In [1]:

```
library(rJava)
library(RWeka)
```

In [2]:

```
adult.train <- read.csv("adult_train.csv")
head(adult.train)
```

age	workclass	education	marital.status	occupation	relationship	race	sex	capital.gain	capital.loss	hours.per.week	native.country	income
59	Private	Some-college	Married-civ-spouse	Adm-clerical	Other-relative	White	Female	0	0	16	United-States	>50K
21	Private	Some-college	Never-married	Adm-clerical	Own-child	Black	Male	0	0	50	United-States	<=50K
38	Private	Bachelors	Divorced	Prof-specialty	Unmarried	Black	Female	15020	0	45	United-States	>50K
33	Private	Some-college	Married-civ-spouse	Handlers-cleaners	Husband	White	Male	0	0	50	United-States	>50K
52	Self-emp-not-inc	HS-grad	Married-civ-spouse	Farming-fishing	Husband	White	Male	0	0	98	United-States	>50K
42	Federal-gov	Bachelors	Married-civ-spouse	Exec-managerial	Husband	White	Male	7298	0	50	United-States	>50K

In [3]:

```
WOW(J48)
```

```
-U      Use unpruned tree.
-O      Do not collapse tree.
-C <pruning confidence>
        Set confidence threshold for pruning.  (default 0.25)
        Number of arguments: 1.
-M <minimum number of instances>
        Set minimum number of instances per leaf.  (default 2)
        Number of arguments: 1.
-R      Use reduced error pruning.
-N <number of folds>
        Set number of folds for reduced error pruning. One fold is used
        as pruning set.  (default 3)
        Number of arguments: 1.
-B      Use binary splits only.
-S      Do not perform subtree raising.
-L      Do not clean up after the tree has been built.
-A      Laplace smoothing for predicted probabilities.
-J      Do not use MDL correction for info gain on numeric attributes.
-Q <seed>
        Seed for random data shuffling (default 1).
        Number of arguments: 1.
-doNotMakeSplitPointActualValue
        Do not make split point actual value.
-output-debug-info
        If set, classifier is run in debug mode and may output
        additional info to the console
-do-not-check-capabilities
        If set, classifier capabilities are not checked before
        classifier is built (use with caution).
-num-decimal-places
        The number of decimal places for the output of numbers in the
        model (default 2).
        Number of arguments: 1.
-batch-size
        The desired batch size for batch prediction (default 100).
        Number of arguments: 1.
```

Problem 1

Using the adult train dataset, train three decision trees - you can use different hyperparameters or/and learning algorithms.

1. Compare the size and depth of the three decision trees.
2. Compare the training errors of the three decision trees.

In [4]:

```
mdl <- J48(income ~ ., data = adult.train)
mdl.2 <- J48(income ~ ., data = adult.train, control = Weka_control(M = 10, C = 0.25, B = F, U = F)) # not binary split
mdl.3 <- J48(income ~ ., data = adult.train, control = Weka_control(M = 10, C = 0.25, B = T, U = F)) # binary split
```

