

DR. ALVIN'S PUBLICATIONS

RANDOM FOREST (CLASSIFICATION)

USING WEKA
DR. ALVIN ANG



1 | PAGE

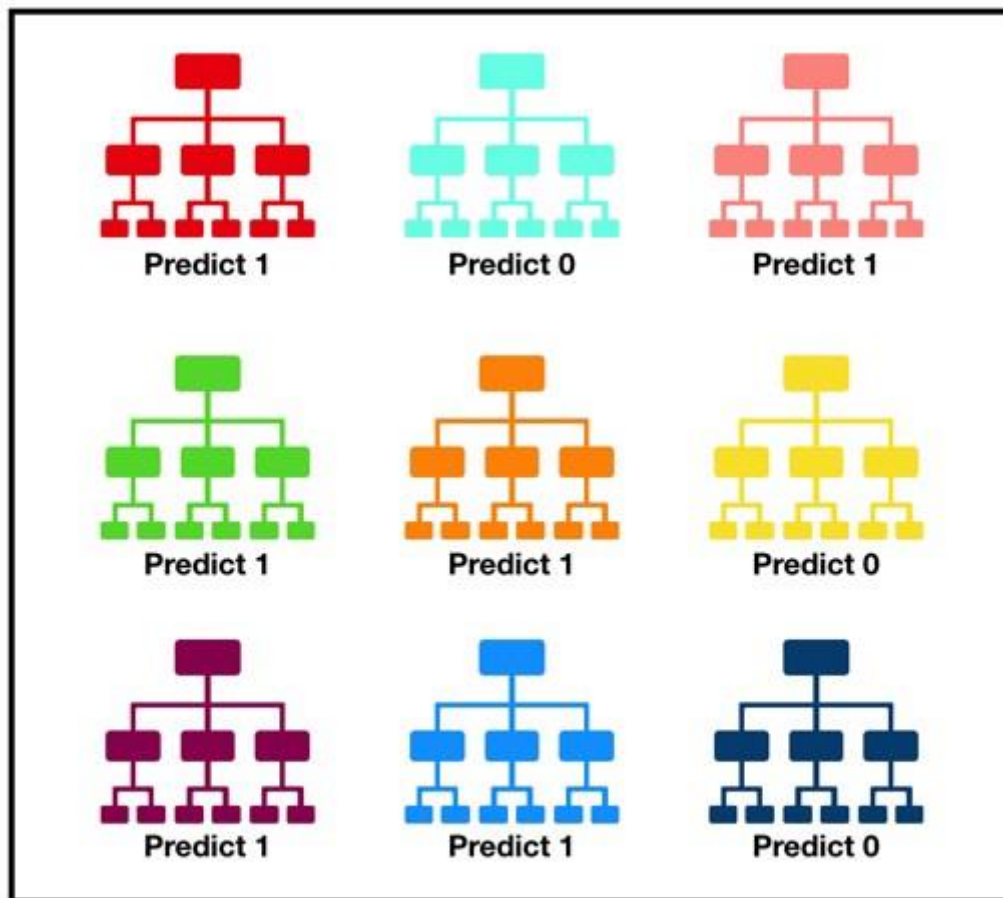
COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

CONTENTS

I. How Do Random Forest Work?	3
II. How are Random Forests Created?	4
A. What is Bagging?	4
B. Step 1: Bootstrapping	5
1. Step 1a: Creating ONE Decision Tree from ONE Bootstrapped Dataset.....	7
C. Step 2: Aggregating	8
III. Random Forest with WEKA	9
A. Step 1: Getting the Dataset	9
B. Step 2: Choosing Random Forest Classifier	10
C. Step 3: Tuning the Hyperparameters	11
D. Step 4: Checking the Random Forest Accuracy	13
E. Step 5: Predict with a New Dataset	14
About Dr. Alvin Ang	19

I. HOW DO RANDOM FOREST WORK?

- Random Forest = an Ensemble of “Trees” (Decision Trees).
- <https://www.alvinang.sg/s/Decision-Tree-Classification-using-WEKA-by-Dr-Alvin-Ang.pdf> → we’ve already learnt how Decision Trees are formed.
- Picture below shows an example of a “Random Forest”.... Made up of multiple Decision Trees.
- Ultimately, the Prediction is made from “Aggregating” all the Trees together.



