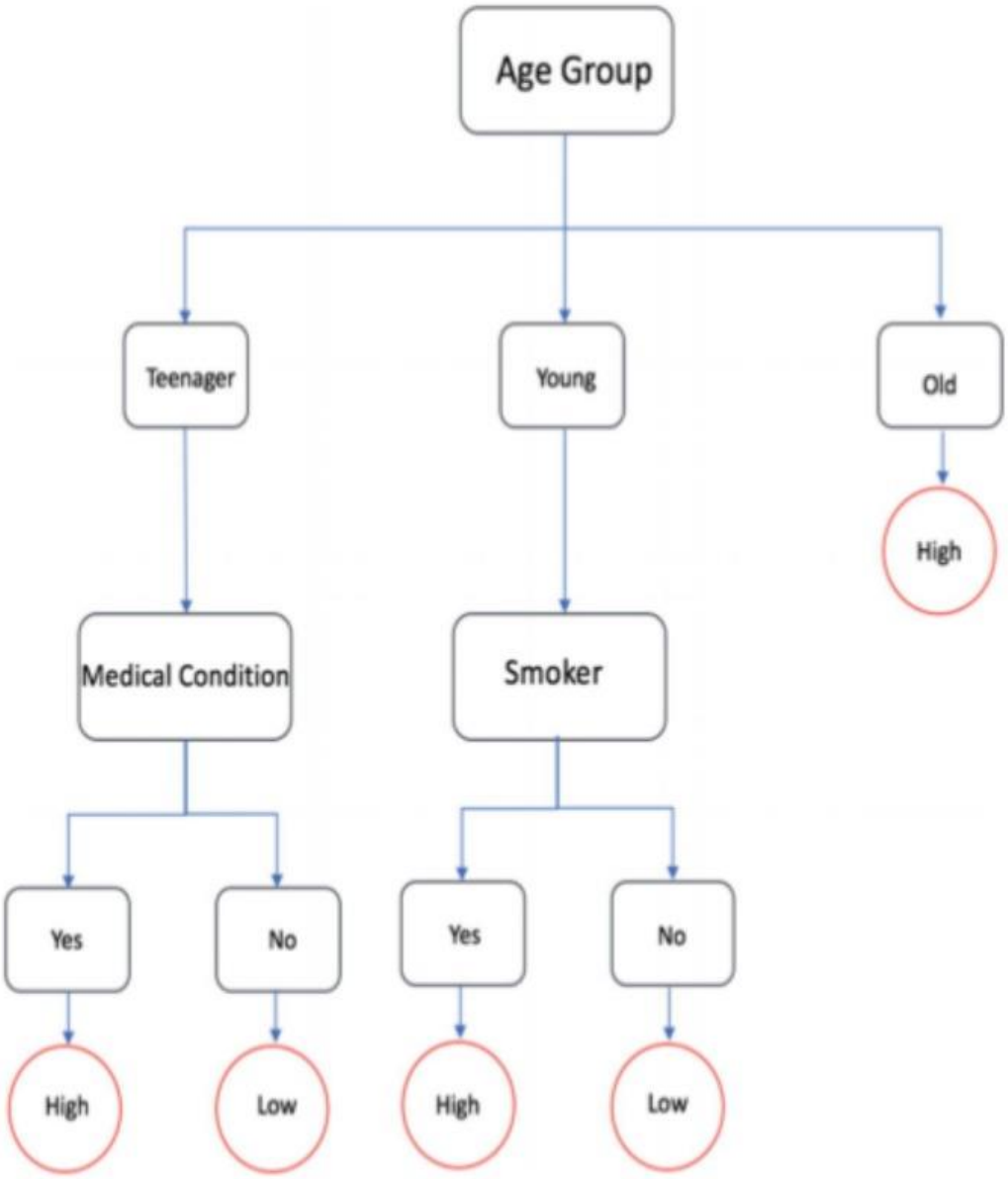




E. STEP 5: VISUALIZE THE TREE

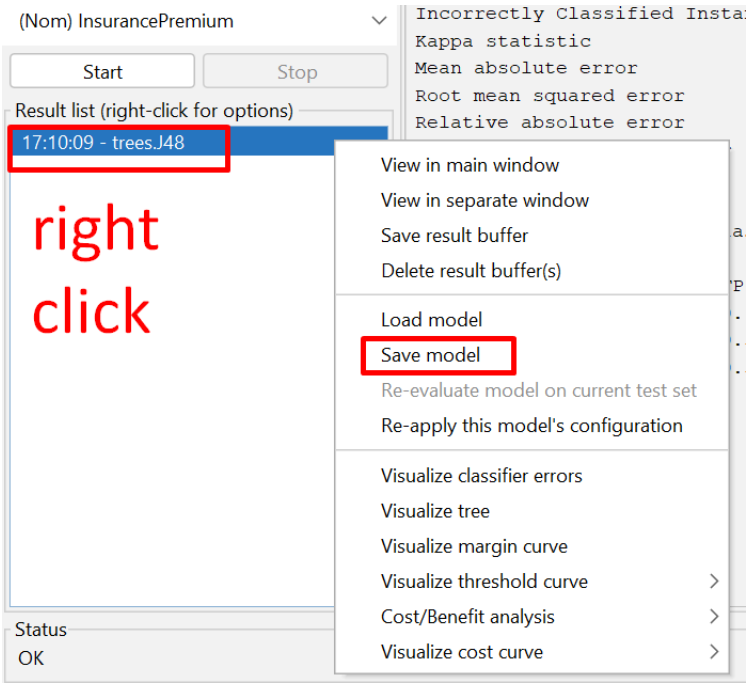


