# Machine Learning for Data Analysis

**Assignment – Week 2**

## **Running a Random Forest**

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

This week's assignment deals with **Random Forests** and interpretation of results.

Random forests are predictive models that allow for a data driven exploration of many explanatory variables in predicting a response or target variable. Random forests provide importance scores for each explanatory variable and also allow you to evaluate any increases in correct classification with the growing of smaller and larger number of trees.

Like decision trees, random forests are a type of data mining algorithm that can select from among a large number of variables, those that are most important in determining the target or response variable to be explained. Also, like decision trees, the target variable in a random forest can be categorical or quantitative. And the group of explanatory variables can be categorical or quantitative, or any combination. Unlike decision trees, however, the results of random forest generalize well to new data since the strongest signals are able to emerge through the growing of many trees. Further, small changes in the data do not impact the results of random forests.

**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.

The **Random Forest** predictive model is used to test and confirm the importance of explanatory variables discovered/chosen by Classification Tree and confirm if models (both, **Random Forest** and **Classification Tree**) are correctly classified.

Predicting of ALCEVER1 (if adolescent ever drank alcohol) and effect and importance of other variables like

**Sample**

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

**Procedure**

Data were during 2004.

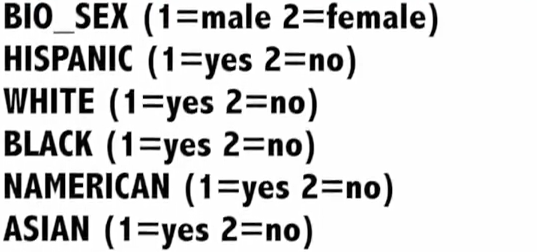
**Measures (current study)**

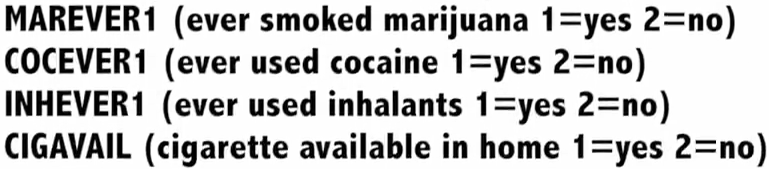
Following **target**, **explanatory binary categorical** and e**xplanatory quantitative** variables used in **Random Forest Tree**:

**Target response variable:**

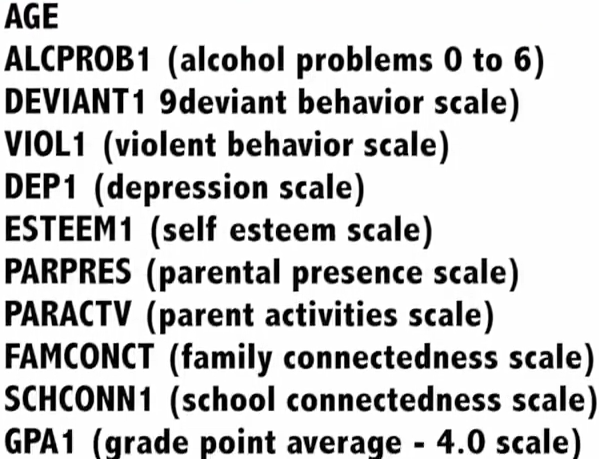
 - Describes if adolescent ever drank alcohol

