Madeline Younes
z5208494

Assignment 2 – Machine Learning

Part 1: Decision Trees

Question 1.1 (5 marks):

In order to make comparisons to the maximum gain tree, the maximum gain tree needs to be generated. This is obtained by dividing the data into positive and negative cases then calculating the gain. Assuming that read is the positive case, the skip and ? cases are negative.

$S(9+, 11-) = 0.9927$
$S_{unknown}(3+, 5-) = 0.9544$
$S_{known}(6+, 6-) = 1$

$$Gain(S, Author) = Entropy(S) - \sum_{v=\{known,unkown\}} \frac{|S_v|}{S}$$
$$= 0.993 - \frac{8}{20} * 0.954 - \frac{12}{20} * 1 = 0.0114$$

$S_{long}(0+, 8-) = 0$
$S_{short}(9+, 3-) = 0.81127$

$$Gain(S, Length) = 0.993 - 0 - \frac{12}{20} * 0.81127$$
$$= 0.506 \ldots$$

$S_{new}(7+, 4-) = 0.94566$
$S_{followup}(2+, 7-) = 0.764204$

$$Gain(S, Thread) = 0.993 - \frac{11}{20} * 0.946 - \frac{9}{20} * 0.764$$
$$= 0.1289 \ldots$$

$S_{home}(4+, 5-) = 0.991076$
$S_{work}(5+, 6-) = 0.9940302$

$$Gain(S, Where\ Read) = 0.993 - \frac{9}{20} * 0.991 - \frac{11}{20} * 0.994$$
$$= 0.00035 \ldots$$

Thereby, *Length* has the largest gain at 0.506. The first split is at the *Length* category.

Since, the *Long* branch has a gain of zero the next split is determined by the maximum gain on the *Short* branch.

$S_{short}(9+, 3-) = 0.81127$

$S_{unknown}(3+, 3-) = 0$
$S_{known}(6+, 0-) = 0$
$Gain(S, Author) = 0$

$S_{new}(7+, 0-) = 0$
$S_{followup}(2+, 3-) = 0.9719$

$$Gain(S, Thread) = 0.811 - 0 - \frac{5}{12} * 0.971$$
$$= 0.406 \ldots$$

$S_{home}(4+, 2-) = 0.91829583$
$S_{real}(5+, 1-) = 0.650022$

$$Gain(S, Where\ Read) = 0.506 - \frac{6}{12} * 0.918 - \frac{6}{12} * 0.650$$
$$= 0.027 \ldots$$

Thereby, *Thread* has the largest gain at 0.406. So, the second split is at the Thread category.

Since the *New* branch has a gain of zero the next split is determined by the maximum gain on the *Followup* branch.

$S_{followup}(2+,3\,-) = 0.971$

$S_{unknown}(0+,3\,-) = 0$
$S_{known}(2+,0\,-) = 0$
$Gain(S, Author) = 0.971 - 0$

$S_{home}(1+,2\,-) = 0.9182$
$S_{work}(1+,1\,-) = 1$

$$\begin{aligned} Gain(S, Author) &= 0.971 - \frac{3}{5} * 0.918 - \frac{2}{5} * 1 \\ &= 0.0202 \ldots \end{aligned}$$

Author clearly has the largest gain at 0.406. So, the third split is at the *Author* category. Thereby, the decision tree is in the order of [*Length*, *Thread* , *Author, WhereRead*]. The expanded version of the tree is depicted in *Figure 1.1.1.*

Figure 1.1.1: Expanded Maximum Gain Tree [*Length*, *Thread* , *Author, WhereRead*]

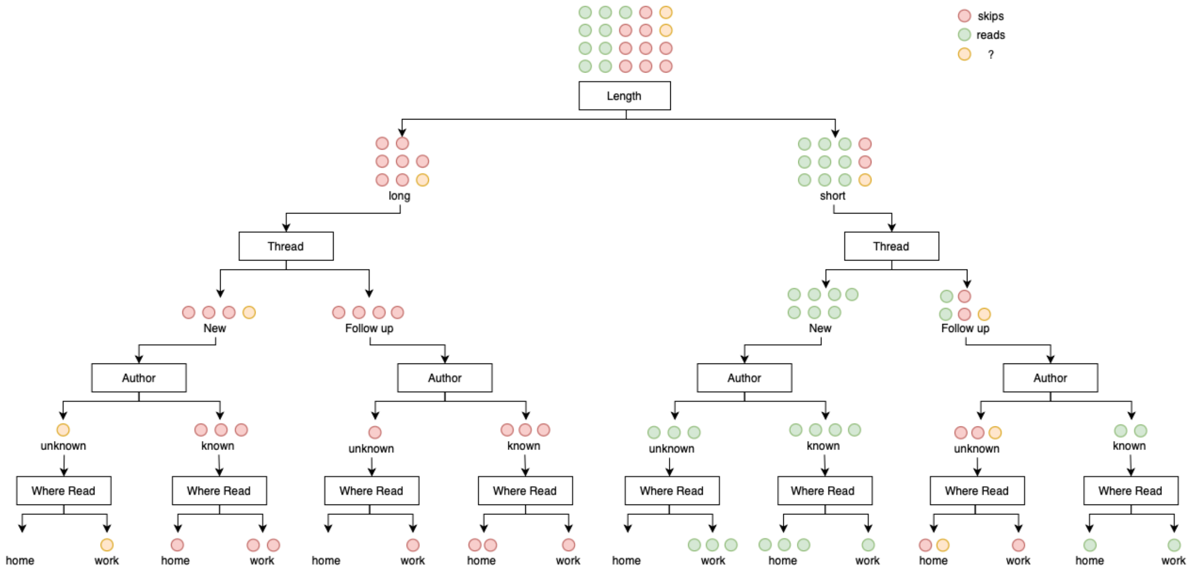
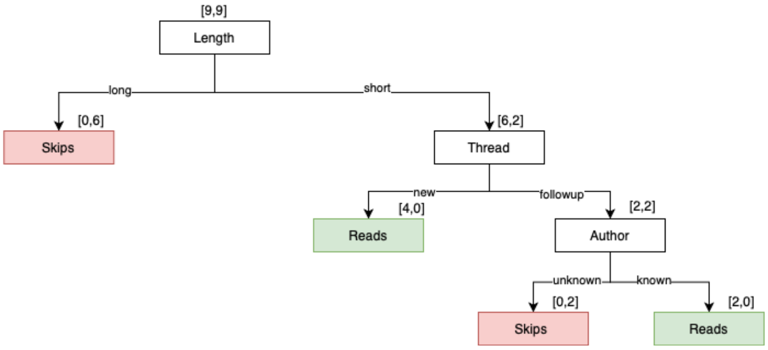


