PRACTICAL APPLICATIONS
MACHINE LEARNING

The works must be structured as a scientific paper with sections like: Introduction, Problem description, Methodology, Results, Discussion, Conclusion, References. **Interpretation of the models** is fundamental, taking into account the variables involved. Note that the quality of writing is important (clarity of presentation, conciseness, lack of typos, etc.).

For Practical Applications 1-3, choose a dataset (public or from the student himself, the latter being a plus). The dataset should be sized to at least 15 variables and 100 observations. Any software may be used for Practical Applications 1-3; although we recommend Weka.

Practical application 1 (non-probabilistic supervised classification):
The non-probabilistic classification algorithms to be used are all seen in class. There will be four analyses: (1) with all original variables; (2) with a univariate filter feature subset selection; (3) with a multivariate filter feature subset selection, (4) with a wrapper feature subset selection. All merit figures should be estimated with an honest method.

Practical application 2 (probabilistic supervised classification):
The probabilistic classification algorithms to be used are all seen in class. The four previous analyses will be also applied here. Metaclassifiers will also be used. All merit figures should be estimated with an honest method.

Practical application 3 (unsupervised classification):
All the unsupervised classification algorithms seen in class will be used. The dataset cannot contain a class variable obviously.

Practical application 4 (probabilistic graphical models):
Choose a dataset with at least 7 variables and 100 observations. Using *BayesFusion* (https://www.bayesfusion.com/) or R (http://www.bnlearn.com/bnrepository/) is recommended. Learn a Bayesian network and show and discuss the relationships found and conditional independencies. Perform inferences (exact and approximate) with the model.

Submissions will include a document (maximum 10 pages, 1 column, 11 pt) and a set of slides (pptx or pdf) to be used in the oral presentation (only some will be selected). The documents will be submitted via Moodle. Otherwise it will not be considered and the student will fail this part of the subject. Please see the Moodle site for the deadlines.

- **Some data repositories** in Internet:
  - Bayesian networks: http://www.cs.huji.ac.il/~galel/Repository/
  - KDNuggets competition site: www.kdnuggets.com/datasets/
  - UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/
  - Kaggle competition data: http://www.kaggle.com/
  - WEKA website: http://www.cs.waikato.ac.nz/ml/weka/datasets.html