**Explanatory variables - Categorical:**



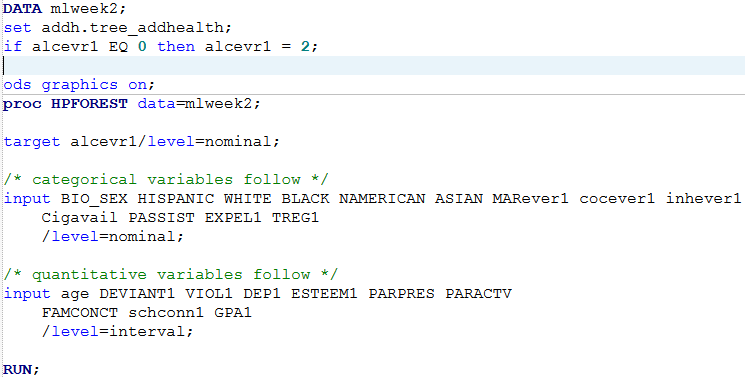


**Explanatory variables - Quantitative:**



**Program code**

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



**Interpretation of results:**

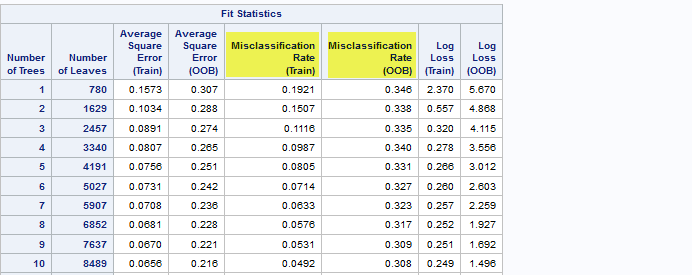
The output shows that **HPFOREST used 6444 observation out of 6504** to satisfy our selection criteria. Based on **GINI** settings, controlling in-bag target variable observation selection, observations with missing target variables values equal to **MISSING** value (.) will not be selected. However, observations with missing explanatory variables value (.) will participate in selection criteria.

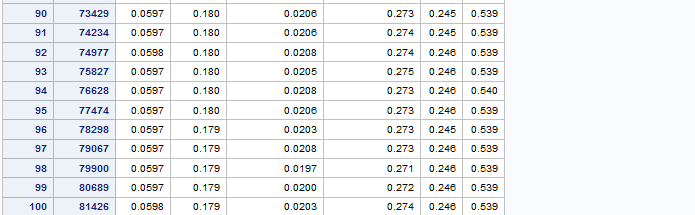
The **Inbag Fraction** has value of 0.6 (default), suggesting that each **inbag** set will represent **60% of data for training set**.

**Variables to try** shows that 5 variables will be selected for each tree and result of **HPFOREST** will be displayed for 100 trees.



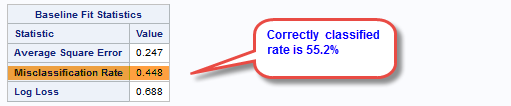
Check of **Misclassification Rates (Train)** and **(OOB)** of first 10 trees (1-10) show that **training set (Train)** and **Out of bag (OOB)** sets are classified correctly with about 80% to 90% for **Training** set and with 65% to 70% for **Out of Bag** sets. Misclassification rates for last 10 trees (90-100) also show high classification rates with not significant fluctuations.

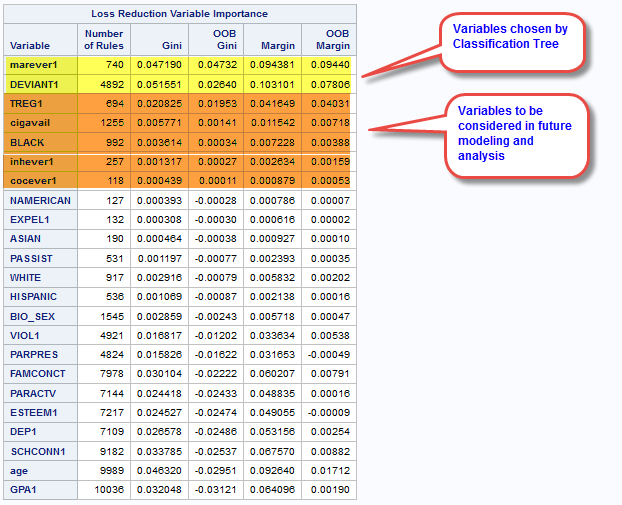




The last but not least output of shows ranking of importance of variables. Based on list of importance we can conclude that variables chosen by **Classification Tree** were also chosen by **Random forest model** but we can **partially rely on Classification Tree results.** Suggesting that the **forest correctly classified 55.2% of the sample** and suggesting that interpretation of a single decision tree may be appropriate.

We can also conclude that variables MAREVER1, DEVIANT1, TREG1 are major confounders. In addition we may also consider to include into regression and other modeling analysis following variables with high importance ranking produced by Random Forest model: TREG1 (regular, non-regular smoker), CIGAVAIL (availability of cigarettes at home), BLACK (race) for specific research, INHEVER1 (ever used inhalants), COCEVER1 (ever used cocaine).





http://mapolarbear-da.blogspot.com/2017/01/ml-assignment-week2.html

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*End\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**Note to remember:**

To summarize, like decision trees, random forests are a type of data mining algorithm that can select from among a large number of variables, those that are most important in determining the target or response variable to be explained. Also, like decision trees, the target variable in a random forest can be categorical or quantitative. And the group of explanatory variables can be categorical or quantitative, or any combination. Unlike decision trees, however, the results of random forest generalize well to new data since the strongest signals are able to emerge through the growing of many trees. Further, small changes in the data do not impact the results of random forests. In my opinion, the main weakness of random forests is simply that results are somewhat less satisfying, since no trees are actually interpreted.