Kyungjin Cho
Stat443

Classin.txt

```
GUIDE      (do not edit this file unless you know what you are doing)
 36.2      (version of GUIDE that generated this file)
 1         (1=model fitting, 2=importance or DIF scoring, 3=data conversion)
"classout.txt"  (name of output file)
 1         (1=one tree, 2=ensemble)
 1         (1=classification, 2=regression, 3=propensity score grouping)
 1         (1=simple model, 2=nearest-neighbor, 3=kernel)
 1         (0=linear 1st, 1=univariate 1st, 2=skip linear, 3=skip linear and interaction)
 1         (0=tree with fixed no. of nodes, 1=prune by CV, 2=by test sample, 3=no pruning)
"rhcdsc1.txt"  (name of data description file)
      10  (number of cross-validations)
 1         (1=mean-based CV tree, 2=median-based CV tree)
   0.500  (SE number for pruning)
 1         (1=estimated priors, 2=equal priors, 3=other priors)
 1         (1=unit misclassification costs, 2=other)
 2         (1=split point from quantiles, 2=use exhaustive search)
 1         (1=default max. number of split levels, 2=specify no. in next line)
 1         (1=default min. node size, 2=specify min. value in next line)
 2         (0=no LaTeX code, 1=tree without node numbers, 2=tree with node numbers)
"class.tex" (latex file name)
 1         (1=color terminal nodes, 2=no colors)
 2         (0=#errors, 1=sample sizes, 2=sample proportions, 3=posterior probs, 4=nothing)
 1         (1=no storage, 2=store fit and split variables, 3=store split variables and values)
 2         (1=do not save fitted values and node IDs, 2=save in a file)
"classpred.txt" (file name for fitted values and node IDs)
 2         (1=do not write R function, 2=write R function)
"class.r" (R code file)
 1         (rank of top variable to split root node)
```

Classout.txt

```
GGG   U   U I DDDD   EEEE
G   G U   U I D   D E
G     U   U I D   D E
G GG U   U I D   D EEE
G   G U   U I D   D E
G   G U   U I D   D E
 GGG    UUU   I DDDD   EEEE
```

GUIDE Classification and Regression Trees and Forests
Version 36.2 (Build date: January 8, 2021)
Compiled with GFortran 10.2.0 on macOS Big Sur 11.1

This software is based upon work supported by the U.S. Army Research Office,
the National Science Foundation and the National Institutes of Health.

This job was started on 02/12/21 at 01:27

Classification tree
Pruning by cross-validation
Data description file: rhcdsc1.txt
Training sample file: rhcdata.txt
Missing value code: NA
Records in data file start on line 2
20 N variables changed to S
D variable is swang1
Number of records in data file: 5735
Length of longest entry in data file: 19
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Number of classes: 2
Training sample class proportions of D variable swang1:

Class	#Cases	Proportion
NoRHC	3551	0.61918047
RHC	2184	0.38081953

Summary information for training sample of size 5735
d=dependent, b=split and fit cat variable using indicator variables,
c=split-only categorical, i=fit-only categorical (via indicators),
s=split-only numerical, n=split and fit numerical, f=fit-only numerical,
m=missing-value flag variable, p=periodic variable, w=weight

