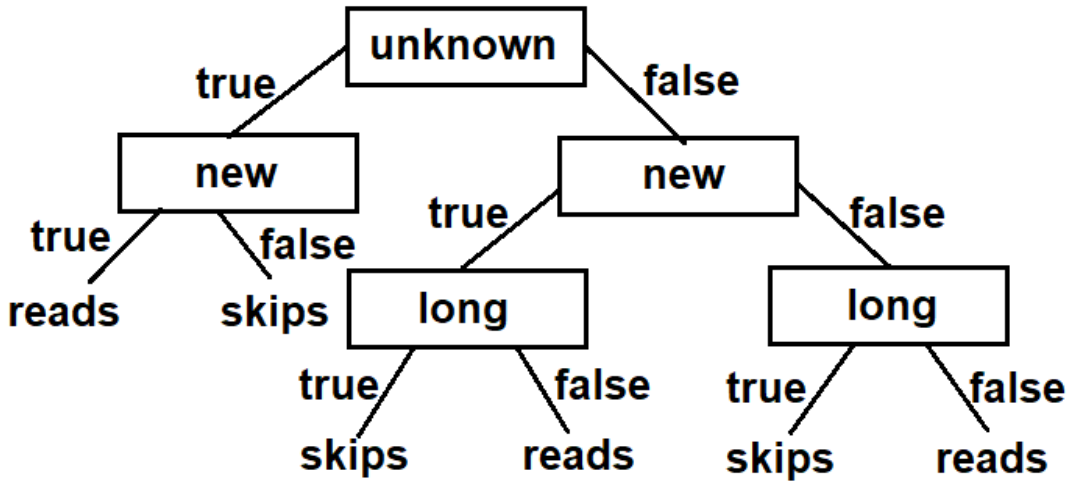
Assignment 2

Zhichao He z5282955

Question 1:

a)



They are different. The function with the maximum information gain split looks like this.

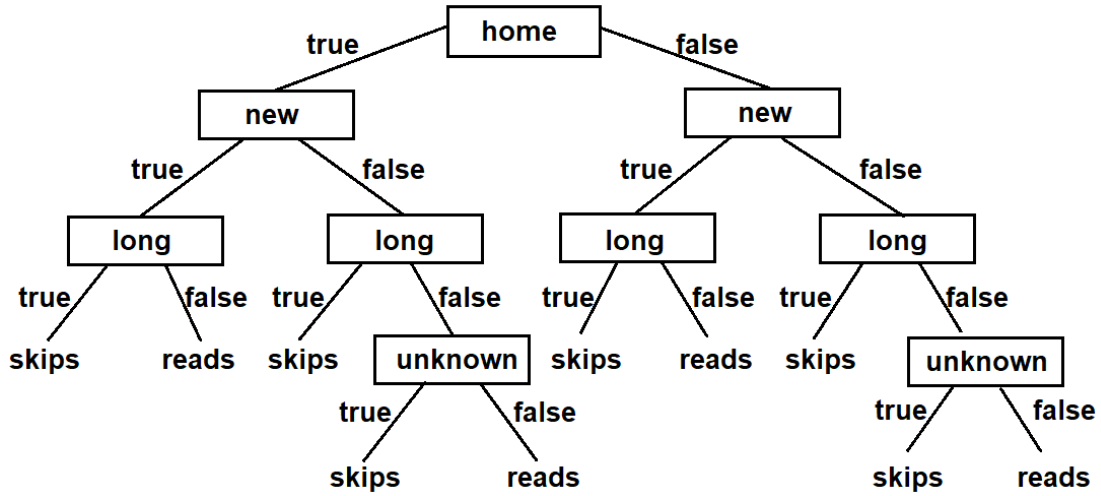
```
if long(e): return skips
else if new(e): return reads
else if unknown(e): return skips
else: return reads                          (function 1)
```

The function from the new decision tree is like

```
if unknown(e): if new(e): return reads
                  else: return skips
else: if new(e): if long(e): return skips
                    else: return reads
              else: if long(e): return skips
                    else: return reads      (function 2)
```

Structurally, they are different functions. And one function cannot be transformed to the other. Additionally, there's a counter example e_{19} from Figure 7.1 which is <*unknown, new, long, work, User_action?*> If we use this to test both functions, the function 1 gives the output User_action = **skips** while function 2 gives **reads**. Thus, they are completely different.

