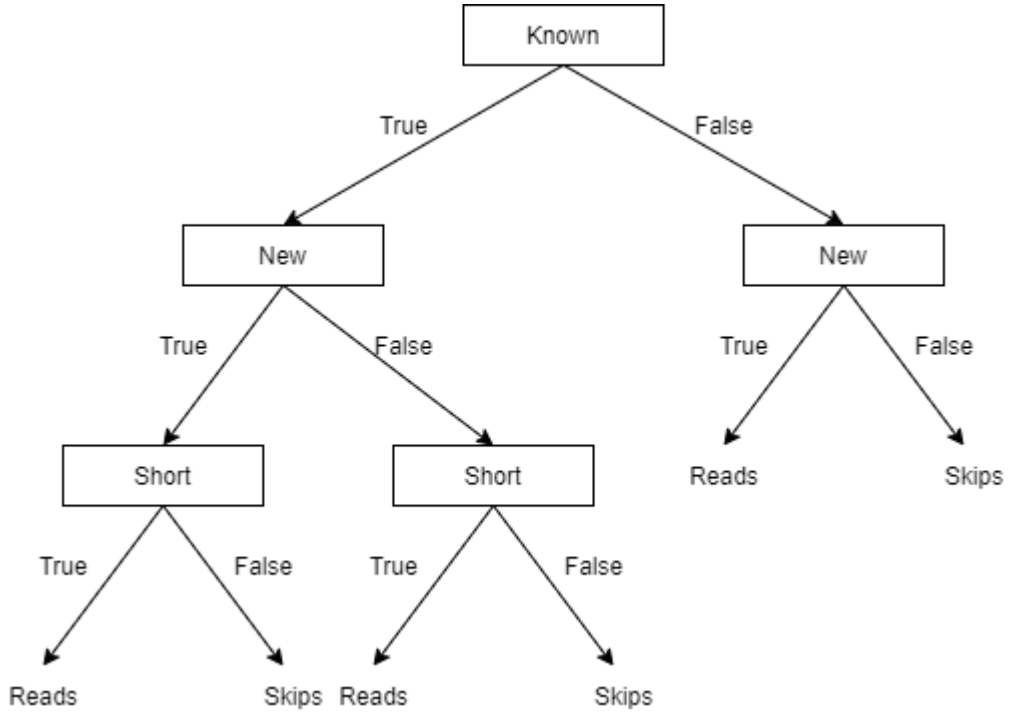
Part 1 – Decision Trees

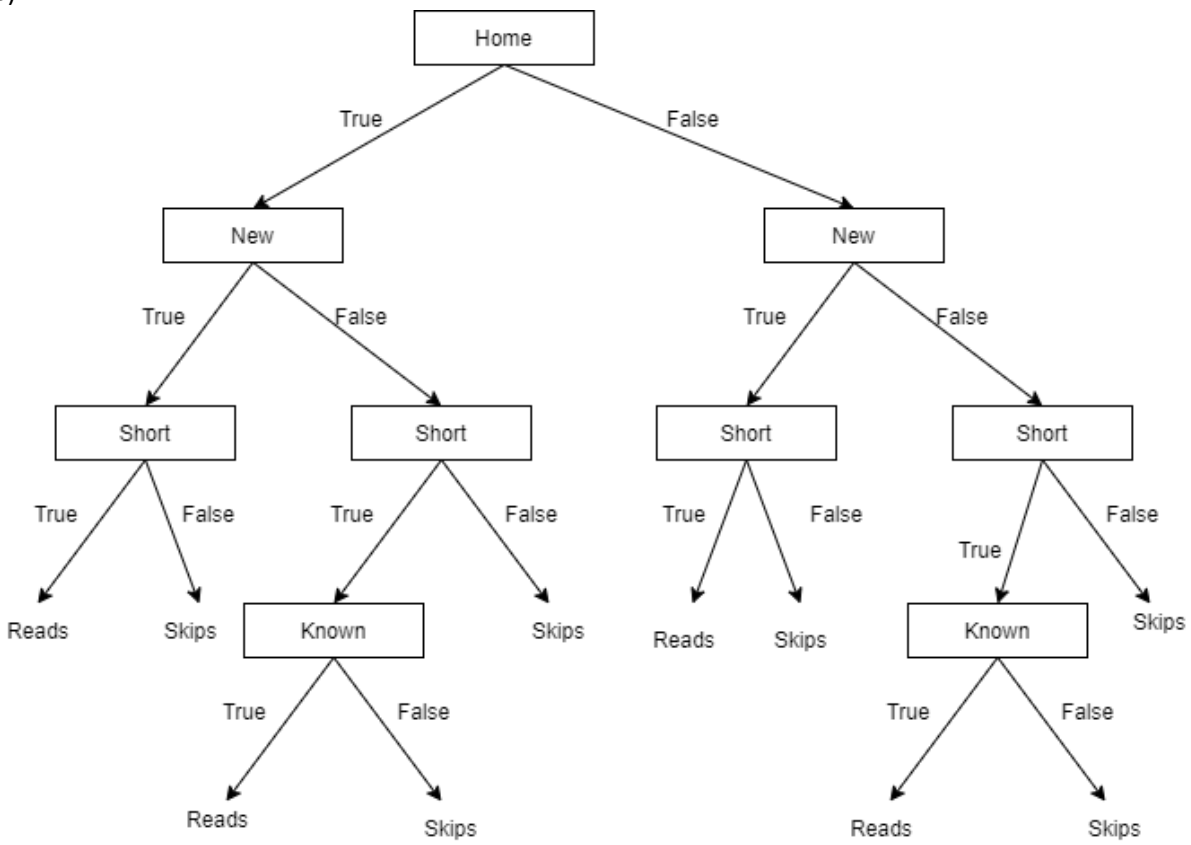
Question 1.1

a)