In [5]:

```
print(mdl)
table(adult.train$income, predict(mdl))
summary(mdl)
```

```
J48 pruned tree
------------------

capital.gain <= 5013
|   age <= 27:  <=50K (663.0/51.0)
|   age > 27
|   |   marital.status =  Divorced
|   |   |   hours.per.week <= 43:  <=50K (306.0/47.0)
|   |   |   hours.per.week > 43
|   |   |   |   workclass =  Federal-gov:  >50K (6.0)
|   |   |   |   workclass =  Local-gov:  <=50K (11.0/2.0)
|   |   |   |   workclass =  Private
|   |   |   |   |   education =  10th:  <=50K (1.0)
|   |   |   |   |   education =  11th:  <=50K (1.0)
|   |   |   |   |   education =  12th:  >50K (2.0)
|   |   |   |   |   education =  1st-4th:  <=50K (0.0)
|   |   |   |   |   education =  5th-6th:  <=50K (1.0)
|   |   |   |   |   education =  7th-8th:  <=50K (2.0)
|   |   |   |   |   education =  9th:  <=50K (0.0)
|   |   |   |   |   education =  Assoc-acdm:  <=50K (3.0/1.0)
|   |   |   |   |   education =  Assoc-voc:  <=50K (3.0/1.0)
|   |   |   |   |   education =  Bachelors:  >50K (20.0/6.0)
|   |   |   |   |   education =  Doctorate:  <=50K (0.0)
|   |   |   |   |   education =  HS-grad:  <=50K (28.0/6.0)
|   |   |   |   |   education =  Masters:  >50K (5.0/1.0)
|   |   |   |   |   education =  Preschool:  <=50K (0.0)
|   |   |   |   |   education =  Prof-school:  <=50K (0.0)
|   |   |   |   |   education =  Some-college:  <=50K (15.0/6.0)
|   |   |   |   workclass =  Self-emp-inc
|   |   |   |   |   age <= 46:  <=50K (5.0/1.0)
|   |   |   |   |   age > 46:  >50K (4.0/1.0)
|   |   |   |   workclass =  Self-emp-not-inc:  >50K (13.0/3.0)
|   |   |   |   workclass =  State-gov:  <=50K (3.0/1.0)
|   |   |   |   workclass =  Without-pay:  <=50K (0.0)
|   |   marital.status =  Married-AF-spouse:  >50K (5.0/1.0)
|   |   marital.status =  Married-civ-spouse
|   |   |   capital.loss <= 1762
|   |   |   |   education =  10th
|   |   |   |   |   workclass =  Federal-gov:  <=50K (0.0)
|   |   |   |   |   workclass =  Local-gov:  <=50K (0.0)
|   |   |   |   |   workclass =  Private:  <=50K (22.0/7.0)
|   |   |   |   |   workclass =  Self-emp-inc:  <=50K (1.0)
|   |   |   |   |   workclass =  Self-emp-not-inc:  >50K (2.0)
|   |   |   |   |   workclass =  State-gov:  >50K (2.0)
|   |   |   |   |   workclass =  Without-pay:  <=50K (0.0)
|   |   |   |   education =  11th:  <=50K (32.0/7.0)
|   |   |   |   education =  12th:  <=50K (9.0/4.0)
|   |   |   |   education =  1st-4th:  <=50K (4.0/2.0)
|   |   |   |   education =  5th-6th:  <=50K (12.0/2.0)
|   |   |   |   education =  7th-8th:  <=50K (37.0/8.0)
|   |   |   |   education =  9th:  <=50K (19.0/6.0)
|   |   |   |   education =  Assoc-acdm:  >50K (56.0/16.0)
|   |   |   |   education =  Assoc-voc
|   |   |   |   |   race =  Amer-Indian-Eskimo:  <=50K (2.0)
|   |   |   |   |   race =  Asian-Pac-Islander:  <=50K (5.0/2.0)
|   |   |   |   |   race =  Black
|   |   |   |   |   |   age <= 40:  <=50K (3.0)
|   |   |   |   |   |   age > 40:  >50K (4.0)
|   |   |   |   |   race =  Other:  >50K (0.0)
|   |   |   |   |   race =  White:  >50K (74.0/22.0)
|   |   |   |   education =  Bachelors:  >50K (379.0/63.0)
|   |   |   |   education =  Doctorate:  >50K (37.0/2.0)
|   |   |   |   education =  HS-grad
|   |   |   |   |   hours.per.week <= 39:  <=50K (67.0/21.0)
|   |   |   |   |   hours.per.week > 39
|   |   |   |   |   |   race =  Amer-Indian-Eskimo:  <=50K (4.0)