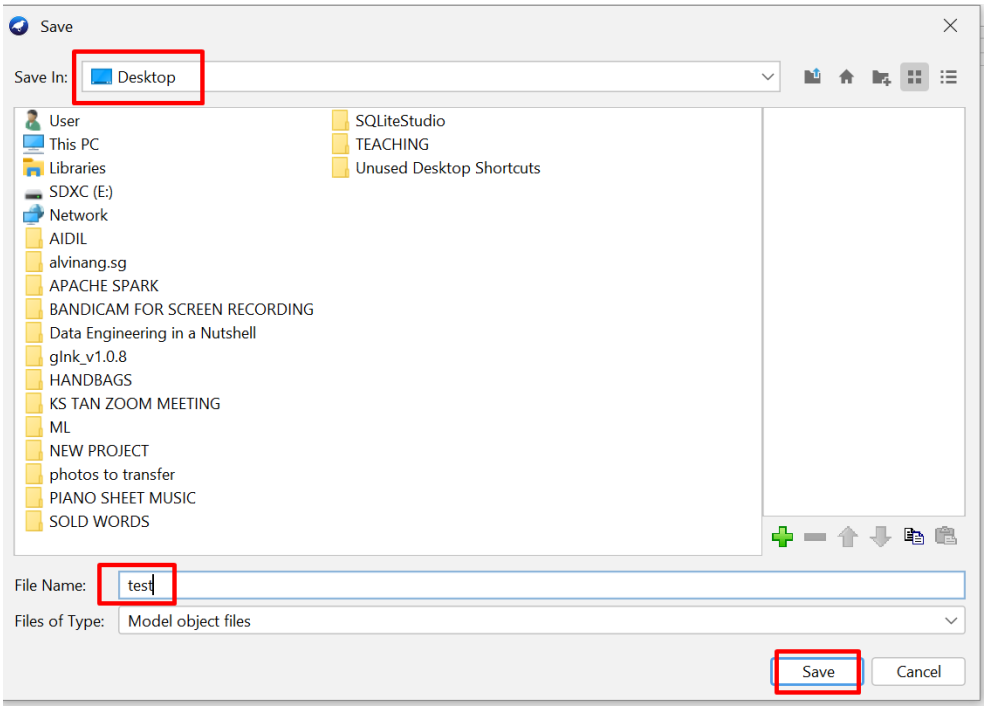




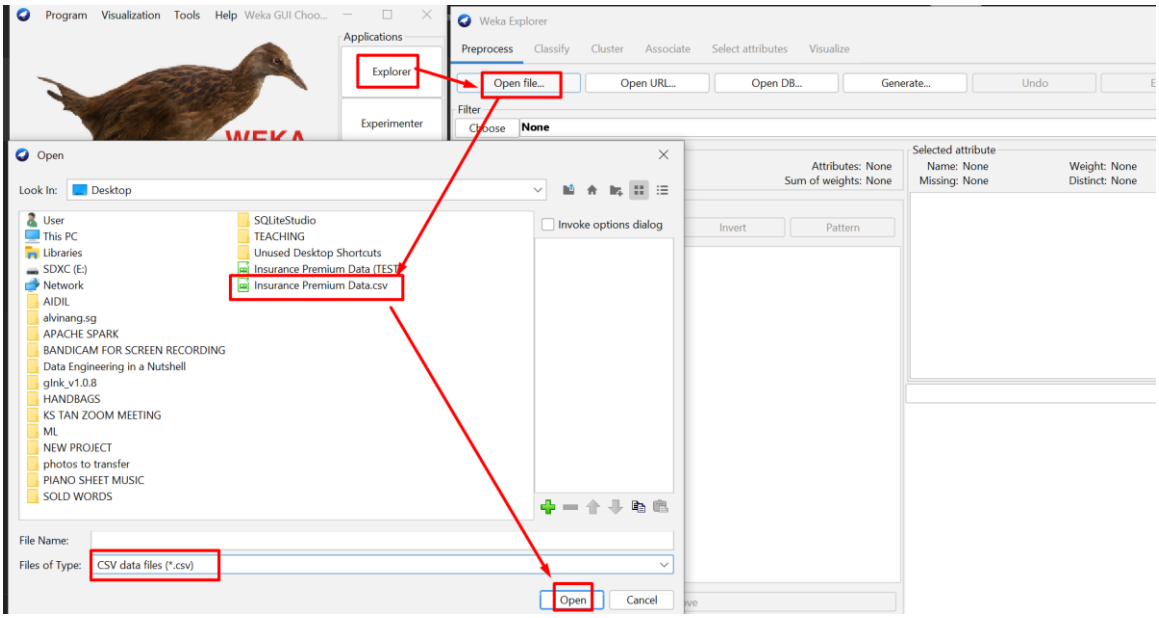
It's the same Tree as we drew earlier!

