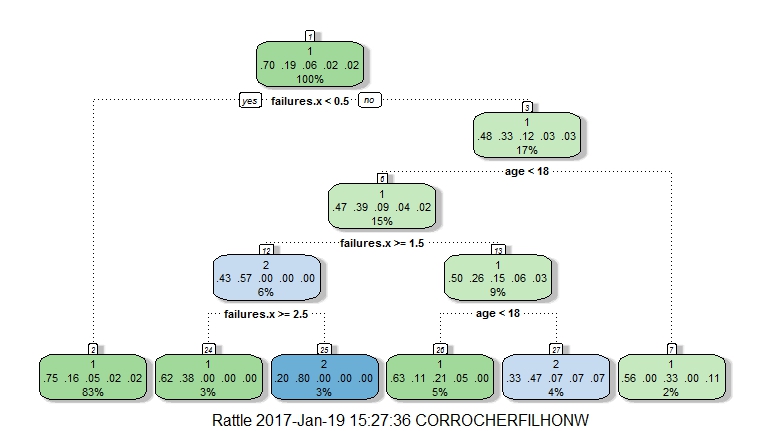
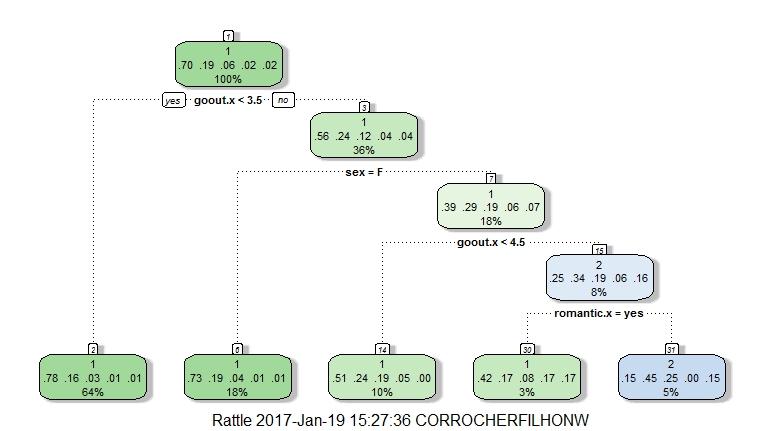


