Substance Abuse and Comorbidity Decision Trees

Phillip Thompkins

University of Maryland, Baltimore County

CMSC478: Machine Learning

Fall 2014

Introduction/Motivation

As an undergraduate majoring in Computer Science and Psychology, I have found that it can be difficult to combine both disciplines. I sought to do so with this project, by attempting to apply concepts from machine learning to some topic in clinical psychology. While searching through data listings, I came across a website for the Substance Abuse and Mental Health Data Archive (SAMHDA) and its National Survey on Drug Use and Health (NSDUH) 2013 dataset (1). My coursework in psychology supported the notion that a variety of mental illnesses were comorbid with, or also present alongside, substance abuse disorders (2, 6, 7). Additionally, my research into the topic to determine the viability of this project showed me that there is also a possibility of mental disorders being induced by substance abuse (3). Substance-related disorders are divided by the Diagnostic and Statistical Manual of Mental Disorders - Fourth Edition (DSM-IV) into “substance abuse” and “substance dependence” (4).

Substance abuse is characterized in the DSM-IV (4) by substance use that leads to “clinically significant impairment or distress,” and one of the following within a 12-month period:

- Substance use leading to failure to fulfill role obligations at work, school, or at home (absences, neglect, poor performance).
- Recurrent substance use in physically hazardous situations (like driving).
- Recurrent substance-related legal problems (arrests).
- Continued substance use despite persistent social or interpersonal problems caused by or exacerbated by substance use (arguments, physical fights).

Substance dependence is characterized within the DSM-IV (4) as substance use that leads to “clinically significant impairment or distress,” and three or more of the following within a 12-month period:

- Tolerance: Need for more of substance for desired effect, or diminished effect with continued use of same amount of substance.
- Withdrawal: Characteristic withdrawal symptoms, or taking same or similar substance to avoid withdrawal symptoms.
- Taking the substance often in larger amounts over a longer period of time than intended.
- Having a persistent desire or unsuccessful efforts to cut down or control substance use.

Additionally, a distinction must be made between “heavy” and “binge” drinking, as either can be at play in instances of alcohol abuse or dependence. The National Institute on Alcohol Abuse and Alcoholism defines heavy drinking as “5 or more drinks on the same occasion on each of 5 or more days in the past 30 days” and binge drinking as “drinking 5 or more alcoholic drinks on the same occasion on at least one day in the past 30 days” (8). So, binge drinking occurs on the way to heavy drinking.

Problem Definition

The problem is defined as an attempt to use machine learning concepts, namely information gain and decision trees, to explore the comorbidity between substance abuse/dependence and other mental disorders. A main research question being asked is whether or not certain drugs lead to higher rates of substance use/dependence comorbidity.

Proposed Method

The NSDUH 2013 data was examined by forming a decision tree within WEKA. Results of variables relative to outcome variables were compiled using SPSS and entered into a .txt file. The attributes that the software can split the tree upon will be determined by calculating the information gain of the variables, and keeping the top 20 variables for our outcomes. A Ruby program will be composed to take in the dataset as a file and count all occurrences of relevant responses. Participants who did not respond to a given survey item will not be counted in the information gain for that item.

Intuition

This approach, creating a decision tree, is likely to work because the data is being examined with a particular focus on which substances lead to more comorbidities. Being able to divide the data based upon which variables carry the strongest split between results is pretty much what a decision tree is built for. Ruby is being used to calculate the information gain because it has utilities (hashes, easy file IO) that would make calculating the information gain quite straightforward.

Experiments

The base dataset held 55,160 responders, each of which held 3,141 variables; that is 173,257,560 individual pieces of data. While examining that much data would produce robust results, it is not necessary to go to such lengths. A wide variety of variables are not necessarily relevant to the question that is being asked or the focus on mental disorders. Variables were hand-selected based on relevance from the following sections of the dataset's codebook: Core Substance Abuse, Substance Dependence/Abuse, Mental Health, Adult Depression, Adolescent Depression, Consumption of Alcohol, and Core Demographics. Variables were only included if responses were discrete values, and recodings of existing variables were not included. The number of variables was condensed from 3,141 to 342.

Our outcome variables are SMIDA2_U (Serious Mental Illness and Substance Dependence or Abuse) and AMIDA2_U (Any Mental Illness and Substance Dependence or Abuse). A "serious mental illness" is defined by United States law as any of the following: schizophrenia, paranoid disorders, bipolar disorders, major depressive disorders, schizoaffective disorders, pervasive developmental disorders, obsessive-compulsive disorders, depression in childhood and adolescence, panic disorder, post-traumatic stress disorder, bulimia nervosa, and anorexia nervosa (5). "Non-serious mental illness" includes things like personality disorders, dysthymia, seasonal affective disorder, anxiety disorders, attention deficit/hyperactivity disorder, and social phobia (5).

There is a wide enough variety of mental disorders in either category that a separate tree will be made examining the "Serious Mental Illness" and "Any Mental Illness" variables mentioned above. Additionally, people under the age of 18 cannot be diagnosed with most mental disorders. So, all participants under the age of 18 are excluded from this analysis, bringing our number of relevant responders from to 37,424. This reduces our total dataset down to 13,285,520 individual items, 7.3% of the initial dataset.

Observations

The variables with the highest information gain for AMIDA2_U (*n*=20) and SMIDA2_U (*n*=21) are displayed below in Tables 1 and 2, respectively. The Python code (Appendix 1) returned the variables with the top 25 information gains, but results were pruned based on redundancies. For example, the variables *HEALTH* and *HEALTH2* both pertained to the exact same question, one was just recoded to condense two of the initial response conditions. Another example is the *CATAG* variables; *CATAGE* broke reported age into categories, but other variables, *CATAG3, CATAG5, and CATAG7*, broke the reported age into more categories. Specific descriptions and response options for all of these variables are outlined in Appendix 2.