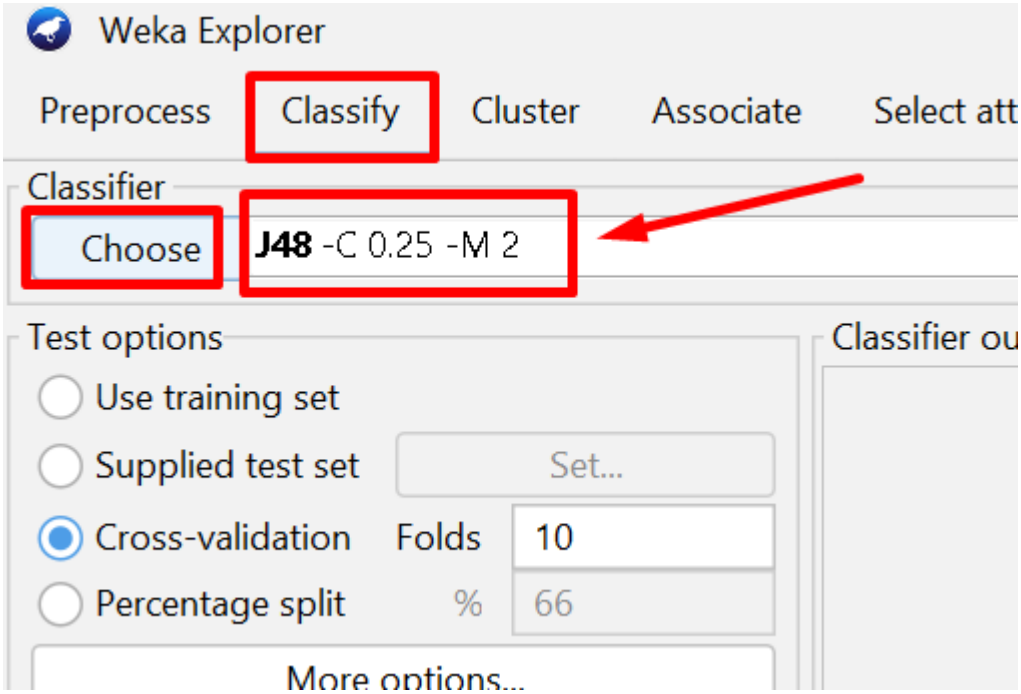
F. STEP 6: SAVE THE TRAINED MODEL

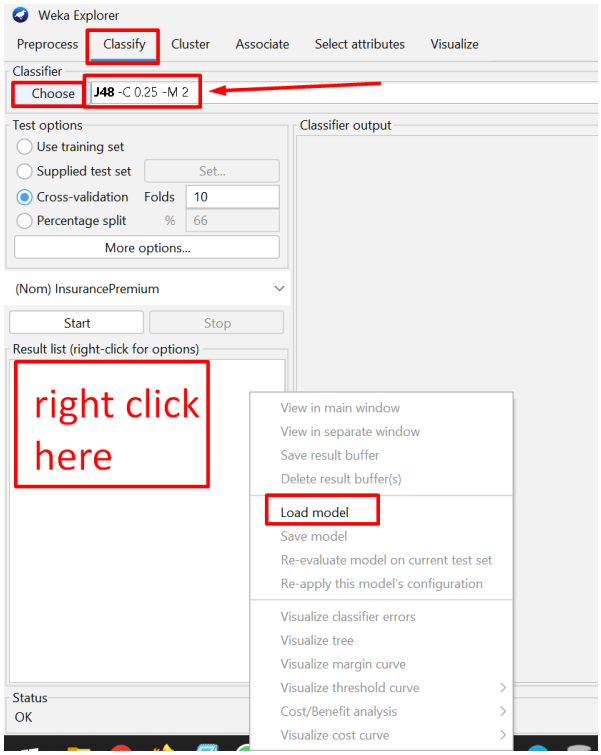


