Assignment on Decision Trees & Ensembles

Machine Learning

Prof. Dr. Marc Pouly & Dr. Tim vor der Brück
Lucerne University of Applied Sciences and Arts

Deadline: 15 May, 2018.

Exercise 1: Theory Questions (1 Point)

Fill in the online theory quiz on ILIAS.

Acceptance criterion: This is an individual exercise. You can participate in the quiz as many times as you want. ILIAS registers your best result over all runs. You passed this exercise if your best score reaches at least 80%.

Exercise 2: Decision Trees & Random Forests (2 Points)

This is a team exercise; hand in your solution as a Jupyter notebook by e-mail to Marc Pouly no later than the above deadline. We award 1 point for an implementation with a decision tree and 1 point for comparison with random forests.

A competition on Kaggle is concerned about the alcohol consumption of secondary school students. Based on various information the task is simply to predict whether a student is a frequent drinker or not. In a publication available here, the authors used decision trees in Weka and KNIME for their analysis and reported an accuracy of 92%. However, there is no indication in the paper that random forests have been evaluated as well. The following data description table is borrowed from the paper cited above:

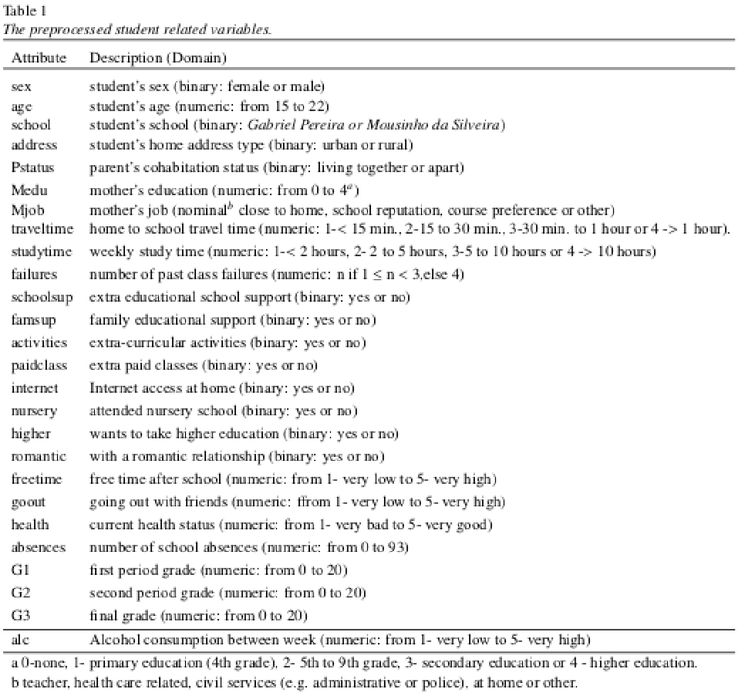

Table 1
The preprocessed student related variables.

Attribute	Description (Domain)
sex	student's sex (binary: female or male)
age	student's age (numeric: from 15 to 22)
school	student's school (binary: *Gabriel Pereira or Mousinho da Silveira*)
address	student's home address type (binary: urban or rural)
Pstatus	parent's cohabitation status (binary: living together or apart)
Medu	mother's education (numeric: from 0 to 4[a])
Mjob	mother's job (nominal[b] close to home, school reputation, course preference or other)
traveltime	home to school travel time (numeric: 1-< 15 min., 2-15 to 30 min., 3-30 min. to 1 hour or 4 -> 1 hour).
studytime	weekly study time (numeric: 1-< 2 hours, 2- 2 to 5 hours, 3-5 to 10 hours or 4 -> 10 hours)
failures	number of past class failures (numeric: n if $1 \leq n < 3$,else 4)
schoolsup	extra educational school support (binary: yes or no)
famsup	family educational support (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
paidclass	extra paid classes (binary: yes or no)
internet	Internet access at home (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
freetime	free time after school (numeric: from 1- very low to 5- very high)
goout	going out with friends (numeric: ffrom 1- very low to 5- very high)
health	current health status (numeric: from 1- very bad to 5- very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)
alc	Alcohol consumption between week (numeric: from 1- very low to 5- very high)

a 0-none, 1- primary education (4th grade), 2- 5th to 9th grade, 3- secondary education or 4 - higher education.
b teacher, health care related, civil services (e.g. administrative or police), at home or other.

Implement your own decision tree predictor in Python and investigate the potential to improve accuracy using random forests.