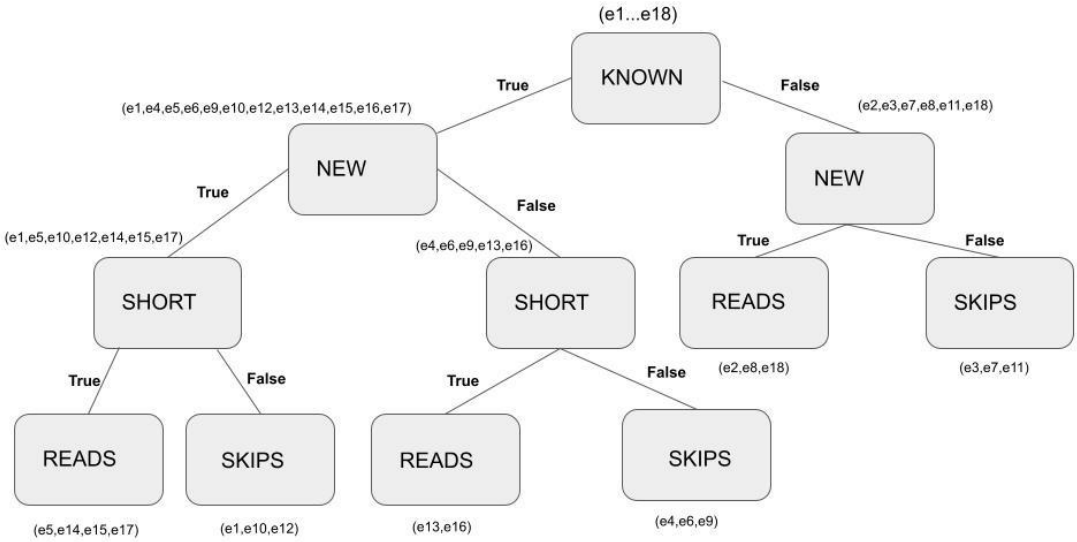
Z5260130

COMP3411 Assignment 2 Roy Chen

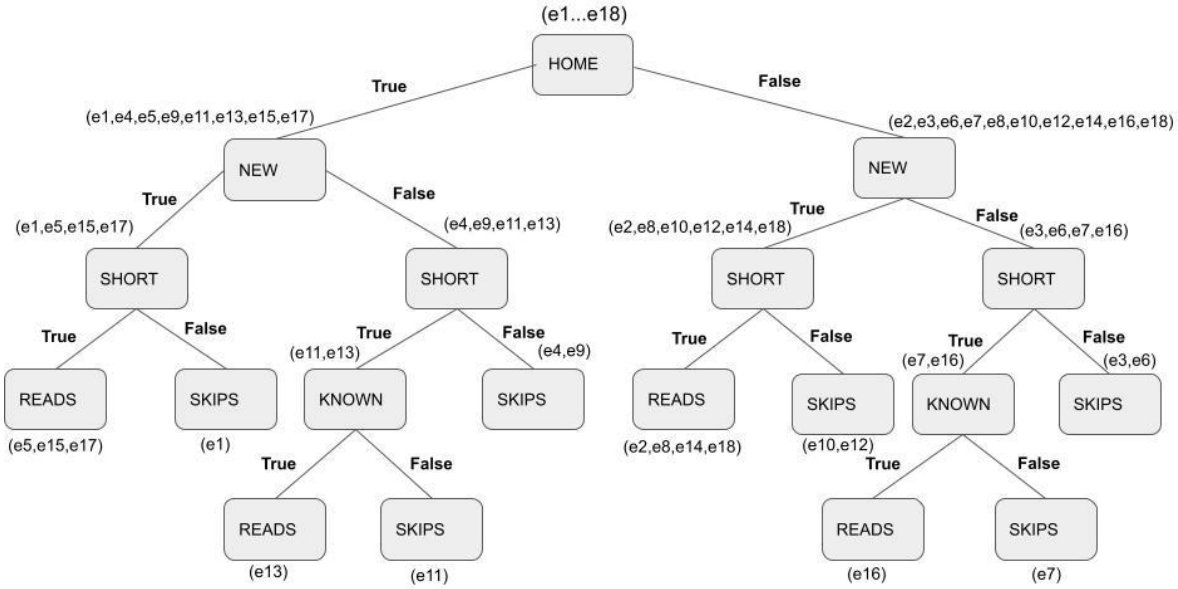
Question 1.1



A) The tree derived from the algorithm represents a different function than that found with the maximum information gain split. This can be determined by analysing the data provided on the table, we can see that there was 6 duplicated results {(e5,e14,e15,e17), (e1,e10,e12), (e13,e16), (e4,e6,e9), (e2,e8,e18), (e3,e7,e11)}. As there are 6 duplicated results out of the 18 provided in the table, this means that there were only 12 unique combinations. As the maximum possible combinations is 16 (2^4) and how we can see that the 12 combinations obtained from the maximum information gain split algorithm we can safely assume that these 12 combinations are unique and would produce equivalent outputs. In order for the two functions to be the same we must determine if the remaining test cases produce equal results. From the tables down below, the first table contains the 4 test cases we are testing and is derived from the decision tree shown above. The second table shows the results of the maximum information gain split algorithm. Evidently they yield different answers hence we can clearly say they represent different functions.