|   |   |   |   |   |   race =  Asian-Pac-Islander:  <=50K (14.0/3.0)
|   |   |   |   |   |   race =  Black:  <=50K (27.0/13.0)
|   |   |   |   |   |   race =  Other:  <=50K (1.0)
|   |   |   |   |   |   race =  White
|   |   |   |   |   |   |   occupation =  Adm-clerical:  >50K (44.0/11.0)
|   |   |   |   |   |   |   occupation =  Armed-Forces:  >50K (0.0)
|   |   |   |   |   |   |   occupation =  Craft-repair
|   |   |   |   |   |   |   |   age <= 40:  <=50K (45.0/19.0)
|   |   |   |   |   |   |   |   age > 40:  >50K (68.0/22.0)
|   |   |   |   |   |   |   occupation =  Exec-managerial:  >50K (58.0/8.0)
|   |   |   |   |   |   |   occupation =  Farming-fishing
|   |   |   |   |   |   |   |   capital.gain <= 1086
|   |   |   |   |   |   |   |   |   age <= 36:  <=50K (7.0/1.0)
|   |   |   |   |   |   |   |   |   age > 36:  >50K (20.0/8.0)
|   |   |   |   |   |   |   |   capital.gain > 1086:  <=50K (2.0)
|   |   |   |   |   |   |   occupation =  Handlers-cleaners:  >50K (9.0/2.0)
|   |   |   |   |   |   |   occupation =  Machine-op-inspct
|   |   |   |   |   |   |   |   relationship =  Husband:  >50K (47.0/20.0)
|   |   |   |   |   |   |   |   relationship =  Not-in-family:  >50K (0.0)
|   |   |   |   |   |   |   |   relationship =  Other-relative:  >50K (0.0)
|   |   |   |   |   |   |   |   relationship =  Own-child:  >50K (0.0)
|   |   |   |   |   |   |   |   relationship =  Unmarried:  >50K (0.0)
|   |   |   |   |   |   |   |   relationship =  Wife:  <=50K (4.0/1.0)
|   |   |   |   |   |   |   occupation =  Other-service
|   |   |   |   |   |   |   |   relationship =  Husband:  <=50K (14.0/2.0)
|   |   |   |   |   |   |   |   relationship =  Not-in-family:  <=50K (0.0)
|   |   |   |   |   |   |   |   relationship =  Other-relative:  <=50K (0.0)
|   |   |   |   |   |   |   |   relationship =  Own-child:  <=50K (0.0)
|   |   |   |   |   |   |   |   relationship =  Unmarried:  <=50K (0.0)
|   |   |   |   |   |   |   |   relationship =  Wife:  >50K (3.0/1.0)
|   |   |   |   |   |   |   occupation =  Priv-house-serv:  <=50K (1.0)
|   |   |   |   |   |   |   occupation =  Prof-specialty:  >50K (13.0/4.0)
|   |   |   |   |   |   |   occupation =  Protective-serv
|   |   |   |   |   |   |   |   workclass =  Federal-gov:  >50K (0.0)
|   |   |   |   |   |   |   |   workclass =  Local-gov:  >50K (11.0/2.0)
|   |   |   |   |   |   |   |   workclass =  Private:  <=50K (6.0/1.0)
|   |   |   |   |   |   |   |   workclass =  Self-emp-inc:  >50K (0.0)
|   |   |   |   |   |   |   |   workclass =  Self-emp-not-inc:  >50K (0.0)
|   |   |   |   |   |   |   |   workclass =  State-gov:  <=50K (2.0/1.0)
|   |   |   |   |   |   |   |   workclass =  Without-pay:  >50K (0.0)
|   |   |   |   |   |   |   occupation =  Sales
|   |   |   |   |   |   |   |   workclass =  Federal-gov:  >50K (0.0)
|   |   |   |   |   |   |   |   workclass =  Local-gov:  >50K (0.0)
|   |   |   |   |   |   |   |   workclass =  Private:  >50K (35.0/12.0)
|   |   |   |   |   |   |   |   workclass =  Self-emp-inc:  <=50K (2.0/1.0)
|   |   |   |   |   |   |   |   workclass =  Self-emp-not-inc:  <=50K (6.0/2.0)
|   |   |   |   |   |   |   |   workclass =  State-gov:  <=50K (1.0)
|   |   |   |   |   |   |   |   workclass =  Without-pay:  >50K (0.0)
|   |   |   |   |   |   |   occupation =  Tech-support:  >50K (11.0/2.0)
|   |   |   |   |   |   |   occupation =  Transport-moving
