Question 1: Decision Trees

Before we start, we can simplify the representation of the data by eliminating duplicated data. Then construct a "truth table" to enumerate all the possible data to compare out put of different decision trees. Figure 1.3 clearly shows the comparison among the tree generated by the maximum information gain split and those obtained from question a) and b).

a) Figure1.1.1 is the tree found when the features are in the order [*Author, Thread, Length, WhereRead*]. This tree can be simplified to Figure 1.2. It represents the different function from that found with the maximum information gain split as for any features, these two trees can't give identical UserActions as output. For example, decision for example e_{19} <*unknown, new, long, work*> given by this tree is *reads*, but that given by the tree found with the maximum information gain split is *skips*.

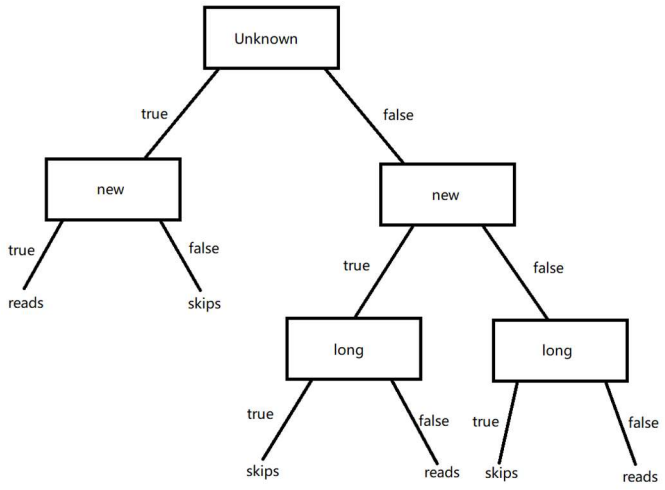


Figure 1.1.1

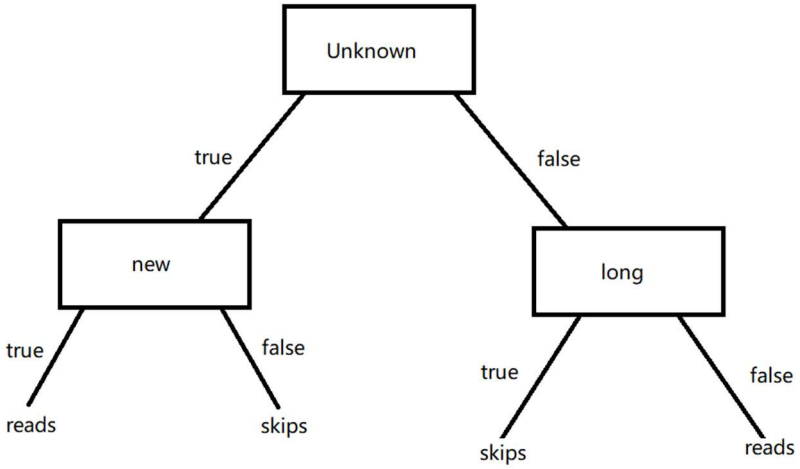


Figure 1.1.2

b) Figure 1.2.1 is the tree found when the features are in the order [*WhereRead, Thread, Length, Author*]. This tree can be simplified to Figure 1.2.2, which represents the same function as that found with the maximum information gain split because it has identical output according to the "truth table" in Figure 1.3. Hence, it represent a different function with the one given for the preceding part as that function is different from the one found with the maximum information gain split.