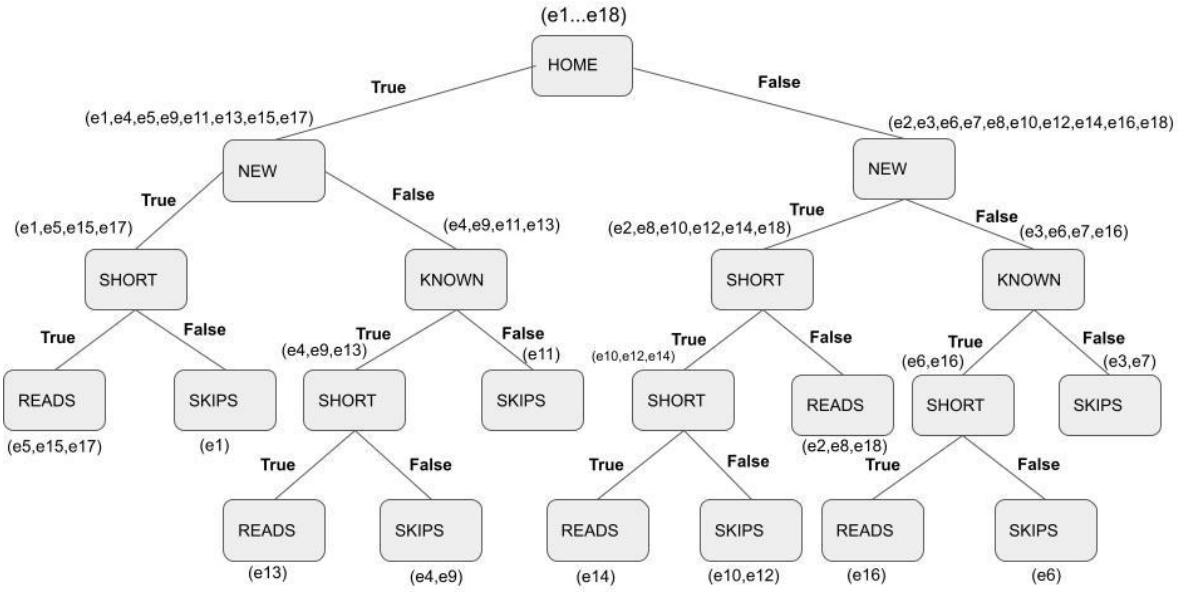
Case	Author	Thread	Length	Where_read	User_action
1	unknown	new	long	home	reads
2	unknown	new	long	work	reads
3	unknown	new	short	home	reads
4	unknown	followup	long	home	skips

Case	Author	Thread	Length	Where_read	User_action
1	unknown	new	long	home	Skips
2	unknown	new	long	work	Skips
3	unknown	new	short	home	Reads
4	unknown	followup	long	home	skips



B) In order to determine if the tree for the new order represents a similar function to the maximum information gain split or the one in the preceding part, we will test the remaining 4 unique inputs as we did in part a. As seen from the results derived by the new tree in the table below and comparing it to the previous results it is evident that the new decision tree produces the same outputs as the maximum information gain split function. The new tree also provides different results when compared the tree in 1(A). In conclusion from the equivalent results with the maximum information gain tree and differences with the tree in 1(A) we can say that these two decision trees indeed do represent the same function.

