Machine learning for data analysis

Assignment 1

Running a classification tree

I chose **addhealth** as my data set and general health (H1GH1) as response variable. H1GH1 is a categorical variable with 5 levels. So I collapsed it into two levels and create a new variable named *health*. *Health* equal 1 means this participant is healthy, while *health* equal 0 means this participant is not healthy.

The HPSPLIT Procedure

Performance Information	
Execution Mode	Single-Machine
Number of Threads	2

Data Access Information			
Data	Engine	Role	Path
WORK.NEW	V9	Input	On Client

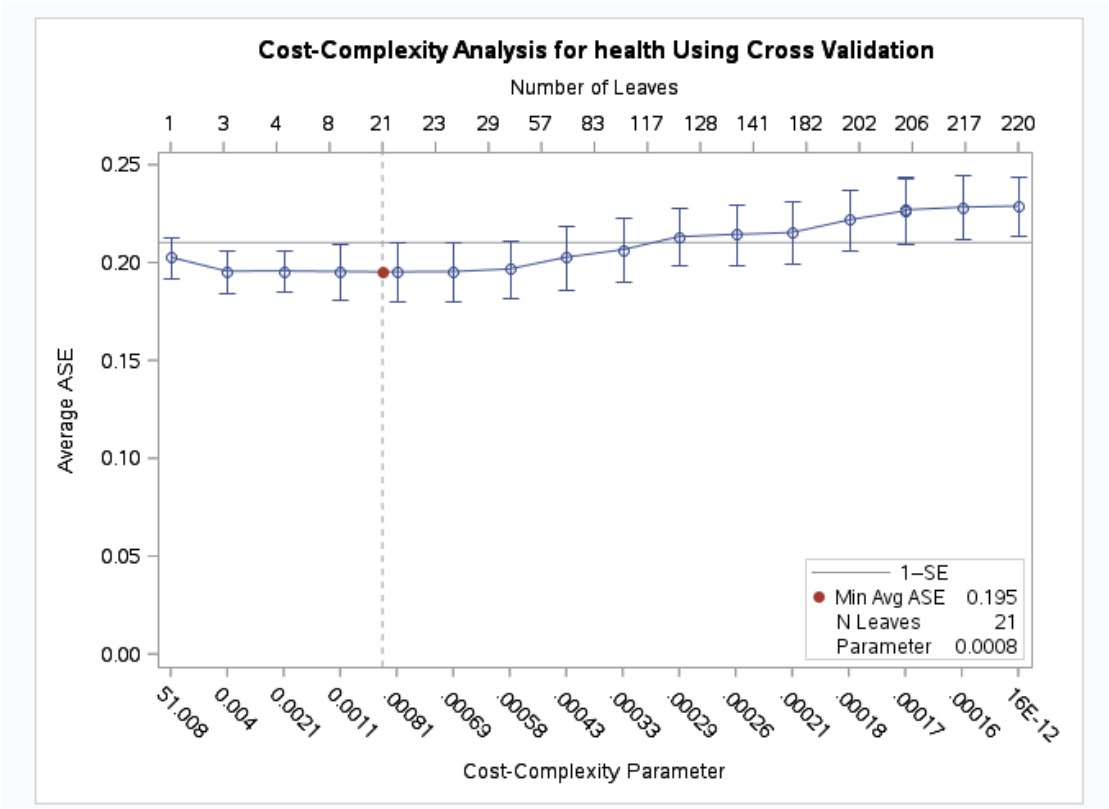
Model Information	
Split Criterion Used	Entropy
Pruning Method	Cost-Complexity
Subtree Evaluation Criterion	Cost-Complexity
Number of Branches	2
Maximum Tree Depth Requested	10
Maximum Tree Depth Achieved	10
Tree Depth	6
Number of Leaves Before Pruning	330
Number of Leaves After Pruning	15
Model Event Level	0