b)

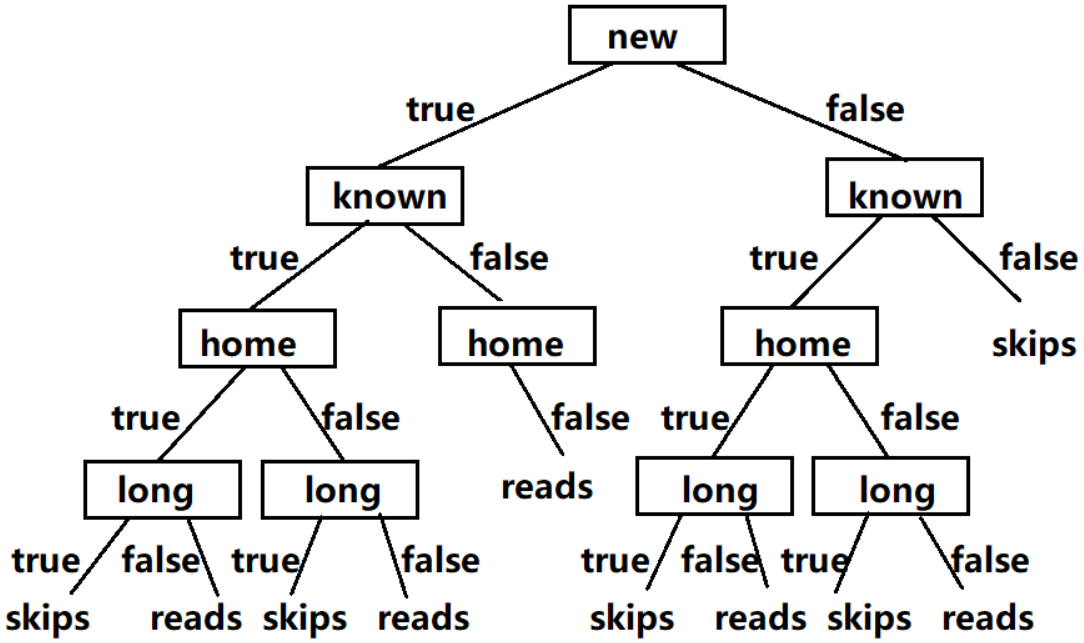


This tree represents the following function 3 which is substantially the same as function 1.

```
if home(e): if new(e): if long(e): return skips
                                else: return reads
                       else: if long(e): return skips
                                  else: if unknown(e): return skips
                                             else: return reads
else: if new(e): if long(e): return skips
                             else: return reads
                  else: if long(e): return skips
                                else: if unknown(e): return skips
                                           else: return false   (function 3)
```

Though the structure of the 2 functions look different, they are substantially identical. As we can transform one function to the other. In the new decision tree, we can see that whatever values $home(e)$ and $new(e)$ are, if $long(e)$ is true, the result is **skips**. This corresponds to 1_{st} line of function 1. Similarly, whatever value $home(e)$ is, if $long(e)$ is false and $new(e)$ is true, the result is **reads.** Corresponding to 2_{nd} line of function 1. Whatever value $home(e)$ is, if $long(e)$ is false and $new(e)$ is false and $unknown(e)$ is true, the result is **skips**. Corresponding to 3_{rd} line of function 1. Finally, whatever value $home(e)$ is, if $long(e)$, $new(e)$ and $unknown(e)$ are all false, the result is **reads**. Corresponding to 4_{th} line of function 1. Thus, they are the same function.

c)
Yes, there is such a tree.



This tree is generated if the order of features is [*Thread, Author, Where, Length*]. This tree correctly classifies the training examples as each example can obtain the correct User_action using this tree. And the tree represents different functions from those 2 in a) and b) Considering a test case <*unknown, new, home, long, User_action?*>, the tree in a) gives **reads**, the tree in b) gives **skips**, but the tree in c) doesn't define this scenario because there is no such kind of data to generate the that branch of the tree. Thus, although the new tree correctly classifies the training data, the function represented is different from the preceding ones.

Question 2:

I used Weka to create the decision tree. As Weka only accepts .arff format files, I combined adult.data and adult.test to make a complete data file and add labels @relation, @attribute and @data to make it a correct .arff file. Before training and testing, I split the data into first 2/3 to be training data and the rest 1/3 to be testing data.