Column	Name		Minimum	Maximum	#Codes/ Levels/ Periods	#Missing
2	cat1	c			9	
3	cat2	c			6	4535
4	ca	c			3	
10	cardiohx	c			2	
11	chfhx	c			2	
12	dementhx	c			2	
13	psychhx	c			2	
14	chrpulhx	c			2	
15	renalhx	c			2	
16	liverhx	c			2	
17	gibledhx	c			2	
18	malighx	c			2	
19	immunhx	c			2	
20	transhx	c			2	
21	amihx	c			2	
22	age	s	18.04	101.8		
23	sex	c			2	
24	edu	s	0.000	30.00		
29	aps1	s	3.000	147.0		
30	scoma1	s	0.000	100.0		
31	meanbp1	s	10.00	259.0		80
32	wblc1	s	0.000	192.0		
33	hrt1	s	8.000	250.0		159
34	resp1	s	2.000	100.0		136
35	temp1	s	27.00	43.00		
36	pafi1	s	11.60	937.5		
37	alb1	s	0.3000	29.00		
38	hema1	s	2.000	66.19		
39	bili1	s	0.9999E-01	58.20		
40	crea1	s	0.9999E-01	25.10		
41	sod1	s	101.0	178.0		
42	pot1	s	1.100	11.90		
43	paco21	s	1.000	156.0		
44	ph1	s	6.579	7.770		
45	swang1	d			2	
46	wtkilo1	s	19.50	244.0		515
47	dnr1	c			2	
48	ninsclas	c			6	
49	resp	c			2	
50	card	c			2	
51	neuro	c			2	

```
   52  gastr     c                      2
   53  renal     c                      2
   54  meta      c                       2
   55  hema      c                       2
   56  seps      c                      2
   57  trauma    c                        2
   58  ortho     c                      2
   60  urin1     s    0.000      9000.            3028
   61  race      c                      3
   62  income    c                        4

   Total  #cases w/   #missing
  #cases    miss. D  ord. vals   #X-var   #N-var   #F-var   #S-var
   5735         0      3443       13       0        0       20
  #P-var   #M-var   #B-var   #C-var   #I-var
     0        0        0       30       0
Number of cases used for training: 5735
Number of split variables: 50
Number of cases excluded due to 0 weight or missing D: 0

Pruning by v-fold cross-validation, with v = 10
Selected tree is based on mean of CV estimates
Number of SE's for pruned tree: 0.5000

Simple node models
Estimated priors
Unit misclassification costs
Univariate split highest priority
Interaction and linear splits 2nd and 3rd priorities
Split values for N and S variables based on exhaustive search
Maximum number of split levels: 15
Minimum node sample size: 57
Top-ranked variables and chi-squared values at root node
    1  0.3346E+03   cat1
    2  0.2728E+03   aps1
    3  0.2430E+03   crea1
    4  0.2402E+03   meanbp1
    5  0.2023E+03   pafi1
    6  0.1482E+03   neuro
    7  0.1247E+03   alb1
    8  0.1178E+03   card
    9  0.1077E+03   hema1
   10  0.9651E+02   wtkilo1
   11  0.9475E+02   resp
```

12	0.7634E+02	seps
13	0.7589E+02	cat2
14	0.6675E+02	bili1
15	0.6475E+02	dnr1
16	0.5661E+02	paco21
17	0.4780E+02	chrpulhx
18	0.4191E+02	hrt1
19	0.4063E+02	transhx
20	0.3675E+02	ninsclas
21	0.3393E+02	dementhx
22	0.3110E+02	ph1
23	0.2956E+02	resp1
24	0.2602E+02	psychhx
25	0.2088E+02	income
26	0.2022E+02	gastr
27	0.1927E+02	renal
28	0.1845E+02	cardiohx
29	0.1630E+02	urin1
30	0.1563E+02	sod1
31	0.1469E+02	age
32	0.1366E+02	malighx
33	0.1240E+02	wblc1
34	0.1206E+02	edu
35	0.1200E+02	ca
36	0.1168E+02	sex
37	0.8807E+01	immunhx
38	0.7795E+01	amihx
39	0.6616E+01	chfhx
40	0.6411E+01	gibledhx
41	0.5011E+01	hema
42	0.4201E+01	scoma1
43	0.3175E+01	liverhx
44	0.3055E+01	pot1
45	0.1861E+01	temp1
46	0.1376E+01	renalhx
47	0.1052E+01	meta
48	0.6357E+00	race

Size and CV mean cost and SE of subtrees:

Tree	#Tnodes	Mean Cost	SE(Mean)	BSE(Mean)	Median Cost	BSE(Median)
1	72	3.259E-01	6.189E-03	5.481E-03	3.235E-01	5.839E-03
2	71	3.259E-01	6.189E-03	5.481E-03	3.235E-01	5.839E-03
3	70	3.259E-01	6.189E-03	5.481E-03	3.235E-01	5.839E-03
4	69	3.259E-01	6.189E-03	5.481E-03	3.235E-01	5.839E-03