|   |   |   |   |   |   |   |   age <= 51
|   |   |   |   |   |   |   |   |   hours.per.week <= 57:  <=50K (27.0/11.0)
|   |   |   |   |   |   |   |   |   hours.per.week > 57:  >50K (6.0/2.0)
|   |   |   |   |   |   |   |   age > 51:  >50K (7.0)
|   |   |   |   education =  Masters:  >50K (144.0/15.0)
|   |   |   |   education =  Preschool:  <=50K (4.0)
|   |   |   |   education =  Prof-school
|   |   |   |   |   age <= 61:  >50K (39.0/4.0)
|   |   |   |   |   age > 61:  <=50K (3.0)
|   |   |   |   education =  Some-college
|   |   |   |   |   occupation =  Adm-clerical
|   |   |   |   |   |   workclass =  Federal-gov:  >50K (4.0/1.0)
|   |   |   |   |   |   workclass =  Local-gov:  >50K (3.0)
|   |   |   |   |   |   workclass =  Private
|   |   |   |   |   |   |   hours.per.week <= 39:  >50K (6.0)
|   |   |   |   |   |   |   hours.per.week > 39
|   |   |   |   |   |   |   |   hours.per.week <= 42:  <=50K (10.0/2.0)
|   |   |   |   |   |   |   |   hours.per.week > 42:  >50K (3.0)
|   |   |   |   |   |   workclass =  Self-emp-inc:  >50K (0.0)
|   |   |   |   |   |   workclass =  Self-emp-not-inc:  >50K (0.0)
|   |   |   |   |   |   workclass =  State-gov:  >50K (0.0)
|   |   |   |   |   |   workclass =  Without-pay:  >50K (0.0)
|   |   |   |   |   occupation =  Armed-Forces:  >50K (0.0)
|   |   |   |   |   occupation =  Craft-repair:  >50K (64.0/25.0)
|   |   |   |   |   occupation =  Exec-managerial:  >50K (53.0/10.0)
|   |   |   |   |   occupation =  Farming-fishing:  <=50K (13.0/1.0)
|   |   |   |   |   occupation =  Handlers-cleaners:  >50K (5.0/2.0)
|   |   |   |   |   occupation =  Machine-op-inspct
|   |   |   |   |   |   capital.gain <= 1471
|   |   |   |   |   |   |   age <= 38:  >50K (7.0/2.0)
|   |   |   |   |   |   |   age > 38:  <=50K (5.0/1.0)
|   |   |   |   |   |   capital.gain > 1471:  <=50K (2.0/1.0)
|   |   |   |   |   occupation =  Other-service:  >50K (6.0/1.0)
|   |   |   |   |   occupation =  Priv-house-serv:  >50K (0.0)
|   |   |   |   |   occupation =  Prof-specialty:  >50K (24.0/5.0)
|   |   |   |   |   occupation =  Protective-serv
|   |   |   |   |   |   age <= 57:  >50K (14.0/3.0)
|   |   |   |   |   |   age > 57:  <=50K (4.0)
|   |   |   |   |   occupation =  Sales:  >50K (48.0/17.0)
|   |   |   |   |   occupation =  Tech-support:  >50K (21.0/2.0)
|   |   |   |   |   occupation =  Transport-moving:  <=50K (21.0/9.0)
|   |   |   capital.loss > 1762
|   |   |   |   capital.loss <= 1980:  >50K (146.0/1.0)
|   |   |   |   capital.loss > 1980
|   |   |   |   |   capital.loss <= 2057:  <=50K (6.0)
|   |   |   |   |   capital.loss > 2057:  >50K (23.0/1.0)
|   |   marital.status =  Married-spouse-absent:  <=50K (31.0/5.0)
|   |   marital.status =  Never-married
|   |   |   relationship =  Husband:  <=50K (0.0)
|   |   |   relationship =  Not-in-family
|   |   |   |   education =  10th:  <=50K (3.0/1.0)
|   |   |   |   education =  11th:  <=50K (1.0)
|   |   |   |   education =  12th:  <=50K (2.0)
|   |   |   |   education =  1st-4th:  <=50K (2.0)
|   |   |   |   education =  5th-6th:  <=50K (2.0)
|   |   |   |   education =  7th-8th:  <=50K (4.0/1.0)
|   |   |   |   education =  9th:  <=50K (2.0)
|   |   |   |   education =  Assoc-acdm:  <=50K (11.0/3.0)
|   |   |   |   education =  Assoc-voc:  <=50K (15.0/2.0)
|   |   |   |   education =  Bachelors
|   |   |   |   |   hours.per.week <= 42:  <=50K (35.0/4.0)