Table 1. The twenty variables with highest Information Gain relative to AMIDA2_U.

AGE2	IREDUC2	SEXRACE	DSTEFF30	HEALTH
DSTRST30	DSTNRV30	EDUCCAT2	CATAG7	BINGEHVY
DSTNGD30	DSTHOP30	DSTCHR30	CIGALCMO	NEWRACE2
SEXAGE	PREGAGE2	IRMARIT	ADTMTHLP	ADRXHLP

Table 2. The twenty-one variables with the highest Information Gain relative to SMIDA2_U.

AGE2	HEALTH	DSTEFF30	DSTNRV30	DSTRST30
EDUCCAT2	CATAG7	BINGEHVY	DSTNGD30	DSTHOP30
DEPNDALC	DSTCHR30	CIGALCMO	NEWRACE2	SEXAGE
PREGAGE2	IRMARIT	ADTMTHLP	SEXRACE	ADRXHLP
IREDUC2				

These sets of variables were input into the WEKA software to create a pair of J48 decision trees with SMIDAU_2 and AMIDAU_2 as the roots, which are displayed in Appendices 3 and 4. Both trees seem to have been built very well, as the accuracies for both the AMI and SMI trees were 95% and 98%, respectively. This is not surprising, as the list of variables that were retrieved were selected based on what would be the most beneficial to include in a decision tree. However, the two trees were of different sizes; the AMI trees had 157 leaves, while the SMI tree had only 77. The SMI tree is both smaller and more accurate.

The contents of both decision trees, while similar due to the shared pool of variables, show an interesting comparison. First, let us examine the AMI tree, whose first two branching variables are consistently ADTMTHLP, which asks if mental health treatment or counseling have been helpful to responders, and then BINGEHVY, which asks about alcohol use in the past month. If responders were not receiving mental health treatment or were seeing extreme benefits from their mental health treatment, the tree instantly predicted the AMI variable as "no." In cases where other responses to ADTMTHLP were reported, the tree deepened, looking at BINGEHVY. The tree generally deepened in situations where responders reported they regularly engaged in heavy binge drinking or heavy non-binge drinking, while non-drinkers were predicted as "no" responses relative to AMI. From there, depending on the branch, the tree would generally examine SEXAGE (variable combining age and sex information), DSTNGD30 (feeling worthless), and DSTNRV30 (feeling nervous) in some order, with other variables occasionally thrown in like health or marital status. In general, responders seemed more like to score as a "yes" on the AMIDAU_2 variable (any mental illness and substance abuse/dependence) if they drink relatively often, if they feel nervous or worthless, and if they're ages 18-25. Interestingly, this entire set of variables overlaps with the concept of a stereotypical college student: alcohol is part of college culture, most college students stress over grades and careers resulting in feelings of worthlessness and nervousness, and college students are often between the ages of 18 and 25.

While not focused on sex and age, the serious mental illness (SMI) decision tree paints a somewhat similar picture. First, it branches on DEPNDALC (alcohol dependence), defaulting to a “no” response if responders were deemed to not have any kind of dependence. Then, for the responders with an alcohol dependence, the tree branches on ADTMTHLP, the usefulness of treatment. As with DEPNDALC, responders who were not receiving mental health treatment or counseling were generally given a “no” response to the SMI variable. Generally, if responders felt worthless (DSTNGD30), hopeless (DSTHOP30), nervous (DSTNRV30), restless (DSTRST30), or that they could not be cheered up (DSTCHR30) some, most, or all of the time in the last 30 days, they were much more likely to be correctly labelled as a “yes” relative to the SMI variable. Additional variables examined in specific branches included health and if responders smoked cigarettes and drank alcohol. Responders who reported being in better health were more often given a “yes” for SMI, as were those who both smoked cigarettes and drank alcohol.

Conclusions

Based upon these decision trees, it can be said that someone is more likely to suffer from a mental illness alongside substance abuse or dependence if they are seeking mental health treatment, generally feel nervous or worthless, and if they engage in binge or heavy drinking every once in a while. Of additional note is the increased likelihood of suffering from *any* mental illness if some of the aforementioned qualities are present *and* the person in question is of typical college age. Also, substance abuse/dependence seem to be highly related with alcohol, as it is the only substance-specific variable (DEPNDALC) to survive preprocessing, and other variables (BINGEHVY, CIGALCMO) also relate to alcohol use. To turn this information into something practical, it would seem that people are more likely to depend on alcohol if they're in a pretty negative place mentally and are already starting to drink. So, interventions for these people could come in the form of mental health counseling, whether to deal with the feelings of worthlessness or nervousness, or to develop alternative coping strategies so that alcohol isn't someone's first response. Considering that higher numbers of responders in the dataset were suffering from a substance disorder and something else when mental health treatment was only somewhat working, current treatment methodologies may need to be reexamined to find more helpful approaches.

However, these findings are not without pitfalls from the dataset itself. Many of the variables relevant to substance abuse were focused primarily on alcohol consumption, as opposed to the tens of substances mentioned in one section of the dataset. While most of the substance variables did not have high response rates outside of alcohol, the ones that did have higher response rates could still warrant further investigation. Perhaps one of the biggest concerns I have with the dataset is how the questions about mental health were overwhelmingly directed towards major depressive disorder, to the point where entire chapters of the dataset focused on depression. The dataset has variables for "serious mental illness" and "any mental illness" and yet the entire dataset only seems to care about depression. While depression is considered a “serious mental illness,” it raises questions about how the researchers who collected this data managed to classify people as suffering from mental illness alongside substance abuse if they only really looked at one mental illness.