Case	Author	Thread	Length	Where_read	User_action
1	unknown	new	long	home	Skips
2	unknown	new	long	work	Skips
3	unknown	new	short	home	Reads
4	unknown	followup	long	home	skips



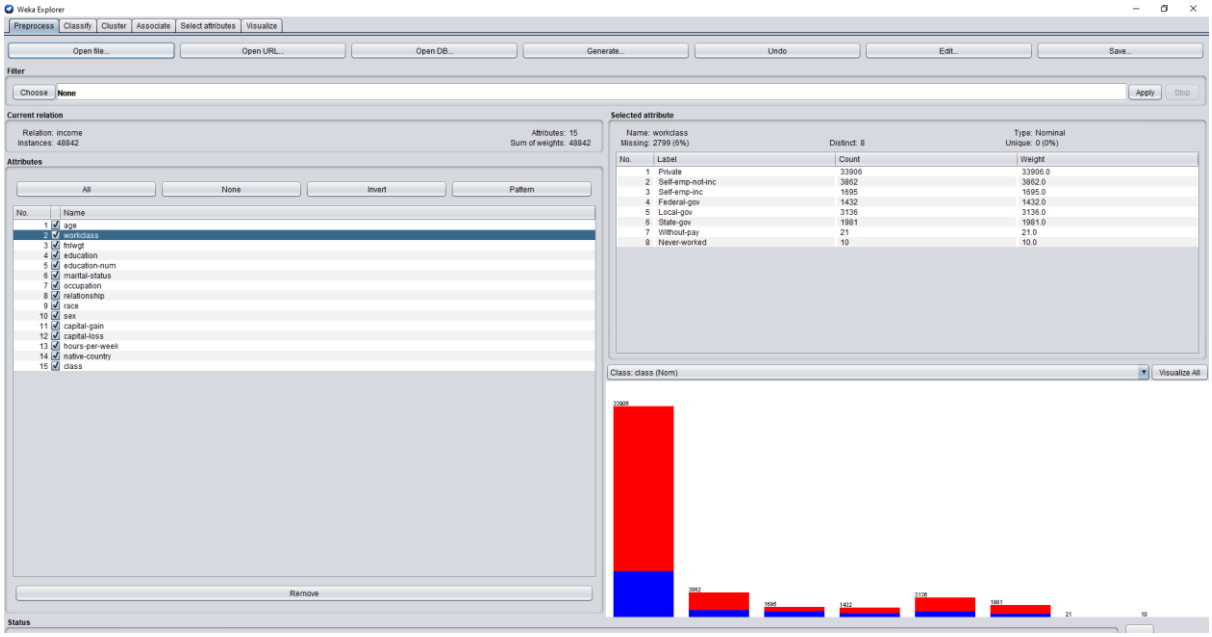
C) In order for a tree to correctly classify the training examples whilst representing a different function than those found by the preceding algorithms needs to satisfy the 12 unique inputs provided whilst providing different outputs for the 4 cases remaining cases. This can be achieved through the order of: [Where_read, Thread, Author, Length].