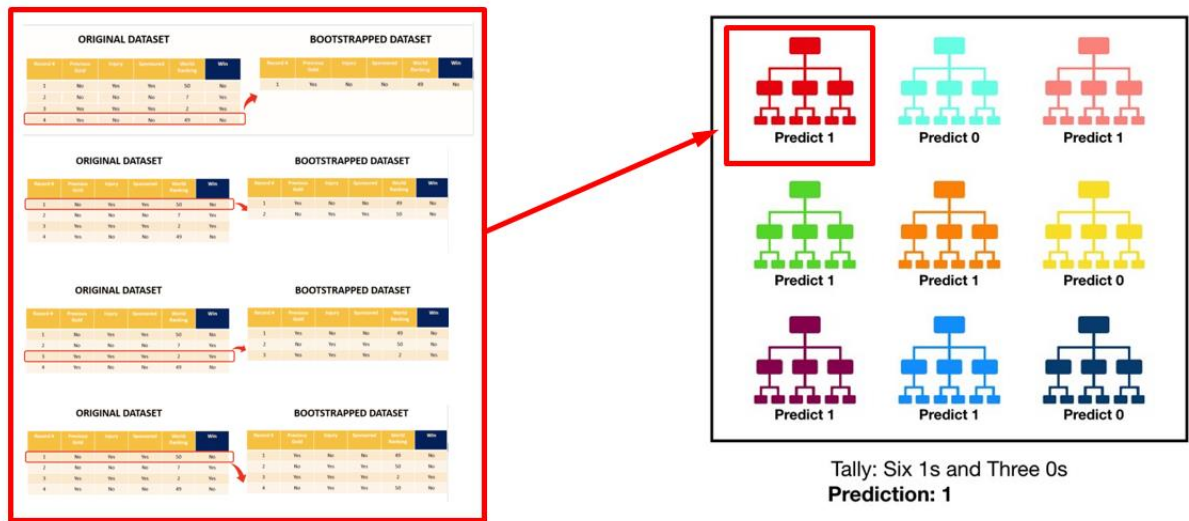
Tally: Six 1s and Three 0s
Prediction: 1

II. HOW ARE RANDOM FORESTS CREATED?

A. WHAT IS BAGGING?

- Random Forests are created through a process called Bagging...

Step 1 → **Bagging**
= 'B' - 'Agging'
= 'Bootstrapping' + 'Aggregating' → Step 2



B. STEP 1: BOOTSTRAPPING

Bootstrapping = Sampling with Replacement

ORIGINAL DATASET						BOOTSTRAPPED DATASET					
Record #	Previous Gold	Injury	Sponsored	World Ranking	Win	Record #	Previous Gold	Injury	Sponsored	World Ranking	Win
1	No	Yes	Yes	50	No						
2	No	No	No	7	Yes						
3	Yes	Yes	Yes	2	Yes						
4	Yes	No	No	49	No						

We are going to Sample this Original Dataset **with replacement**

- Bootstrapping (I believe) means the normal way of “strapping your boots” with your shoelaces.
- You sequentially take out one row of data (from the left side) and insert / strap it to the right side.
- But this could be done ‘randomly’ as opposed to an ordered / sequential manner.

Let’s say we decide to randomly pick the last row first....

ORIGINAL DATASET						BOOTSTRAPPED DATASET					
Record #	Previous Gold	Injury	Sponsored	World Ranking	Win	Record #	Previous Gold	Injury	Sponsored	World Ranking	Win
1	No	Yes	Yes	50	No	1	Yes	No	No	49	No
2	No	No	No	7	Yes						
3	Yes	Yes	Yes	2	Yes						
4	Yes	No	No	49	No						

Subsequently, we randomly choose the 1st row to “bootstrap” to the 2nd sample

ORIGINAL DATASET						BOOTSTRAPPED DATASET					
Record #	Previous Gold	Injury	Sponsored	World Ranking	Win	Record #	Previous Gold	Injury	Sponsored	World Ranking	Win
1	No	Yes	Yes	50	No	1	Yes	No	No	49	No
2	No	No	No	7	Yes	2	No	Yes	Yes	50	No
3	Yes	Yes	Yes	2	Yes						
4	Yes	No	No	49	No						