References

1. Center for Behavioral Health Statistics and Quality. (2013). Results from the 2012 National Survey on Drug Use and Health: Summary of national findings. (HHS Publication No. SMA 13–4795, NSDUH Series H–46). Rockville, MD: Substance Abuse and Mental Health Services Administration.
2. Merikangas, K. R., Mehta, R. L., Molnar, B. E., Walters, E. E., Swendsen, J. D., Aguilar-Gaziola, S., ... & Kessler, R. C. (1998). Comorbidity of substance use disorders with mood and anxiety disorders: results of the International Consortium in Psychiatric Epidemiology. *Addictive behaviors*, *23*(6), 893-907.
3. Schuckit, M. A. (2006). Comorbidity between substance use disorders and psychiatric conditions. *Addiction*, *101*(s1), 76-88.
4. American Psychiatric Association. (1994). Diagnostic and statistical manual of mental disorders (4th ed.). Washington, DC.
5. Serious Vs. Non-Serious Mental Illness. (n.d.). Retrieved from http://www.bcbsil.com/provider/standards/serious_vs_non_serious.html
6. Grant, B. F., Stinson, F. S., Dawson, D. A., Chou, S. P., Dufour, M. C., Compton, W., … & Kaplan, K. (2004). Prevalence and Co-occurrence of Substance Use Disorders and Independent Mood and Anxiety Disorders: Results From the National Epidemiologic Survey on Alcohol and Related Conditions. Archives of general psychiatry, 61(8), 807-816.
7. Grant, B. F. (1995). Comorbidity between DSM-IV drug use disorders and major depression: results of a national survey of adults. Journal of substance abuse, 7(4), 481-497.
8. Drinking Levels Defined. (n.d.). Retrieved from http://www.niaaa.nih.gov/alcohol-health/overview-alcohol-consumption/moderate-binge-drinking

