#1 Session 2 A data challenge

General guidelines:

- Most of the exercises rely on the Knime software. Version 4.3 has been tested for this session.
- Some specific third party packages may be used such as the DeepLearning4J extension to be able to run richer neural network designs and experiments compared to the default tools. **Take care of potential libraries conflicts, for instance, adding Anaconda to your $PATH can make Knime crash suddenly when using the DL4J plugin.**
- **Keep trace of all your experiments**. To this end, use sources versioning or accurate file organization. Never loose or overwrite a previous exercise nor trial by the next one.

The data challenge:

A client is looking for a data scientist team. He then proposes a competition on a data classification problem recruiting interesting specialists. Challengers are given a data set with few knowledge. The only provided elements are:

- The data are vectors of indicators describing human individuals. Each indicator is a scalar or string that refers to a category. Each data sample relates to a single person.
- Each sample is described by numerical values such as "age", "number of years studying", and so on as well as categorical variables as "marital status", "sex", "occupation".
- The target indicator is "bankable", which can be considered as a boolean.
- Only a subset of the annotated data is provided to the challengers, the remaining is a test set that will be provided by the client at the end of the session to evaluate your proposal.
- Challengers may use K Nearest Neighbors (KNN) and/or tree-based and/or neural network based approaches to design the best performing classifier. However, the scientific approach will have to be explained on the last day in a short oral presentation.

Sylvie Galichet and Alexandre Benoit,
LISTIC Lab, Univ. Savoie Mont Blanc, France.

Guidelines, a potential challenge approach:

1. Download the data provided by your potential client:
   bankable_ consumer_ train.csv check the *CSV reader* node to load them.
2. Observe the data to get an idea of its size and statistics (magnitudes, distribution and any other indicator that you think is relevant).
3. Define a set of metrics for performance evaluation.
4. Define a model optimization strategy (init conditions, seeds, (cross)validation process, etc.).
5. Define classifiers based on KNN, trees and/or neural network type. The design should follow a clear strategy. For instance, iterate from simple designs to more complex ones but take the time to find the best hyper parameters tuning of a given model before moving to a more complex one.
6. **take care of data normalization and transformation ; provide appropriate data to your models !** Some approaches, such as neural networks, are sensitive to the input data nature and scaling:
   - Regarding data amplitude normalization, have a look at the *Normalizer* and *Normalizer(apply)* nodes.
   - Regarding conversion from categorical variable, they can be transformed into integer variables using the *Category to Number* node OR as one hot vector using the node *One to Many*.
7. Prepare an experiment tracing journal/log not to loose / overwrite strategies and experiments this will let you improve faster !
8. Optimize models and identify the best configuration with respect to your performance indicators. Identify an approach that will allow you to improve your models. Keep trace of everything, do not loose any strategy/experiment log.
9. Prepare for the challenge results submission :
   - Predict with your best model on the *bankable_ consumer_ test.csv* file that will be provided some hours before the deadline (same format but no label column).
   - Write results in a csv file using the *CSV writer node* following:
     • predicted category column **MUST BE NAMED *prediction***,
     • CSV file **MUST BE NAMED *datachallenge_ name1_ name2.csv*** with *name1* and *name2* the last name of each team member.
   - Prepare a few slides (5 max.) to present your contribution from data analysis to your approach and the obtained performance level.
   - **Submission consists of 3 items : exported workflow, result proposals and slides. Penalization will be applied on incomplete and/or ill-formed items.**

Sylvie Galichet and Alexandre Benoit,
LISTIC Lab, Univ. Savoie Mont Blanc, France.