|   |   |   |   |   hours.per.week > 42
|   |   |   |   |   |   workclass =  Federal-gov:  >50K (1.0)
|   |   |   |   |   |   workclass =  Local-gov:  <=50K (5.0/1.0)
|   |   |   |   |   |   workclass =  Private
|   |   |   |   |   |   |   occupation =  Adm-clerical:  >50K (0.0)
|   |   |   |   |   |   |   occupation =  Armed-Forces:  >50K (0.0)
|   |   |   |   |   |   |   occupation =  Craft-repair:  >50K (0.0)
|   |   |   |   |   |   |   occupation =  Exec-managerial:  >50K (10.0/2.0)
|   |   |   |   |   |   |   occupation =  Farming-fishing:  >50K (0.0)
|   |   |   |   |   |   |   occupation =  Handlers-cleaners:  >50K (0.0)
|   |   |   |   |   |   |   occupation =  Machine-op-inspct:  >50K (0.0)
|   |   |   |   |   |   |   occupation =  Other-service:  >50K (0.0)
|   |   |   |   |   |   |   occupation =  Priv-house-serv:  >50K (0.0)
|   |   |   |   |   |   |   occupation =  Prof-specialty
|   |   |   |   |   |   |   |   race =  Amer-Indian-Eskimo:  <=50K (0.0)
|   |   |   |   |   |   |   |   race =  Asian-Pac-Islander:  <=50K (0.0)
|   |   |   |   |   |   |   |   race =  Black:  >50K (2.0)
|   |   |   |   |   |   |   |   race =  Other:  <=50K (0.0)
|   |   |   |   |   |   |   |   race =  White:  <=50K (10.0/1.0)
|   |   |   |   |   |   |   occupation =  Protective-serv:  >50K (0.0)
|   |   |   |   |   |   |   occupation =  Sales:  >50K (7.0/2.0)
|   |   |   |   |   |   |   occupation =  Tech-support:  >50K (0.0)
|   |   |   |   |   |   |   occupation =  Transport-moving:  >50K (0.0)
|   |   |   |   |   |   workclass =  Self-emp-inc:  <=50K (0.0)
|   |   |   |   |   |   workclass =  Self-emp-not-inc:  <=50K (1.0)
|   |   |   |   |   |   workclass =  State-gov:  <=50K (0.0)
|   |   |   |   |   |   workclass =  Without-pay:  <=50K (0.0)
|   |   |   |   education =  Doctorate:  >50K (11.0)
|   |   |   |   education =  HS-grad:  <=50K (70.0/7.0)
|   |   |   |   education =  Masters:  <=50K (32.0/12.0)
|   |   |   |   education =  Preschool:  <=50K (0.0)
|   |   |   |   education =  Prof-school
|   |   |   |   |   hours.per.week <= 57:  >50K (8.0)
|   |   |   |   |   hours.per.week > 57:  <=50K (3.0/1.0)
|   |   |   |   education =  Some-college
|   |   |   |   |   capital.loss <= 653:  <=50K (43.0/8.0)
|   |   |   |   |   capital.loss > 653:  >50K (4.0/1.0)
|   |   |   relationship =  Other-relative:  <=50K (16.0)
|   |   |   relationship =  Own-child:  <=50K (66.0/3.0)
|   |   |   relationship =  Unmarried:  <=50K (42.0/1.0)
|   |   |   relationship =  Wife:  <=50K (0.0)
|   |   marital.status =  Separated:  <=50K (70.0/10.0)
|   |   marital.status =  Widowed:  <=50K (63.0/12.0)
capital.gain > 5013:  >50K (391.0/5.0)

Number of Leaves  :     177

Size of the tree :      219



          <=50K  >50K
   <=50K   1699   308
   >50K     316  1677
```

```
=== Summary ===

Correctly Classified Instances        3376               84.4    %
Incorrectly Classified Instances       624               15.6    %
Kappa statistic                          0.688 
Mean absolute error                      0.2345
Root mean squared error                  0.3424
Relative absolute error                 46.898  %
Root relative squared error             68.4821 %
Total Number of Instances             4000     

=== Confusion Matrix ===

    a    b   <-- classified as
 1699  308 |    a =  <=50K
  316 1677 |    b =  >50K
```