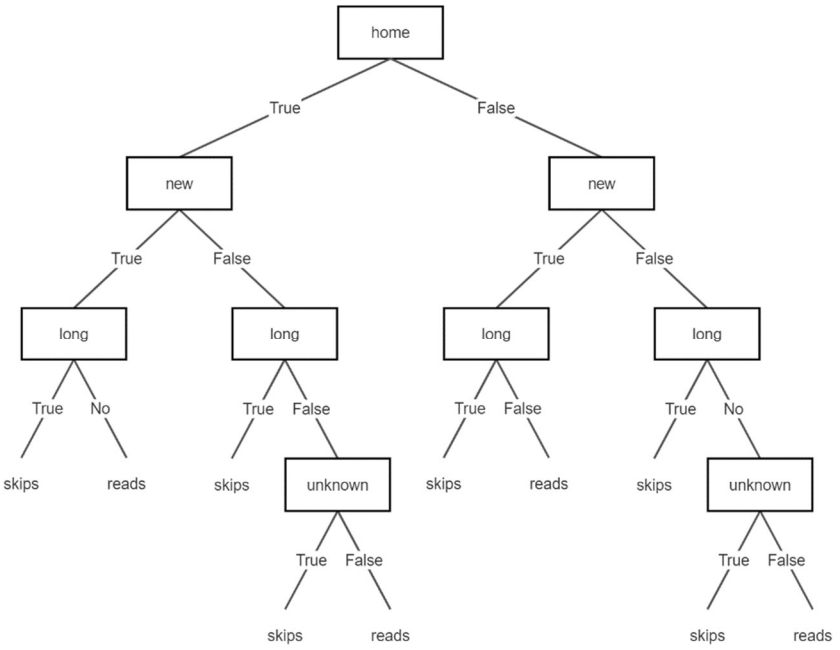


Figure 1.2.1

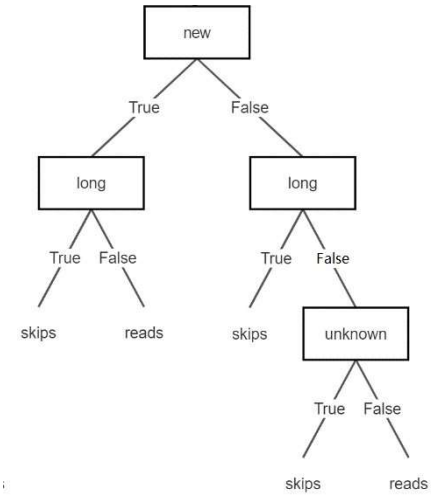


Figure 1.2.2

Example	Author	Thread	Length	Where	Action	maxi	a	b
e4	Known	Foollowup	long	home	skip	skip	skip	skip
e6	Known	Foollowup	long	work	skip	skip	skip	skip
e13	Known	Foollowup	short	home	read	read	read	read
e16	Known	Foollowup	short	work	read	read	read	read
e1	Known	new	long	home	skip	skip	skip	skip
e10	Known	new	long	work	skip	skip	skip	skip
e17	Known	new	short	home	read	read	read	read
e14	Known	new	short	work	read	read	read	read
	Unknown	Foollowup	long	home		skip	skip	skip
e3	Unknown	Foollowup	long	work	skip	skip	skip	skip
e11	Unknown	Foollowup	short	home	skip	skip	skip	skip
e7	Unknown	Foollowup	short	work	skip	skip	skip	skip
	Unknown	new	long	home		skip	read	skip
e19	Unknown	new	long	work		skip	read	skip
	Unknown	new	short	home		read	read	read
e18	Unknown	new	short	work	read	read	read	read

Figure 1.3

c) Figure 1.4 contains a tree that correctly classifies the training examples but represents a different function than those found by the preceding examples. It is obtained by adding branches to Figure 1.1.2. It has exactly the same actions as the training examples but gives different output with any tree mentioned before. For example, decision for <*Unknown, new, short, home*> is *skips* but that of any tree found by the preceding examples is *reads*.

