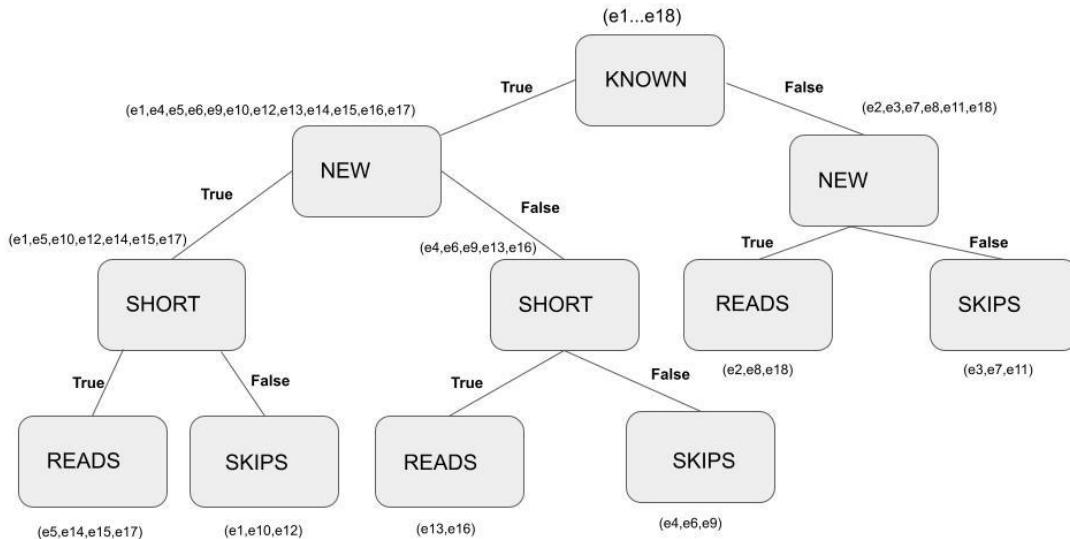


COMP3411 Assignment 2 Roy Chen

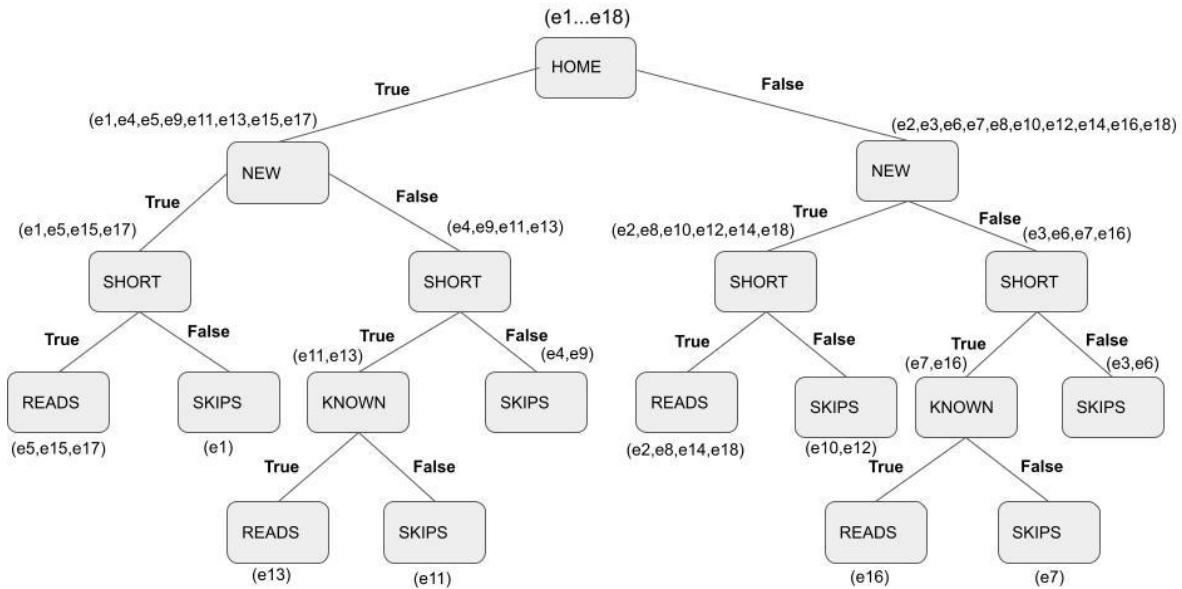
Question 1.1



- 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 were 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.