In [6]:

```
print(mdl.2)
```

```
J48 pruned tree
------------------

capital.gain <= 5013
|   age <= 27:  <=50K (663.0/51.0)
|   age > 27
|   |   marital.status =  Divorced
|   |   |   hours.per.week <= 43:  <=50K (306.0/47.0)
|   |   |   hours.per.week > 43
|   |   |   |   workclass =  Federal-gov:  >50K (6.0)
|   |   |   |   workclass =  Local-gov:  <=50K (11.0/2.0)
|   |   |   |   workclass =  Private
|   |   |   |   |   education =  10th:  <=50K (1.0)
|   |   |   |   |   education =  11th:  <=50K (1.0)
|   |   |   |   |   education =  12th:  >50K (2.0)
|   |   |   |   |   education =  1st-4th:  <=50K (0.0)
|   |   |   |   |   education =  5th-6th:  <=50K (1.0)
|   |   |   |   |   education =  7th-8th:  <=50K (2.0)
|   |   |   |   |   education =  9th:  <=50K (0.0)
|   |   |   |   |   education =  Assoc-acdm:  <=50K (3.0/1.0)
|   |   |   |   |   education =  Assoc-voc:  <=50K (3.0/1.0)
|   |   |   |   |   education =  Bachelors:  >50K (20.0/6.0)
|   |   |   |   |   education =  Doctorate:  <=50K (0.0)
|   |   |   |   |   education =  HS-grad:  <=50K (28.0/6.0)
|   |   |   |   |   education =  Masters:  >50K (5.0/1.0)
|   |   |   |   |   education =  Preschool:  <=50K (0.0)
|   |   |   |   |   education =  Prof-school:  <=50K (0.0)
|   |   |   |   |   education =  Some-college:  <=50K (15.0/6.0)
|   |   |   |   workclass =  Self-emp-inc:  <=50K (9.0/4.0)
|   |   |   |   workclass =  Self-emp-not-inc:  >50K (13.0/3.0)
|   |   |   |   workclass =  State-gov:  <=50K (3.0/1.0)
|   |   |   |   workclass =  Without-pay:  <=50K (0.0)
|   |   marital.status =  Married-AF-spouse:  >50K (5.0/1.0)
|   |   marital.status =  Married-civ-spouse
|   |   |   capital.loss <= 1762
|   |   |   |   education =  10th
|   |   |   |   |   age <= 50:  <=50K (13.0/3.0)
|   |   |   |   |   age > 50:  >50K (14.0/6.0)
|   |   |   |   education =  11th:  <=50K (32.0/7.0)
|   |   |   |   education =  12th:  <=50K (9.0/4.0)
|   |   |   |   education =  1st-4th:  <=50K (4.0/2.0)
|   |   |   |   education =  5th-6th:  <=50K (12.0/2.0)
|   |   |   |   education =  7th-8th:  <=50K (37.0/8.0)
|   |   |   |   education =  9th:  <=50K (19.0/6.0)
|   |   |   |   education =  Assoc-acdm:  >50K (56.0/16.0)
|   |   |   |   education =  Assoc-voc:  >50K (88.0/30.0)
|   |   |   |   education =  Bachelors:  >50K (379.0/63.0)
|   |   |   |   education =  Doctorate:  >50K (37.0/2.0)
|   |   |   |   education =  HS-grad
|   |   |   |   |   hours.per.week <= 39:  <=50K (67.0/21.0)
|   |   |   |   |   hours.per.week > 39
|   |   |   |   |   |   race =  Amer-Indian-Eskimo:  <=50K (4.0)
|   |   |   |   |   |   race =  Asian-Pac-Islander:  <=50K (14.0/3.0)
|   |   |   |   |   |   race =  Black:  <=50K (27.0/13.0)
|   |   |   |   |   |   race =  Other:  <=50K (1.0)
|   |   |   |   |   |   race =  White:  >50K (449.0/172.0)
|   |   |   |   education =  Masters:  >50K (144.0/15.0)
|   |   |   |   education =  Preschool:  <=50K (4.0)
|   |   |   |   education =  Prof-school:  >50K (42.0/7.0)
|   |   |   |   education =  Some-college:  >50K (313.0/109.0)
|   |   |   capital.loss > 1762:  >50K (175.0/8.0)
|   |   marital.status =  Married-spouse-absent:  <=50K (31.0/5.0)
|   |   marital.status =  Never-married:  <=50K (408.0/83.0)
|   |   marital.status =  Separated:  <=50K (70.0/10.0)
|   |   marital.status =  Widowed:  <=50K (63.0/12.0)
capital.gain > 5013:  >50K (391.0/5.0)

Number of Leaves  :     53

Size of the tree :      64
```

