HW1 (Classification)

(A) Decision Trees (on paper)
You are given the following training dataset that has to do with mushrooms and their characteristics (Colour, Height, Stripes, Texture). We wish to examine whether it is possible to use these characteristics in order to classify a mushroom as poisonous or non-poisonous. For the mushrooms in the given dataset the characteristic Poisonous (class variable) has been set by a botanist.

Colour	Height	Stripes	Texture	Poisonous
Purple	Tall	Yes	Rough	Yes
Purple	Tall	Yes	Smooth	Yes
Red	Short	Yes	Hairy	No
Blue	Short	No	Smooth	No
Blue	Short	Yes	Hairy	Yes
Red	Tall	No	Hairy	No
Blue	Tall	Yes	Smooth	Yes
Blue	Short	Yes	Smooth	Yes
Blue	Tall	No	Hairy	No
Blue	Short	Yes	Rough	Yes
Red	Short	No	Smooth	No
Purple	Short	No	Hairy	Yes
Red	Tall	Yes	Rough	No
Purple	Tall	Yes	Hairy	Yes
Purple	Tall	No	Rough	No
Purple	Tall	No	Smooth	No

(i) Which characteristic is used as the root of the decision tree according to Hunt's algorithm? Explain why you chose that characteristic (by computing the gini index for every scenario). **To avoid excessive computations, for all the categorical variables examine only multi-way splits (i.e., for Colour and Texture split into 3 children)**. How many instances of the training dataset are wrongly classified if we use a tree with one node only (the root)?

(ii) Verify your result in WEKA using algorithm J48 and choosing "Use training set" in "Test options" (we provide the file mushrooms.csv). Which characteristic does algorithm J48 choose as root of the decision tree? Give the full tree WEKA creates with the default values of the algorithm parameters.

(iii) Given the following test dataset:

Colour	Height	Stripes	Texture	Poisonous
Purple	Tall	Yes	Rough	Yes
Red	Tall	Yes	Smooth	No
Red	Short	No	Hairy	Yes
Blue	Short	No	Smooth	No

What is the accuracy achieved by the tree you created in (a) and by the tree created by WEKA in (b)? (the answer requires checking each given instance of the test dataset on the respective tree).

(B) Case study using WEKA

File **car.arff** contains the evaluation of 1728 cars and their ranking in four categories(column class): unacc, acc, good and vgood. It also contains the following car characteristics:

- buying (in thousands of euros)
- maint (in thousands of euros)
- doors
- persons
- lug_boot (in liters)
- safety (1-low, 2-average, 3-high)

Use 10-fold cross validation and experiment with different values for the minNumObj parameter of algorithm J48 (decision tree).

Choose the best model and report its accuracy and confusion matrix. Comment on the car categories that are not satisfactorily predicted. Using the decision tree, report one rule you consider strong for each car category. A rule is a path from the root of the tree to a leaf and describes the characteristics of a large number of instances belonging to the category that is the majority in the leaf.