Appendix 1: Python Code

```
# Phillip Thompkins (phith1@umbc.edu)
# CMSC478 Fall 2014
# Final Project - Information Gain Calculator
# Takes in a file and calcs the IG for each variable

from math import log
import sys
import operator

# checks for valid number of command line args
if len(sys.argv) != 2:
  print "Improper number of command line args."

TOTAL_PEOPLE = 37424.0
AMI_YES = 1835.0
AMI_NO = 35589.0
AMI_ENTROPY = 0 - ((AMI_YES / TOTAL_PEOPLE) * log((AMI_YES / TOTAL_PEOPLE),2)) - ((AMI_NO /
TOTAL_PEOPLE) * log((AMI_NO / TOTAL_PEOPLE),2))
SMI_YES = 554.0
SMI_NO = 36870.0
SMI_ENTROPY = 0 - ((SMI_YES / TOTAL_PEOPLE) * log((SMI_YES / TOTAL_PEOPLE),2)) - ((SMI_NO /
TOTAL_PEOPLE) * log((SMI_NO / TOTAL_PEOPLE),2))

# we need something to hold our variables and their i-gain
# let's use a hash table!
variables = {}

# command line should be 1) location of file
# first, we open up the file and analyze it, using open-uri
with open(str(sys.argv[1]), "r") as theFile:
    for line in theFile:
            vars = line.split(',')
            # if standard yes/no question...
            varName = vars[0]
            ya = 0.0
            na = 0.0

            if len(vars) == 5:
                    ya -= ((float(vars[1]) / SMI_YES) * log((float(vars[1]) / SMI_YES),2))
                    ya -= ((float(vars[2]) / SMI_YES) * log((float(vars[2]) / SMI_YES),2))
                    na -= ((float(vars[3]) / SMI_NO) * log((float(vars[3]) / SMI_NO),2))
                    na -= ((float(vars[4]) / SMI_NO) * log((float(vars[4]) / SMI_NO),2))

            # if not a yes/no question...
            if len(vars) == 7:
                    ya -= 0 - ((float(vars[1]) / SMI_YES) * log((float(vars[1]) / SMI_YES),2))
                    ya -= 0 - ((float(vars[2]) / SMI_YES) * log((float(vars[2]) / SMI_YES),2))
                    ya -= 0 - ((float(vars[3]) / SMI_YES) * log((float(vars[3]) / SMI_YES),2))
                    na -= 0 - ((float(vars[4]) / SMI_NO) * log((float(vars[4]) / SMI_NO),2))
                    na -= 0 - ((float(vars[5]) / SMI_NO) * log((float(vars[5]) / SMI_NO),2))
                    na -= 0 - ((float(vars[6]) / SMI_NO) * log((float(vars[6]) / SMI_NO),2))
```

```
if len(vars) == 9:
        ya -= 0 - ((float(vars[1]) / SMI_YES) * log((float(vars[1]) / SMI_YES),2))
        ya -= 0 - ((float(vars[2]) / SMI_YES) * log((float(vars[2]) / SMI_YES),2))
        ya -= 0 - ((float(vars[3]) / SMI_YES) * log((float(vars[3]) / SMI_YES),2))
        ya -= 0 - ((float(vars[4]) / SMI_YES) * log((float(vars[4]) / SMI_YES),2))
        na -= 0 - ((float(vars[5]) / SMI_NO) * log((float(vars[5]) / SMI_NO),2))
        na -= 0 - ((float(vars[6]) / SMI_NO) * log((float(vars[6]) / SMI_NO),2))
        na -= 0 - ((float(vars[7]) / SMI_NO) * log((float(vars[7]) / SMI_NO),2))
        na -= 0 - ((float(vars[8]) / SMI_NO) * log((float(vars[8]) / SMI_NO),2))

if len(vars) == 11:
        ya -= 0 - ((float(vars[1]) / SMI_YES) * log((float(vars[1]) / SMI_YES),2))
        ya -= 0 - ((float(vars[2]) / SMI_YES) * log((float(vars[2]) / SMI_YES),2))
        ya -= 0 - ((float(vars[3]) / SMI_YES) * log((float(vars[3]) / SMI_YES),2))
        ya -= 0 - ((float(vars[4]) / SMI_YES) * log((float(vars[4]) / SMI_YES),2))
        ya -= 0 - ((float(vars[5]) / SMI_YES) * log((float(vars[5]) / SMI_YES),2))
        na -= 0 - ((float(vars[6]) / SMI_NO) * log((float(vars[6]) / SMI_NO),2))
        na -= 0 - ((float(vars[7]) / SMI_NO) * log((float(vars[7]) / SMI_NO),2))
        na -= 0 - ((float(vars[8]) / SMI_NO) * log((float(vars[8]) / SMI_NO),2))
        na -= 0 - ((float(vars[9]) / SMI_NO) * log((float(vars[9]) / SMI_NO),2))
        na -= 0 - ((float(vars[10]) / SMI_NO) * log((float(vars[10]) / SMI_NO),2))

if len(vars) == 15:
        ya -= 0 - ((float(vars[1]) / SMI_YES) * log((float(vars[1]) / SMI_YES),2))
        ya -= 0 - ((float(vars[2]) / SMI_YES) * log((float(vars[2]) / SMI_YES),2))
        ya -= 0 - ((float(vars[3]) / SMI_YES) * log((float(vars[3]) / SMI_YES),2))
        ya -= 0 - ((float(vars[4]) / SMI_YES) * log((float(vars[4]) / SMI_YES),2))
        ya -= 0 - ((float(vars[5]) / SMI_YES) * log((float(vars[5]) / SMI_YES),2))
        ya -= 0 - ((float(vars[6]) / SMI_YES) * log((float(vars[6]) / SMI_YES),2))
        ya -= 0 - ((float(vars[7]) / SMI_YES) * log((float(vars[7]) / SMI_YES),2))
        na -= 0 - ((float(vars[8]) / SMI_NO) * log((float(vars[8]) / SMI_NO),2))
        na -= 0 - ((float(vars[9]) / SMI_NO) * log((float(vars[9]) / SMI_NO),2))
        na -= 0 - ((float(vars[10]) / SMI_NO) * log((float(vars[10]) / SMI_NO),2))
        na -= 0 - ((float(vars[11]) / SMI_NO) * log((float(vars[11]) / SMI_NO),2))
        na -= 0 - ((float(vars[12]) / SMI_NO) * log((float(vars[12]) / SMI_NO),2))
        na -= 0 - ((float(vars[13]) / SMI_NO) * log((float(vars[13]) / SMI_NO),2))
        na -= 0 - ((float(vars[14]) / SMI_NO) * log((float(vars[14]) / SMI_NO),2))

if len(vars) == 23:
        ya -= 0 - ((float(vars[1]) / SMI_YES) * log((float(vars[1]) / SMI_YES),2))
        ya -= 0 - ((float(vars[2]) / SMI_YES) * log((float(vars[2]) / SMI_YES),2))
        ya -= 0 - ((float(vars[3]) / SMI_YES) * log((float(vars[3]) / SMI_YES),2))
        ya -= 0 - ((float(vars[4]) / SMI_YES) * log((float(vars[4]) / SMI_YES),2))
        ya -= 0 - ((float(vars[5]) / SMI_YES) * log((float(vars[5]) / SMI_YES),2))
        ya -= 0 - ((float(vars[6]) / SMI_YES) * log((float(vars[6]) / SMI_YES),2))
        ya -= 0 - ((float(vars[7]) / SMI_YES) * log((float(vars[7]) / SMI_YES),2))
        ya -= 0 - ((float(vars[8]) / SMI_YES) * log((float(vars[8]) / SMI_YES),2))
        ya -= 0 - ((float(vars[9]) / SMI_YES) * log((float(vars[9]) / SMI_YES),2))
        ya -= 0 - ((float(vars[10]) / SMI_YES) * log((float(vars[10]) / SMI_YES),2))
        ya -= 0 - ((float(vars[11]) / SMI_YES) * log((float(vars[11]) / SMI_YES),2))
        na -= 0 - ((float(vars[12]) / SMI_NO) * log((float(vars[12]) / SMI_NO),2))
        na -= 0 - ((float(vars[13]) / SMI_NO) * log((float(vars[13]) / SMI_NO),2))
        na -= 0 - ((float(vars[14]) / SMI_NO) * log((float(vars[14]) / SMI_NO),2))
```

```
                na -= 0 - ((float(vars[15]) / SMI_NO) * log((float(vars[15]) / SMI_NO),2))
                na -= 0 - ((float(vars[16]) / SMI_NO) * log((float(vars[16]) / SMI_NO),2))
                na -= 0 - ((float(vars[17]) / SMI_NO) * log((float(vars[17]) / SMI_NO),2))
                na -= 0 - ((float(vars[18]) / SMI_NO) * log((float(vars[18]) / SMI_NO),2))
                na -= 0 - ((float(vars[19]) / SMI_NO) * log((float(vars[19]) / SMI_NO),2))
                na -= 0 - ((float(vars[20]) / SMI_NO) * log((float(vars[20]) / SMI_NO),2))
                na -= 0 - ((float(vars[21]) / SMI_NO) * log((float(vars[21]) / SMI_NO),2))
                na -= 0 - ((float(vars[22]) / SMI_NO) * log((float(vars[22]) / SMI_NO),2))
        iGain = SMI_ENTROPY - (ya + na)
        variables[varName] = iGain

theFile.close()

# so at this point we're done with our hash and now we have to...
# sort the hash and get the top i-gains!
sorted_vars = sorted(variables.items(), key=operator.itemgetter(1))
for item in sorted_vars[-25:]:
    print "%s: %d" % (item)
```