Number of Observations Read	6337
Number of Observations Used	4117

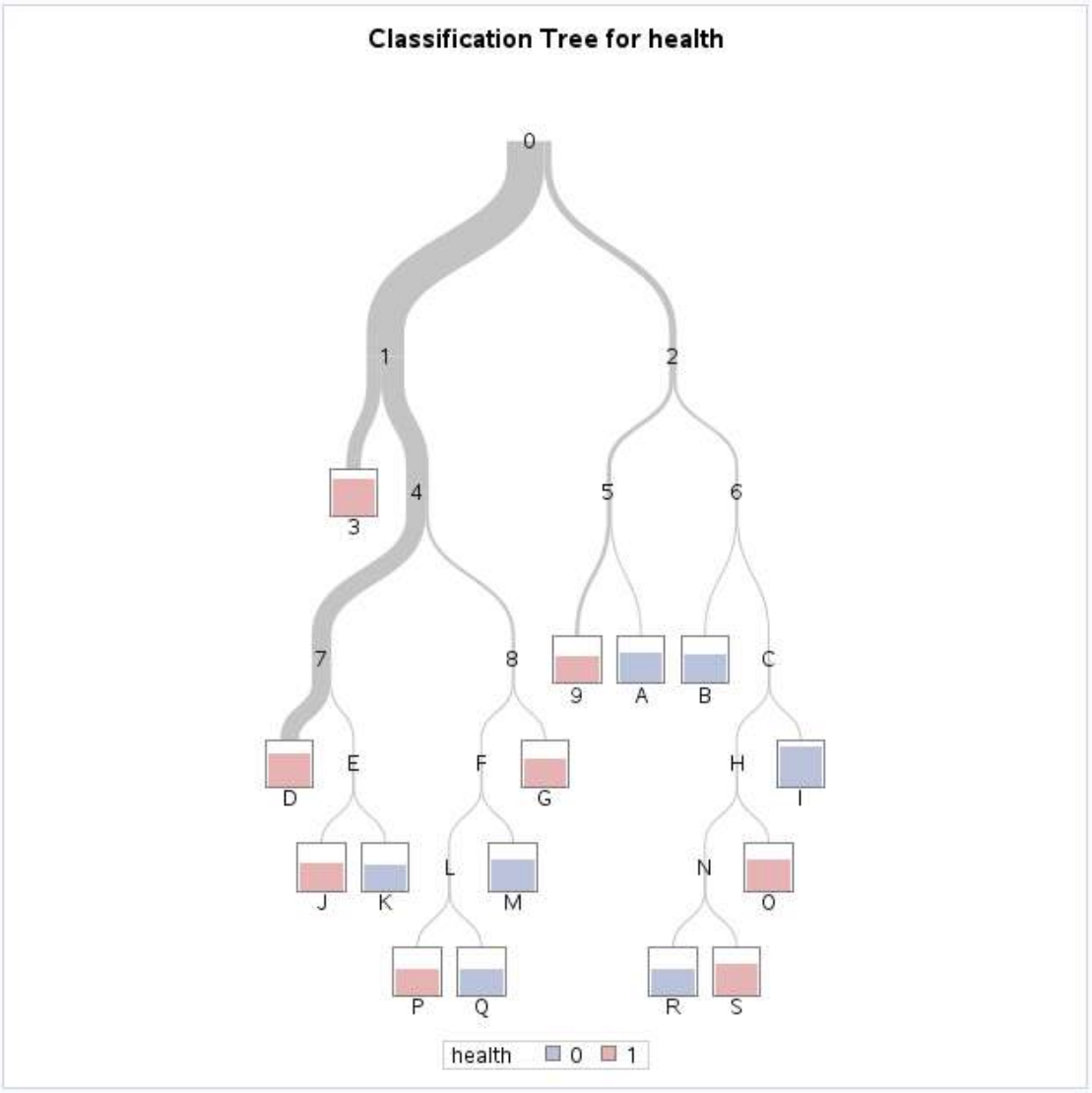
This is the result of ***hpsplit*** procedure. I used *prune costcomplexity* statement in my code to pruning a large tree