Then I used J48 algorithm to build a few decision trees. I changed some parameters and compare the results in the table.

Algorithm	J48						
ConfiFactor	-	0.75	0.50	0.25	0.1	0.25	0.25
MinNumObj	2	2	2	2	2	3	4
Model time	2.68s	73.68s	3.16s	5.1s	3.3s	2.72s	2.35s
Test time	0.07s	0.03s	0.02s	0.1s	0.01s	0.02s	0.02s
# leaves	9129	8025	2386	696	205	503	401
Tree size	10657	9365	2946	911	275	651	530
Accuracy	84.1%	84.1%	85.3%	85.9%	85.5%	85.8%	85.7%

ConfidenceFactor is the confidence factor for pruning (smaller values incur more pruning). The tree become smaller as this parameter decreases, but the accuracy isn't affected much.

MinNumObj is the minimum number of instances per leaf. The tree become smaller as this parameter increases but doesn't affect the accuracy much.

Considering the accuracy and efficiency, the optimal choice of parameters is the shaded one which has relatively high accuracy and the shortest test time. The decision tree of it is below.

```
capital-gain <= 6849
|   matital-status = Married-civ-spouse
|   |   capital-loss <= 1844
|   |   |   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
|   |   |   |   |   |   |   |   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
|   |   |   |   |   |   |   |   |   |   workclass = Private
|   |   |   |   |   |   |   |   |   |   |   fnlwgt <= 152968
|   |   |   |   |   |   |   |   |   |   |   |   age <= 49
|   |   |   |   |   |   |   |   |   |   |   |   |   capital-gain <= 1506: <=50K (170.22/48.51)
|   |   |   |   |   |   |   |   |   |   |   |   |   capital-gain > 1506
|   |   |   |   |   |   |   |   |   |   |   |   |   |   capital-gain <= 3273: >50K (4.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   capital-gain > 3273: <=50K (7.0/2.0)
|   |   |   |   |   |   |   |   |   |   |   |   age > 49
|   |   |   |   |   |   |   |   |   |   |   |   |   hours-per-week <= 51
|   |   |   |   |   |   |   |   |   |   |   |   |   |   age <= 59: >50K (42.37/14.2)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   age > 59: <=50K (12.37/2.51)
|   |   |   |   |   |   |   |   |   |   |   |   |   hours-per-week > 51: <=50K (6.17/0.17)
|   |   |   |   |   |   |   |   |   |   |   fnlwgt > 152968
|   |   |   |   |   |   |   |   |   |   |   |   hours-per-week <= 47
|   |   |   |   |   |   |   |   |   |   |   |   |   education = Bachelors: <=50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   education = Some-college: <=50K (208.45/89.54)
|   |   |   |   |   |   |   |   |   |   |   |   |   education = 11th: <=50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   education = HS-grad: <=50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   education = Prof-school: <=50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   education = Assoc-acdm: <=50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   education = Assoc-voc
|   |   |   |   |   |   |   |   |   |   |   |   |   |   fnlwgt <= 192344: <=50K (24.17/6.17)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   fnlwgt > 192344: >50K (41.17/14.17)
|   |   |   |   |   |   |   |   |   |   |   |   |   education = 9th: <=50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   education = 7th-8th: <=50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   education = 12th: <=50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   education = Masters: <=50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   education = 1st-4th: <=50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   education = 10th: <=50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   education = Doctorate: <=50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   education = 5th-6th: <=50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   education = Preschool: <=50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   hours-per-week > 47