Appendix 2: Decision Tree Variables

AGE2	Age. Single ages and ranges as age increases.
ADTMTHLP	How much has treatment/counseling helped with depression in the last 12 months? 1-5 point Likert scale.
ADRXHLP	How much has prescription medicine for mood helped in the last 12 months? 1-5 point Likert scale.
BINGEHVY	Level of alcohol use in past month. Heavy binge, Nonheavy binge, Non-binge alcohol consumption, no alcohol consumed.
CATAG7	Age, broken into seven small categories.
CIGALCMO	Past month use of both cigarettes and alcohol; four options corresponding to two-variable boolean OR.
DEPNDALC	Alcohol dependence in the past year; binary yes/no response.
DSTCHR30	How often have you felt that nothing could cheer you up in the past month? 1-5 point Likert scale.
DSTEFF30	How often did you feel that everything was an effort in the past month? 1-5 point Likert scale.
DSTHOP30	How often did you feel hopeless in the past month? 1-5 point Likert scale.
DSTNGD30	How often have you felt down, no good, or worthless in the past month? 1-5 point Likert scale.
DSTNRV30	How often have you felt nervous in the past month? 1-5 point Likert scale.
DSTRST30	How often have you felt restless in the past month? 1-5 point Likert scale.
EDUCCAT2	Education. Less than high school, high school, some college, college graduate.
HEALTH	Overall health score. Excellent, Very Good, Good, Fair, Poor.
IREDUC2	Education variable recoded.
IRMARIT	Marital status. Married, Widowed, Divorced/Separated, Never Married.
NEWRACE2	Race/Ethnicity.
PREGAGE2	Pregnancy age; unclear if for parent or responder.
SEXAGE	Combines gender/age. M/F 12-17 and 18-25, then “Otherwise” category.
SEXRACE	Combines gender/race. M/F White, Black, Hispanic, Other Races.