In [7]:

```
print(mdl.3)
```

```
J48 pruned tree
------------------

marital.status =  Married-civ-spouse
|   capital.gain <= 5013.0
|   |   native.country =  Mexico:  <=50K (29.0/5.0)
|   |   native.country !=  Mexico
|   |   |   education =  7th-8th:  <=50K (38.0/8.0)
|   |   |   education !=  7th-8th
|   |   |   |   education =  11th:  <=50K (39.0/9.0)
|   |   |   |   education !=  11th
|   |   |   |   |   capital.loss <= 1762.0
|   |   |   |   |   |   education =  Doctorate:  >50K (37.0/2.0)
|   |   |   |   |   |   education !=  Doctorate
|   |   |   |   |   |   |   education =  Masters:  >50K (145.0/16.0)
|   |   |   |   |   |   |   education !=  Masters
|   |   |   |   |   |   |   |   education =  Bachelors:  >50K (394.0/68.0)
|   |   |   |   |   |   |   |   education !=  Bachelors
|   |   |   |   |   |   |   |   |   occupation =  Exec-managerial:  >50K (163.0/32.0)
|   |   |   |   |   |   |   |   |   occupation !=  Exec-managerial
|   |   |   |   |   |   |   |   |   |   education =  Prof-school:  >50K (39.0/6.0)
|   |   |   |   |   |   |   |   |   |   education !=  Prof-school
|   |   |   |   |   |   |   |   |   |   |   occupation =  Tech-support:  >50K (48.0/8.0)
|   |   |   |   |   |   |   |   |   |   |   occupation !=  Tech-support
|   |   |   |   |   |   |   |   |   |   |   |   occupation =  Prof-specialty
|   |   |   |   |   |   |   |   |   |   |   |   |   age <= 31.0:  <=50K (10.0/3.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   age > 31.0:  >50K (55.0/9.0)
|   |   |   |   |   |   |   |   |   |   |   |   occupation !=  Prof-specialty
|   |   |   |   |   |   |   |   |   |   |   |   |   age <= 28.0:  <=50K (97.0/26.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   age > 28.0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   occupation =  Farming-fishing:  <=50K (54.0/17.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   occupation !=  Farming-fishing
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   workclass =  Federal-gov:  >50K (26.0/6.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   workclass !=  Federal-gov
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   education =  9th:  <=50K (15.0/5.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   education !=  9th
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age <= 59.0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   race =  Asian-Pac-Islander
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   education =  HS-grad:  <=50K (12.0/2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   education !=  HS-grad:  >50K (11.0/4.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   race !=  Asian-Pac-Islander
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   hours.per.week <= 55.0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   occupation =  Transport-moving:  <=50K (66.0/27.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   occupation !=  Transport-moving
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   workclass =  Local-gov:  >50K (33.0/8.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   workclass !=  Local-gov
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   occupation =  Handlers-cleaners:  >50K (14.0/4.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   occupation !=  Handlers-cleaners
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   occupation =  Protective-serv:  <=50K (12.0/4.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   occupation !=  Protective-serv
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age <= 44.0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   relationship =  Wife:  >50K (41.0/16.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   relationship !=  Wife
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   hours.per.week <= 39.0:  <=50K (21.0/3.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   hours.per.week > 39.0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   occupation =  Sales:  >50K (33.0/14.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   occupation !=  Sales
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   workclass =  Private:  <=50K (140.0/62.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   workclass !=  Private:  >50K (10.0/4.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age > 44.0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   relationship =  Wife:  <=50K (23.0/9.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   relationship !=  Wife:  >50K (159.0/51.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   hours.per.week > 55.0:  >50K (64.0/15.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age > 59.0:  <=50K (60.0/19.0)
|   |   |   |   |   capital.loss > 1762.0:  >50K (176.0/6.0)
|   capital.gain > 5013.0:  >50K (309.0/1.0)
marital.status !=  Married-civ-spouse
|   capital.gain <= 4650.0
|   |   hours.per.week <= 43.0:  <=50K (1174.0/99.0)
|   |   hours.per.week > 43.0
|   |   |   workclass =  Federal-gov:  >50K (18.0/4.0)
|   |   |   workclass !=  Federal-gov
|   |   |   |   relationship =  Own-child:  <=50K (36.0/1.0)
|   |   |   |   relationship !=  Own-child
|   |   |   |   |   age <= 25.0:  <=50K (31.0/1.0)
|   |   |   |   |   age > 25.0
|   |   |   |   |   |   capital.loss <= 1092.0
|   |   |   |   |   |   |   education =  HS-grad:  <=50K (65.0/12.0)
|   |   |   |   |   |   |   education !=  HS-grad
|   |   |   |   |   |   |   |   occupation =  Exec-managerial
|   |   |   |   |   |   |   |   |   sex =  Female:  <=50K (25.0/11.0)
|   |   |   |   |   |   |   |   |   sex !=  Female:  >50K (20.0/4.0)
|   |   |   |   |   |   |   |   occupation !=  Exec-managerial
|   |   |   |   |   |   |   |   |   sex =  Female:  <=50K (57.0/13.0)
|   |   |   |   |   |   |   |   |   sex !=  Female
|   |   |   |   |   |   |   |   |   |   education =  Some-college:  <=50K (25.0/5.0)
|   |   |   |   |   |   |   |   |   |   education !=  Some-college
|   |   |   |   |   |   |   |   |   |   |   marital.status =  Divorced:  >50K (17.0/6.0)
|   |   |   |   |   |   |   |   |   |   |   marital.status !=  Divorced
|   |   |   |   |   |   |   |   |   |   |   |   occupation =  Sales:  >50K (10.0/3.0)
|   |   |   |   |   |   |   |   |   |   |   |   occupation !=  Sales:  <=50K (36.0/15.0)
|   |   |   |   |   |   capital.loss > 1092.0:  >50K (23.0/7.0)
|   capital.gain > 4650.0:  >50K (90.0/4.0)

Number of Leaves  :     46

Size of the tree :      91
```