Figure 1.1.2: Simplified Maximum Gain Tree [*Length*, *Thread* , *Author, WhereRead*]



a) Suppose you change the algorithm to always select the first element of the list of features. What tree is found when the features are in the order [*Author, Thread, Length, WhereRead*]? Does this tree represent a different function than that found with the maximum information gain split? Explain.

The tree generated from the order [*Author, Thread, Length, WhereRead*] is depicted in *Figure 1.1.3* and is then simplified to the tree is *Figure 1.1.4*. Using e_{19} and e_{20} as test inputs for the Tree generated by [*Author, Thread, Length, WhereRead*], the *User_action* is *reads* and *skips* respectively. Whilst the output for the maximum information split tree *(Figure 1.1.2)* is *skip* and *skip*. Hence, the function is different to the maximum gain split.

Also, by looking at the function rules for the tree [*Author, Thread, Length, WhereRead*] which are:

Skips = known ***and*** new ***and*** long
Skips = known ***and*** followup ***and*** long
Skips = unknown ***and*** followup
Reads = known ***and*** followup ***and*** short
Read = unknown ***and*** new
Reads = known ***and*** new ***and*** short

And comparing them to the function rules of the maximum gain tree which are:

Skips = long
Skips = short ***and*** followup ***and*** unknown
Reads = short ***and*** new
Reads = short ***and*** followup ***and*** known

It can be confirmed that the two functions are different.

Figure 1.1.3: Expanded Tree [*Author, Thread, Length, WhereRead*]

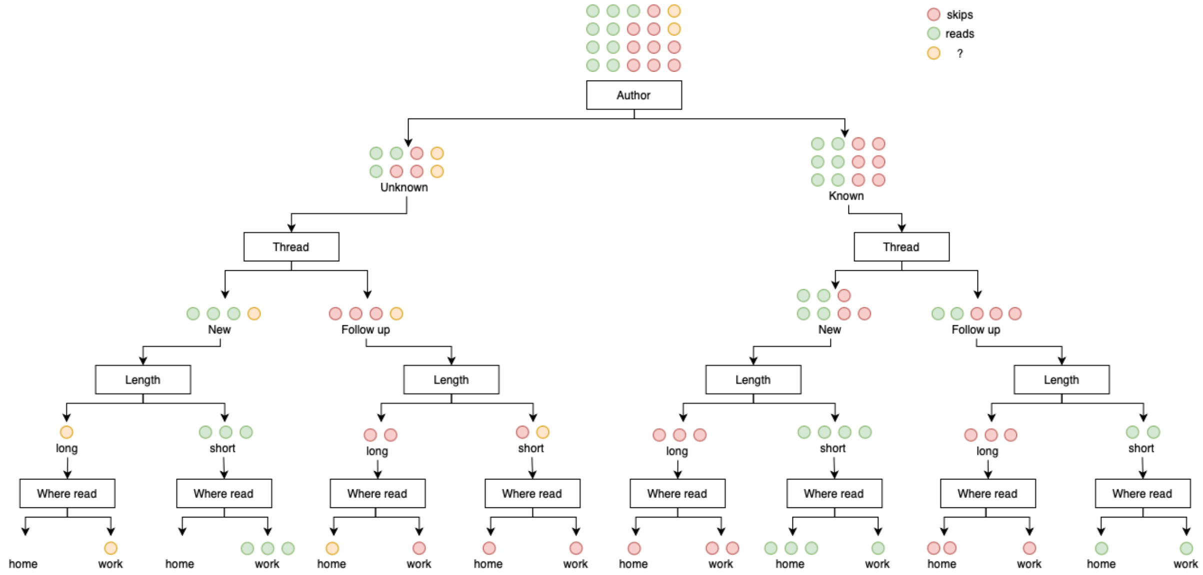
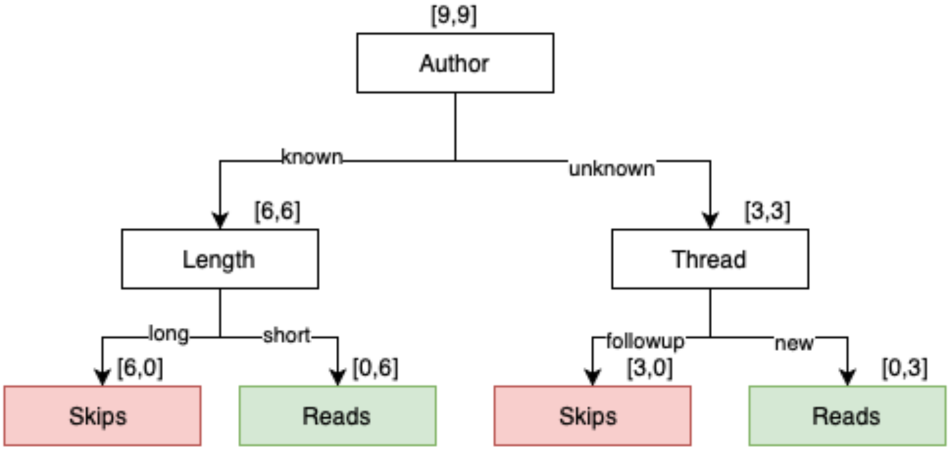


