Working Guide

Take a dataset:

Data must contain at least 500-1000 observations with some continuous and categorical variables (5 *may* be enough).

For example:

- From the *UC Irvine Machine Learning Repository*

  http://archive.ics.uci.edu/ml/

- Or from *Kaggle*

  https://www.kaggle.com/datasets

- Or from R libraries.

- Or from other sources that you **want**…

Make a *basic descriptive study*:

(i) **Frequency tables** for at least two of the continuous variables.

See e.g., as **tutorials** for basic frequency tables:

http://www.statcan.gc.ca/edu/power-pouvoir/ch8/5214814-eng.htm

https://cran.r-project.org/web/packages/agricolae/vignettes/tutorial.pdf

https://cran.r-project.org/web/packages/fdth/fdth.pdf

(ii) Calculate measures of centrality, variability, and shape (skewness and kurtosis). **Interpret** results.

(iii) Take one of the categorical variables and create **groups** based on it.
For *example*: suppose that you have a variable named *gender* and a variable *salary*. You may compare and study the differences in salaries between women and men.

(iv) For the continuous variables: make histograms, density plots, normal probability plots (QQ), box plots and other ones as you may consider. Discuss the normality of data based on graphs.

(v) Then, repeat the previous plots for each group studied in (iii) and compare the results among them. For *example*: Are there differences between women and men?

(vi) Take a categorical variable and show the frequency table. Take two categorical variables and show the descriptive contingency table. Make mosaic plots and explain the results.

In all cases (iv-vi) it is **advisable** to use **advanced** options based on `ggplot` and/or `lattice` libraries.

Part 2

Using the variables of your dataset, apply the library `caret` or/and `H2O` for analyzing possible relationships between a categorical (preferably dichotomic) variable and other variables of your dataset.

Split the data set into a *training* set and a *testing* set.

Use three or four techniques included in `caret` and/or `H2O`

Recommended: take (if possible) any other techniques different from the ones shown in class.

Make predictions of the data set labelled as the *testing* set.

Show and interpret the *confusion matrix*.

Repeat the previous task by using an *ensemble* of the previous classifiers.

Compare and **interpret** the obtained results.

Note: Do not worry if the results are *not* very satisfactory. It is just an exercise...