Experiment 5: Decision Tree Induction

Aim: Generate a Decision Tree by using J48 algorithm.

DESCRIPTION:

Decision tree learning is one of the most widely used and practical methods for inductive inference over supervised data. It represents a procedure for classifying categorical database on their attributes. This representation of acquired knowledge in tree form is intuitive and easy to assimilate by humans.

ILLUSTRATION:

Build a decision tree for the following data

AGE	INC OM E	STUD ENT	CREDIT_RATIN G	BUYS_COMPUTER
Youth	High	No	Fair	No
Youth	High	No	Excellent	No
Middle aged	High	No	Fair	Yes
Senior	Medium	No	Fair	Yes
Senior	Low	Yes	Fair	Yes
Senior	Low	Yes	Excellent	No
Middle aged	Medium	Yes	Excellent	Yes
Youth	Low	No	Fair	No
Youth	Medium	Yes	Fair	Yes
Senior	Medium	Yes	Fair	Yes
Youth	Medium	Yes	Excellent	Yes
Middle aged	Medium	No	Excellent	Yes
Middle aged	High	Yes	Fair	Yes
Senior	Medium	No	Excellent	No

The entropy is a measure of the uncertainty associated with a random variable. As uncertainty increases, so does entropy, values range from [0-1] to present the entropy of information

Entropy (D) = $\sum_{j=1}^{c} -p \log_2 p$
Information gain is used as an attribute selection measure; pick the attribute having the highest information gain, the gain is calculated by:
Gain (D, A) = Entropy(D) - $\sum_{j=1}^{c}(|Dj|/|D|)Entropy(D)$
Where, D: A given data partition A: Attribute
V: Suppose we were partition the tuples in D on some attribute A having v distinct values D is split into v partition or subsets, (D1, D2….. Dj) , where Dj contains those tuples in D that have outcome Aj of A.

Class P: buys_computer=”yes”
Class N: buys_computer=”no”

Entropy (D) = -9/14log (9/14)-5/15log (5/14) =0.940
Compute the expected information requirement for each attribute start with the attribute age Gain (age, D)
= Entropy (D) - $\sum_{youth,middle-aged,senior}^{\cdots} \left(\frac{Sv}{S}\right) \text{Entropy}(Sv)$
= Entropy (D) – 5/14Entropy(Syouth)-4/14Entropy(Smiddle-aged)-5/14Entropy(Ssenior)
= 0.940-0.694
=0.246

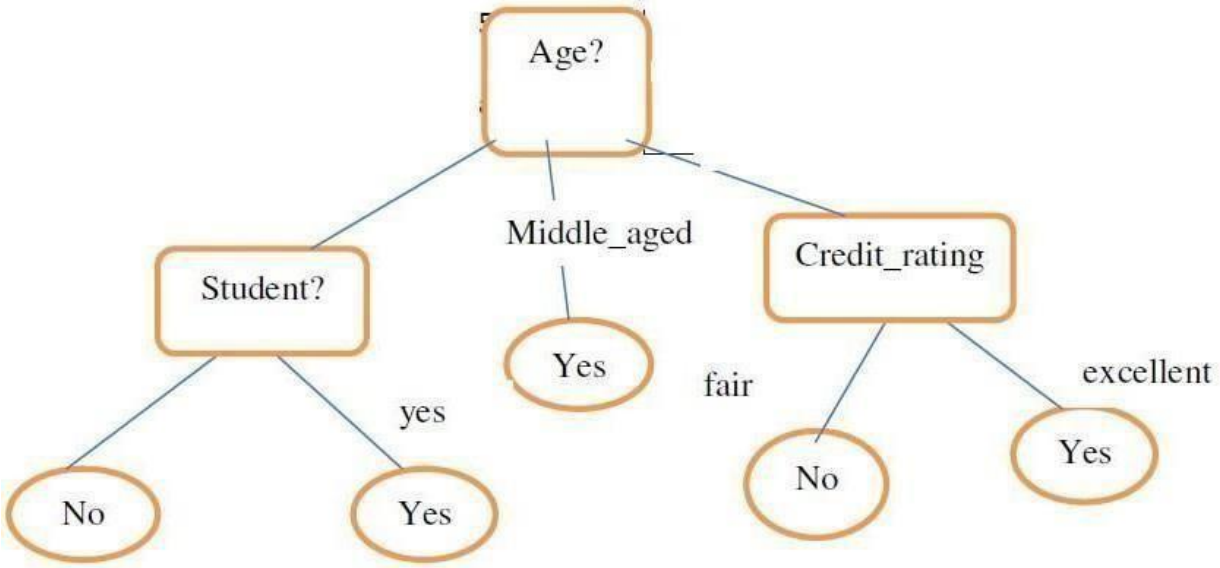
Similarly, for other attributes,
Gain (Income, D) =0.029
Gain (Student, D) = 0.151
Gain (credit_rating, D) = 0.048

Income	Student	Credit_rating	Class
High	No	Fair	No
High	No	Excellent	No
Medium	No	Fair	No
Low	Yes	Fair	Yes
medium	Yes	excellent	yes

Now, calculating information gain for subtable (age<=30)
I The attribute age has the highest information gain and therefore becomes the splitting
* attribute at the root node of the decision tree. Branches are grown for each outcome of age. These tuples are shown partitioned accordingly.