Figure 1.1.4: Simplified Tree [*Author, Thread, Length, WhereRead*]



b) What tree is found when the features are in the order [*WhereRead, Thread, Length, Author*]? Does this tree represent a different function than that found with the maximum information gain split or the one given for the preceding part? Explain.

The tree generated from the order [*WhereRead, Thread, Length, Author*] is depicted in *Figure 1.1.4* and the simplified version of the tree is *in Figure 1.1.5.* This tree although seems different to the simplified maximum gain tree in *Figure 1.1.2* produces the same output for the e_{19} and e_{20} test cases given. Thereby, is a similar function to the maximum gain tree.

Table 1.1.1 Comparison of User Result for Section B and Maximum Gain Tree

Test Case	User Result	Maximum Gain Tree	Section B Tree
e_{19}	skips	skips	skips
e_{20}	skips	skips	skips

Both of the tree's functions can be simplified to the rules:

Skips = long
Skips = short ***and*** followup ***and*** unknown
Reads = short ***and*** new
Reads = short ***and*** followup ***and*** known

Thereby, confirming that both trees share the same function.

Figure 1.1.4: Expanded Tree [*WhereRead, Thread, Length, Author]*

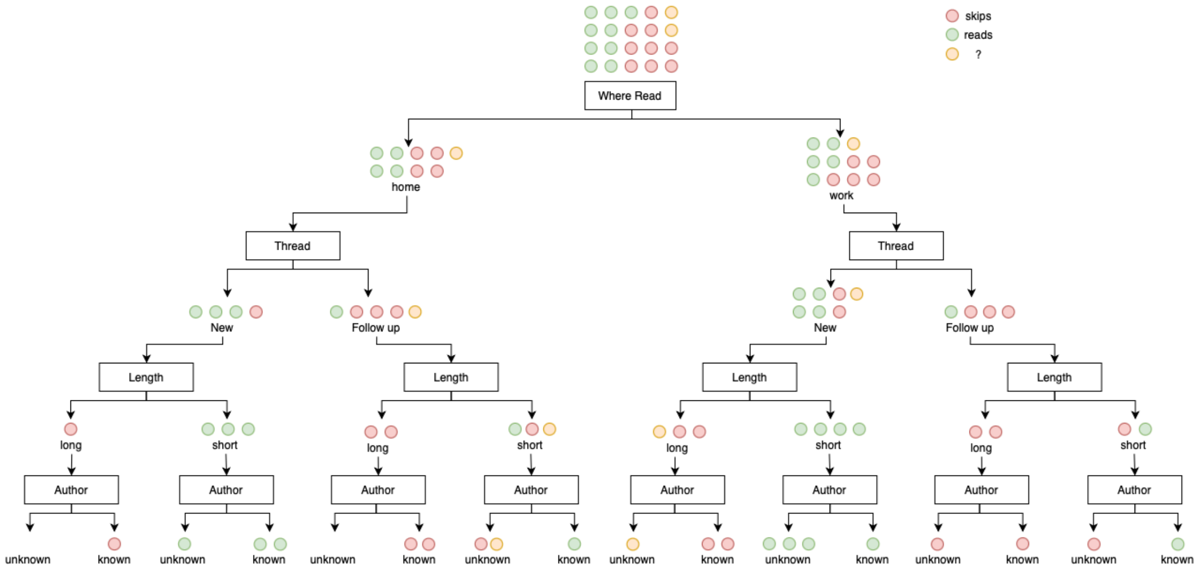
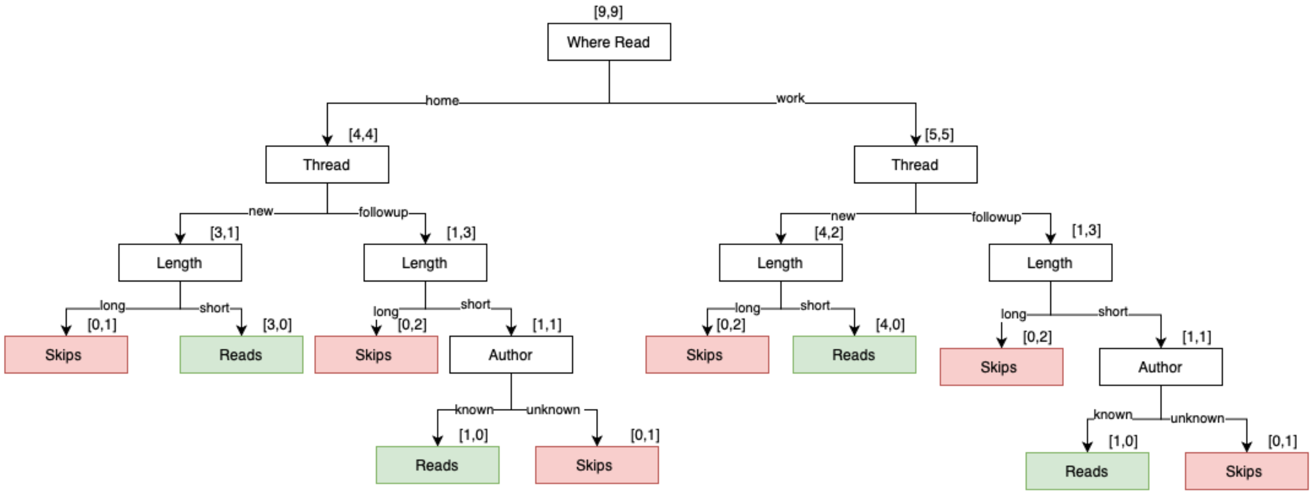


