SOME GUIDELINES FOR THE WRITTEN DOCUMENT

The written document to be presented should present the following sections. For each section there is a short description of what should appear there.

1- Description of the original data:

Where was it obtained (web page link), description of the problem on hands: target of classification, original number of examples, original number of columns meaning and kind of values, impact of missing values, etc.

2- Description of pre-processing of data:

Kind of preprocessing done to the original data: Simplification of data set? Removing of examples? Removing of columns? Replacing/removing missing values? Simplification of values? Normalization? Remember that, when applicable, you should describe the procedures performed to do that.

3- Splitting procedure of data set into train and validation data set.

Description of the procedure followed in order to obtain a representative validation data set.

4- Execution of different machine learning methods:

For each method you should say which kind of test method to compute accuracy you use. Cross-validation? K-fold cross-validation? How many folders? Why that number? Which is the best accuracy and parameters you achieved with cross-validation? In addition, for each method you should describe and discus some particular issues:

a) *Naïve Bayes:* Think about hypothesis of independence of variables. Do you have enough number of elements to obtain reliable probabilities? Keep that information for the discussion section.

b) *K-NN:* Description of procedure followed for choosing the best k-parameter. Show a graph with varying k. Have you adjusted other parameters as distance measure? Have you considered removal of irrelevant features if accuracy is poor compared with other approaches? (remember that k-nn is sensible to irrelevant features when computing distance to closest examples)

c) *Decision Trees:* Discussion of choice of pruning threshold. Discuss interpretability of the obtained DT. Show some of the most relevant rules. Discuss how + and – examples are mixed in leaves in order to estimate the reliability of the tree.

d) *SVMs:* Discussion of procedure followed to select the different parameters for the SVM. Show number of support vectors.

e) *Meta-learning algorithms:* Performance of Adaboost and Bagging using at least Decision stumps and decision trees.

5- Comparison and conclusions.

Comparison and discussion of results of the different machine learning methods on the *validation data-set*. Which is the best method when testing on the validation data set? Write a comparative table. Is there an explanation for that (some hypothesis applicable, etc.)? Are in general accuracy on validation data set similar to the obtained with cross-validation? There are for some methods huge differences? If that's the case, why do you think that happens? Final personal evaluation of which is the best method you consider and why.