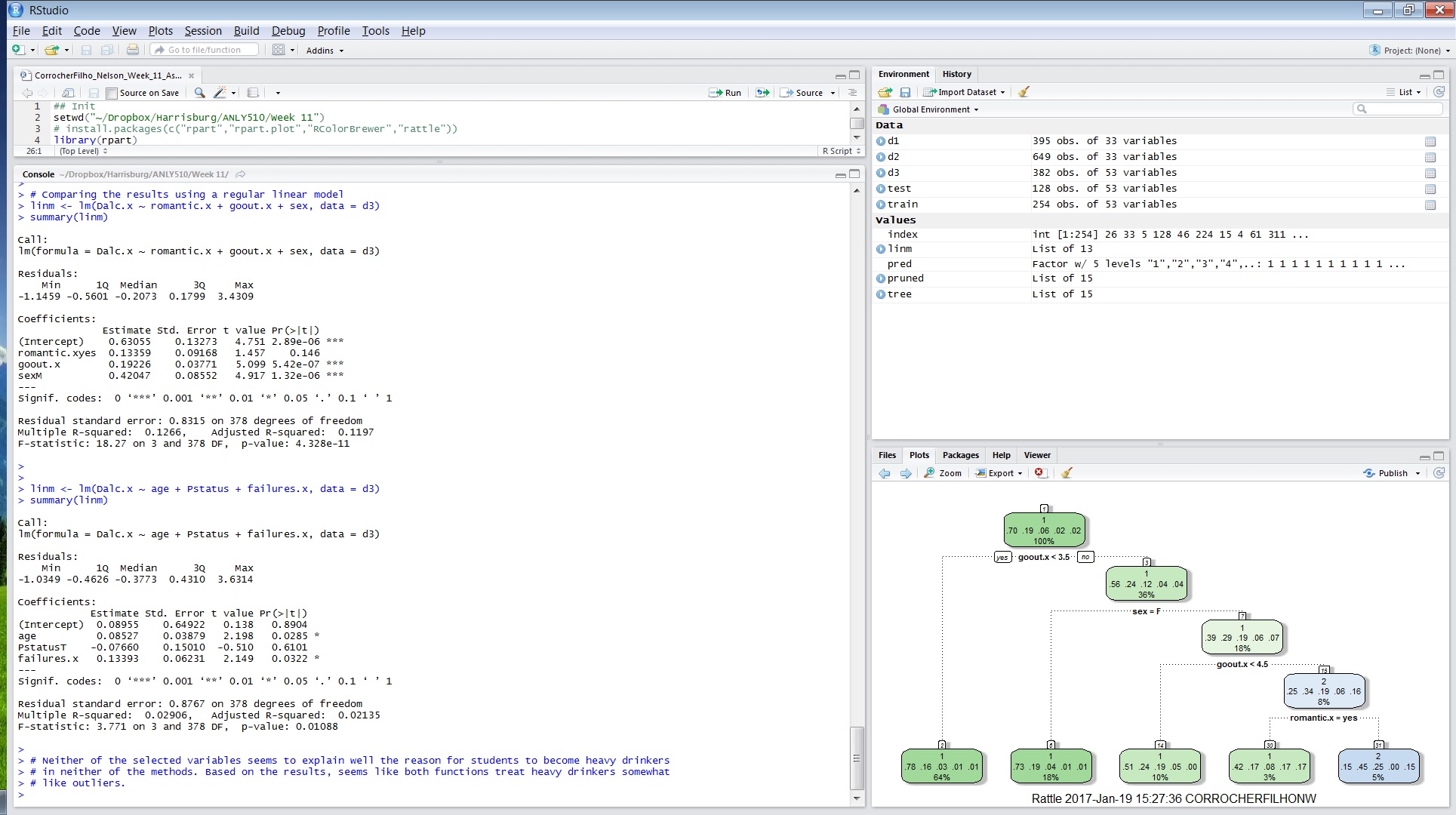
This pictures is just a basic initialization of R (best to check the R file attached).



In this decision tree, I tried to classify students based on the amount of alcohol drank daily using three variables: age, parental status and school failures. Considering the alcohol drank daily is an ordinal variable that goes from 1 to 5, from 1 being a no drinker to 5 being a heavy drinker. As we can see here, the method couldn’t classify heavy drinkers (>3), even though they are around 8% on the data set. However, failures and age seems to have some classification power on no drinker (1) vs light drinkers (2).



As before, I’ve tried to classify this daily alcoholic consumption based on another three variables: gender, going out with friends and having a romantic partner. The tree suggests that going out, being male and having a romantic partner together increase the chance of students being light drinkers. One important note: The fourth leave from left to right concentrates most of the heavy drinkers but is classified as a no drinker. There is probably another variables that could further break the classification between light and heavy drinker.



Finally, I’ve tried multivariate regression model to capture the explanation power of the same variables I’ve used above. Again, it seems that these variables don’t have a strong explanation power, at least to detect heavy drinkers.

I normally select the method that seems to be more fit for the analysis I’m trying to do, and select a second method to verify if it is aligned to what I found out using the first tool. For example, I could use a decision tree to find if a person has a significant risk of having cancer based on some variables. Then, to verify it, I could use a cluster with two centroids to check how far one group is from the other. If the first tool would give me, let’s say, 10/90% sample separation, I would expect the cluster groups to be apart from each other (gets difficult to visualize when many variables are involved, but it’s okay for only a few variables).

Regarding the tool of choice, a consider three mainly variables to select a tool: Correctness, Interface and how professional the graphs look.

* It seems weird to say correctness as, in theory, any mature software should return the right values. However, I had problem with SAS in the past because it carried a bug for years in which it returns a wrong KS test.
* Interface is good because some statistical function, even though they can be adjusted by tenths of parameters, only two or three are relevant for regular use and good interfaces don’t force you to fill all of them, automatically providing reasonable numbers for these other parameters. This can speed things up a lot, especially when you are learning a new methodology.
* Professionally looking graphs is self-explanatory: sometimes you just want to plug a graph directly into PowerPoint instead of having to work on its cosmetics for you to show it to executives. This also speedy things up drastically.