Next, we randomly choose the 3rd row to “bootstrap” to the 3rd sample

ORIGINAL DATASET						BOOTSTRAPPED DATASET					
Record #	Previous Gold	Injury	Sponsored	World Ranking	Win	Record #	Previous Gold	Injury	Sponsored	World Ranking	Win
1	No	Yes	Yes	50	No	1	Yes	No	No	49	No
2	No	No	No	7	Yes	2	No	Yes	Yes	50	No
3	Yes	Yes	Yes	2	Yes	3	Yes	Yes	Yes	2	Yes
4	Yes	No	No	49	No						

Somehow, lastly, we randomly still choose back the 1st row to “bootstrap” to the 4th sample

ORIGINAL DATASET						BOOTSTRAPPED DATASET					
Record #	Previous Gold	Injury	Sponsored	World Ranking	Win	Record #	Previous Gold	Injury	Sponsored	World Ranking	Win
1	No	Yes	Yes	50	No	1	Yes	No	No	49	No
2	No	No	No	7	Yes	2	No	Yes	Yes	50	No
3	Yes	Yes	Yes	2	Yes	3	Yes	Yes	Yes	2	Yes
4	Yes	No	No	49	No	4	No	Yes	Yes	50	No

Notice that we didn’t touch the 2nd row at all!

This is possible in Bootstrapping since its purely random!

In essence, we have created a “Bootstrapped” Dataset FROM the “Original” Dataset....

1. STEP 1A: CREATING ONE DECISION TREE FROM ONE BOOTSTRAPPED DATASET...

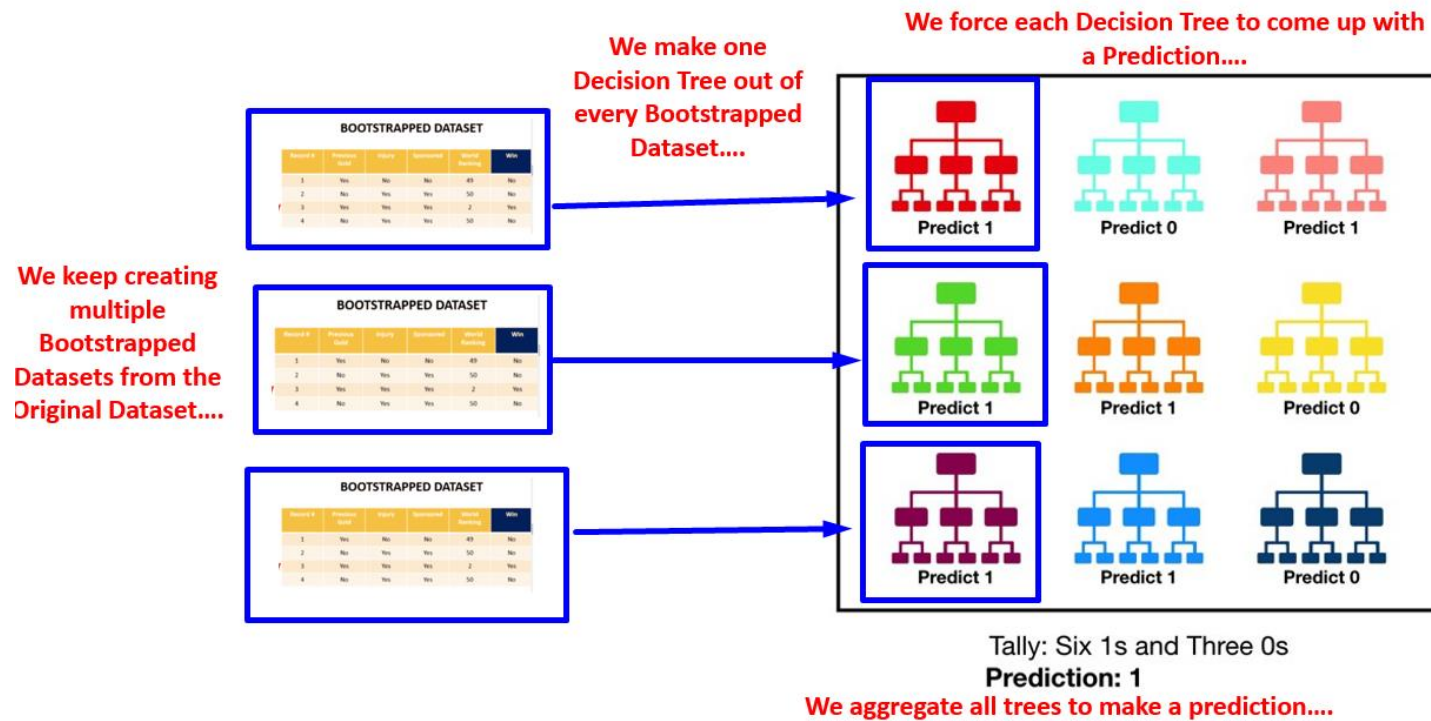
Creating ONE Decision Tree from a Bootstrapped Dataset....