Figure 1.1.5: Simplified Tree [*WhereRead, Thread, Length, Author]*



c) Is there a tree that correctly classifies the training examples but represents a different function than those found by the preceding algorithms? If so, give it. If not, explain why.

There is another tree which can be obtained removing the cases (8, 9, 12, 15, 17,18) which are repeated examples in the training data. By removing the duplicates a tree of the order [Author, Where Read, Length, Thread] is produced shown in *Figure 1.1.6*. By generating an additional testcase as seen in *Table 1.1.2* which was not included in the training data further testing can be performed on all the trees. Using the additional test case it can be seen that the new tree performs differently to the previous trees as seen in *Table 1.1.3.*

Table 1.1.2: Test Cases

Test Case	*Author*	*Thread*	*Length*	*Where_read*	*User_Action*
e_{19}	*Unknown*	*New*	*Long*	*Work*	*?*
e_{20}	*Unknown*	*Followup*	*Short*	*Home*	*?*
e_{21}	*Unknown*	*New*	*Short*	*Home*	*?*

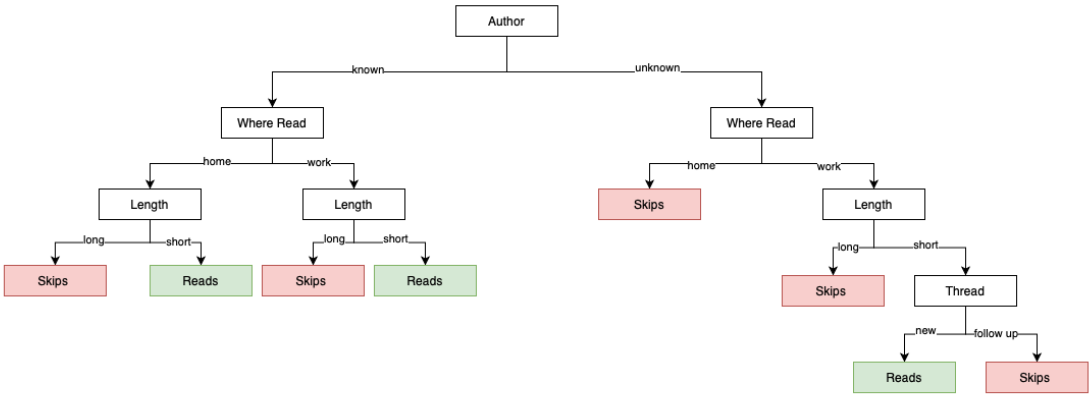
Table 1.1.2: User Action based on Decision Tree Functions

Test Case	*Section A Tree*	*Section B Tree/Maximum Gain Tree*	*Section C Tree*
e_{19}	*Read*	*Skip*	*Skip*
e_{20}	*Skip*	*Skip*	*Skip*
e_{21}	*Read*	*Read*	*Skip*

The Tree with the order [Author, Where Read, Length, Thread] producing the rules:
Skips = long
Skips = unknown ***and*** followup
Reads = known ***and*** short
Reads = unknown ***and*** new

These rules contradict the rule Read = short ***and*** new present in the section A and section B trees thereby, confirming that the tree produced from the order [Author, Where Read, Length, Thread] is a new function.