|   |   |   |   |   |   |   |   |   |   |   |   |   education = Bachelors: >50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   education = Some-college: >50K (69.88/25.2)
|   |   |   |   |   |   |   |   |   |   |   |   |   education = 11th: >50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   education = HS-grad: >50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   education = Prof-school: >50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   education = Assoc-acdm: >50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   education = Assoc-voc: <=50K (20.68/7.68)
|   |   |   |   |   |   |   |   |   |   |   |   |   education = 9th: >50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   education = 7th-8th: >50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   education = 12th: >50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   education = Masters: >50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   education = 1st-4th: >50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   education = 10th: >50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   education = Doctorate: >50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   education = 5th-6th: >50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   education = Preschool: >50K (0.0)
|   |   |   |   |   |   |   |   |   |   workclass = Self-emp-not-inc: <=50K (102.1/23.75)
|   |   |   |   |   |   |   |   |   |   workclass = Self-emp-inc
|   |   |   |   |   |   |   |   |   |   |   hours-per-week <= 59
|   |   |   |   |   |   |   |   |   |   |   |   hours-per-week <= 42: <=50K (11.37/3.1)
|   |   |   |   |   |   |   |   |   |   |   |   hours-per-week > 42: >50K (7.14/1.07)
|   |   |   |   |   |   |   |   |   |   |   hours-per-week > 59: <=50K (9.05/1.03)
|   |   |   |   |   |   |   |   |   |   workclass = Federal-gov: <=50K (22.46/8.16)
```

```
|   |   |   |   |   |   |   |   |   |   workclass = Local-gov
|   |   |   |   |   |   |   |   |   |   |   fnlwgt <= 188709: <=50K (10.34/0.12)
|   |   |   |   |   |   |   |   |   |   |   fnlwgt > 188709: >50K (19.27/8.17)
|   |   |   |   |   |   |   |   |   |   workclass = State-gov: <=50K (11.23/5.08)
|   |   |   |   |   |   |   |   |   |   workclass = Without-pay: <=50K (0.0)
|   |   |   |   |   |   |   |   |   |   workclass = Never-worked: <=50K (0.0)
|   |   |   |   |   |   |   |   |   occupation = Other-service: <=50K (144.99/32.06)
|   |   |   |   |   |   |   |   |   occupation = Sales
|   |   |   |   |   |   |   |   |   |   workclass = Private
|   |   |   |   |   |   |   |   |   |   |   fnlwgt <= 89259: <=50K (67.76/17.19)
|   |   |   |   |   |   |   |   |   |   |   fnlwgt > 89259
|   |   |   |   |   |   |   |   |   |   |   |   age <= 40: <=50K (119.89/49.47)
|   |   |   |   |   |   |   |   |   |   |   |   age > 40: >50K (147.25/56.46)
|   |   |   |   |   |   |   |   |   |   workclass = Self-emp-not-inc
|   |   |   |   |   |   |   |   |   |   |   fnlwgt <= 345734: <=50K (60.29/17.42)
|   |   |   |   |   |   |   |   |   |   |   fnlwgt > 345734: >50K (7.1/0.02)
|   |   |   |   |   |   |   |   |   |   workclass = Self-emp-inc
|   |   |   |   |   |   |   |   |   |   |   hours-per-week <= 42
|   |   |   |   |   |   |   |   |   |   |   |   hours-per-week <= 39: >50K (4.1/1.05)
|   |   |   |   |   |   |   |   |   |   |   |   hours-per-week > 39: <=50K (11.69/2.16)
|   |   |   |   |   |   |   |   |   |   |   hours-per-week > 42: >50K (42.41/12.2)
|   |   |   |   |   |   |   |   |   |   workclass = Federal-gov: <=50K (0.0)
|   |   |   |   |   |   |   |   |   |   workclass = Local-gov: <=50K (1.02/0.01)
|   |   |   |   |   |   |   |   |   |   workclass = State-gov: <=50K (2.04/1.01)
|   |   |   |   |   |   |   |   |   |   workclass = Without-pay: <=50K (0.0)
|   |   |   |   |   |   |   |   |   |   workclass = Never-worked: <=50K (0.0)
