SY19
TP7: Tree-based and ensemble methods

1 `Expenditure and Default` dataset

The dataset `Expenditure and Default` contains data about the default behavior and monthly expenditure behavior of a large sample (13,444 observations) of credit card users. The variables are :

- Cardhldr = Dummy variable, 1 if application for credit card accepted, 0 if not
- Default = 1 if defaulted 0 if not (observed when Cardhldr = 1, 10,499 observations),
- Age = Age in years plus twelfths of a year,
- Adepcnt = 1 + number of dependents,
- Acadmos = months living at current address,
- Majordrg = Number of major derogatory reports,
- Minordrg = Number of minor derogatory reports,
- Ownrent = 1 if owns their home, 0 if rent
- Income = Monthly income (divided by 10,000),
- Selfempl = 1 if self employed, 0 if not,
- Inc_per = Income divided by number of dependents,
- Exp_Inc = Ratio of monthly credit card expenditure to yearly income,
- Spending = Average monthly credit card expenditure (for Cardhldr = 1),
- Logspend = Log of spending.

1. Split the data into a training set of 10,000 and a test set.
2. Build a classification tree to predict variable `Cardhldr` from the other variables (excluding `Default`, `Exp_Inc`, `Spending` and `Logspend`). Represent this tree graphically. Compute the corresponding confusion matrix and error rate.
3. Optimally prune the tree and represent the obtained pruned tree. Compute the corresponding confusion matrix and error rate. Did pruning improve the performance ? The interpretability ?
4. Plot of ROC curve of the pruned tree classifier.

5. Compute the corresponding confusion matrix and error rate of bagged decision decision tree classifiers and random forests. Plot the variable importance measures from the random forest classifier.
6. Compare the results to those obtained using logistic regression and a GAM.

2 `Boston` data

1. Build a regression tree to predict variable `medv` in the `Boston` dataset.
2. Does bagging improve the performance ? Build and evaluate a random forest regressor for this problem.
3. Compare the performance to those of different variants of linear regression (least squares, ridge, lasso, variable selection). Compare the variables in the regression tree to those selected using the lasso or stepwise selection methods.
4. Compare the performances to those of a GAM.