Figure 1.1.5: Simplified Tree [Author, Where Read, Length, Thread]



Question 1.2 (5 marks):

The goal of a decision tree is to formulate a good generalisation model that is neither underfitted or overfitted to the given data. Pruning is a technique used to ensure that data is not overfitted in the models. Pruning can be achieved in Weka by varying the MinNumber, the confidenceFactor and enabling subtreeRaising. The MinNumber sets the minimum number of cases that reach a leaf, limiting the number of splits to prevent the nodes from becoming too small. The confidenceFactor sets the value for the statical test that is performed on the leaves at each level of the tree once it is constructed working from bottom to top. The subtreeRaising enables to the program to prune the interior node raising the subtree up one stage.

There are several different ways to generate a decision tree in Weka using supervised machine learning. The test and training data can be combined in one large data document to utilise cross-validation and percentage split methods. Both these methods split the data given into training data and test data based on the folds and percentage split settings. Alternatively, a specified test and training data set can be used. In all methods prior to generating the model, a header specifying the classification of the data *(Figure 1.2.3)* is added and the .data files are converted to .arff files to be read by Weka. Whilst experimenting with the settings the accuracy is determined using the percentage next to Correctly Classified Instances which is calculated using the formula: Accuracy = Correctly Classified Instances/(Correctly Classified Instances + Incorrectly Classified Instances).

In Tables 1.2.1 and 1.2.1 the effects of varying the folds and percentage split can be observed respectively. Varying the folds inconsistently affected the accuracy as the folds
Whilst the increasing the percentage split seemed to improve the accuracy up to 80%, before it dropped off. Although, this increase in accuracy may have been due to overfitting the data or not providing enough data for the model to test against. Thereby, to obtain consistent results set training and test data files were used.

Observing *Table 1.2.3* it is seen that varying the confidence factor increases the size of the tree and it reaches peak accuracy at 0.1, with the following results being over fitted to the data. In *Table 1.2.4*, the MinNumber of objects is varied observing the affect of forcing the tree size on the accuracy. As the MinNumber is increased the tree size decreases as expected, the highest level of accuracy is obtained when the MinNumber is at 10. Combining the results from the tables, the MinNumber of objects is set to 10 and the confidence factor is tuned in order to obtain the greatest accuracy in *Table 1.2.5*. The Subtree Raising is also enabled to allow for interior node pruning and the highest accuracy obtained is 86.7453 %.

Table 1.2.6 depicts the affect of utilising reduced error pruning and varying the number of folds used in cross validation within the J48 settings is experimented with although a result with a higher accuracy and smaller tree size than that with the settings (-B -C 0.05 -M 10) is not obtained.

Hence, the tree with the greatest levels of accuracy and smallest sized tree can be modelled with the settings (-B -C 0.05 -M 10). It's accuracy is shown in *Figure 1.2.1* and the tree visualisation is shown in *Figure 1.2.2*.

<u>Utilising One Data File</u>

The Base settings were set to (-S -R -B -N 3 -Q 1 -M 2), the percentage and number of folds split were varied in their respective tables.

Table 1.2.1: Varying the Number of folds

No. fold	Size of Tree	Accuracy
5	911	86.141%
10	911	86.0939%
20	911	86.7453%

Table 1.2.2: Varying the percentage split

Split	Size of Tree	Accuracy
20%	911	85.1129%
50%	911	85.6763%
60%	911	85.8883 %
66%	911	85.8786%
70%	911	86.2008%
80%	911	86.1384%
90%	911	85.4218%

Utilising separate Training and Test Data Files

Table 1.2.3: Varying the Confidence factor

Settings	Size of Tree	Accuracy
-S -B -C 0.001 -M 2	65	85.9898 %
-S -B -C 0.05 -M 2	235	86.6163 %
-S -B -C 0.1 -M 2	247	86.604 %
-S -B -C 0.25 -M 2	725	86.4873 %
-S -B -C 0.7 -M 2	3977	84.4236 %

Table 1.2.4: Varying the MinNumber of Objects

Settings	Size of Tree	Accuracy
-S -B -C 0.05 -M 100	71	86.3338 %
-S -B -C 0.05 -M 50	77	86.4259 %
-S -B -C 0.05 -M 15	123	86.6839 %
-S -B -C 0.05 -M 10	141	86.7207 %
-S -B -C 0.05 -M 5	175	86.6839 %