|   |   |   |   |   |   |   |   |   occupation = Exec-managerial
|   |   |   |   |   |   |   |   |   |   workclass = Private: >50K (346.13/132.59)
|   |   |   |   |   |   |   |   |   |   workclass = Self-emp-not-inc: <=50K (68.41/21.5)
|   |   |   |   |   |   |   |   |   |   workclass = Self-emp-inc: >50K (91.89/32.22)
|   |   |   |   |   |   |   |   |   |   workclass = Federal-gov: >50K (25.53/6.34)
|   |   |   |   |   |   |   |   |   |   workclass = Local-gov: >50K (40.84/19.54)
|   |   |   |   |   |   |   |   |   |   workclass = State-gov
|   |   |   |   |   |   |   |   |   |   |   age <= 45: <=50K (12.17/3.04)
|   |   |   |   |   |   |   |   |   |   |   age > 45: >50K (11.32/3.19)
|   |   |   |   |   |   |   |   |   |   workclass = Without-pay: >50K (0.0)
|   |   |   |   |   |   |   |   |   |   workclass = Never-worked: >50K (0.0)
|   |   |   |   |   |   |   |   |   occupation = Prof-specialty: >50K (258.32/113.43)
|   |   |   |   |   |   |   |   |   occupation = Handlers-cleaners: <=50K (92.91/18.68)
|   |   |   |   |   |   |   |   |   occupation = Machine-op-inspct: <=50K (197.06/64.45)
|   |   |   |   |   |   |   |   |   occupation = Adm-clerical
|   |   |   |   |   |   |   |   |   |   workclass = Private: <=50K (201.14/80.48)
|   |   |   |   |   |   |   |   |   |   workclass = Self-emp-not-inc: >50K (4.08/1.05)
|   |   |   |   |   |   |   |   |   |   workclass = Self-emp-inc: >50K (1.02/0.01)
|   |   |   |   |   |   |   |   |   |   workclass = Federal-gov: >50K (55.14/18.73)
|   |   |   |   |   |   |   |   |   |   workclass = Local-gov
|   |   |   |   |   |   |   |   |   |   |   sex = Female: <=50K (15.11/5.04)
|   |   |   |   |   |   |   |   |   |   |   sex = Male: >50K (9.39/3.26)
|   |   |   |   |   |   |   |   |   |   workclass = State-gov: <=50K (14.29/5.1)
|   |   |   |   |   |   |   |   |   |   workclass = Without-pay: <=50K (0.0)
|   |   |   |   |   |   |   |   |   |   workclass = Never-worked: <=50K (0.0)
|   |   |   |   |   |   |   |   |   occupation = Farming-fishing: <=50K (146.01/30.07)
|   |   |   |   |   |   |   |   |   occupation = Transport-moving: <=50K (201.14/75.48)
|   |   |   |   |   |   |   |   |   occupation = Priv-house-serv: <=50K (0.0)
|   |   |   |   |   |   |   |   |   occupation = Protective-serv
|   |   |   |   |   |   |   |   |   |   race = White
|   |   |   |   |   |   |   |   |   |   |   workclass = Private: <=50K (19.39/2.13)
|   |   |   |   |   |   |   |   |   |   |   workclass = Self-emp-not-inc: >50K (1.02/0.01)
```

```
|   |   |   |   |   |   |   |   |   |   |   workclass = Self-emp-inc: >50K (1.02/0.01)
|   |   |   |   |   |   |   |   |   |   |   workclass = Federal-gov
|   |   |   |   |   |   |   |   |   |   |   |   age <= 46: <=50K (2.03/0.01)
|   |   |   |   |   |   |   |   |   |   |   |   age > 46: >50K (2.06/0.03)
|   |   |   |   |   |   |   |   |   |   |   workclass = Local-gov
|   |   |   |   |   |   |   |   |   |   |   |   education = Bachelors: >50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   education = Some-college
|   |   |   |   |   |   |   |   |   |   |   |   |   age <= 53
|   |   |   |   |   |   |   |   |   |   |   |   |   |   age <= 38: <=50K (32.35/13.03)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   age > 38
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   fnlwgt <= 121253: <=50K (8.09/2.03)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   fnlwgt > 121253: >50K (28.21/4.15)
|   |   |   |   |   |   |   |   |   |   |   |   |   age > 53: <=50K (5.77/0.27)
|   |   |   |   |   |   |   |   |   |   |   |   education = 11th: >50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   education = HS-grad: >50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   education = Prof-school: >50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   education = Assoc-acdm: >50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   education = Assoc-voc