Appendix 3: AMIDAU_2 Decision Tree

```
J48 pruned tree
------------------

ADTMTHLP = legit skip: No (34749.0/1286.0)
ADTMTHLP = some
|  BINGEHVY = Did not use in past month: No (250.0/27.0)
|  BINGEHVY = Binge but not Heavy Use
|  |  SEXAGE = male 18 to 25: Yes (25.0/6.0)
|  |  SEXAGE = otherwise: No (54.0/20.0)
|  |  SEXAGE = female 18 to 25
|  |  |  DSTCHR30 = none of the time: No (4.0/1.0)
|  |  |  DSTCHR30 = a little of the time: No (12.0/1.0)
|  |  |  DSTCHR30 = some of the time: Yes (14.0/5.0)
|  |  |  DSTCHR30 = most of the time
|  |  |  |  DSTNGD30 = a little of the time: Yes (0.0)
|  |  |  |  DSTNGD30 = none of the time: Yes (0.0)
|  |  |  |  DSTNGD30 = some of the time: No (2.0)
|  |  |  |  DSTNGD30 = most of the time: Yes (3.0)
|  |  |  |  DSTNGD30 = all of the time: Yes (0.0)
|  |  |  |  DSTNGD30 = dont know: Yes (0.0)
|  |  |  |  DSTNGD30 = refused to answer: Yes (0.0)
|  |  |  |  DSTNGD30 = no answer: Yes (0.0)
|  |  |  |  DSTNGD30 = bad data: Yes (0.0)
|  |  |  DSTCHR30 = all of the time: No (1.0)
|  |  |  DSTCHR30 = dont know: No (0.0)
|  |  |  DSTCHR30 = refused to answer: No (0.0)
|  |  |  DSTCHR30 = no answer: No (0.0)
|  |  |  DSTCHR30 = bad data: No (0.0)
|  BINGEHVY = Past month but not Binge: No (156.0/20.0)
|  BINGEHVY = Heavy Alcohol Use: Yes (56.0/14.0)
ADTMTHLP = a little
|  BINGEHVY = Did not use in past month: No (157.0/21.0)
|  BINGEHVY = Binge but not Heavy Use
|  |  DSTNRV30 = a little of the time
|  |  |  SEXAGE = male 18 to 25: No (2.0)
|  |  |  SEXAGE = otherwise: Yes (5.0)
|  |  |  SEXAGE = female 18 to 25
|  |  |  |  IREDUC2 = college freshman: Yes (3.0)
|  |  |  |  IREDUC2 = 12th grade: Yes (1.0)
|  |  |  |  IREDUC2 = college senior or grad: Yes (0.0)
|  |  |  |  IREDUC2 = 11th grade: No (2.0/1.0)
|  |  |  |  IREDUC2 = college soph or junior: No (3.0)
|  |  |  |  IREDUC2 = 10th grade: Yes (0.0)
|  |  |  |  IREDUC2 = 9th grade: Yes (0.0)
|  |  |  |  IREDUC2 = 6th grade: Yes (0.0)
|  |  |  |  IREDUC2 = 8th grade: Yes (0.0)
|  |  |  |  IREDUC2 = 5th grade or less: Yes (0.0)
|  |  |  |  IREDUC2 = 7th grade: Yes (0.0)
|  |  DSTNRV30 = none of the time: No (5.0)
|  |  DSTNRV30 = some of the time
|  |  |  IRMARIT = never been married
|  |  |  |  CIGALCMO = Cig and No Alc: Yes (0.0)
|  |  |  |  CIGALCMO = Alc and No Cig: Yes (13.0/2.0)
|  |  |  |  CIGALCMO = Cig and Alc: No (15.0/6.0)
|  |  |  |  CIGALCMO = No Cig or Alc: Yes (0.0)
|  |  |  IRMARIT = married: No (8.0/3.0)
|  |  |  IRMARIT = widowed: No (0.0)
|  |  |  IRMARIT = divorced or separated: No (5.0)
|  |  DSTNRV30 = most of the time
|  |  |  HEALTH = excellent: Yes (1.0)
|  |  |  HEALTH = very good: No (7.0)
|  |  |  HEALTH = good
|  |  |  |  DSTNGD30 = a little of the time: No (2.0/1.0)
|  |  |  |  DSTNGD30 = none of the time: No (1.0)
|  |  |  |  DSTNGD30 = some of the time: No (2.0)
|  |  |  |  DSTNGD30 = most of the time: Yes (5.0)
|  |  |  |  DSTNGD30 = all of the time: Yes (0.0)
|  |  |  |  DSTNGD30 = dont know: Yes (0.0)
|  |  |  |  DSTNGD30 = refused to answer: Yes (0.0)
|  |  |  |  DSTNGD30 = no answer: Yes (0.0)
|  |  |  |  DSTNGD30 = bad data: Yes (0.0)
|  |  |  HEALTH = poor: No (2.0)
|  |  |  HEALTH = fair: No (2.0)
|  |  |  HEALTH = 94.0: No (0.0)
|  |  |  HEALTH = 97.0: No (0.0)
|  |  DSTNRV30 = all of the time: Yes (4.0/1.0)
|  |  DSTNRV30 = dont know: No (0.0)
|  |  DSTNRV30 = refused to answer: No (0.0)
|  |  DSTNRV30 = no answer: No (0.0)
|  |  DSTNRV30 = bad data: No (0.0)
|  BINGEHVY = Past month but not Binge: No (98.0/15.0)
|  BINGEHVY = Heavy Alcohol Use
|  |  DSTNGD30 = a little of the time: Yes (5.0)
|  |  DSTNGD30 = none of the time
|  |  |  IRMARIT = never been married: No (4.0)
|  |  |  IRMARIT = married: Yes (3.0/1.0)
|  |  |  IRMARIT = widowed: No (0.0)
|  |  |  IRMARIT = divorced or separated: No (0.0)
|  |  DSTNGD30 = some of the time
|  |  |  CATAG7 = 18 to 20: Yes (4.0)
|  |  |  CATAG7 = 35 and up: No (2.0/1.0)
|  |  |  CATAG7 = 26 to 34: No (3.0)
|  |  |  CATAG7 = 21 to 25: No (8.0/3.0)
|  |  DSTNGD30 = most of the time
|  |  |  CIGALCMO = Cig and No Alc: Yes (0.0)
|  |  |  CIGALCMO = Alc and No Cig
|  |  |  |  DSTNRV30 = a little of the time: No (0.0)
|  |  |  |  DSTNRV30 = none of the time: No (0.0)
|  |  |  |  DSTNRV30 = some of the time: No (2.0)
|  |  |  |  DSTNRV30 = most of the time: Yes (2.0)
|  |  |  |  DSTNRV30 = all of the time: No (0.0)