Table 1.2.5: Varying the Confidence Factor and Enabling Subtree Raising

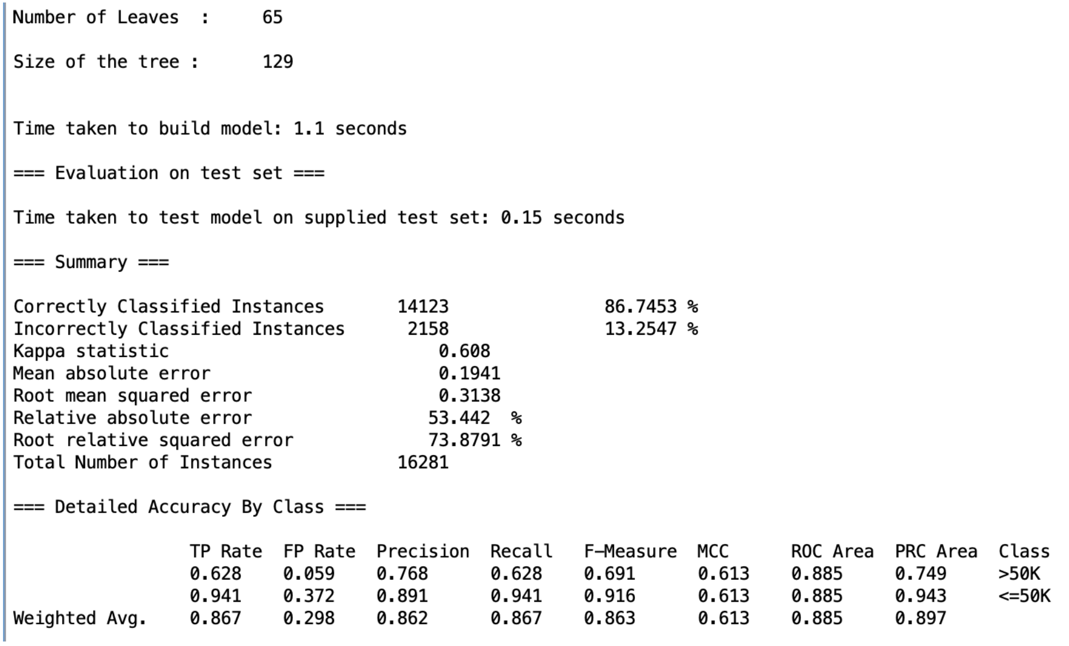
Settings	Size of Tree	Accuracy
-B -C 0.001 -M 10	38	85.6827 %
-B -C 0.025 -M 10	127	86.7146 %
-B -C 0.05 -M 10	129	86.7453 %
-B -C 0.1 -M 10	143	86.6839 %
-B -C 0.25 -M 10	297	86.1328 %

Table 1.2.6: Varying the number of folds with Reduced Error Pruning Enabled

Settings	Size of Tree	Accuracy
-S -R -B -N 3 -Q 1 -M 10	221	86.1986 %
-S -R -B -N 5 -Q 1 -M 10	281	86.2478 %
-S -R -B -N 7 -Q 1 -M 10	257	86.5119 %

-S -R -B -N 10 -Q 1 -M 10	369	86.2232 %
-S -R -B -N 100 -Q 1 -M 10	143	86.1618 %

Figure 1.2.1: Tree Size and Detailed Accuracy Printout



```
Number of Leaves  :     65

Size of the tree :      129


Time taken to build model: 1.1 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.15 seconds

=== Summary ===

Correctly Classified Instances       14123               86.7453 %
Incorrectly Classified Instances      2158               13.2547 %
Kappa statistic                          0.608
Mean absolute error                      0.1941
Root mean squared error                  0.3138
Relative absolute error                 53.442  %
Root relative squared error             73.8791 %
Total Number of Instances            16281

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.628    0.059    0.768      0.628    0.691      0.613    0.885     0.749     >50K
                 0.941    0.372    0.891      0.941    0.916      0.613    0.885     0.943     <=50K
Weighted Avg.    0.867    0.298    0.862      0.867    0.863      0.613    0.885     0.897
```