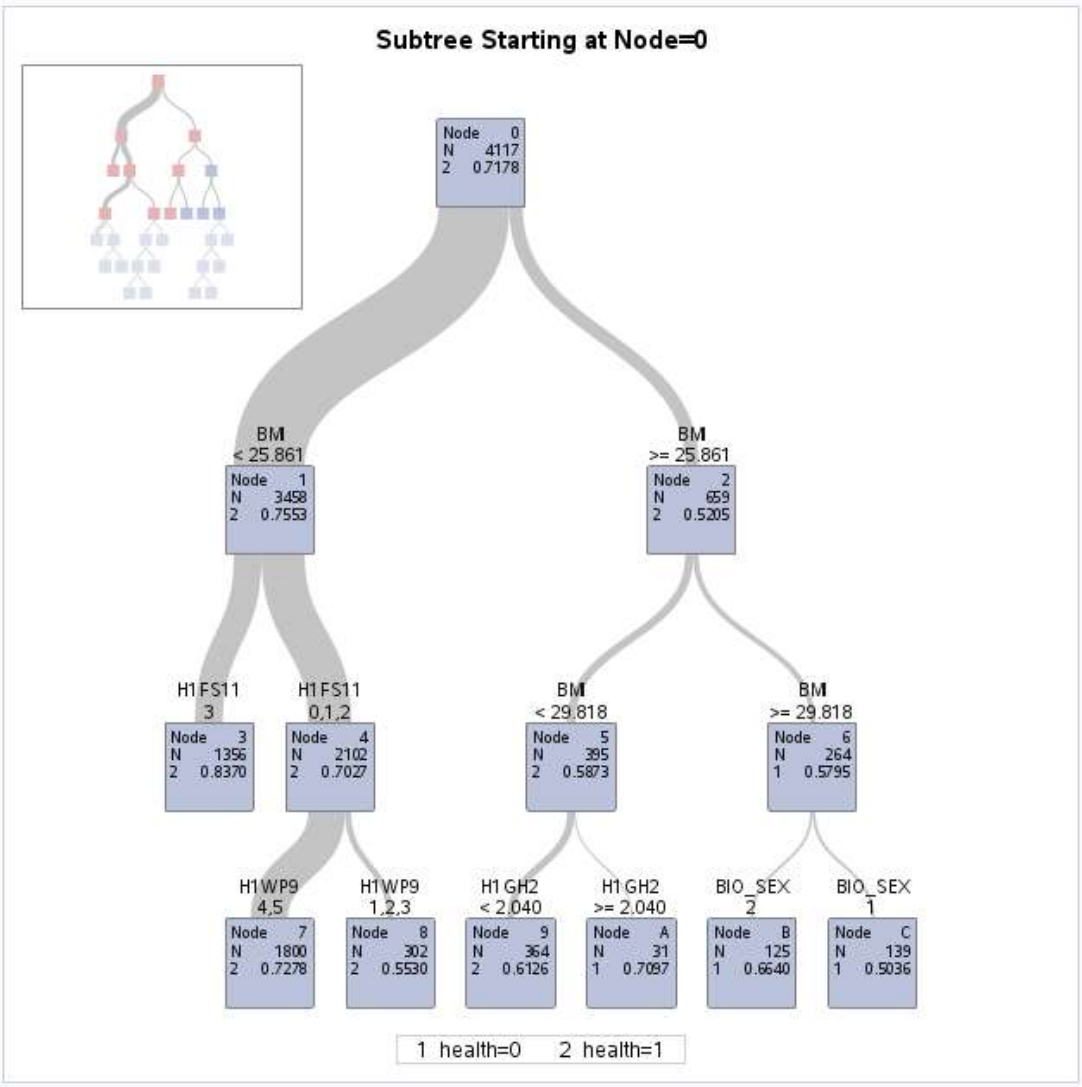
into a smaller tree. As we can see in the model information table, the number of leaves before pruning is 330, while the number of leaves after pruning is only 15, which is much smaller than 330. This result indentifies the function of prune statement.

This plot shows the average line for the standard error.



This is the general model and the final classification tree.





This shows how well the tree fit the model.