Income=”high” S11=0, S12=2
I=0
Income=”medium” S21=1 S22=1
I (S21, S23) = 1
Income=”low” S31=1 S32=0
I=0
Entropy for income
E(income) = (2/5)(0) + (2/5)(1) + (1/5)(0) = 0.4
Gain(income) = 0.971 - 0.4 = 0.571

Similarly, Gain(student)=0.971
Gain(credit)=0.0208
Gain(student) is highest ,



A decision tree for the concept buys_computer, indicating whether a customer at All Electronics is likely to purchase a computer. Each internal (non-leaf) node represents a test on an attribute. Each leaf node represents a class (either buys_computer=”yes” or buys_computer=”no”.

first create a csv file for the above problem,the csv file for the above problem will look like the rows and columns in the above figure. This file is written in excel sheet.

Clipboard Font

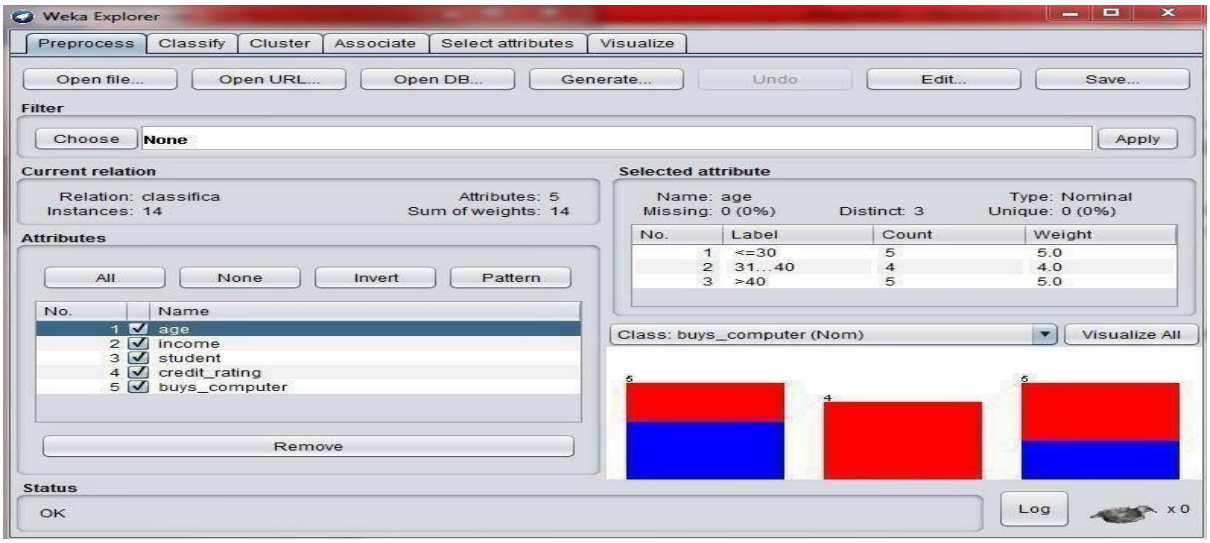
A1 f_x age

	A	B	C	D	E	F
1	age	income	student	credit_rat	buys_computer	
2	<=30	high	no	fair	no	
3	<=30	high	no	excellent	no	
4	31...40	high	no	fair	yes	
5	>40	medium	no	fair	yes	
6	>40	low	yes	fair	yes	
7	>40	low	yes	excellent	no	
8	31...40	low	yes	excellent	yes	
9	<=30	medium	no	fair	no	
10	<=30	low	yes	fair	yes	
11	>40	medium	yes	fair	yes	
12	<=30	medium	yes	excellent	yes	
13	31...40	medium	no	excellent	yes	
14	31...40	high	yes	fair	yes	
15	>40	medium	no	excellent	no	
16						

Procedure for running the rules in weka:

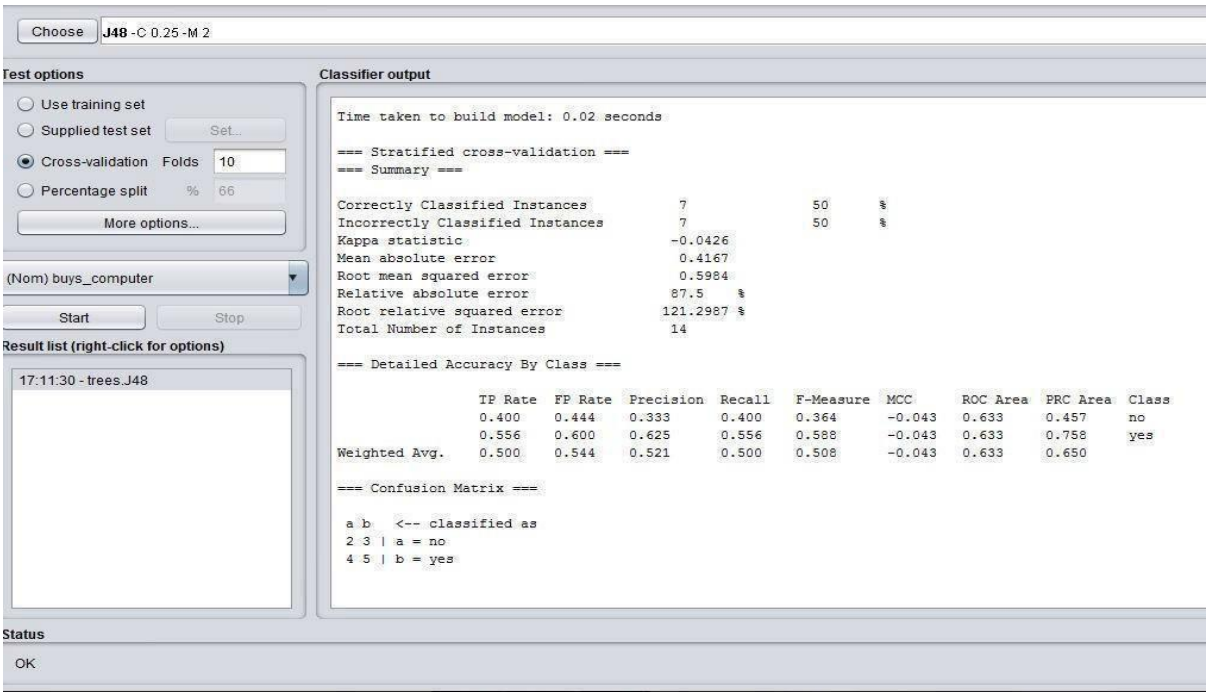
Step 1:

Open weka explorer and open the file and then select all the item sets. The figure gives a better understanding of how to do that.



Step2:

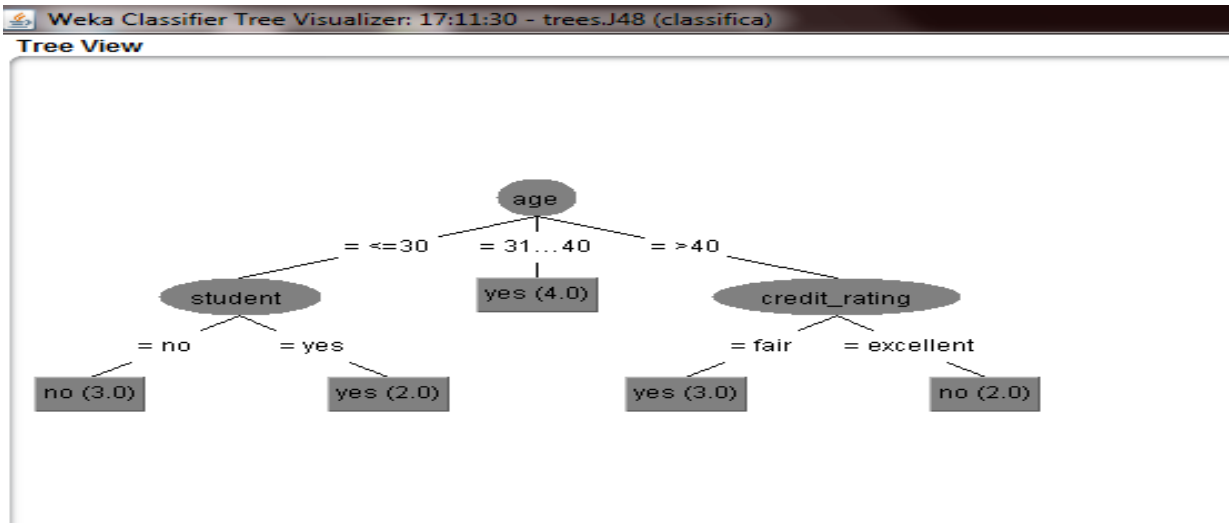
Now select the classify tab in the tool and click on start button and then we can see the result of the problem as below



Step3:

Check the main result which we got manually and the result in weka by right clicking on the result and visualizing the tree.

The visualized tree in weka is as shown below:



Conclusion:

The solution what we got manually and the weka both are same.

Exercise:

1.Apply decision tree algorithm to book a table in a hotel/ book a train ticket/ movie ticket.