- Now that we have created our 1st Bootstrapped Dataset, we use it to identify the IMPORTANT KEY Features
- Then, we create ONE Decision Tree out of it.
- This process of Creating a Decision Tree out from a dataset has already been explained here....<https://www.alvinang.sg/s/Decision-Tree-Classification-using-WEKA-by-Dr-Alvin-Ang.pdf>



C. STEP 2: AGGREGATING

This is the Process of Aggregating...



Note: Each “Tree” is also known as ONE “Bag” or ONE “Classifier”

III. RANDOM FOREST WITH WEKA

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

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

Random Forest continues off from Decision Tree..... here....

<https://www.alvinang.sg/s/Decision-Tree-Classification-using-WEKA-by-Dr-Alvin-Ang.pdf>

A. STEP 1: GETTING THE DATASET

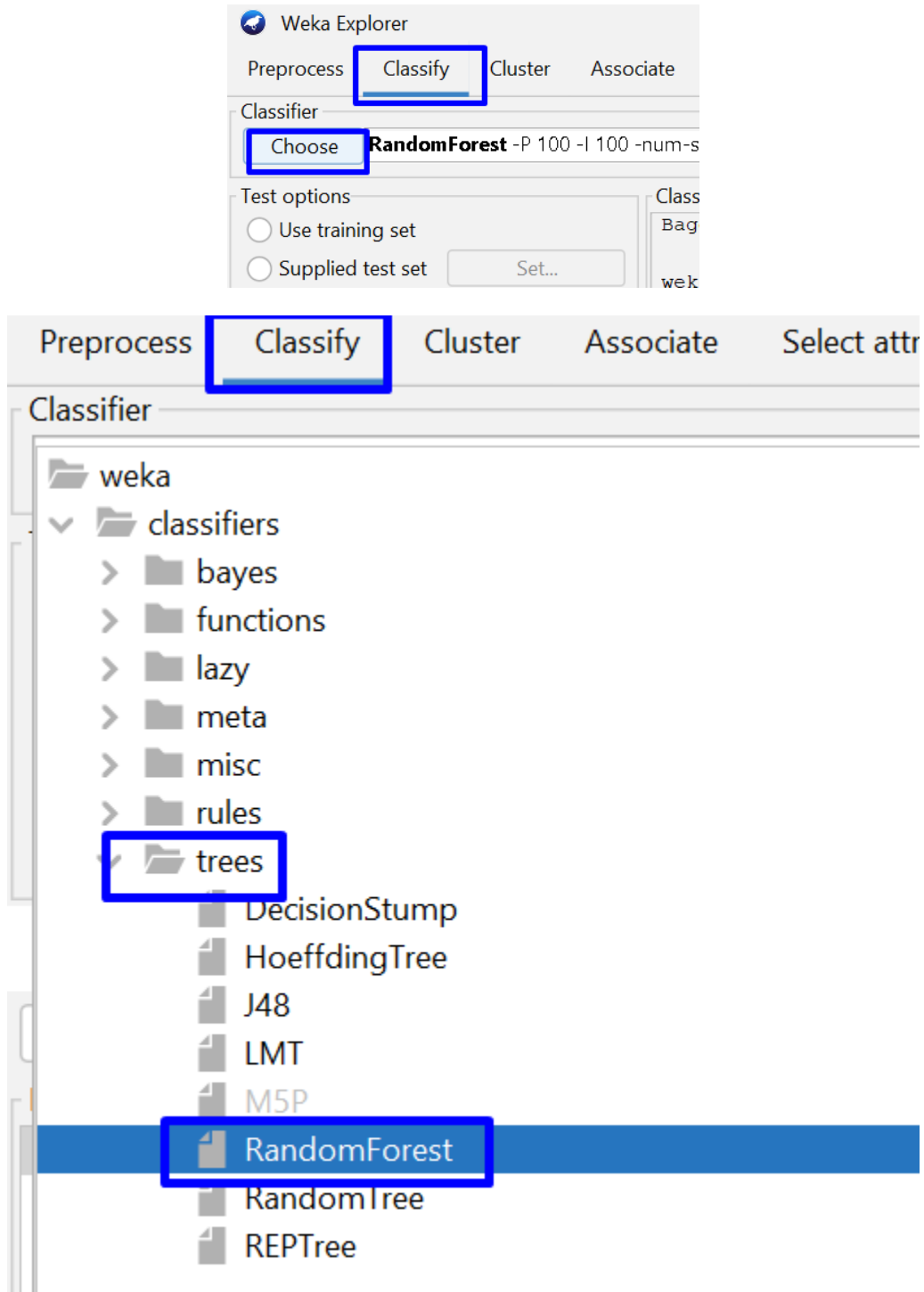
The screenshot shows the Weka GUI with the 'Open file...' dialog box open. The dialog box displays the desktop contents, and the file 'Insurance Premium Data.csv' is selected. The background shows the Weka Explorer interface with a table of data and a bar chart. The table has columns 'No.', 'Label', and 'Count'. The bar chart shows two bars, one blue and one red, representing the distribution of the 'Label' variable.