|   |   |   |   |   |   |   |   |   |   |   |   |   fnlwgt <= 112772: <=50K (2.12/0.06)
|   |   |   |   |   |   |   |   |   |   |   |   |   fnlwgt > 112772: >50K (15.29/2.12)
|   |   |   |   |   |   |   |   |   |   |   |   education = 9th: >50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   education = 7th-8th: >50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   education = 12th: >50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   education = Masters: >50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   education = 1st-4th: >50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   education = 10th: >50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   education = Doctorate: >50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   education = 5th-6th: >50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   education = Preschool: >50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   workclass = State-gov
|   |   |   |   |   |   |   |   |   |   |   |   age <= 41: <=50K (9.13/0.03)
|   |   |   |   |   |   |   |   |   |   |   |   age > 41: >50K (15.35/6.22)
|   |   |   |   |   |   |   |   |   |   |   workclass = Without-pay: <=50K (0.0)
|   |   |   |   |   |   |   |   |   |   |   workclass = Never-worked: <=50K (0.0)
|   |   |   |   |   |   |   |   |   |   race = Asian-Pac-Islander: <=50K (2.18/0.14)
|   |   |   |   |   |   |   |   |   |   race = Amer-Indian-Eskimo: <=50K (1.05)
|   |   |   |   |   |   |   |   |   |   race = Other: >50K (2.05/0.05)
|   |   |   |   |   |   |   |   |   |   race = Black
|   |   |   |   |   |   |   |   |   |   |   fnlwgt <= 154897: <=50K (3.14/0.05)
|   |   |   |   |   |   |   |   |   |   |   fnlwgt > 154897: >50K (11.09/3.05)
|   |   |   |   |   |   |   |   |   occupation = Armed-Forces: >50K (1.02/0.01)
|   |   |   |   |   |   |   |   capital-loss > 1510: <=50K (43.0/1.0)
|   |   |   |   capital-gain > 5060
|   |   |   |   |   capital-gain <= 6612: >50K (111.0)
|   |   |   |   |   capital-gain > 6612: <=50K (5.0)
|   |   |   education-num > 11
|   |   |   |   hours-per-week <= 30
|   |   |   |   |   sex = Female
|   |   |   |   |   |   race = White: >50K (145.0/54.0)
|   |   |   |   |   |   race = Asian-Pac-Islander: <=50K (6.0/3.0)
|   |   |   |   |   |   race = Amer-Indian-Eskimo: <=50K (1.0)
|   |   |   |   |   |   race = Other: <=50K (1.0)
|   |   |   |   |   |   race = Black: <=50K (6.0/1.0)
|   |   |   |   |   sex = Male: <=50K (386.0/111.0)
|   |   |   |   hours-per-week > 30
|   |   |   |   |   age <= 33
|   |   |   |   |   |   age <= 25: <=50K (121.0/28.0)
|   |   |   |   |   |   age > 25
```

```
|   |   |   |   |   |   |   education-num <= 12: <=50K (129.0/40.0)
|   |   |   |   |   |   |   education-num > 12
|   |   |   |   |   |   |   |   relationship = Wife: >50K (120.0/36.0)
|   |   |   |   |   |   |   |   relationship = Own-child: <=50K (2.0)
|   |   |   |   |   |   |   |   relationship = Husband
|   |   |   |   |   |   |   |   |   age <= 28: <=50K (166.0/75.0)
|   |   |   |   |   |   |   |   |   age > 28: >50K (542.0/229.0)
|   |   |   |   |   |   |   |   relationship = Not-in-family: >50K (0.0)
|   |   |   |   |   |   |   |   relationship = Other-relative: <=50K (10.0/4.0)
|   |   |   |   |   |   |   |   relationship = Unmarried: >50K (0.0)
|   |   |   |   |   age > 33
|   |   |   |   |   |   occupation = Tech-support: >50K (128.28/38.19)
|   |   |   |   |   |   occupation = Craft-repair
|   |   |   |   |   |   |   education = Bachelors: >50K (142.11/66.16)
|   |   |   |   |   |   |   education = Some-college: >50K (0.0)
|   |   |   |   |   |   |   education = 11th: >50K (0.0)
|   |   |   |   |   |   |   education = HS-grad: >50K (0.0)
|   |   |   |   |   |   |   education = Prof-school: <=50K (5.39/2.22)
