# 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**

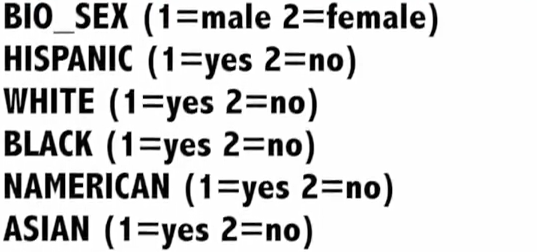
Data were during 2004.

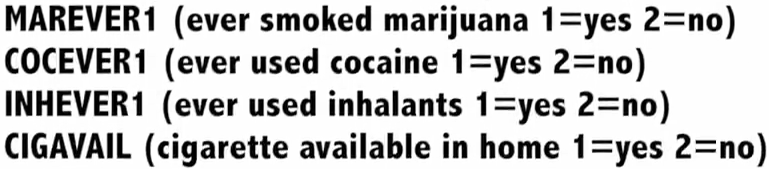
**Measures (current study)**

**Target response variable:**

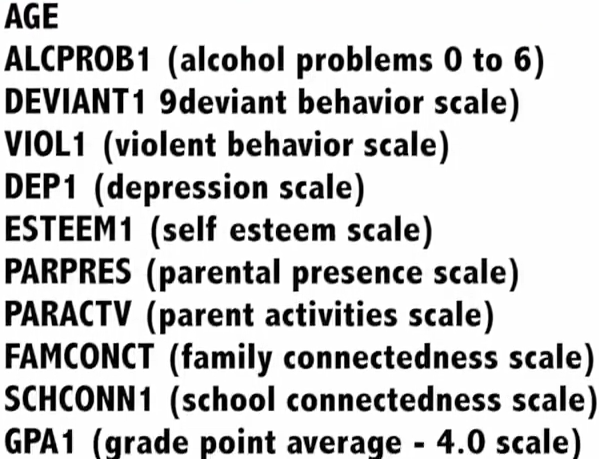


**Explanatory variables - Categorical:**



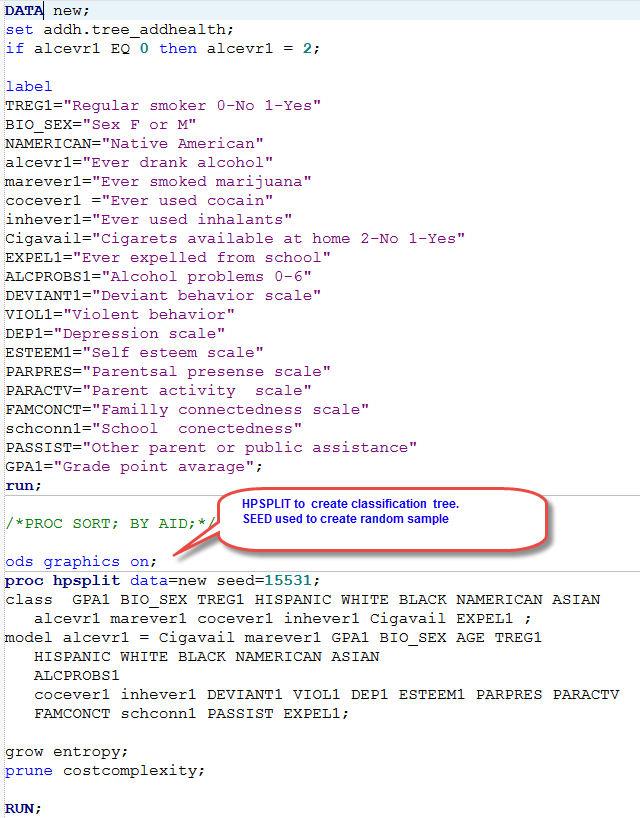


**Explanatory variables - Quantitative:**



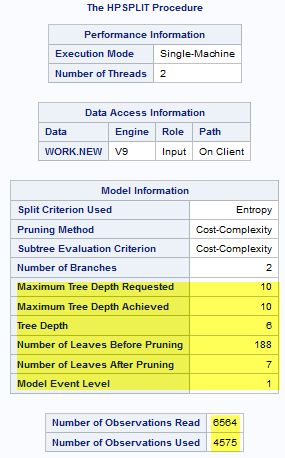
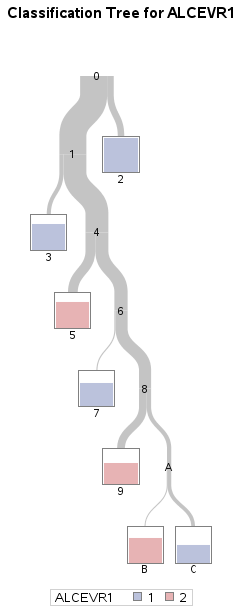
**Program code**

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

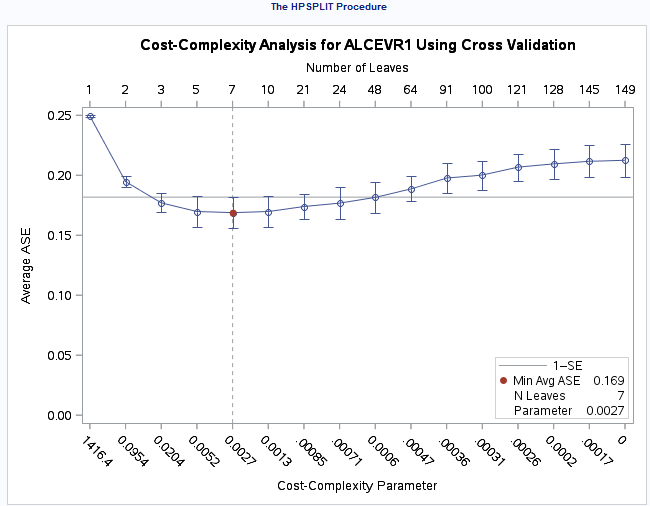