Figure 1.2.2: Tree Generated

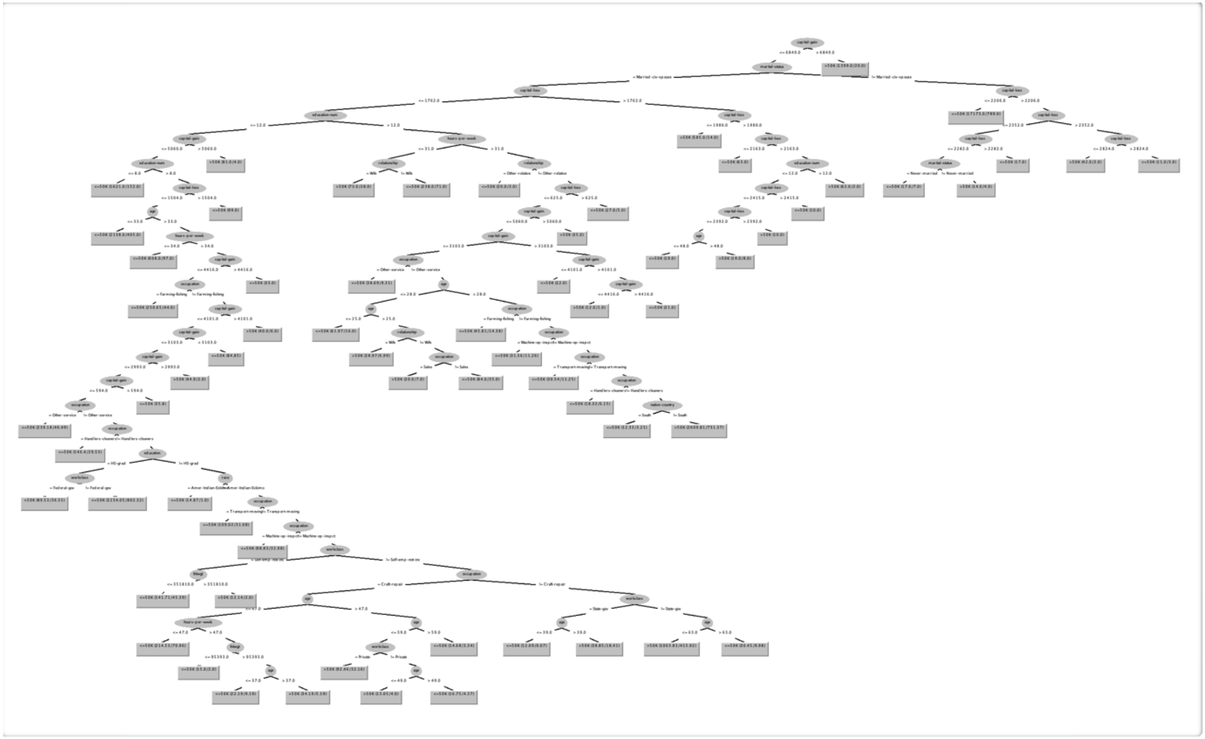


Figure 1.2.3: ARFF File headers

```
@RELATION adult

@ATTRIBUTE age numeric
@ATTRIBUTE workclass {Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked}
@ATTRIBUTE fnlwgt numeric
@ATTRIBUTE education {Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th,
Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool}
@ATTRIBUTE education-num numeric
@ATTRIBUTE marital-status {Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-
AF-spouse}
@ATTRIBUTE occupation {Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners,
Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces}
@ATTRIBUTE relationship {Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried}
@ATTRIBUTE race {White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black}
@ATTRIBUTE sex {Female, Male}
@ATTRIBUTE capital-gain numeric
@ATTRIBUTE capital-loss numeric
@ATTRIBUTE hours-per-week numeric
@ATTRIBUTE native-country {United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc),
India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland,
France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand,
Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands}
@ATTRIBUTE class {>50K, <=50K}

@DATA
```