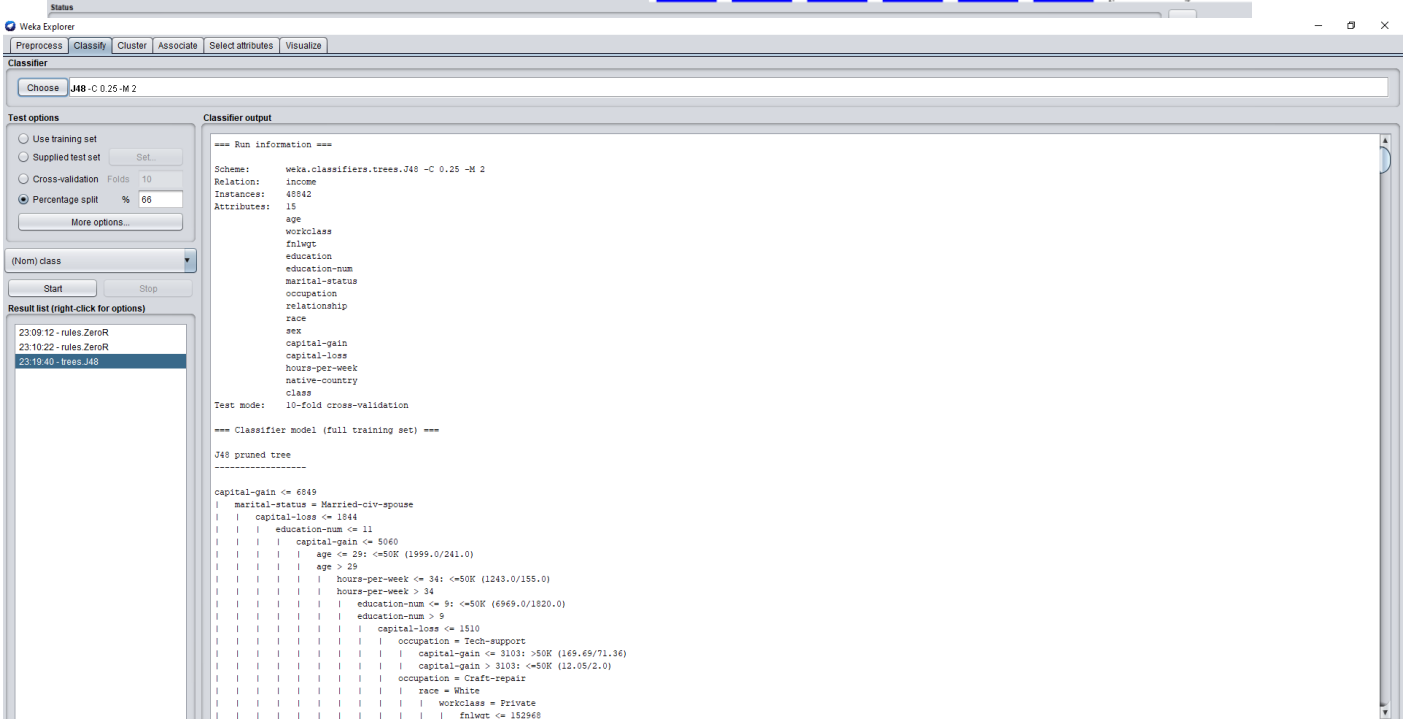
After analysing the results deduced in the table below, we can see that the outputs do not match with any of the preceding outputs by the trees. In conclusion, it is clear that the tree correctly encompasses the training examples whilst it represents a different function.

Case	Author	Thread	Length	Where_read	User_action
1	unknown	new	long	home	skips
2	unknown	new	long	work	reads
3	unknown	new	short	home	reads
4	unknown	followup	long	home	skips