Problem 2

Select any two out of the tree decision trees. Perform k-fold cross validation on both of them. Compare the results. Which one is more likely to perform better on the test dataset?

In [8]:

```
evaluate_Weka_classifier(mdl.2, class = T, numFolds = 10)
evaluate_Weka_classifier(mdl.3, class = T, numFolds = 10)
```

```
=== 10 Fold Cross Validation ===

=== Summary ===

Correctly Classified Instances        3188               79.7    %
Incorrectly Classified Instances       812               20.3    %
Kappa statistic                          0.5941
Mean absolute error                      0.2809
Root mean squared error                  0.3824
Relative absolute error                 56.1726 %
Root relative squared error             76.4777 %
Total Number of Instances             4000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.775    0.181    0.812      0.775    0.793      0.595    0.864     0.838      <=50K
                 0.819    0.225    0.783      0.819    0.801      0.595    0.864     0.860      >50K
Weighted Avg.    0.797    0.203    0.798      0.797    0.797      0.595    0.864     0.849

=== Confusion Matrix ===

    a    b   <-- classified as
 1556  451 |    a =  <=50K
  361 1632 |    b =  >50K
=== 10 Fold Cross Validation ===

=== Summary ===

Correctly Classified Instances        3224               80.6    %
Incorrectly Classified Instances       776               19.4    %
Kappa statistic                          0.612
Mean absolute error                      0.26
Root mean squared error                  0.3719
Relative absolute error                 52.0044 %
Root relative squared error             74.3734 %
Total Number of Instances             4000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.791    0.179    0.817      0.791    0.804      0.612    0.882     0.869      <=50K
                 0.821    0.209    0.796      0.821    0.808      0.612    0.882     0.881      >50K
Weighted Avg.    0.806    0.194    0.806      0.806    0.806      0.612    0.882     0.875

=== Confusion Matrix ===

    a    b   <-- classified as
 1587  420 |    a =  <=50K
  356 1637 |    b =  >50K
```

Problem 3

Perform holdout cross validations on the two classifiers selected in Problem 2.

1. Compare the results between the two classifiers.
2. Does the results from holdout cross validation agree with k-fold cross validation?

Problem 4

Load the adult test dataset, and predict the classes in the testing dataset. Compare the performance of the two classifiers. Present the ROC curve and confusion matrix.

In [9]:

```r
adult.test <- read.csv("adult_test.csv")
p.2 <- predict(mdl.2, newdata = adult.test, type = c("class"))
p.3 <- predict(mdl.3, newdata = adult.test, type = c("class"))
```

In [10]:

```r
## accuracy of model p.2
accuracy_p_2 = sum(adult.test$income == p.2)/length(p.2)

accuracy_p_2
```

```
0.769
```

In [11]:

```r
## accuracy of model p.3
accuracy_p_3 = sum(adult.test$income == p.3)/length(p.3)
accuracy_p_3
```

```
0.6
```

In [ ]:

```r
```