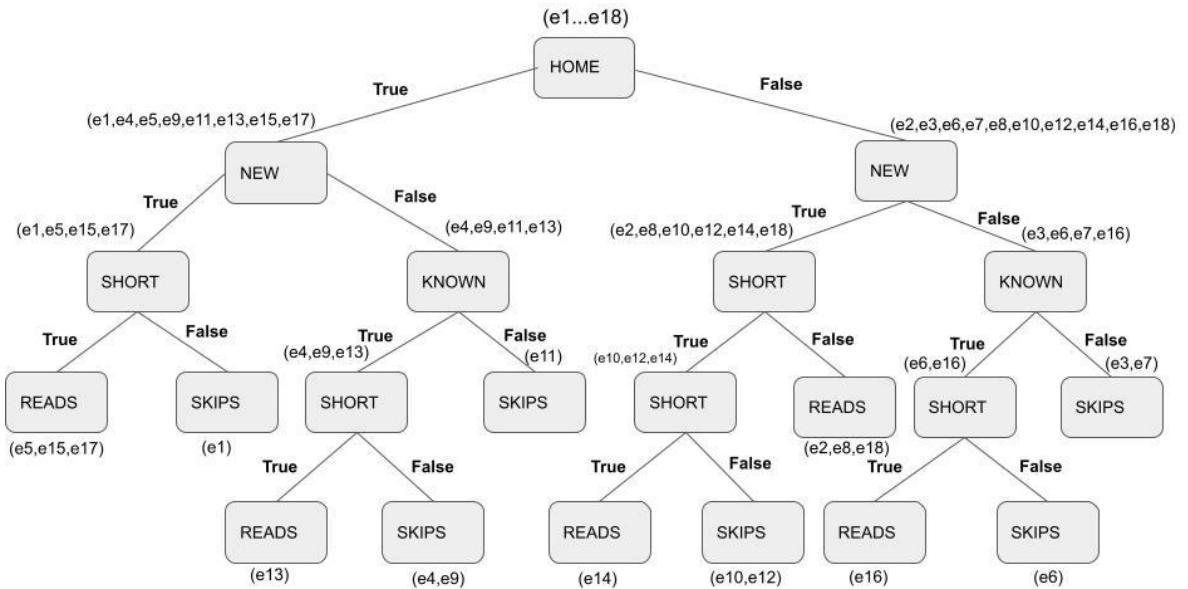
Case	Author	Thread	Length	Where_read	User_action
1	unknown	new	long	home	reads
2	unknown	new	long	work	reads
3	unknown	new	short	home	reads
4	unknown	followup	long	home	skips

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

Question 1.2

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

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

Filter Choose None Apply Stop

Current relation Relation: income Instances: 48842 Attributes: 15 Sum of weights: 49842

Attributes

All None Invert Pattern

No.	Name
1	age
2	workclass
3	fnlwgt
4	education
5	education-num
6	marital-status
7	occupation
8	relationship
9	race
10	sex
11	capital-gain
12	capital-loss
13	hours-per-week
14	native-country
15	class

Selected attribute Name: workclass Missing: 2799 (6%) Distinct: 8 Type: Nominal Unique: 0 (0%)

No.	Label	Count	Weight
1	Private	33906	33906.0
2	Self-emp-not-inc	3862	3862.0
3	Self-emp-inc	1895	1895.0
4	Local-gov	1432	1432.0
5	Local-gov	3136	3136.0
6	State-gov	1981	1981.0
7	Without-pay	21	21.0
8	Never-worked	10	10.0

Class: class (Nom) Visualize All

Status

Preprocess Classify Cluster Associate Select attributes Visualize

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)

- 23:00:12 -rules ZeroR
- 23:10:22 -rules ZeroR
- 23:19:40 -trees J48

Classifier output

```

*** Run information ***
Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: income
Instances: 48842
Attributes: 15
Nominal classes: 1

Test mode: 10-fold cross-validation

*** Classifier Model (full training set) ***
J48 pruned tree
-----
capital-gain <= 6849
| marital-status = Married-civ-spouse
| capital-loss <= 3844
| | education-num <= 11
| | | capital-gain <= 5060
| | | | age <= 29; <=50K (1999.0/241.0)
| | | | age > 29
| | | | | hours-per-week <= 34; <=50K (1243.0/155.0)
| | | | | hours-per-week > 34
| | | | | | education-num <= 9; <=50K (6969.0/1820.0)
| | | | | | | education-num > 9; <=50K (6969.0/1820.0)
| | | | | | | | capital-loss <= 1510
| | | | | | | | occupation = Tech-support
| | | | | | | | | capital-gain <= 3103; >50K (169.69/71.36)
| | | | | | | | | capital-gain > 3103; <=50K (12.05/2.0)
| | | | | | | | | occupation = Craft-repair
| | | | | | | | | race = White
| | | | | | | | | | workclass = Private
| | | | | | | | | | fnlwgt <= 152968

```

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 - nules.ZeroR
- 23:10:22 - nules.ZeroR
- 23:19:40 - trees.J48**

Number of Leaves : 696

Size of the tree : 911

Time taken to build model: 1.45 seconds

*** Stratified cross-validation ***

*** Summary ***

	Correctly Classified Instances	86.0939 %
Incorrectly Classified Instances	6792	13.9061 %
Hinges statistic		0.587
Mean absolute error		0.1963
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 ***

	TF 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	0.950	0.951		<=50K
Weighted Avg.	0.861	0.318	0.855	0.861	0.855	0.890	0.895		

*** Confusion Matrix ***

	a	b	<- classified as
7009	4679	1	a = >50K
2114	35041	1	b = <=50K

Status

OK

Log x 0