|   |   |   |   |   |   |   education = Assoc-acdm
|   |   |   |   |   |   |   |   age <= 37: <=50K (15.06/3.0)
|   |   |   |   |   |   |   |   age > 37
|   |   |   |   |   |   |   |   |   fnlwgt <= 153052: <=50K (18.11/5.06)
|   |   |   |   |   |   |   |   |   fnlwgt > 153052: >50K (26.06/5.06)
|   |   |   |   |   |   |   education = Assoc-voc: >50K (0.0)
|   |   |   |   |   |   |   education = 9th: >50K (0.0)
|   |   |   |   |   |   |   education = 7th-8th: >50K (0.0)
|   |   |   |   |   |   |   education = 12th: >50K (0.0)
|   |   |   |   |   |   |   education = Masters: >50K (20.11/8.5)
|   |   |   |   |   |   |   education = 1st-4th: >50K (0.0)
|   |   |   |   |   |   |   education = 10th: >50K (0.0)
|   |   |   |   |   |   |   education = Doctorate: <=50K (1.22/0.11)
|   |   |   |   |   |   |   education = 5th-6th: >50K (0.0)
|   |   |   |   |   |   |   education = Preschool: >50K (0.0)
|   |   |   |   |   |   occupation = Other-service: <=50K (57.01/10.49)
|   |   |   |   |   |   occupation = Sales: >50K (541.61/188.0)
|   |   |   |   |   |   occupation = Exec-managerial: >50K (1180.96/289.91)
|   |   |   |   |   |   occupation = Prof-specialty: >50K (1446.68/373.37)
|   |   |   |   |   |   occupation = Handlers-cleaners: <=50K (31.56/10.27)
|   |   |   |   |   |   occupation = Machine-op-inspct: <=50K (51.92/22.44)
|   |   |   |   |   |   occupation = Adm-clerical: >50K (240.26/98.22)
|   |   |   |   |   |   occupation = Farming-fishing
|   |   |   |   |   |   |   workclass = Private: <=50K (20.36/6.17)
|   |   |   |   |   |   |   workclass = Self-emp-not-inc: <=50K (41.74/10.36)
|   |   |   |   |   |   |   workclass = Self-emp-inc: >50K (6.11/0.06)
|   |   |   |   |   |   |   workclass = Federal-gov: <=50K (2.04/1.02)
|   |   |   |   |   |   |   workclass = Local-gov: <=50K (0.0)
|   |   |   |   |   |   |   workclass = State-gov: >50K (1.02/0.01)
|   |   |   |   |   |   |   workclass = Without-pay: <=50K (1.02/0.01)
|   |   |   |   |   |   |   workclass = Never-worked: <=50K (0.0)
|   |   |   |   |   |   occupation = Transport-moving: <=50K (54.98/20.47)
|   |   |   |   |   |   occupation = Priv-house-serv: <=50K (2.04/0.02)
|   |   |   |   |   |   occupation = Protective-serv
|   |   |   |   |   |   |   age <= 42
|   |   |   |   |   |   |   |   capital-gain <= 1506: <=50K (30.15/11.06)
|   |   |   |   |   |   |   |   capital-gain > 1506: >50K (2.02/0.02)
|   |   |   |   |   |   |   age > 42: >50K (44.19/7.59)
|   |   |   |   |   |   occupation = Armed-Forces: >50K (1.02/0.01)
|   |   capital-loss > 1844
```

```
|   |   |   capital-loss <= 1980: >50K (857.0/18.0)
|   |   |   capital-loss > 1980
|   |   |   |   capital-loss <= 2163: <=50K (104.0)
|   |   |   |   capital-loss > 2163
|   |   |   |   |   capital-loss <= 2415
|   |   |   |   |   |   capital-loss <= 2377
|   |   |   |   |   |   |   age <= 64: <=50K (38.0/4.0)
|   |   |   |   |   |   |   age > 64: >50K (30.0/3.0)
|   |   |   |   |   |   capital-loss > 2377: >50K (82.0)
|   |   |   |   |   capital-loss > 2415: <=50K (14.0)
|   matital-status = Divorced: <=50K (6454.0/498.0)
|   matital-status = Never-married
|   |   capital-loss <= 2206: <=50K (15843.0/495.0)
|   |   capital-loss > 2206
|   |   |   capital-loss <= 2377: <=50K (38.0/9.0)
|   |   |   capital-loss > 2377: >50K (27.0/1.0)
|   matital-status = Separated: <=50K (1505.0/76.0)
|   matital-status = Widowed
|   |   capital-loss <= 2205: <=50K (1460.0/82.0)
|   |   capital-loss > 2205
|   |   |   race = White: >50K (23.0/9.0)
|   |   |   race = Asian-Pac-Islander: >50K (0.0)
|   |   |   race = Amer-Indian-Eskimo: >50K (0.0)
|   |   |   race = Other: >50K (0.0)
|   |   |   race = Black: <=50K (2.0)
|   matital-status = Married-spouse-absent: <=50K (613.0/44.0)
|   matital-status = Married-AF-spouse: <=50K (35.0/12.0)
capital-gain > 6849: >50K (2055.0/28.0)
```