**Interpretation of results:**

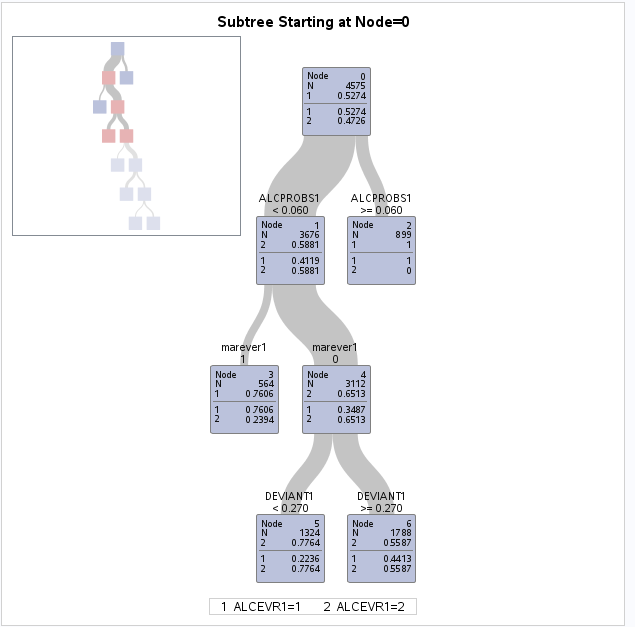
The output shows that program rad 6564 total number of observations from which 4575 observation with complete data satisfied our selection criteria. Observations with missing values in at least one variable were dropped and did not participate in our analysis. The ‘Model Information’ table also shows 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’.



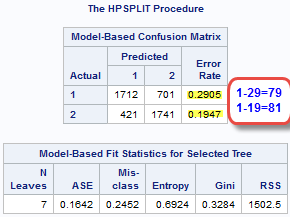
**Analysis of created Classification Tree:**



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).