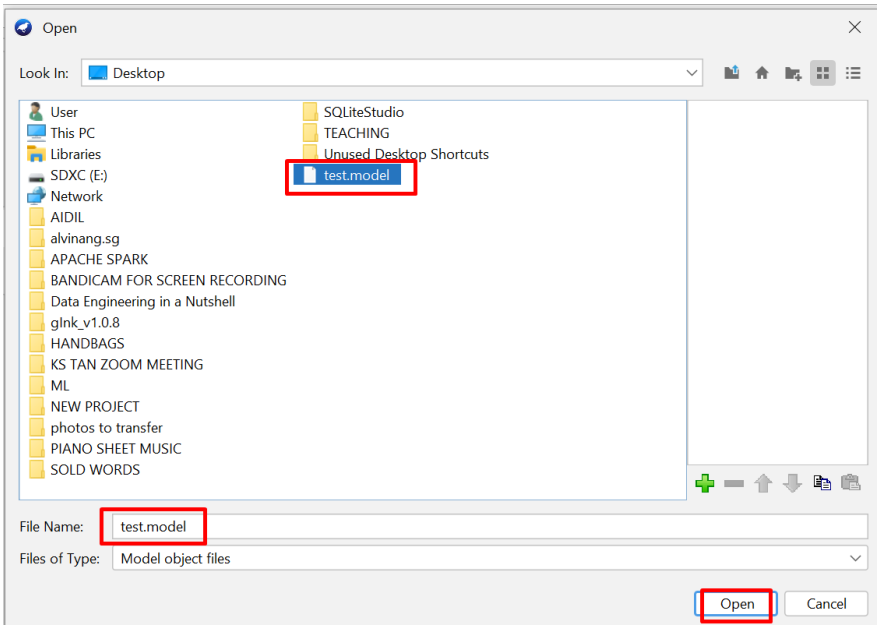


G. STEP 7: LOAD THE SAVED TRAINED MODEL





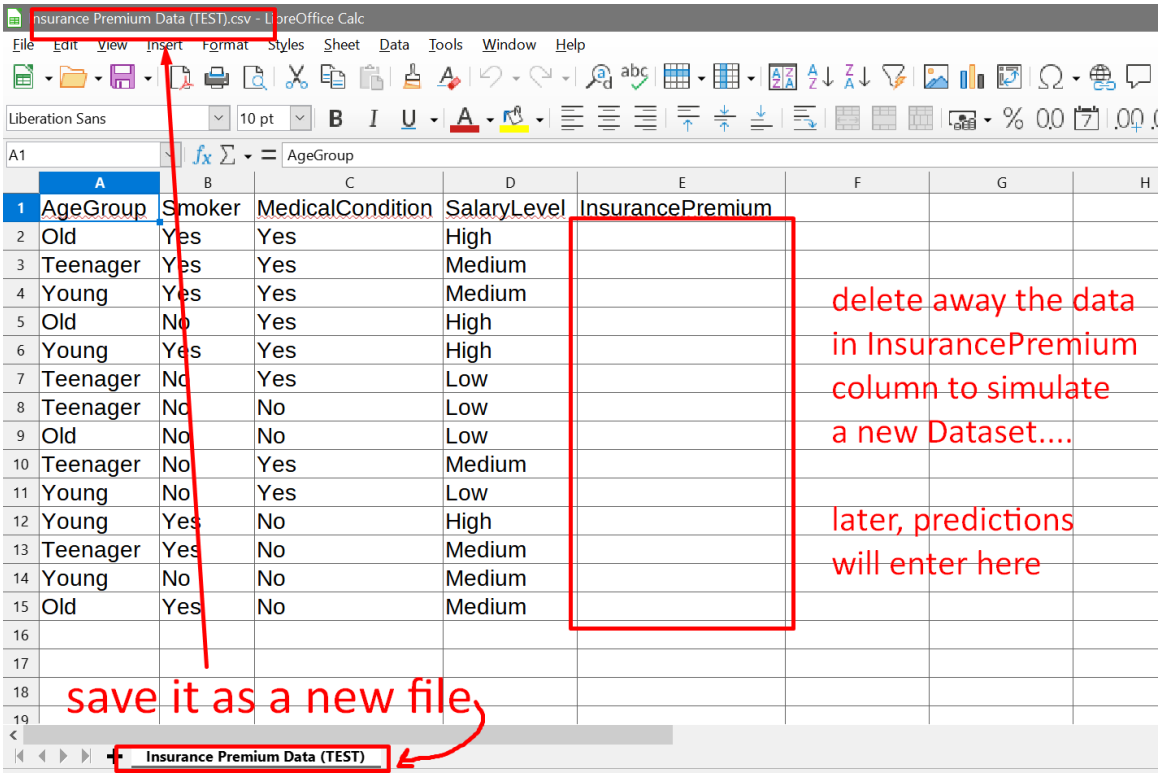




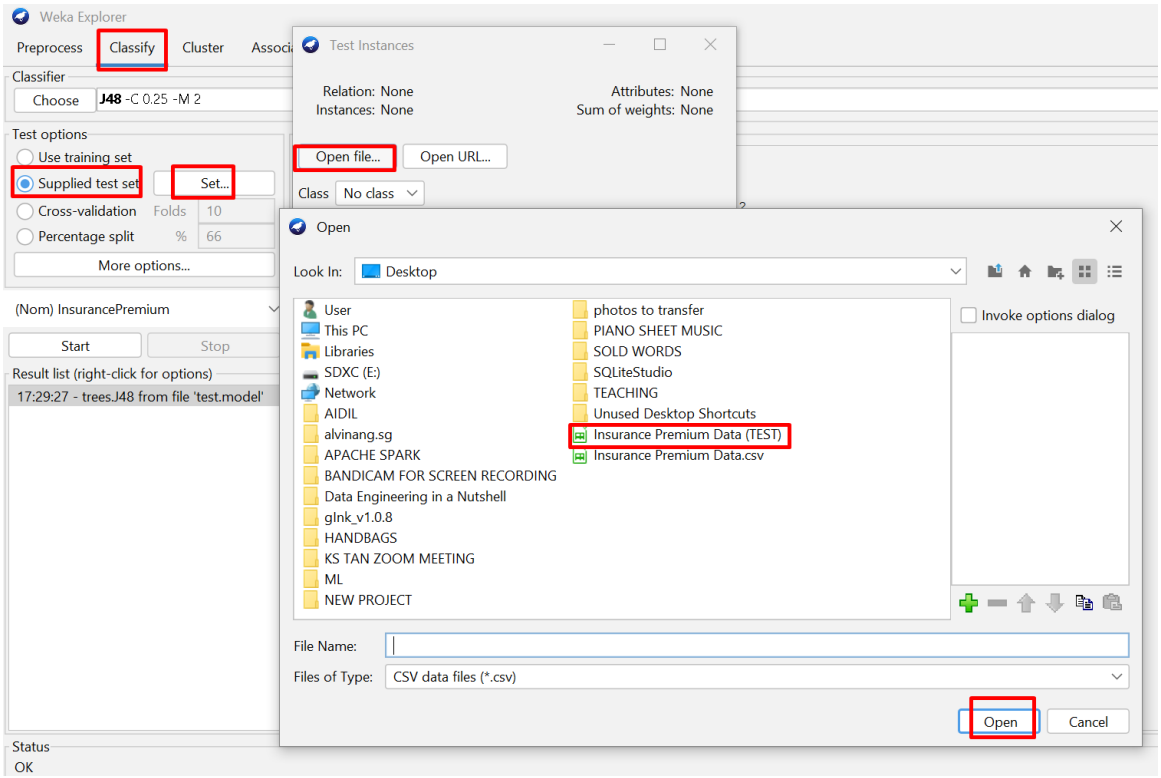