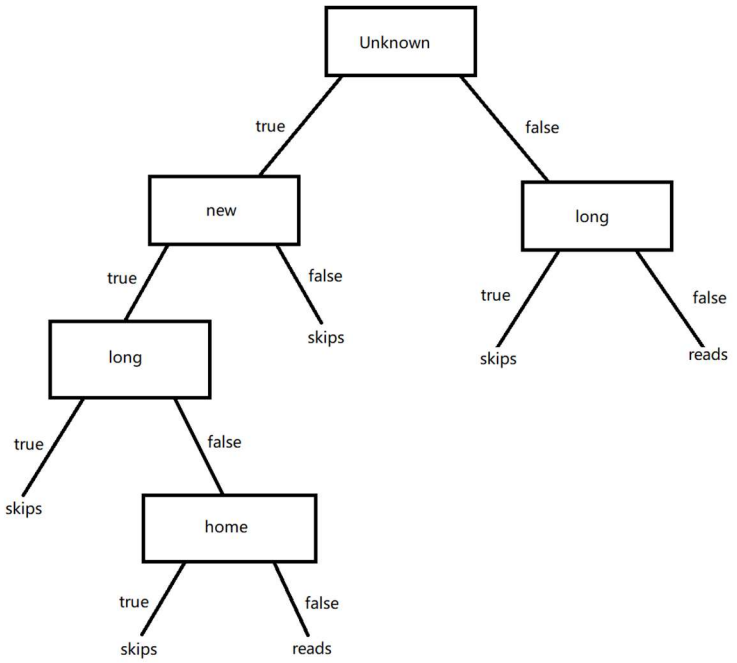


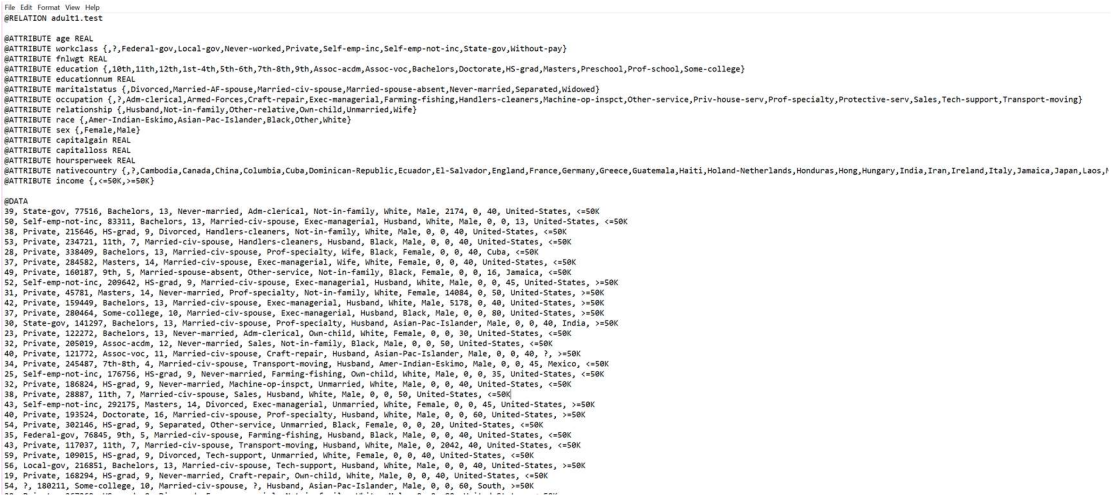
Figure 1.4

Question 2: Decision Trees

I tried sklearn and Weka as tools and finally choose Weka to show my work as there is a lower possibility to get a wrong model using a GUI program as a beginner and it is more intuitive to show the steps.

Firstly, I prepared the data to train by the following steps:

1. Download the data from http://archive.ics.uci.edu/ml/machine-learning-databases/adult/
2. Combine .data and .test together according to the requirement of the instruction on WebCMS3.
3. Unify the dataset by eliminating '.' at the end of some rows and Converting '>50K' to '>=50K'
4. Add header to each column according to adult.names
5. Convert the file type CSV to ARFF, result is shown in Figure 2.1



```
File  Edit  Format  View  Help
@RELATION adult1.test

@ATTRIBUTE age REAL
@ATTRIBUTE workclass {,?,Federal-gov,Local-gov,Never-worked,Private,Self-emp-inc,Self-emp-not-inc,State-gov,Without-pay}
@ATTRIBUTE fnlwgt REAL
@ATTRIBUTE education {,10th,11th,12th,1st-4th,5th-6th,7th-8th,9th,Assoc-acdm,Assoc-voc,Bachelors,Doctorate,HS-grad,Masters,Preschool,Prof-school,Some-college}
@ATTRIBUTE educationnum REAL
@ATTRIBUTE maritalstatus {,Divorced,Married-AF-spouse,Married-civ-spouse,Married-spouse-absent,Never-married,Separated,Widowed}
@ATTRIBUTE occupation {,?,Adm-clerical,Armed-Forces,Craft-repair,Exec-managerial,Farming-fishing,Handlers-cleaners,Machine-op-inspct,Other-service,Priv-house-serv,Prof-specialty,Protective-serv,Sales,Tech-support,Transport-moving}
@ATTRIBUTE relationship {,Husband,Not-in-family,Other-relative,Own-child,Unmarried,Wife}
@ATTRIBUTE race {,Amer-Indian-Eskimo,Asian-Pac-Islander,Black,Other,White}
@ATTRIBUTE sex {,Female,Male}
@ATTRIBUTE capitalgain REAL
@ATTRIBUTE capitalloss REAL
@ATTRIBUTE hoursperweek REAL
@ATTRIBUTE nativecountry {,?,Cambodia,Canada,China,Columbia,Cuba,Dominican-Republic,Ecuador,El-Salvador,England,France,Germany,Greece,Guatemala,Haiti,Holand-Netherlands,Honduras,Hong,Hungary,India,Iran,Ireland,Italy,Jamaica,Japan,Laos,
@ATTRIBUTE income {,<=50K,>=50K}

@DATA
39, State-gov, 77516, Bachelors, 13, Never-married, Adm-clerical, Not-in-family, White, Male, 2174, 0, 40, United-States, <=50K
50, Self-emp-not-inc, 83311, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, White, Male, 0, 0, 13, United-States, <=50K
38, Private, 215646, HS-grad, 9, Divorced, Handlers-cleaners, Not-in-family, White, Male, 0, 0, 40, United-States, <=50K
53, Private, 234721, 11th, 7, Married-civ-spouse, Handlers-cleaners, Husband, Black, Male, 0, 0, 40, United-States, <=50K
28, Private, 338409, Bachelors, 13, Married-civ-spouse, Prof-specialty, Wife, Black, Female, 0, 0, 40, Cuba, <=50K
37, Private, 284582, Masters, 14, Married-civ-spouse, Exec-managerial, Wife, White, Female, 0, 0, 40, United-States, <=50K
49, Private, 160187, 9th, 5, Married-spouse-absent, Other-service, Not-in-family, Black, Female, 0, 0, 16, Jamaica, <=50K
52, Self-emp-not-inc, 209642, HS-grad, 9, Married-civ-spouse, Exec-managerial, Husband, White, Male, 0, 0, 45, United-States, >=50K
31, Private, 45781, Masters, 14, Never-married, Prof-specialty, Not-in-family, White, Female, 14084, 0, 50, United-States, >=50K
42, Private, 159449, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, White, Male, 5178, 0, 40, United-States, >=50K
37, Private, 280464, Some-college, 10, Married-civ-spouse, Exec-managerial, Husband, Black, Male, 0, 0, 80, United-States, >=50K
30, State-gov, 141297, Bachelors, 13, Married-civ-spouse, Prof-specialty, Husband, Asian-Pac-Islander, Male, 0, 0, 40, India, >=50K
23, Private, 122272, Bachelors, 13, Never-married, Adm-clerical, Own-child, White, Female, 0, 0, 30, United-States, <=50K
32, Private, 205019, Assoc-acdm, 12, Never-married, Sales, Not-in-family, Black, Male, 0, 0, 50, United-States, <=50K
40, Private, 121772, Assoc-voc, 11, Married-civ-spouse, Craft-repair, Husband, Asian-Pac-Islander, Male, 0, 0, 40, ?, >=50K
34, Private, 245487, 7th-8th, 4, Married-civ-spouse, Transport-moving, Husband, Amer-Indian-Eskimo, Male, 0, 0, 45, Mexico, <=50K
25, Self-emp-not-inc, 176756, HS-grad, 9, Never-married, Farming-fishing, Own-child, White, Male, 0, 0, 35, United-States, <=50K
32, Private, 186824, HS-grad, 9, Never-married, Machine-op-inspct, Unmarried, White, Male, 0, 0, 40, United-States, <=50K
38, Private, 28887, 11th, 7, Married-civ-spouse, Sales, Husband, White, Male, 0, 0, 50, United-States, <=50K
43, Self-emp-not-inc, 292175, Masters, 14, Divorced, Exec-managerial, Unmarried, White, Female, 0, 0, 45, United-States, >=50K
40, Private, 193524, Doctorate, 16, Married-civ-spouse, Prof-specialty, Husband, White, Male, 0, 0, 60, United-States, >=50K
54, Private, 302146, HS-grad, 9, Separated, Other-service, Unmarried, Black, Female, 0, 0, 20, United-States, <=50K
35, Federal-gov, 76845, 9th, 5, Married-civ-spouse, Farming-fishing, Husband, Black, Male, 0, 0, 40, United-States, <=50K
43, Private, 117037, 11th, 7, Married-civ-spouse, Transport-moving, Husband, White, Male, 0, 2042, 40, United-States, <=50K
59, Private, 109015, HS-grad, 9, Divorced, Tech-support, Unmarried, White, Female, 0, 0, 40, United-States, <=50K
56, Local-gov, 216851, Bachelors, 13, Married-civ-spouse, Tech-support, Husband, White, Male, 0, 0, 40, United-States, >=50K
19, Private, 168294, HS-grad, 9, Never-married, Craft-repair, Own-child, White, Male, 0, 0, 40, United-States, <=50K
54, ?, 180211, Some-college, 10, Married-civ-spouse, ?, Husband, Asian-Pac-Islander, Male, 0, 0, 60, South, >=50K
```

