Decision Trees

You are given the following training dataset that has to do with the travel means people use to go to their jobs. For each person the following attributes are recorded: gender, whether s/he is car owner, the cost s/he is willing to pay in euros/km, and the income. We would like to examine whether these attributes can be used to predict the transportation mode (class attribute) a person uses. The values of the class attribute in the given dataset have been given by people who filled out a questionnaire.

Training dataset (file **transportation_train.arff**)

gender	**car**	**travel_cost**	**income**	**transportation_mode**
m	n	cheap	low	bus
m	y	cheap	medium	bus
f	y	cheap	medium	train
m	n	cheap	low	bus
m	y	cheap	medium	bus
m	n	cheap	medium	train
f	y	standard	medium	train
f	y	expensive	high	car
m	y	expensive	medium	car
f	y	expensive	high	car

(A) Execute Hunt's algorithm by hand using Gini index as the measure of impurity:

1. Which attribute will be chosen as the root of the decision tree? Explain by computing Gini index for all possible scenarios. For the categorical attributes travel_cost and income try multi-way splits only (i.e., split into three children).
2. If we terminate the algorithm after the creation of the root of the tree, how many instances are wrongly classified?

(B) Test your solution with WEKA:

1. Which is the attribute chosen as the root of the tree by algorithm J48?
2. Give the full tree build by WEKA using the default parameter values of J48.
3. Give the full tree when minNumObj=1.

(Γ) You are given the following test dataset (file **transportation_test.arff**):

gender	**car**	**travel_cost**	**income**	**transportation_mode**
f	y	cheap	high	train
m	n	standard	medium	bus
f	n	expensive	medium	car
f	n	expensive	high	car

Report on the accuracy achieved by the single node tree you created by hand and the WEKA trees of cases B2 and B3.