H. STEP 8: USING THE TREE TO MAKE A PREDICTION

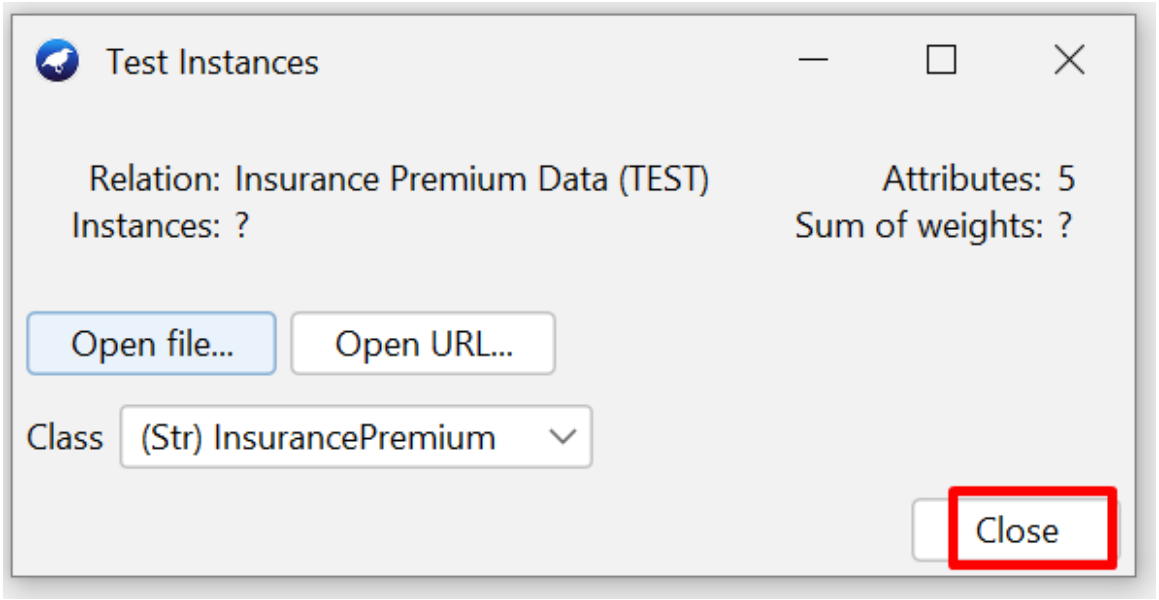
Go here and download the data again:

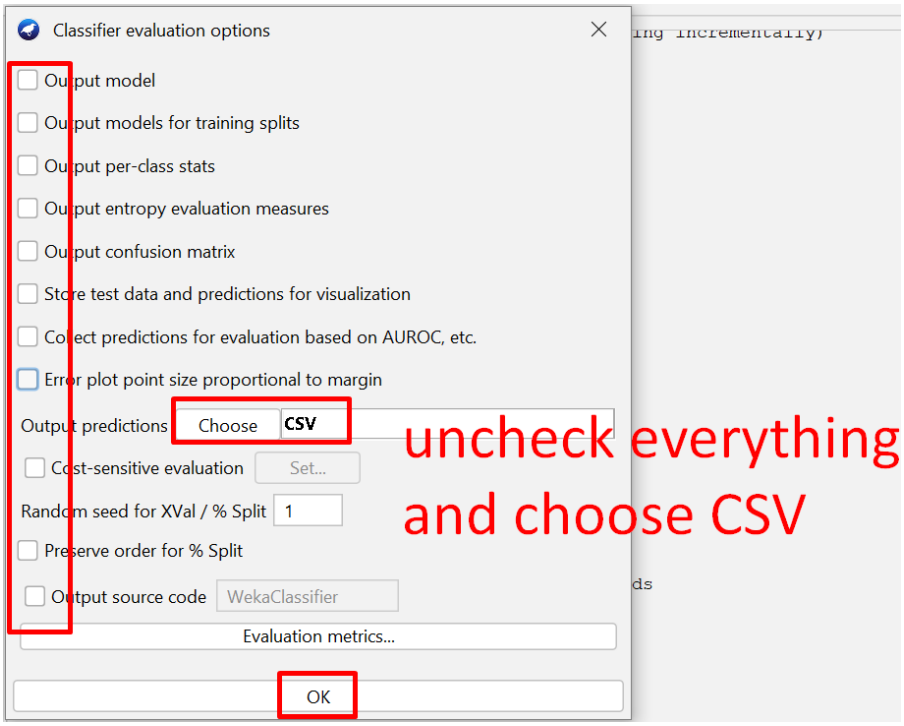
https://www.alvinang.sg/s/Insurance-Premium-Datacsv.csv