Yes, this tree represents a different function than the maximum information gain tree. For example: with case e19 (unknown, new, long, work) from figure 7.1 (the image containing the cases) – the maximum information gain tree predicts skip whereas the tree above predicts read.

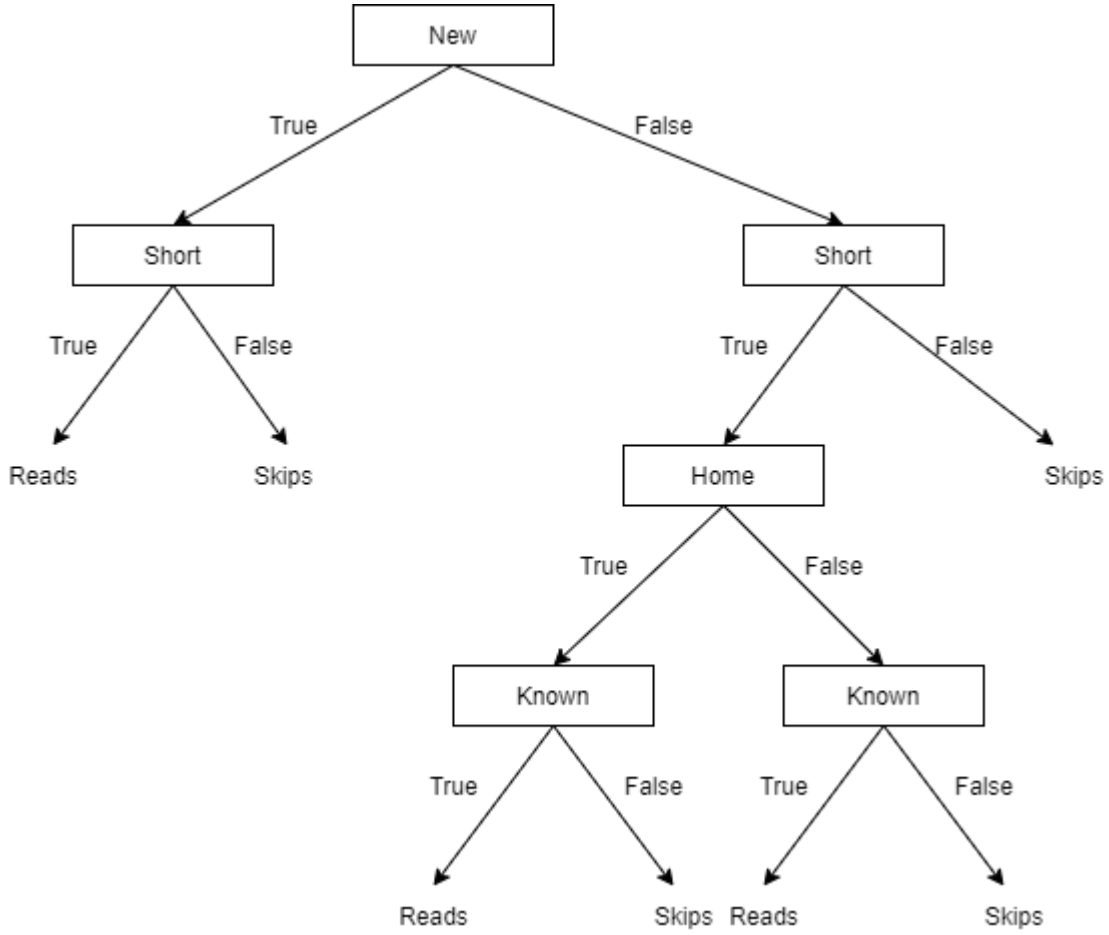
b)



No, this tree represents the same function than the maximum information gain tree. For the cases e19 and e20, this tree gives the same prediction as the maximum information gain tree. Since this tree is the same function as the maximum information gain tree – which is a different function to the tree in (a) – this tree is also a different function to the tree in part (a).

c) Yes, the tree below is a tree which represents a different function to the trees above and correctly classifies the training examples. The following tree is an example. This tree predicts e19 from the cases as skip whereas the tree from part (a) predicts read.