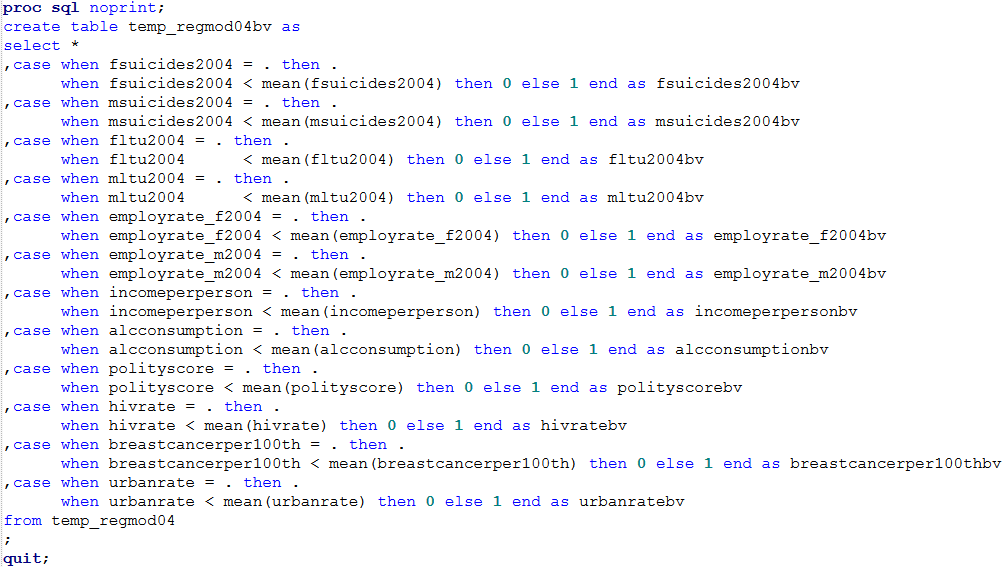
In addition, Model-Based Confusion Matrix table also shows that values of ALCEVR1 variable are correctly classified, having 79% for ever tried alcohol (1) and 81% for never tried alcohol respectively.



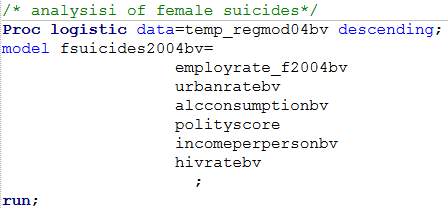
End;

The logistics regression requires to binary variables only, meaning both response and explanatory variables must have values of 1 (as positive answer) or 0 (as negative answer). The GAPMINDER contains only quantities variables and because of that we have to bin each variable into 2 categories (1 and 0), which is very artificial, but we still need to run this study to complete assignment.

Program code to create binary variables out of quantitative:



**Code to run logistics regression** with ALL listed binary variables against female suicides ratings:



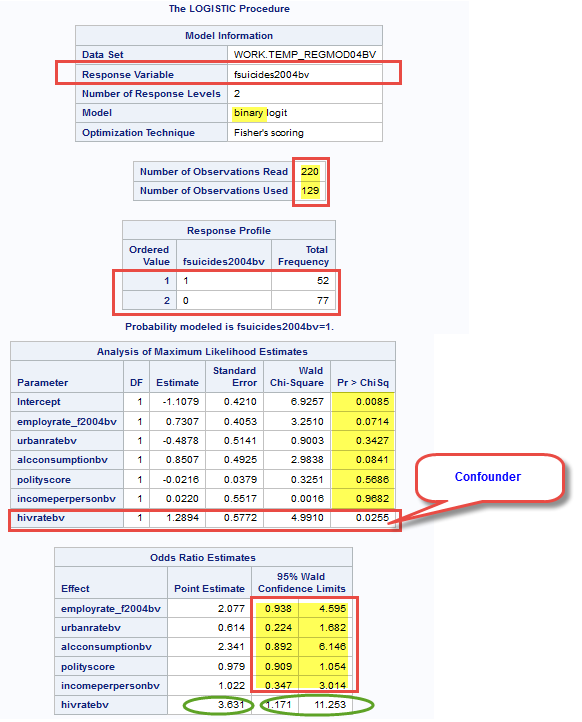
**Logistics Regression Model outputs and interpretations**

* **Female suicides ratings model – with ALL binary explanatory variables (finder confounder)**

As we can see total of **220 observations with 129 observations of complete data** were used in our model. It looks like 1/3rd of observations containing one or more observations with missing data were dropped, decreasing **variation of model sample size**. Parameter estimates and P-value table shows that there is only one variable: **HIVRATEBV with significant estimate value of 1.29 (beta= 1.29) and signioficant p-value of 0.0255** in the model, **all other variables have non sognigicant p-value > 0.05**. Based on results of logistics regression model we can conclude that **HIVRATEBV variable is the major confounder**.