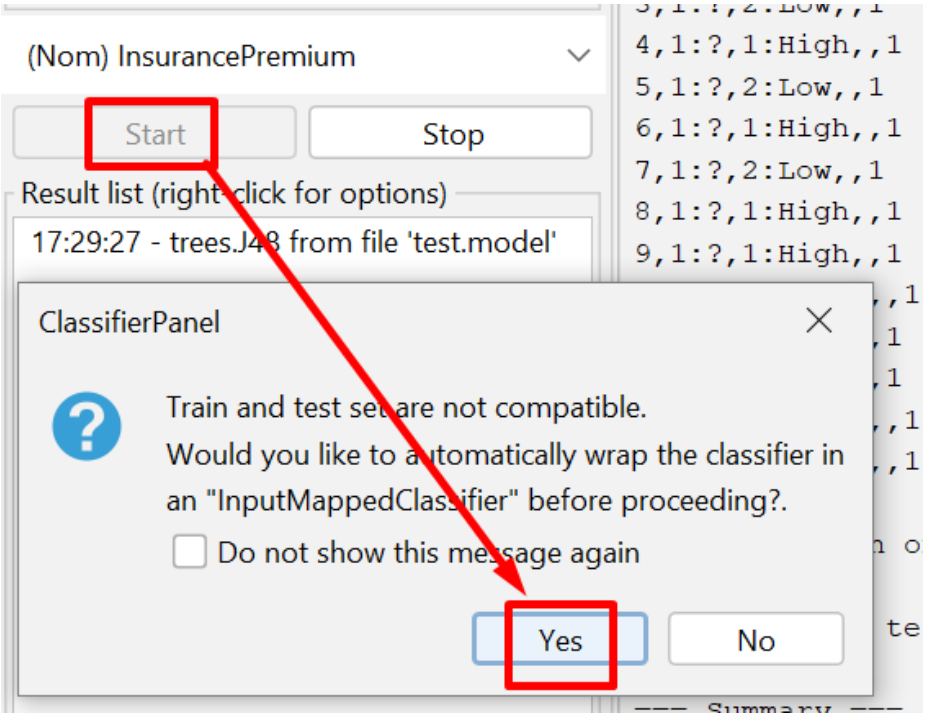


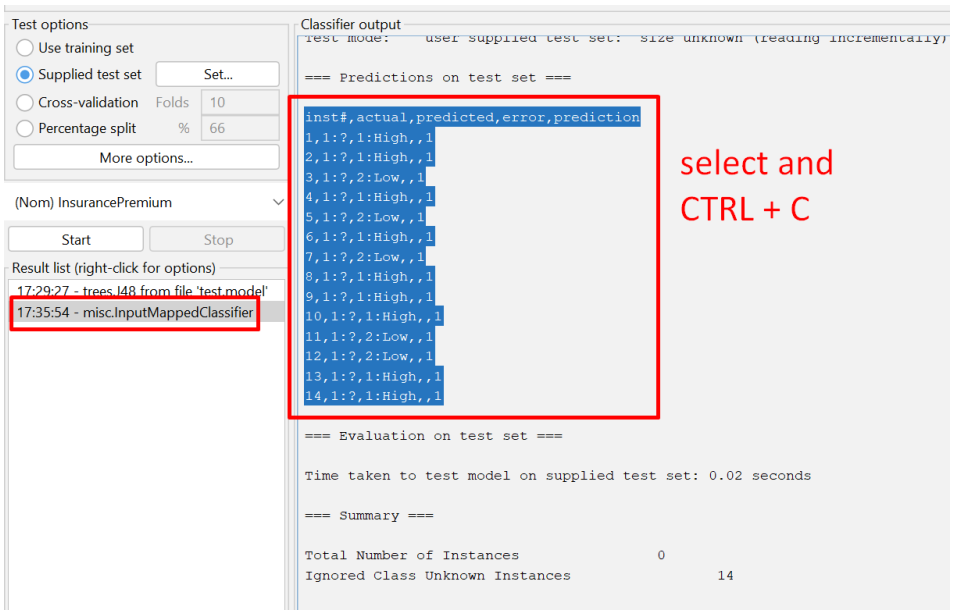
	A	B	C	D	E	F	G	H
1	AgeGroup	Smoker	MedicalCondition	SalaryLevel	InsurancePremium			
2	Old	Yes	Yes	High				
3	Teenager	Yes	Yes	Medium				
4	Young	Yes	Yes	Medium				
5	Old	No	Yes	High				
6	Young	Yes	Yes	High				
7	Teenager	No	Yes	Low				
8	Teenager	No	No	Low				
9	Old	No	No	Low				
10	Teenager	No	Yes	Medium				
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				
16								
17								
18								