|  |  |  |  DSTNRV30 = dont know: No (0.0)
|  |  |  |  DSTNRV30 = refused to answer: No (0.0)
|  |  |  |  DSTNRV30 = no answer: No (0.0)
|  |  |  |  DSTNRV30 = bad data: No (0.0)
|  |  |  CIGALCMO = Cig and Alc: Yes (8.0)
|  |  |  CIGALCMO = No Cig or Alc: Yes (0.0)
|  |  DSTNGD30 = all of the time: Yes (11.0/2.0)
|  |  DSTNGD30 = dont know: Yes (0.0)
|  |  DSTNGD30 = refused to answer: Yes (0.0)
|  |  DSTNGD30 = no answer: Yes (0.0)
|  |  DSTNGD30 = bad data: Yes (0.0)
ADTMTHLP = not at all
|  BINGEHVY = Did not use in past month: No (146.0/16.0)
|  BINGEHVY = Binge but not Heavy Use
|  |  DSTNRV30 = a little of the time: No (15.0/3.0)
|  |  DSTNRV30 = none of the time: No (8.0)
|  |  DSTNRV30 = some of the time
|  |  |  EDUCCAT2 = some college
|  |  |  |  DSTRST30 = a little of the time: Yes (2.0)
|  |  |  |  DSTRST30 = most of the time: No (0.0)
|  |  |  |  DSTRST30 = none of the time: No (2.0/1.0)
|  |  |  |  DSTRST30 = some of the time: No (5.0)
|  |  |  |  DSTRST30 = dont know: No (0.0)
|  |  |  |  DSTRST30 = all of the time: No (0.0)
|  |  |  |  DSTRST30 = refused to answer: No (0.0)
|  |  |  |  DSTRST30 = no answer: No (0.0)
|  |  |  |  DSTRST30 = bad data: No (0.0)
|  |  |  EDUCCAT2 = high school grad: Yes (6.0/1.0)
|  |  |  EDUCCAT2 = college grad: No (7.0/1.0)
|  |  |  EDUCCAT2 = less than high school: Yes (4.0)
|  |  DSTNRV30 = most of the time
|  |  |  DSTNGD30 = a little of the time: Yes (0.0)
|  |  |  DSTNGD30 = none of the time: No (1.0)
|  |  |  DSTNGD30 = some of the time: No (2.0/1.0)
|  |  |  DSTNGD30 = most of the time: Yes (7.0)
|  |  |  DSTNGD30 = all of the time: No (4.0/1.0)
|  |  |  DSTNGD30 = dont know: Yes (0.0)
|  |  |  DSTNGD30 = refused to answer: Yes (0.0)
|  |  |  DSTNGD30 = no answer: Yes (0.0)
|  |  |  DSTNGD30 = bad data: Yes (0.0)
|  |  DSTNRV30 = all of the time
|  |  |  PREGAGE2 = 18 to 25
|  |  |  |  IRMARIT = never been married: Yes (6.0/1.0)
|  |  |  |  IRMARIT = married: Yes (1.0)
|  |  |  |  IRMARIT = widowed: Yes (0.0)
|  |  |  |  IRMARIT = divorced or separated: No (3.0)
|  |  |  PREGAGE2 = otherwise: No (2.0)
|  |  |  PREGAGE2 = 26 to 44: Yes (4.0)
|  |  DSTNRV30 = dont know: No (0.0)
|  |  DSTNRV30 = refused to answer: No (0.0)
|  |  DSTNRV30 = no answer: No (0.0)
|  |  DSTNRV30 = bad data: No (0.0)
|  BINGEHVY = Past month but not Binge: No (90.0/10.0)
|  BINGEHVY = Heavy Alcohol Use
|  |  DSTNRV30 = a little of the time: No (6.0/1.0)
|  |  DSTNRV30 = none of the time: No (2.0/1.0)
|  |  DSTNRV30 = some of the time
|  |  |  IRMARIT = never been married: Yes (7.0/1.0)
|  |  |  IRMARIT = married: No (2.0)
|  |  |  IRMARIT = widowed: Yes (0.0)
|  |  |  IRMARIT = divorced or separated: No (1.0)
|  |  DSTNRV30 = most of the time: Yes (8.0)
|  |  DSTNRV30 = all of the time: Yes (6.0)
|  |  DSTNRV30 = dont know: Yes (0.0)
|  |  DSTNRV30 = refused to answer: Yes (0.0)
|  |  DSTNRV30 = no answer: Yes (0.0)
|  |  DSTNRV30 = bad data: Yes (0.0)
ADTMTHLP = no answer: No (312.0/16.0)
ADTMTHLP = extremely: No (365.0/73.0)
ADTMTHLP = a lot
|  BINGEHVY = Did not use in past month: No (266.0/28.0)
|  BINGEHVY = Binge but not Heavy Use: No (145.0/48.0)
|  BINGEHVY = Past month but not Binge: No (192.0/22.0)
|  BINGEHVY = Heavy Alcohol Use
|  |  SEXAGE = male 18 to 25: Yes (10.0/1.0)
|  |  SEXAGE = otherwise
|  |  |  DSTNRV30 = a little of the time: No (8.0/1.0)
|  |  |  DSTNRV30 = none of the time: No (1.0)
|  |  |  DSTNRV30 = some of the time: No (7.0)
|  |  |  DSTNRV30 = most of the time: Yes (3.0)
|  |  |  DSTNRV30 = all of the time: Yes (3.0/1.0)
|  |  |  DSTNRV30 = dont know: No (0.0)
|  |  |  DSTNRV30 = refused to answer: No (0.0)
|  |  |  DSTNRV30 = no answer: No (0.0)
|  |  |  DSTNRV30 = bad data: No (0.0)
|  |  SEXAGE = female 18 to 25: Yes (13.0/5.0)
ADTMTHLP = dont know: No (19.0/2.0)
ADTMTHLP = refused to answer: No (12.0/2.0)

Number of Leaves  :     157

Size of the tree :      187


Time taken to build model: 6.66 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        35750               95.5269 %
Incorrectly Classified Instances       1674                4.4731 %
Kappa statistic                        0.1851
Mean absolute error                    0.0827
Root mean squared error                0.2034
Relative absolute error               88.6971 %
Root relative squared error           94.1901 %
Total Number of Instances             37424
```