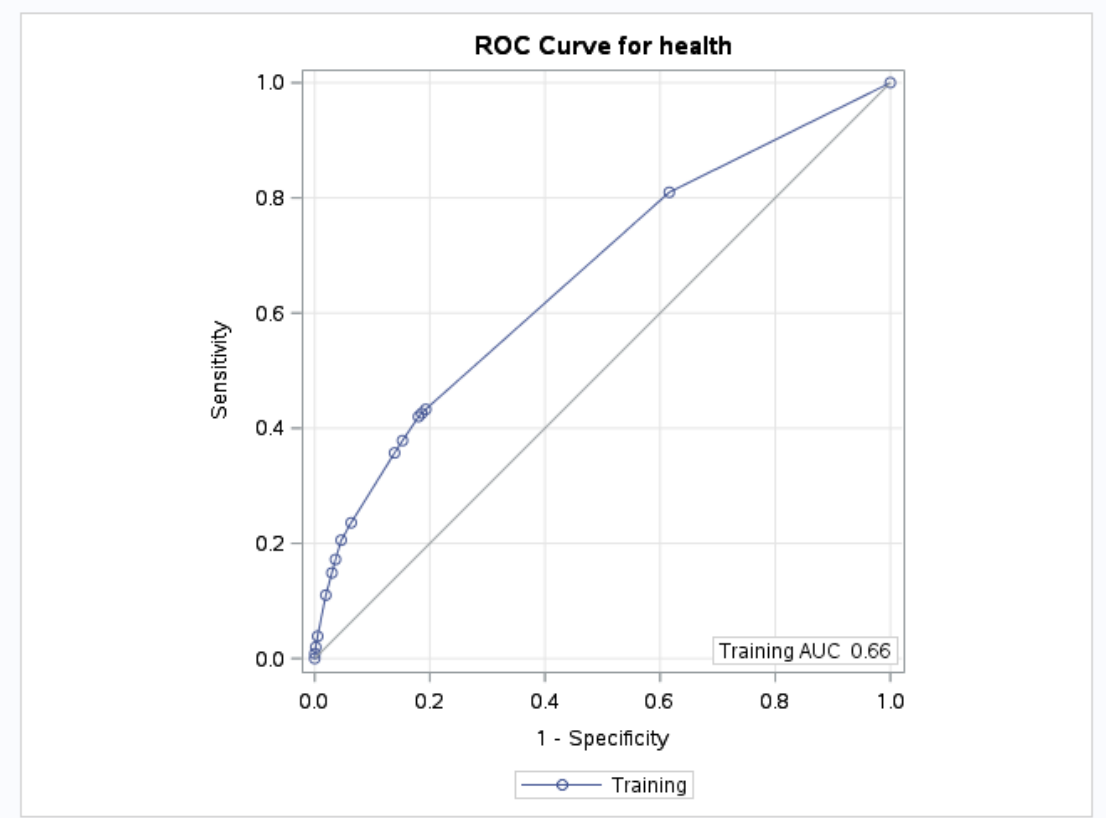
The HPSPLIT Procedure

Model-Based Confusion Matrix

Actual	Predicted 0	Predicted 1	Error Rate
0	239	923	0.7943
1	136	2819	0.0460

Model-Based Fit Statistics for Selected Tree

N Leaves	ASE	Mis-class	Sensitivity	Specificity	Entropy	Gini	RSS	AUC
15	0.1841	0.2572	0.2057	0.9540	0.7963	0.3682	1515.8	0.6635



This table shows the most important variables to determine the response variable health. It shows BMI (body mass index) contributes the most to the health.

Variable Importance				
		Training		
Variable	Variable Label	Relative	Importance	Count
BMI		1.0000	8.5911	3
H1FS11	feeling happy	0.6809	5.8493	3
H1GH2	frequency of headache	0.5078	4.3625	4
H1WP9	how close with mother	0.4627	3.9753	1
H1GI20	grade	0.2864	2.4607	2
BIO_SEX	gender	0.2142	1.8403	1

Code:

```
LIBNAME mydata "/courses/d1406ae5ba27fe300" access=readonly;
data new; set mydata.addhealth_pds;

if H1GI20=97 then delete; if H1GI20=99 then delete; if H1GI20=96 then delete;
if H1GI20=98 then delete;
if H1GH1=6 then H1GH1=.; if H1GH1=8 then H1GH1=.;
if H1GH2=6 then H1GH2=.; if H1GH2=8 then H1GH2=.;
if H1GH6=6 then H1GH6=.; if H1GH6=8 then H1GH6=.;
if H1GH59A=96 then H1GH59A=.; if H1GH59A=98 then H1GH59A=.; if H1GH59A=99 then H1GH59A=.;
if H1GH59B=96 then H1GH59B=.; if H1GH59B=98 then H1GH59B=.; if H1GH59B=99 then H1GH59B=.;
if H1GH60=996 then H1GH60=.; if H1GH60=998 then H1GH60=.; if H1GH60=999 then H1GH60=.;
if H1FS11=6 then H1FS11=.; if H1FS11=8 then H1FS11=.;
if H1WP9=6 then H1WP9=.; if H1WP9=7 then H1WP9=.;
if H1WP9=8 then H1WP9=.; if H1WP9=9 then H1WP9=.;
if H1WP13=6 then H1WP13=.; if H1WP13=7 then H1WP13=.;
if H1WP13=8 then H1WP13=.; if H1WP13=9 then H1WP13=.;


H1GH59=H1GH59A * 12 + H1GH59B;/*add a new variable*/
BMI=H1GH60 * 0.454/(H1GH59 * 0.0254)**2;/*body mass index*/

if H1GH1 <= 2 then health = 1;
else health = 0;

label AID="respondent ID"
      BIO_SEX="gender"
      H1GI20="grade"

      H1GH1="general health"
      H1GH2="frequency of headache"
      H1GH6="frequnecy of feeling weak"
      H1GH59A="height in feet"
      H1GH59B="height in inch"
      H1GH60="weigt (pound)"
      H1FS11="feeling happy"
      H1WP9="how close with mother"
      H1WP13="how close with father"
      H1GH59="height (inch)";


proc sort; by AID;


ods graphics on;
proc hpsplit seed = 10000;
class health BIO_SEX H1FS11 H1WP9 H1WP13;
model health = BIO_SEX BMI H1FS11 H1WP9 H1WP13 H1GI20 H1GH2;
grow entropy;
prune costcomplexity;
run;
```