# Machine Learning for Data Analysis

**Assignment – Week 1**

**Running a Classification Tree**

**By** [mapolarbear@gmail.com](mailto:mapolarbear@gmail.com)

This week's assignment is about to run first Classification Tree and interpret results.

This week’s assignment involves decision trees, and more specifically, classification trees. Decision trees are predictive models that allow for a data driven exploration of nonlinear relationships and interactions among many explanatory variables in predicting a response or target variable. When the response variable is categorical (two levels), the model is a called a classification tree. Explanatory variables can be either quantitative, categorical or both. Decision trees create segmentations or subgroups in the data, by applying a series of simple rules or criteria over and over again which choose variable constellations that best predict the response (i.e. target) variable.

**About My research**

For research purposes of Machine Learning course we are advised to use ADDHEALTH modified dataset and I’m interested in how substances like alcohol, marijuana, smoking etc. Affecting adolescent life and what we can predict and up to what level. I will be using ADDHEALTH for the first time, for my previous courses I was using GAPMINDER, so, this is bit new for me.

**Sample**

ADDHEALTH - The sample to be used represent adolescent various data collected in 2004.

**Procedure**

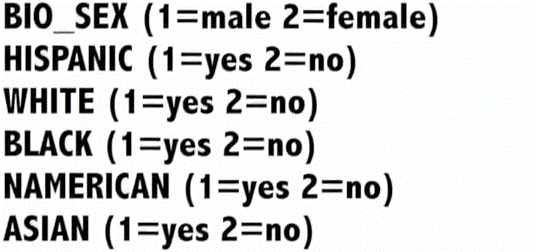
Data were during 2004.

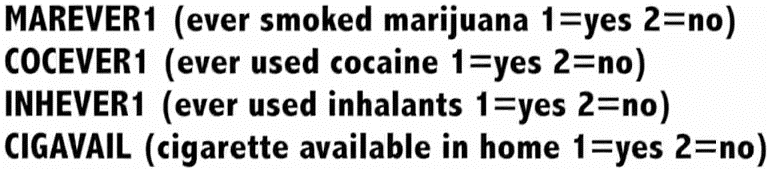
**Measures (current study)**

**Target response variable:**

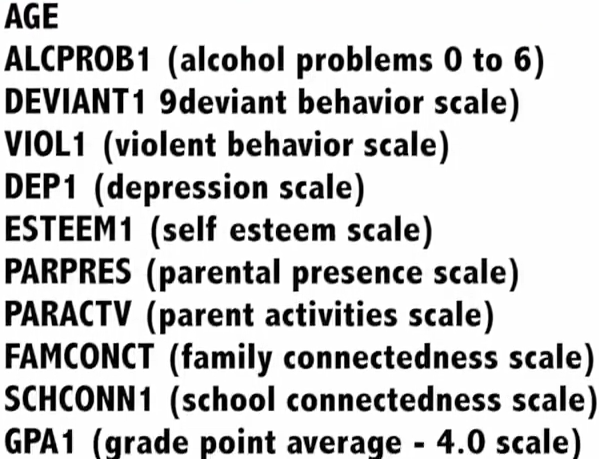
****

**Explanatory variables - Categorical:**

****

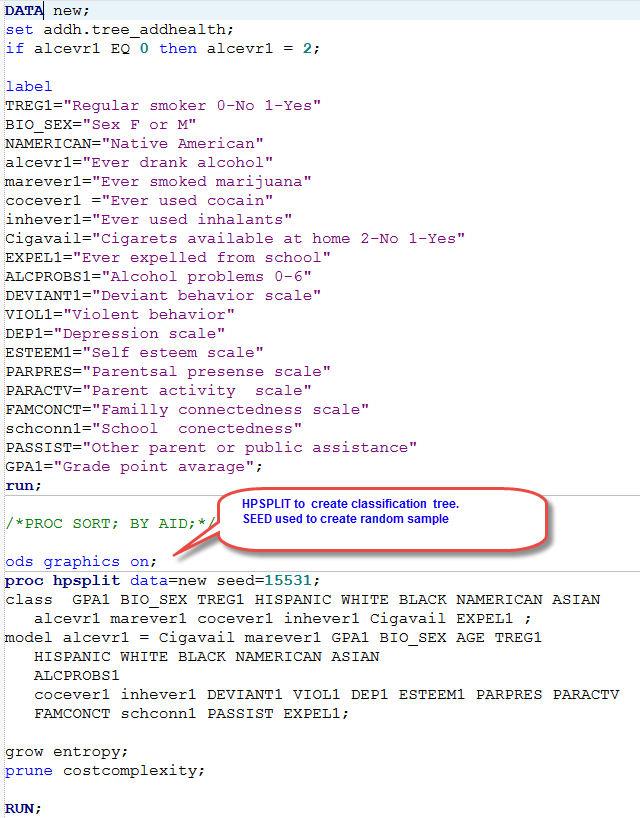
****

**Explanatory variables - Quantitative:**



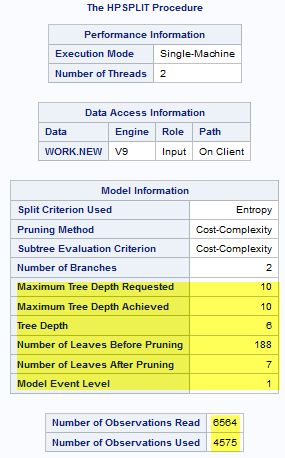
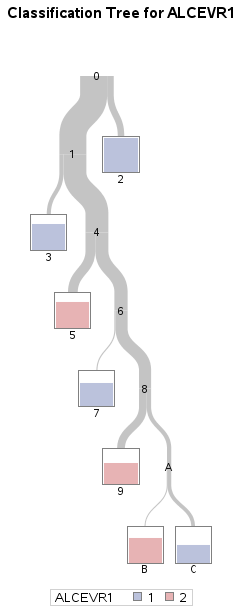
**Program code**

The response variable ALCEVR1 value of 2 was assigned as ‘NEVER DRUNC ALCOHOL’ replacing value of 0 (zero).