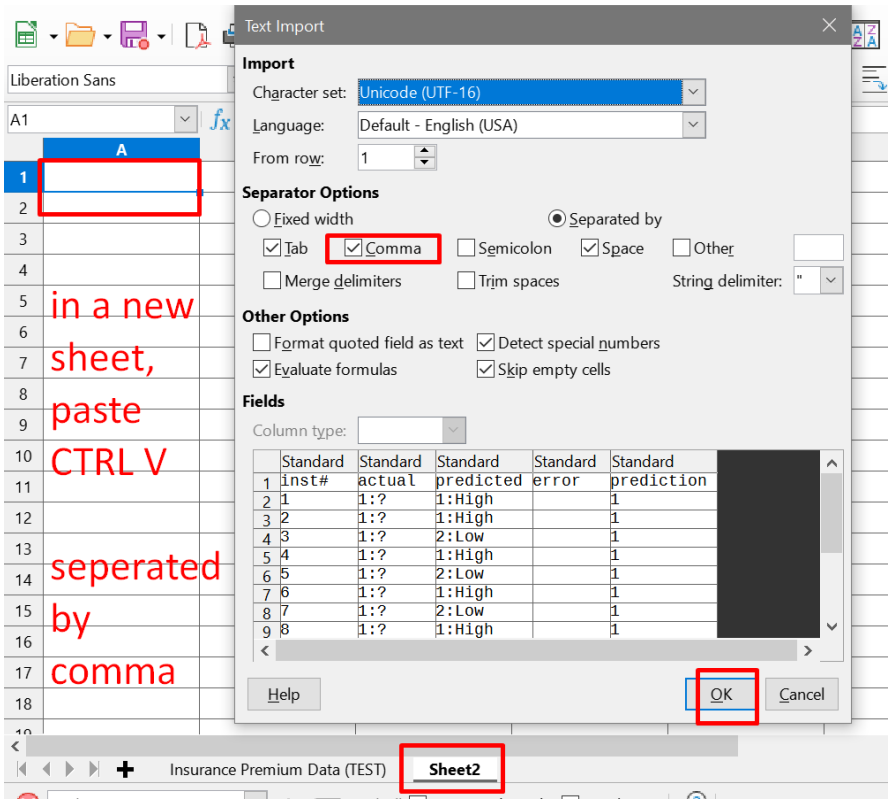


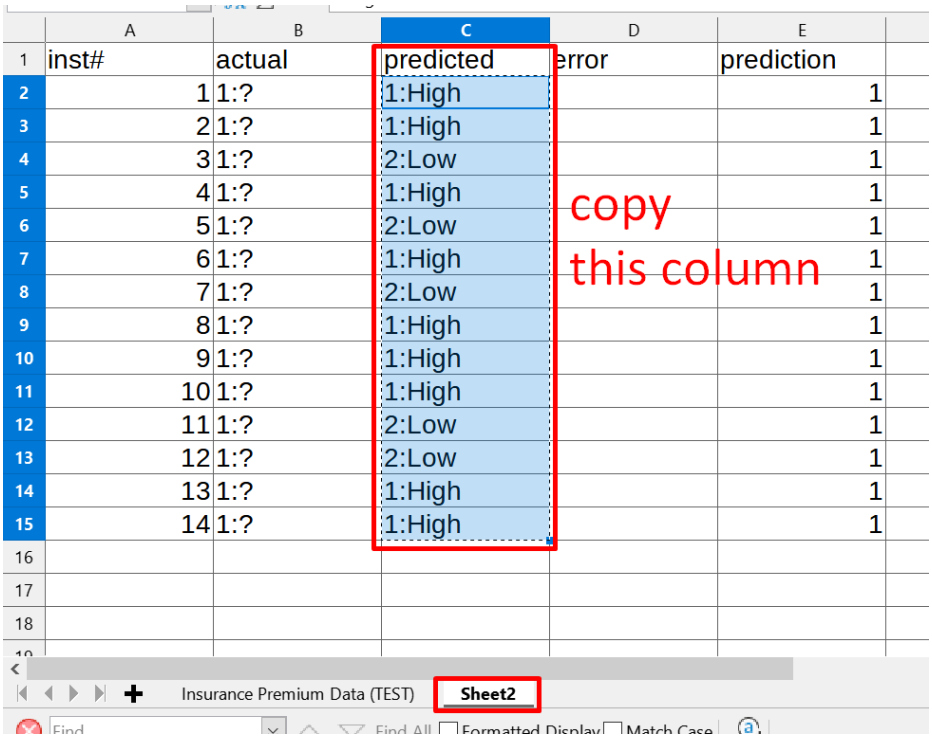


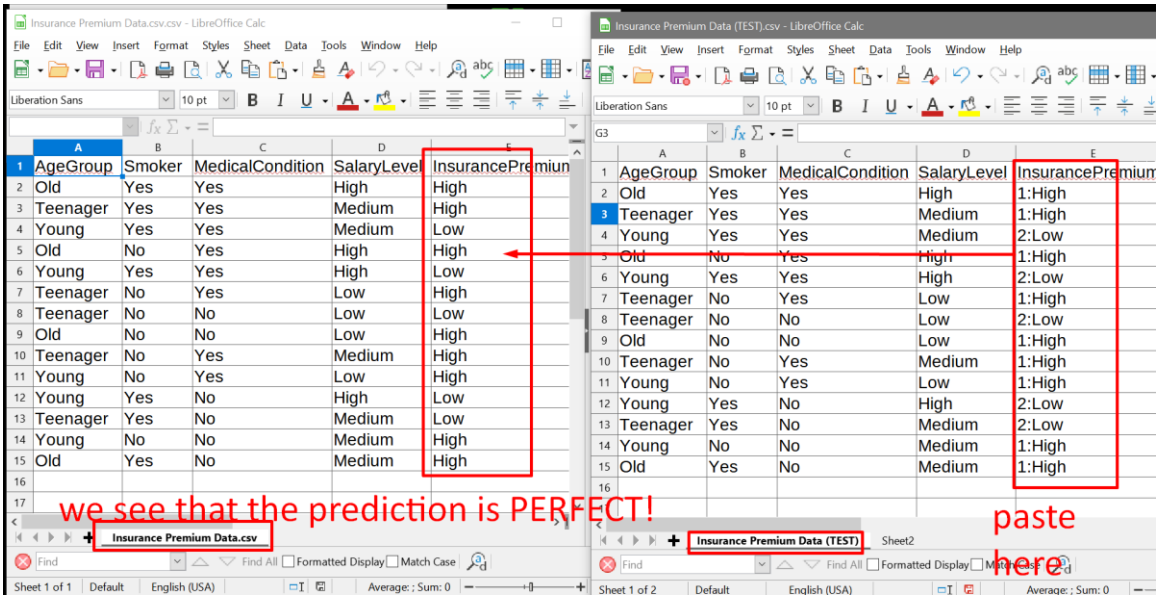












III. CONCLUSION

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.