No.	Label	Count
1	Age Group	4
2	Teenager	5
3	Young	5

Class: InsurancePremium (Nom)

we will be using back the same dataset that was used to create our Decision Tree...

B. STEP 2: CHOOSING RANDOM FOREST CLASSIFIER



C. STEP 3: TUNING THE HYPERPARAMETERS

double click the box

% size of EACH BOOTSTRAPPED dataset
100% means the new Bootstrapped Dataset will be SAME SIZE as the Original Dataset
Say Dataset has 100 rows
50% means new Bootstrapped Dataset will only be half the size (50 rows) of the Original Dataset

represents Number of Trees or Number of Bags
just leave it as default of 100 trees
of course, the more trees the higher the accuracy

weka.classifiers.trees.RandomForest

Capabilities

bagSizePercent 100

batchSize 100

breakTiesRandomly False

calcOutOfBag False

computeAttributeImportance False

debug False

doNotCheckCapabilities False

maxDepth 0

numDecimalPlaces 2

numExecutionSlots 1

numFeatures 0

numIterations 100

outputOutOfBagComplexityStatistics False

printClassifiers False

seed 1

storeOutOfBagPredictions False

Open... Save... OK Cancel

Log x 0

represents the No. of Features / Columns to consider at each Split

e.g. the Insurance Dataset has 5 columns (5 Features)

4 Variables/Columns predicting the Target (Insurance Premium)

if numFeatures = 2, means even if there are more important Features to be split, it will only pick the top 2 at each node

by default is left at 0 meaning, it will automatically select the right number of features to split

PRC Area	Class
0.898	High
0.814	Low
0.868	Low

D. STEP 4: CHECKING THE RANDOM FOREST ACCURACY

Random Forest wins Decision Tree in terms of Accuracy!!!!

Weka Explorer Classifier: Choose **RandomForest** P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Test options: Use training set, Supplied test set, Cross-validation Folds: 10, Percentage split %: 66