The tree is below:



Question 1.2 – Weka

Firstly, I converted adult.data into csv using excel. I then removed the quotation marks (“) at the start and end of every line. I then converted the csv file to arff using weka. I then edited the arff file to put in the correct attributes. I then created the tree using:

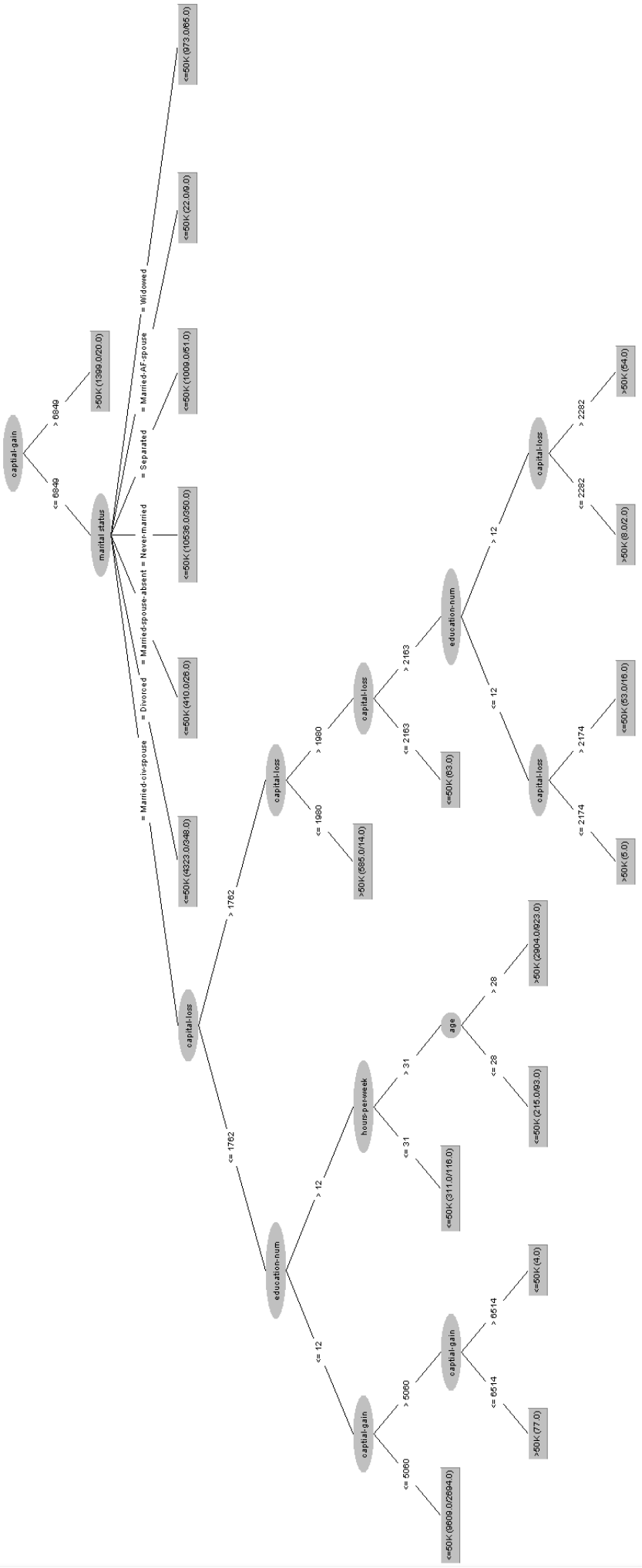
J48 with default settings (confidence factor = 0.25) and using cross validation with 10 folds
Confidence factor is the value used for pruning – the smaller this value the more pruning. With these settings a tree is produced with: 564 leaves, and size of tree 710. The accuracy rate – that is percentage of cases classified correctly is: 86.2469%

After this, I reduced the confidence factor to 0.1, 0.01, 0.001, 0.0001 and finally to 0.00001.

With confidence factor 0.00001 a tree is produced with 19 leaves and size of tree: 32. The accuracy rate is: 85.3655%

Now I have drastically reduced the size of the tree while only reducing the accuracy rate by less than 1%.

This tree is shown below:


capital-gain
> 6849
<= 6849
>50K (1399.0/20.0)
marital status
= Married-civ-spouse
= Divorced
= Never-married
= Separated
= Widowed
= Married-AF-spouse
<=50K (973.0/65.0)
<=50K (22.0/9.0)
<=50K (1009.0/51.0)
<=50K (10536.0/350.0)
<=50K (410.0/28.0)
<=50K (4323.0/348.0)
capital-loss
> 1762
<= 1762
capital-loss
> 1980
<= 1980
>50K (686.0/14.0)
capital-loss
> 2163
<= 2163
<=50K (63.0)
education-num
> 12
<= 12
capital-loss
> 2282
<= 2282
>50K (64.0)
>50K (8.0/2.0)
capital-loss
> 2174
<= 2174
<=50K (53.0/16.0)
>50K (5.0)
education-num
> 12
<= 12
hours-per-week
> 31
<= 31
<=50K (311.0/116.0)
age
> 28
<= 28
>50K (2904.0/923.0)
<=50K (215.0/93.0)
capital-gain
> 5060
<= 5060
<=50K (9609.0/2684.0)
capital-gain
> 6514
<= 6514
<=50K (4.0)
>50K (77.0)