README File for Journal Supplemental Materials

ARTICLE INFORMATION

Journal name: Jahrbücher für Nationalökonomie und Statistik (JBNST)

Paper: Early prediction of university dropouts - a random forest approach

Authors: Andreas Behr, Marco Giese, Herve D. Teguim K., Katja Theune

DEPOSIT INFORMATION

Total Number of Files: 3

Description: Information provided in the three following files is needed to make use of the data. The paper addresses the implementation of a random forest model for predicting student dropout.
The model is based on conditional inference trees and conditional inference forests and is implemented in the statistical program R (R version 3.4.1). Packages to load are dplyr, ROCR, party, and caret (see R Codes). Estimates are calculated using data of the fifth cohort of the German National Educational Panel Study (NEPS) (see Data Access). These data contain 17,910 students and almost 3,000 variables covering a wide range of different aspects of student background and the course of study (see Data Description).
Setting requirements are: 100 trees, $[\sqrt{p}]$ variables randomly sampled as candidates from the total number of features p. Performances using 10-fold cross-validation, in which each procedure is repeated 20 times, achieve an AUC of 0.86. Important predictors derived from the conditional variable importance are the final grade at secondary school and also determinants associated with student satisfaction and their subjective academic self-concept and self-assessment.
In the file "Data Description" we describe which variables are used in this study, including the original variable name, the new variable name in our dataset, the wave (we generally used the first available wave, since we are interested in dropout prediction as early as possible), variable label and the range of values of the variables (this is additionally explained in more detail in the codebook which is available on the NEPS homepage). Since we are not allowed to hand out the data and the one to one R-code of our data pre-processing, we provide this detailed data description, in which everything needed to re-construct the data is described.

Filenames:

- **Data Access:** provides information necessary for data access.
- **Data Description:** provides a detailed description of the data with information about the variables used.

- **R Codes:** provides detailed R codes required to obtain the results. Packages, setting parameters and variables needed for the implementation of the model are explained.

Reference to the accepted paper (APA):
Behr, A., Giese, M., Teguim K., Herve D., Theune, K., "Early Prediction of University Dropout - A Random Forest Approach", Journal of Economics and Statistics, forthcoming.