PQHS 471 (2019) Final

The dataset is about bankruptcy prediction of Polish companies. The data contains financial rates from a year of forecasting period. The class label indicates bankruptcy status after a year. The original data are from the UCI Machine Learning repository. The description of all the 64 financial variables are at https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data

For the purpose of this exam, we only use the 5th year data. I removed 10 variables that have the most missing data, and then 153 companies that had missing data for some of the remaining variables. The final data for this exam has two sets: The training set `trainset` contains 3838 companies (with 262 bankrupt) and the test set `testset` contains 1919 companies (142 bankrupt).

What to do:

1. Supervised learning:
   - Build prediction models using random forests, boosting, and SVM. I would like to see that you have made an effort to improve your models' performance. Turn in a summary of what you have tried, your final models, and the accuracy rate on the test sets.
2. Unsupervised learning: Suppose we only have the financial variables and do not yet have the information on the bankruptcy status.
   - Would the companies be clustered so that when the bankruptcy status becomes available later on, most would-bankrupt companies are clustered together?
   - Would the MDS technique be useful for separating the would-bankrupt companies from the other companies?

What to turn in:

1. Your code, with comments, as a file;
2. Your results in a report (in words, graphs or tables)

The final is due by the midnight of Saturday, May 4, 2019.