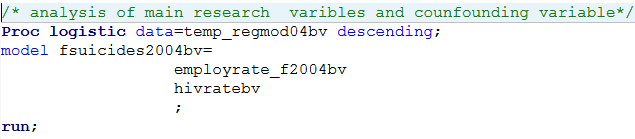
**Odds Ratio estimates**

In general terms, odds ratio is probability of event occuring in one group comparing to probability of event occuring in another group. Based on Odds Ratio Estimates table we can conclude that **females with (1) HIVRATEBV estimate of 3.63 are 3.6 times more likely to commit suicide than without (0).** The **95% Confidence limits of Odds Ration Estimates** also shows that in **95%, or samples will be in between values 1.17 and 11.25**. Another interpretation and **conclusion of odds ration limits** can be: **females with HIV in countries with HIV ratings from 1.17 to 11.25 are more likely commit suicide that without HIV.**



* **Female suicides ratings model – with MAIN binary explanatory variables and confounder**

**Program code:**



As we can see total of **220 observations with 139 observations of complete data** were used in our model. It looks like 1/3rd of observations containing one or more observations with missing data were dropped, decreasing **variation of model sample size**. Parameter estimates and P-value table shows that both variables **employrate\_f2004bv** and **hivratebv** are independently significant:

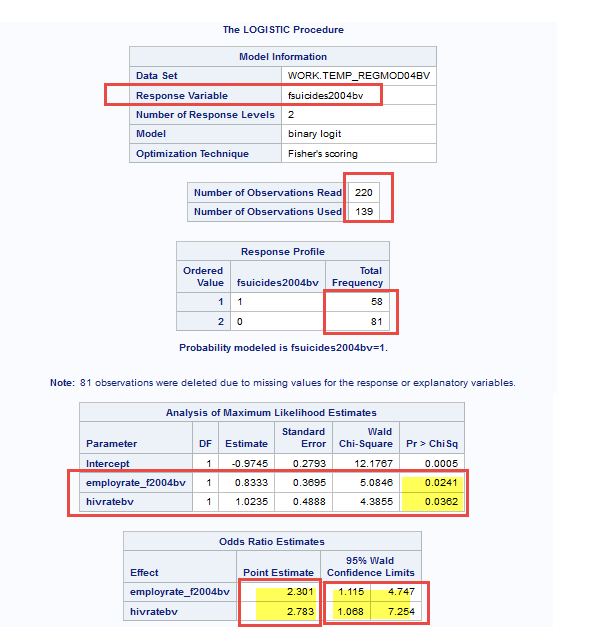
**employrate\_f2004bv with significant estimate value of 0.83 (beta= o.83) and significant p-value of 0.0241**  **hivratebv with significant estimate value of 1.02 (beta= 1.02) and significant p-value of 0.0362**

We can ignore NULL hypothesis and accept alternate hypothesis that

**Odds Ratio estimates**

Based on Odds Ratio Estimates table we can conclude that **females in countries with Employment rate of 2.3 are 2.3 times more likely commit suicides.** We can also create prediction that at the same time, females in countries **with (1) HIV rate estimate of 2.3 are 2.3 times more likely to commit suicide than without (0).** The **95% Confidence limits of Odds Ration Estimates** also shows that in **95%, or samples will be in between values 1.16 to 4.74 and 1.1 to 7.25**. But because of overlapping sample confidence limits, we cannot say that HIV rate is more related to female suicide rates that employment rate. Another interpretation and conclusion of odds ration limits can be: **Controlling female employment rate and HIV rate can affect female suicide rate in countries with female employment rates in between 1.1 to 4.5 and with HIV in countries with HIV ratings from 1.1 to 7.25.**

**Conclusion: Current results support previous finding and we can conclude that there is correlation between female suicides rate and employment rate. We can also conclude that the HIV rate also affects female suicide rating in corresponding country.**



\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* end of week 4 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*