Figure 2. 1 First several rows of the data

Secondly, I imported the data to Weka, shown in Figure 2.2. Then I chose J48 to classify the dataset with Percentage Split of 66%. The reason why I chose 66% split is that the size of training set is two times larger than that of testing dataset. So, we could use 1x3 cross validation to split the combined file. Figure 2.3 is the result I got with 85% of correctly

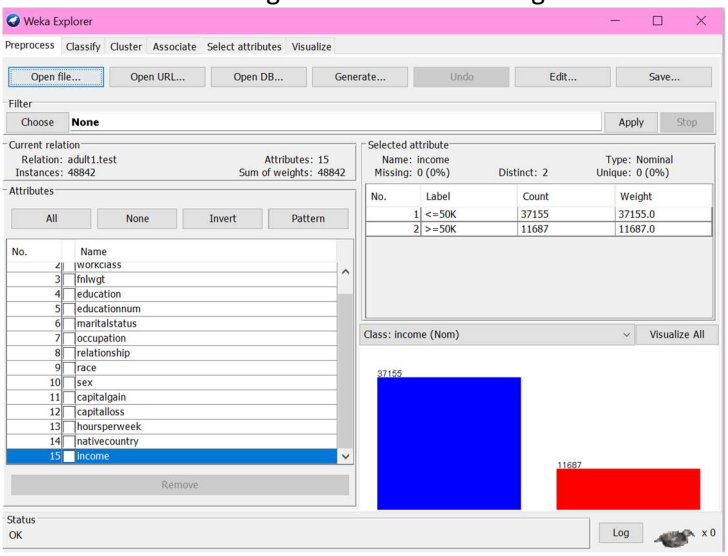


Figure 2.2

classified Instances.