Appendix 4: SMIDAU_2 Decision Tree

```
J48 pruned tree
------------------

DEPNDALC = no: No (35715.0/298.0)
DEPNDALC = yes
|   ADTMTHLP = legit skip: No (1435.0/110.0)
|   ADTMTHLP = some
|   |   DSTNGD30 = a little of the time
|   |   |   DSTHOP30 = a little of the time
|   |   |   |   BINGEHVY = Did not use in past month: Yes (2.0)
|   |   |   |   BINGEHVY = Binge but not Heavy Use: No (3.0/1.0)
|   |   |   |   BINGEHVY = Past month but not Binge: No (0.0)
|   |   |   |   BINGEHVY = Heavy Alcohol Use: No (1.0)
|   |   |   DSTHOP30 = none of the time: No (3.0)
|   |   |   DSTHOP30 = some of the time: Yes (4.0)
|   |   |   DSTHOP30 = most of the time: No (0.0)
|   |   |   DSTHOP30 = all of the time: No (1.0)
|   |   |   DSTHOP30 = dont know: No (0.0)
|   |   |   DSTHOP30 = refused to answer: No (0.0)
|   |   |   DSTHOP30 = no answer: No (0.0)
|   |   |   DSTHOP30 = bad data: No (0.0)
|   |   DSTNGD30 = none of the time: No (6.0/1.0)
|   |   DSTNGD30 = some of the time: Yes (24.0/8.0)
|   |   DSTNGD30 = most of the time: Yes (21.0/3.0)
|   |   DSTNGD30 = all of the time: Yes (12.0)
|   |   DSTNGD30 = dont know: Yes (0.0)
|   |   DSTNGD30 = refused to answer: Yes (0.0)
|   |   DSTNGD30 = no answer: Yes (0.0)
|   |   DSTNGD30 = bad data: Yes (0.0)
|   ADTMTHLP = a little
|   |   DSTNRV30 = a little of the time: No (10.0/4.0)
|   |   DSTNRV30 = none of the time: No (1.0)
|   |   DSTNRV30 = some of the time: No (17.0/5.0)
|   |   DSTNRV30 = most of the time: Yes (15.0/4.0)
|   |   DSTNRV30 = all of the time: Yes (6.0/1.0)
|   |   DSTNRV30 = dont know: Yes (0.0)
|   |   DSTNRV30 = refused to answer: Yes (0.0)
|   |   DSTNRV30 = no answer: Yes (0.0)
|   |   DSTNRV30 = bad data: Yes (0.0)
|   ADTMTHLP = not at all
|   |   CIGALCMO = Cig and No Alc: No (3.0)
|   |   CIGALCMO = Alc and No Cig: Yes (8.0/1.0)
|   |   CIGALCMO = Cig and Alc
|   |   |   BINGEHVY = Did not use in past month: Yes (0.0)
|   |   |   BINGEHVY = Binge but not Heavy Use: Yes (14.0/3.0)
|   |   |   BINGEHVY = Past month but not Binge: Yes (1.0)
|   |   |   BINGEHVY = Heavy Alcohol Use: No (8.0/2.0)
|   |   CIGALCMO = No Cig or Alc: Yes (0.0)
|   ADTMTHLP = no answer: No (15.0/1.0)
|   ADTMTHLP = extremely
|   |   HEALTH = excellent
|  |  |  DSTRST30 = a little of the time: No (1.0)
|  |  |  DSTRST30 = most of the time: Yes (1.0)
|  |  |  DSTRST30 = none of the time: Yes (2.0)
|  |  |  DSTRST30 = some of the time: No (5.0)
|  |  |  DSTRST30 = dont know: No (0.0)
|  |  |  DSTRST30 = all of the time: No (1.0)
|  |  |  DSTRST30 = refused to answer: No (0.0)
|  |  |  DSTRST30 = no answer: No (0.0)
|  |  |  DSTRST30 = bad data: No (0.0)
|  |  HEALTH = very good: No (14.0/6.0)
|  |  HEALTH = good
|  |  |  IREDUC2 = college freshman: Yes (2.0)
|  |  |  IREDUC2 = 12th grade: Yes (4.0/1.0)
|  |  |  IREDUC2 = college senior or grad
|  |  |  |  PREGAGE2 = 18 to 25: No (2.0)
|  |  |  |  PREGAGE2 = otherwise: Yes (0.0)
|  |  |  |  PREGAGE2 = 26 to 44: Yes (3.0)
|  |  |  IREDUC2 = 11th grade: Yes (0.0)
|  |  |  IREDUC2 = college soph or junior: No (4.0)
|  |  |  IREDUC2 = 10th grade: Yes (0.0)
|  |  |  IREDUC2 = 9th grade: Yes (0.0)
|  |  |  IREDUC2 = 6th grade: Yes (0.0)
|  |  |  IREDUC2 = 8th grade: Yes (0.0)
|  |  |  IREDUC2 = 5th grade or less: Yes (0.0)
|  |  |  IREDUC2 = 7th grade: Yes (0.0)
|  |  HEALTH = poor: No (2.0/1.0)
|  |  HEALTH = fair: Yes (3.0)
|  |  HEALTH = 94.0: No (0.0)
|  |  HEALTH = 97.0: No (0.0)
|  ADTMTHLP = a lot
|  |  DSTCHR30 = none of the time: No (6.0/1.0)
|  |  DSTCHR30 = a little of the time: No (20.0/5.0)
|  |  DSTCHR30 = some of the time: Yes (16.0/6.0)
|  |  DSTCHR30 = most of the time: Yes (9.0/4.0)
|  |  DSTCHR30 = all of the time: Yes (2.0)
|  |  DSTCHR30 = dont know: No (0.0)
|  |  DSTCHR30 = refused to answer: No (0.0)
|  |  DSTCHR30 = no answer: No (0.0)
|  |  DSTCHR30 = bad data: No (0.0)
|  ADTMTHLP = dont know: No (2.0/1.0)
|  ADTMTHLP = refused to answer: No (0.0)

Number of Leaves  :     77

Size of the tree :     90


Time taken to build model: 7.75 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         36957               98.7521 %
Incorrectly Classified Instances         467                1.2479 %
```

Kappa statistic	0.3315
Mean absolute error	0.0234
Root mean squared error	0.1081
Relative absolute error	80.1036 %
Root relative squared error	89.5392 %
Total Number of Instances	37424