Classifier output:

```

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-
Time taken to build model: 0.05 seconds
==== Stratified cross-validation ====
==== Summary ====
Correctly Classified Instances      11      78.5714 %
Incorrectly Classified Instances     3      21.4286 %
Kappa statistic                    0.5116
Mean absolute error                 0.3553
Root mean squared error            0.3984
Relative absolute error             74.6181 %
Root relative squared error        80.7461 %
Total Number of Instances         14
==== Detailed Accuracy By Class ====
          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC
          0.889   0.400   0.800   0.889   0.842   0.
          0.600   0.111   0.750   0.600   0.667   0.
Weighted Avg.   0.786   0.297   0.782   0.786   0.779   0.
==== Confusion Matrix ====

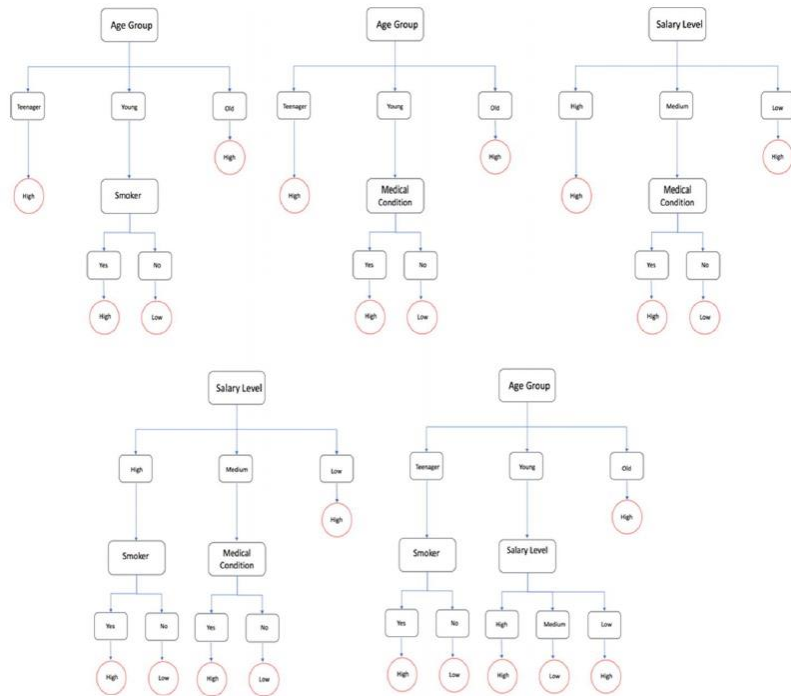
```

Single Decision Tree Accuracy:

```

Classifier output
MedicalCondition = yes: High (3.0)
MedicalCondition = No: Low (2.0)
Group = Young
Smoker = Yes: Low (3.0)
Smoker = No: High (2.0)
Number of Leaves : 5
Number of the tree : 8
Time taken to build model: 0.01 seconds
==== Stratified cross-validation ====
==== Summary ====
Correctly Classified Instances      6      42.8571 %
Incorrectly Classified Instances     8      57.1429 %
Kappa statistic                    -0.1429
Mean absolute error                 0.4881
Root mean squared error            0.6307
Relative absolute error             102.5 %
Root relative squared error        127.8423 %
Total Number of Instances         14
==== Detailed Accuracy By Class ====
          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC
          0.444   0.600   0.571   0.444   0.500  -0.14
          0.400   0.556   0.286   0.400   0.333  -0.14
Weighted Avg.   0.429   0.584   0.469   0.429   0.440  -0.14

```



Above shows a possible Random Forest made up of multiple trees.

E. STEP 5: PREDICT WITH A NEW DATASET

Go here and download the data again:

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

insurance Premium Data (TEST).csv - LibreOffice Calc

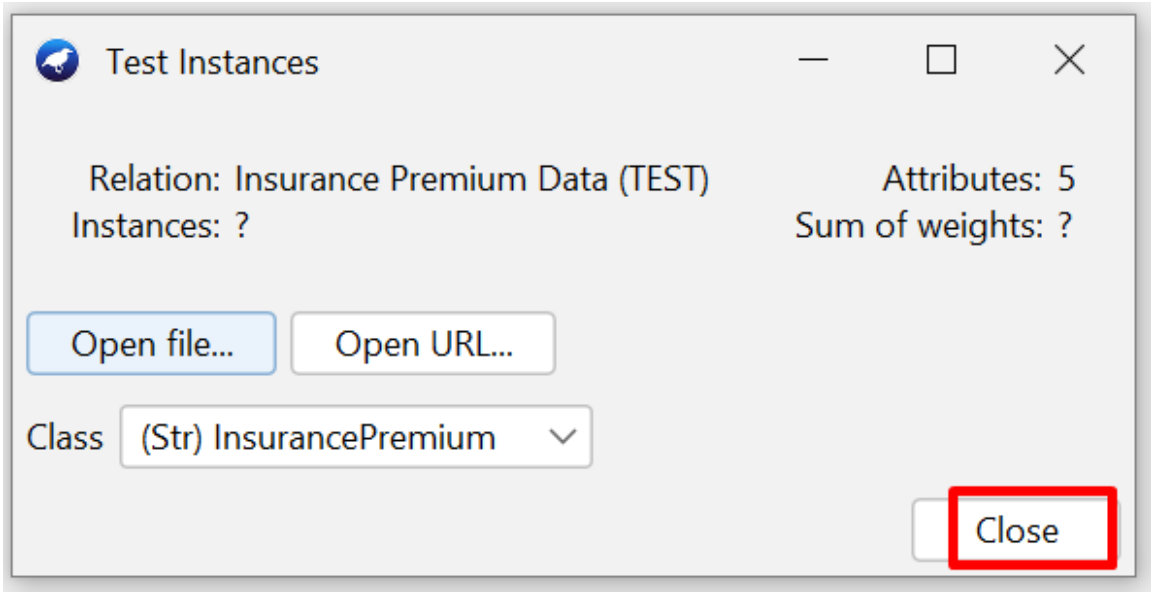
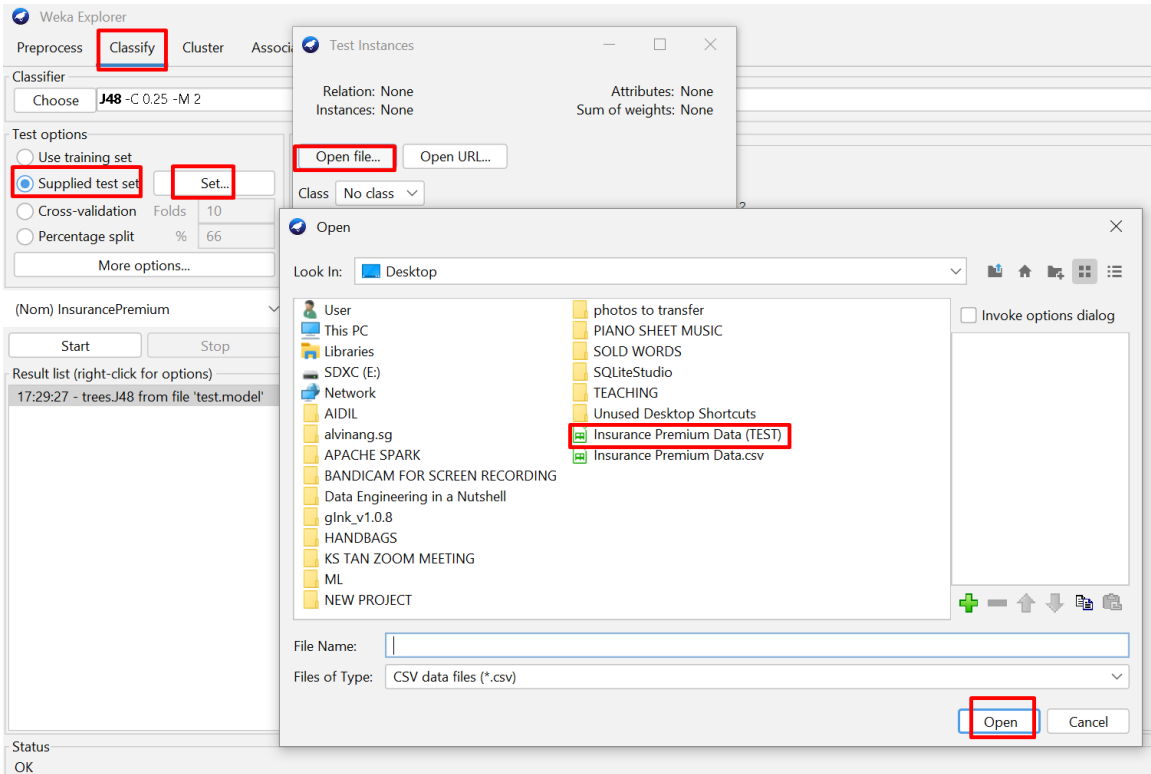
	A	B	C	D	E	F	G	H
1	AgeGroup	Smoker	MedicalCondition	SalaryLevel	InsurancePremium			
2	Old	Yes	Yes	High				
3	Teenager	Yes	Yes	Medium				
4	Young	Yes	Yes	Medium				
5	Old	No	Yes	High				
6	Young	Yes	Yes	High				
7	Teenager	No	Yes	Low				
8	Teenager	No	No	Low				
9	Old	No	No	Low				
10	Teenager	No	Yes	Medium				
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								
19								