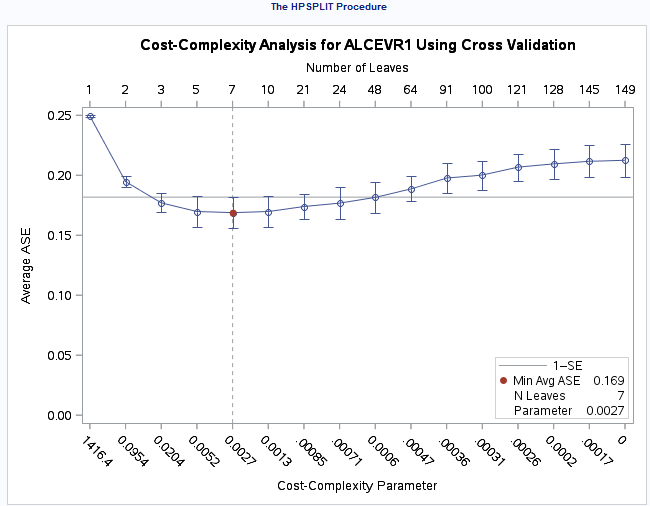


**Interpretation of results:**

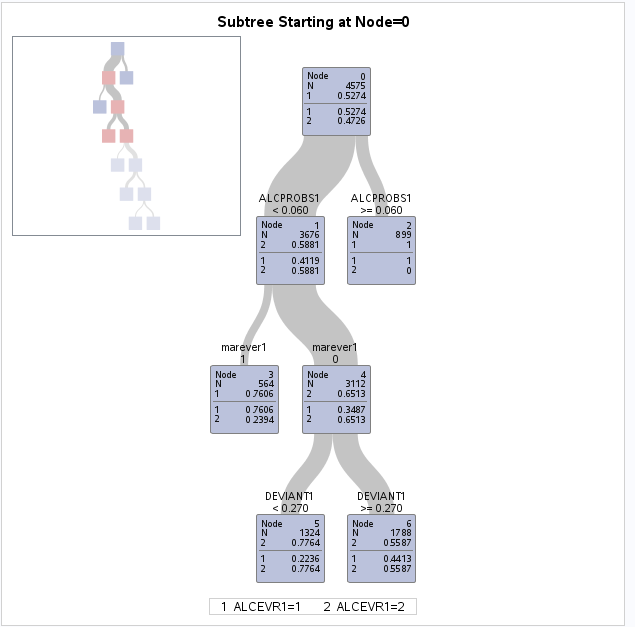
The output shows that program rad 6564 total number of observations from which 4575 observation with complete data satisfied our selection criteria. Observations with missing values in at least one variable were dropped and did not participate in our analysis. The ‘Model Information’ table also shows that program created 2 (two) branches, for values 1 and 2 of ALCEVR1. The maximum level/depth of tree ached 10 with 188 leaves pages out of which only 6 nodes/levels with 7 leaves pages were chosen for model event level 1 (value of ACLEVR1 response variable).

The Cost-Complexity chart shoes that ASE – 1 method was applied to evaluate standard error (ASA) and have values ASA=0.17, and Levels=7, which corresponds to ‘Number of Leaves after Pruning’.



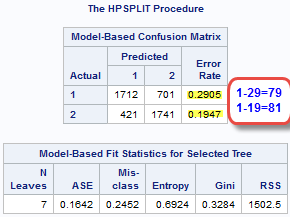
**Analysis of created Classification Tree:**



We can interpret diagram and say that adolescents, who never tried alcohol ALCEVR1=2 and alcohol problems rate < 0.060 (ALCPROBS1 2=0.5881) will not smoke marijuana (MAREVER1=0, 2=0.65 or 65%) and will less likely to have deviant behavior above 0.270 rate (DEVIANT1 >=0.270), but in contrary, will have likely deviant behavior < 0.270 and will be less deviant.

Attached Variable Importance table also shows relativeness and importance of variables, based on which we may conclude that none of variables where extremely overshadowed, but keep in mind to adjust our future models by adding TREG1 (Regular or not smoker), because the importance of this variable is pretty high: 46% (4.6013).

In addition, Model-Based Confusion Matrix table also shows that values of ALCEVR1 variable are correctly classified, having 79% for ever tried alcohol (1) and 81% for never tried alcohol respectively.



<http://mapolarbear-da.blogspot.com/2017/01/coursera-ml-assignment-week-1.html>

<http://mapolarbear-da.blogspot.com/2017/01/ml-assignment-week1.html>

End;