Figure 1.2.4: J48 printout of Pruned Tree

```
capital-gain <= 6849.0
|   marital-status = Married-civ-spouse
|   |   capital-loss <= 1762.0
|   |   |   education-num <= 12.0
|   |   |   |   capital-gain <= 5060.0
|   |   |   |   |   education-num <= 8.0: <=50K (1621.0/152.0)
|   |   |   |   |   education-num > 8.0
|   |   |   |   |   |   capital-loss <= 1504.0
|   |   |   |   |   |   |   age <= 33.0: <=50K (2138.0/405.0)
|   |   |   |   |   |   |   age > 33.0
|   |   |   |   |   |   |   |   hours-per-week <= 34.0: <=50K (608.0/97.0)
|   |   |   |   |   |   |   |   hours-per-week > 34.0
|   |   |   |   |   |   |   |   |   capital-gain <= 4416.0
|   |   |   |   |   |   |   |   |   |   occupation = Farming-fishing: <=50K (250.65/44.6)
|   |   |   |   |   |   |   |   |   |   occupation != Farming-fishing
|   |   |   |   |   |   |   |   |   |   |   capital-gain <= 4101.0
|   |   |   |   |   |   |   |   |   |   |   |   capital-gain <= 3103.0
|   |   |   |   |   |   |   |   |   |   |   |   |   capital-gain <= 2993.0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   capital-gain <= 594.0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   occupation = Other-service: <=50K (239.18/46.49)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   occupation != Other-service
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   occupation = Handlers-cleaners: <=50K (146.4/29.53)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   occupation != Handlers-cleaners
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   education = HS-grad
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   workclass = Federal-gov: >50K (89.53/36.55)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   workclass != Federal-gov: <=50K (2234.05/802.32)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   education != HS-grad
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   race = Amer-Indian-Eskimo: <=50K (14.87/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   race != Amer-Indian-Eskimo
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   occupation = Transport-moving: <=50K (109.02/31.08)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   occupation != Transport-moving
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   occupation = Machine-op-inspct: <=50K (98.83/32.98)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   occupation != Machine-op-inspct
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   workclass = Self-emp-not-inc
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   fnlwgt <= 351810.0: <=50K (141.71/45.39)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   fnlwgt > 351810.0: >50K (12.14/2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   workclass != Self-emp-not-inc
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   occupation = Craft-repair
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age <= 47.0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   hours-per-week <= 47.0: <=50K
(214.53/70.96)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   hours-per-week > 47.0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   fnlwgt <= 95393.0: <=50K (15.0/2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   fnlwgt > 95393.0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age <= 37.0: <=50K (22.19/9.19)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age > 37.0: >50K (34.19/5.19)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age > 47.0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age <= 59.0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   workclass = Private: >50K (92.46/32.16)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   workclass != Private
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age <= 49.0: >50K (13.05/4.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age > 49.0: <=50K (16.75/4.37)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age > 59.0: <=50K (14.68/3.34)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   occupation != Craft-repair
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   workclass = State-gov: <=50K (50.94/20.51)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   workclass != State-gov
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age <= 63.0: >50K (1063.83/415.92)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   age > 63.0: <=50K (30.45/9.98)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   capital-gain > 594.0: <=50K (35.8)
|   |   |   |   |   |   |   |   |   |   |   |   |   capital-gain > 2993.0: >50K (44.9/2.0)
|   |   |   |   |   |   |   |   |   |   |   |   capital-gain > 3103.0: <=50K (84.85)
|   |   |   |   |   |   |   |   |   |   |   capital-gain > 4101.0: >50K (40.0/6.0)
|   |   |   |   |   |   |   |   |   capital-gain > 4416.0: <=50K (33.0)
|   |   |   |   |   |   capital-loss > 1504.0: <=50K (99.0)
|   |   |   |   capital-gain > 5060.0: >50K (81.0/4.0)
|   |   |   education-num > 12.0
|   |   |   |   hours-per-week <= 31.0
|   |   |   |   |   relationship = Wife: >50K (73.0/28.0)
|   |   |   |   |   relationship != Wife: <=50K (238.0/71.0)
|   |   |   |   hours-per-week > 31.0
|   |   |   |   |   relationship = Other-relative: <=50K (20.0/3.0)
|   |   |   |   |   relationship != Other-relative
|   |   |   |   |   |   capital-loss <= 625.0
|   |   |   |   |   |   |   capital-gain <= 5060.0
|   |   |   |   |   |   |   |   capital-gain <= 3103.0
|   |   |   |   |   |   |   |   |   occupation = Other-service: <=50K (38.69/9.31)
|   |   |   |   |   |   |   |   |   occupation != Other-service
|   |   |   |   |   |   |   |   |   |   age <= 28.0
|   |   |   |   |   |   |   |   |   |   |   age <= 25.0: <=50K (61.97/16.0)
|   |   |   |   |   |   |   |   |   |   |   age > 25.0
|   |   |   |   |   |   |   |   |   |   |   |   relationship = Wife: >50K (28.97/6.99)
|   |   |   |   |   |   |   |   |   |   |   |   relationship != Wife
|   |   |   |   |   |   |   |   |   |   |   |   |   occupation = Sales: >50K (20.0/7.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   occupation != Sales: <=50K (84.0/35.0)
|   |   |   |   |   |   |   |   |   |   age > 28.0
|   |   |   |   |   |   |   |   |   |   |   occupation = Farming-fishing: <=50K (45.81/14.38)
|   |   |   |   |   |   |   |   |   |   |   occupation != Farming-fishing
|   |   |   |   |   |   |   |   |   |   |   |   occupation = Machine-op-inspct: <=50K (31.56/11.26)
|   |   |   |   |   |   |   |   |   |   |   |   occupation != Machine-op-inspct
|   |   |   |   |   |   |   |   |   |   |   |   |   occupation = Transport-moving: <=50K (30.54/11.25)
|   |   |   |   |   |   |   |   |   |   |   |   |   occupation != Transport-moving
|   |   |   |   |   |   |   |   |   |   |   |   |   |   occupation = Handlers-cleaners: <=50K (18.32/6.15)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   occupation != Handlers-cleaners
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   native-country = South: <=50K (12.33/3.21)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   native-country != South: >50K (2609.81/731.37)
|   |   |   |   |   |   |   |   capital-gain > 3103.0
|   |   |   |   |   |   |   |   |   capital-gain <= 4101.0: <=50K (22.0)
|   |   |   |   |   |   |   |   |   capital-gain > 4101.0
|   |   |   |   |   |   |   |   |   |   capital-gain <= 4416.0: >50K (12.0/1.0)
|   |   |   |   |   |   |   |   |   |   capital-gain > 4416.0: <=50K (21.0)
|   |   |   |   |   |   |   capital-gain > 5060.0: >50K (35.0)
|   |   |   |   |   |   capital-loss > 625.0: <=50K (27.0/5.0)
|   |   capital-loss > 1762.0
|   |   |   capital-loss <= 1980.0: >50K (585.0/14.0)
|   |   |   capital-loss > 1980.0
|   |   |   |   capital-loss <= 2163.0: <=50K (63.0)
|   |   |   |   capital-loss > 2163.0
|   |   |   |   |   education-num <= 12.0
|   |   |   |   |   |   capital-loss <= 2415.0
|   |   |   |   |   |   |   capital-loss <= 2392.0
|   |   |   |   |   |   |   |   age <= 48.0: <=50K (19.0)
|   |   |   |   |   |   |   |   age > 48.0: >50K (19.0/8.0)
|   |   |   |   |   |   |   capital-loss > 2392.0: >50K (10.0)
|   |   |   |   |   |   capital-loss > 2415.0: <=50K (10.0)
|   |   |   |   |   education-num > 12.0: >50K (62.0/2.0)
|   marital-status != Married-civ-spouse
|   |   capital-loss <= 2206.0: <=50K (17173.0/789.0)
|   |   capital-loss > 2206.0
|   |   |   capital-loss <= 2352.0
|   |   |   |   capital-loss <= 2282.0
|   |   |   |   |   marital-status = Never-married: <=50K (17.0/7.0)
|   |   |   |   |   marital-status != Never-married: >50K (14.0/4.0)
|   |   |   |   capital-loss > 2282.0: <=50K (17.0)
|   |   |   capital-loss > 2352.0
|   |   |   |   capital-loss <= 2824.0: >50K (42.0/2.0)
|   |   |   |   capital-loss > 2824.0: <=50K (11.0/3.0)
capital-gain > 6849.0: >50K (1399.0/20.0)
```