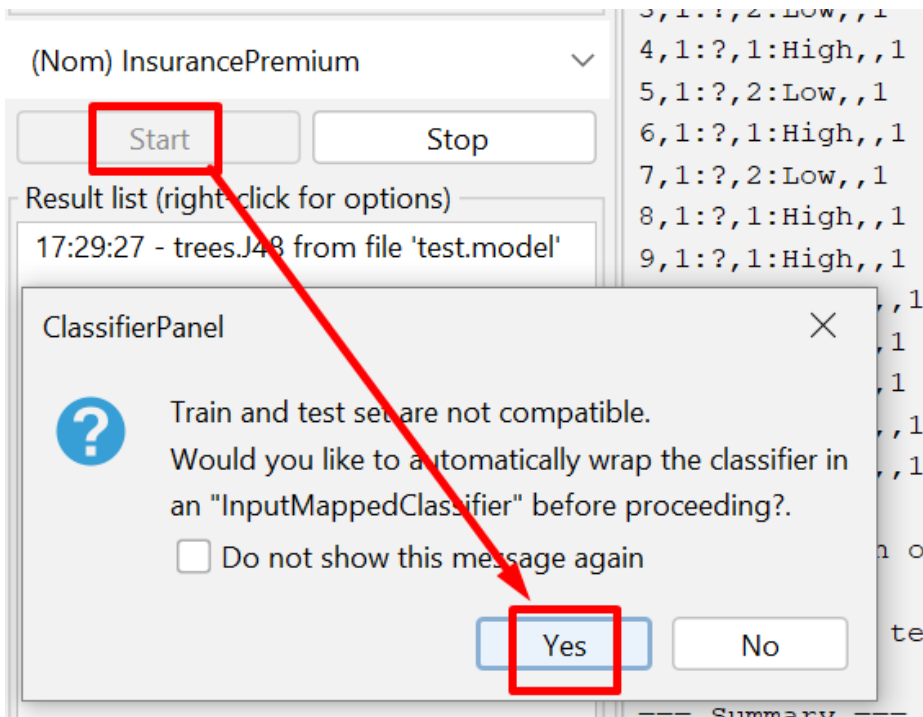
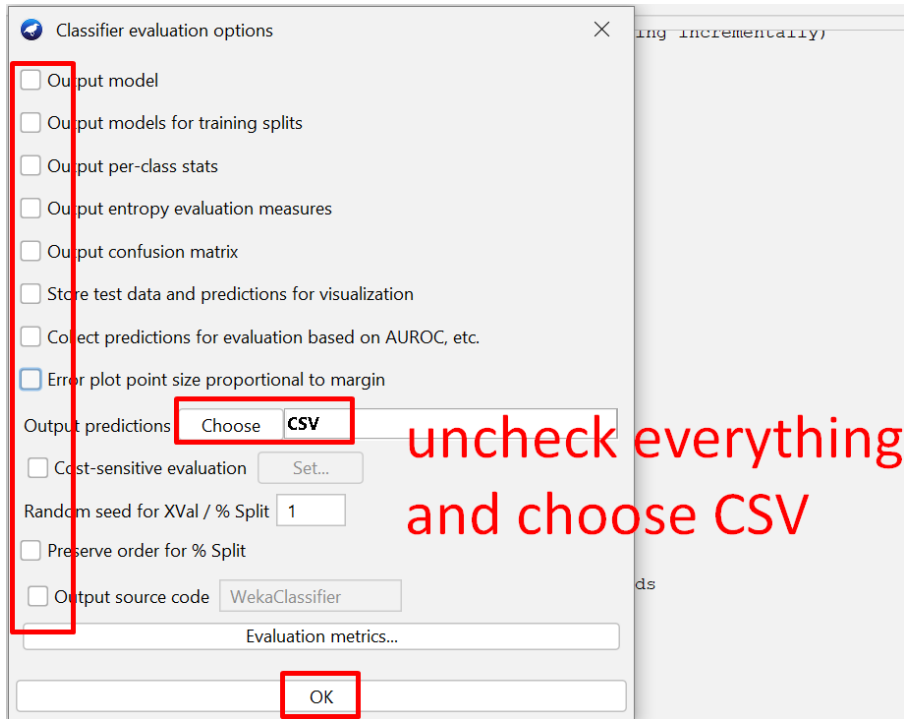
save it as a new file

delete away the data in InsurancePremium column to simulate a new Dataset....

later, predictions will enter here

Insurance Premium Data (TEST)





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.l48 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
```

=== 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

select and CTRL + C

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
16					
17					
18					

copy
this column

we see that the prediction is PERFECT!

paste
here

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.