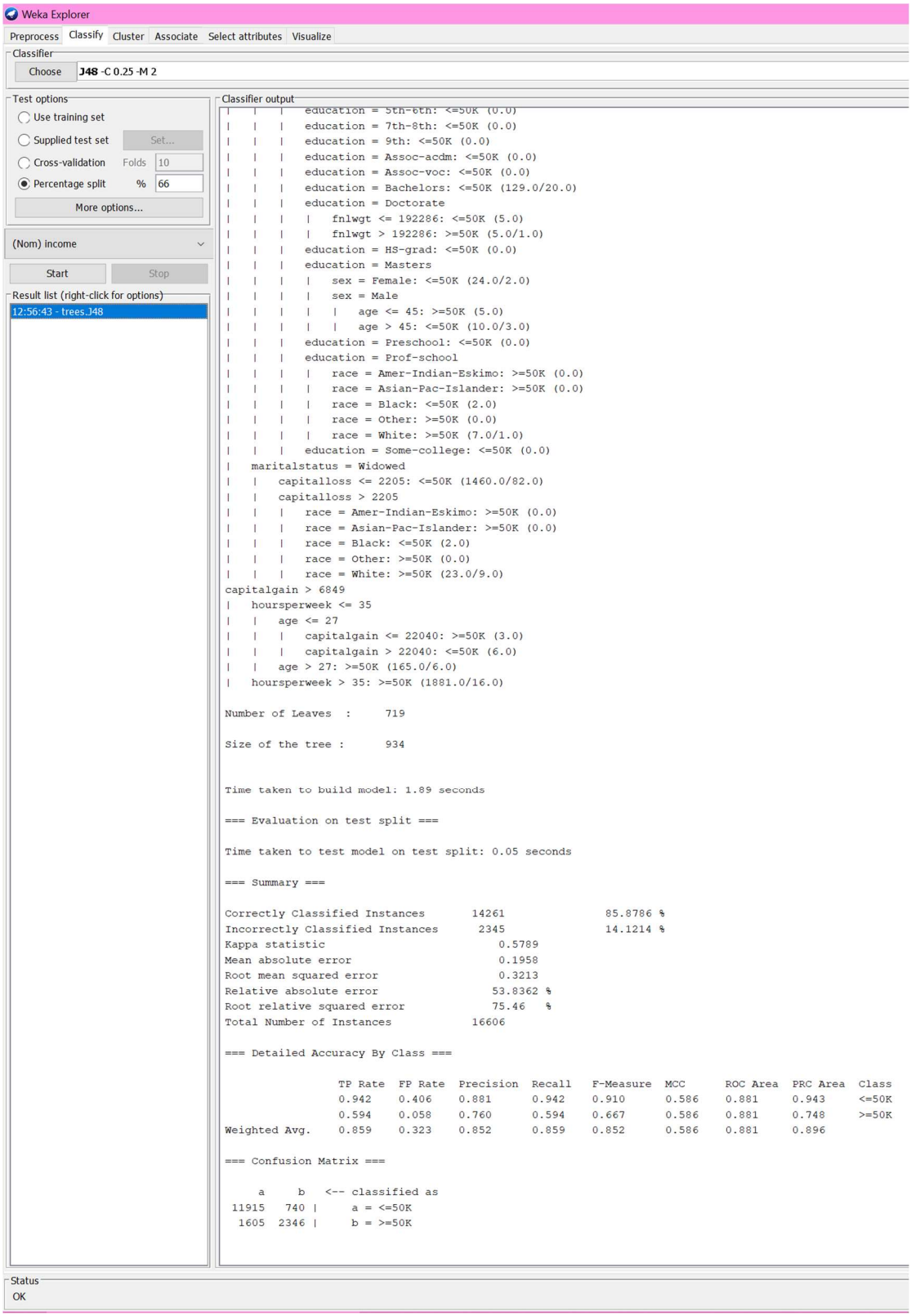


Figure 2.3