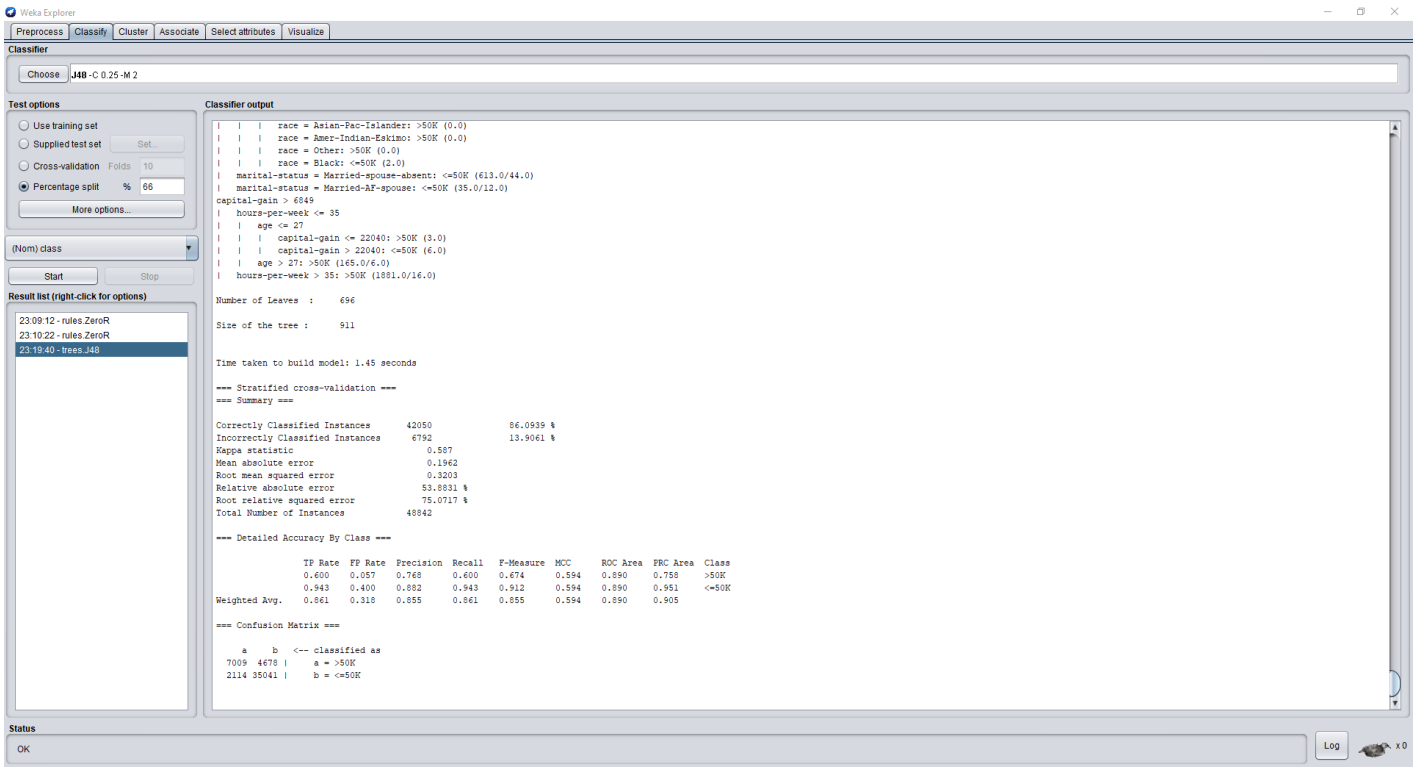
Z5260130

Question 1.2





Z5260130


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
23:09:12 - rules.ZeroR
23:10:22 - rules.ZeroR
23:19:40 - trees.J48
Classifier output
| | | race = Asian-Pac-Islander: >50K (0.0)
| | | race = Amer-Indian-Eskimo: >50K (0.0)
| | | race = Other: >50K (0.0)
| | | race = Black: <=50K (2.0)
| marital-status = Married-spouse-absent: <=50K (613.0/44.0)
| marital-status = Married-AF-spouse: <=50K (35.0/12.0)
capital-gain > 6849
| hours-per-week <= 35
| | age <= 27
| | | capital-gain <= 22040: >50K (3.0)
| | | capital-gain > 22040: <=50K (6.0)
| | age > 27: >50K (165.0/6.0)
| hours-per-week > 35: >50K (1881.0/16.0)
Number of Leaves : 696
Size of the tree : 911
Time taken to build model: 1.45 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 42050 86.0939 %
Incorrectly Classified Instances 6792 13.9061 %
Kappa statistic 0.587
Mean absolute error 0.1962
Root mean squared error 0.3203
Relative absolute error 53.8831 %
Root relative squared error 75.0717 %
Total Number of Instances 48842
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.600 0.057 0.768 0.600 0.674 0.594 0.890 0.758 >50K
0.943 0.400 0.882 0.943 0.912 0.594 0.890 0.951 <=50K
Weighted Avg. 0.861 0.318 0.855 0.861 0.855 0.594 0.890 0.905
=== Confusion Matrix ===
a b <-- classified as
7009 4678 | a = >50K
2114 35041 | b = <=50K
Status
OK
Log
x 0