Then I used REPTree algorithm to build a few other decision trees. I changed some parameters and compare the results in the table.

Algorithm	REPTree						
MaxDepth	noLimit	7	4	2	4	4	4
MinNum	2	2	2	2	3	2	2
numFolds	3	3	3	3	3	4	5
Model time	0.83s	0.7s	0.37s	0.35s	0.77s	0.46s	0.64s
Test time	0.02s	0.02s	0.01s	0.02s	0.01s	0.01s	0.01s
Tree size	2041	1094	238	45	200	217	209
Accuracy	84.9%	85.1%	84.3%	82.7%	84.3%	84.5%	84.4%

MaxDepth is the max depth of the decision tree, as we decrease the MaxDepth, the tree size will decrease, and accuracy will also decrease slightly.

MinNum is the minimum total weight of the instances in a leaf, as it increases, the tree size decreases. But the accuracy doesn't change much.

numFolds determines the amount used for pruning. As it increases,

the tree size decreases slightly. But the accuracy doesn't change much.
Following is a decision tree with the minimum tree size in the table (not the most accurate).

```
relationship = Wife
|   education = Bachelors : >50K (298/90) [166/64]
|   education = Some-college : <=50K (306/132) [162/69]
|   education = 11th : <=50K (36/3) [8/1]
|   education = HS-grad : <=50K (485/153) [239/84]
|   education = Prof-school : >50K (25/4) [13/1]
|   education = Assoc-acdm : >50K (78/36) [30/14]
|   education = Assoc-voc : >50K (76/33) [42/20]
|   education = 9th : <=50K (25/2) [10/1]
|   education = 7th-8th : <=50K (22/1) [7/0]
|   education = 12th : <=50K (14/3) [7/0]
|   education = Masters : >50K (117/20) [63/12]
|   education = 1st-4th : <=50K (1/0) [5/0]
|   education = 10th : <=50K (27/1) [19/2]
|   education = Doctorate : >50K (18/1) [8/2]
|   education = 5th-6th : <=50K (17/4) [4/0]
|   education = Preschool : <=50K (1/0) [2/0]
relationship = Own-child
|   capital-gain < 4718.5 : <=50K (4991/46) [2541/29]
|   capital-gain >= 4718.5 : >50K (24/5) [25/8]
relationship = Husband
|   education = Bachelors : >50K (2375/758) [1261/416]
|   education = Some-college : <=50K (2486/1128) [1172/480]
|   education = 11th : <=50K (322/44) [157/24]
|   education = HS-grad : <=50K (4297/1328) [2091/677]
|   education = Prof-school : >50K (362/53) [196/33]
|   education = Assoc-acdm : >50K (380/194) [201/93]
|   education = Assoc-voc : <=50K (576/250) [307/137]
|   education = 9th : <=50K (186/17) [119/11]
|   education = 7th-8th : <=50K (337/39) [166/17]
|   education = 12th : <=50K (114/26) [52/11]
|   education = Masters : >50K (876/212) [463/113]
|   education = 1st-4th : <=50K (74/6) [37/2]
|   education = 10th : <=50K (303/48) [168/27]
|   education = Doctorate : >50K (256/44) [121/23]
|   education = 5th-6th : <=50K (164/17) [74/5]
|   education = Preschool : <=50K (10/0) [13/0]
relationship = Not-in-family
|   capital-gain < 8296 : <=50K (8237/643) [4007/295]
|   capital-gain >= 8296 : >50K (236/0) [103/1]
relationship = Other-relative : <=50K (1006/33) [500/19]
relationship = Unmarried
|   capital-gain < 7139.5 : <=50K (3352/148) [1695/89]
|   capital-gain >= 7139.5 : >50K (51/3) [27/3]
```

In conclusion, the best accuracy of predicting this dataset is 85.5% using J48 algorithm. Appropriate pruning and limitation of max tree depth will adjust the decision tree to a smaller size and the accuracy

will remain much the same or improve. Increasing MinNumObj in J48 may result in a smaller tree. Increasing MinNum or numFold in REPTree may also result a smaller tree. The best Max tree depth of REPTree is 7.