5	68	3.259E-01	6.189E-03	5.481E-03	3.235E-01	5.839E-03
6	67	3.259E-01	6.189E-03	5.481E-03	3.235E-01	5.839E-03
7	66	3.259E-01	6.189E-03	5.481E-03	3.235E-01	5.839E-03
8	64	3.259E-01	6.189E-03	5.481E-03	3.235E-01	5.839E-03
9	63	3.259E-01	6.189E-03	5.481E-03	3.235E-01	5.839E-03
10	62	3.259E-01	6.189E-03	5.481E-03	3.235E-01	5.839E-03
11	61	3.259E-01	6.189E-03	5.481E-03	3.235E-01	5.839E-03
12	60	3.259E-01	6.189E-03	5.481E-03	3.235E-01	5.839E-03
13	59	3.259E-01	6.189E-03	5.481E-03	3.235E-01	5.839E-03
14	58	3.259E-01	6.189E-03	5.481E-03	3.235E-01	5.839E-03
15	57	3.259E-01	6.189E-03	5.481E-03	3.235E-01	5.839E-03
16	56	3.259E-01	6.189E-03	5.481E-03	3.235E-01	5.839E-03
17	55	3.259E-01	6.189E-03	5.481E-03	3.235E-01	5.839E-03
18	54	3.259E-01	6.189E-03	5.481E-03	3.235E-01	5.839E-03
19	53	3.259E-01	6.189E-03	5.481E-03	3.235E-01	5.839E-03
20	52	3.259E-01	6.189E-03	5.481E-03	3.235E-01	5.839E-03
21	51	3.259E-01	6.189E-03	5.481E-03	3.235E-01	5.839E-03
22	50	3.259E-01	6.189E-03	5.481E-03	3.235E-01	5.839E-03
23	49	3.259E-01	6.189E-03	5.481E-03	3.235E-01	5.839E-03
24	48	3.259E-01	6.189E-03	5.481E-03	3.235E-01	5.839E-03
25	46	3.259E-01	6.189E-03	5.481E-03	3.235E-01	5.839E-03
26	45	3.259E-01	6.189E-03	5.481E-03	3.235E-01	5.839E-03
27	44	3.259E-01	6.189E-03	5.481E-03	3.235E-01	5.839E-03
28	43	3.259E-01	6.189E-03	5.481E-03	3.235E-01	5.839E-03
29	42	3.259E-01	6.189E-03	5.481E-03	3.235E-01	5.839E-03
30	41	3.259E-01	6.189E-03	5.481E-03	3.235E-01	5.839E-03
31	35	3.217E-01	6.168E-03	4.973E-03	3.208E-01	4.978E-03
32	31	3.217E-01	6.168E-03	4.723E-03	3.211E-01	3.516E-03
33	29	3.193E-01	6.156E-03	4.796E-03	3.202E-01	4.218E-03
34	25	3.180E-01	6.150E-03	4.585E-03	3.200E-01	3.392E-03
35	24	3.182E-01	6.151E-03	4.977E-03	3.188E-01	4.340E-03
36	22	3.177E-01	6.148E-03	4.882E-03	3.173E-01	4.888E-03
37	21	3.177E-01	6.148E-03	4.882E-03	3.173E-01	4.888E-03
38	19	3.179E-01	6.149E-03	5.334E-03	3.191E-01	5.314E-03
39+	17	3.172E-01	6.145E-03	3.358E-03	3.139E-01	3.477E-03
40++	12	3.167E-01	6.143E-03	2.739E-03	3.147E-01	2.121E-03
41**	10	3.175E-01	6.147E-03	2.273E-03	3.188E-01	3.560E-03
42	8	3.205E-01	6.162E-03	3.577E-03	3.217E-01	6.541E-03
43	6	3.229E-01	6.175E-03	3.773E-03	3.249E-01	7.965E-03
44	5	3.228E-01	6.174E-03	3.471E-03	3.249E-01	5.539E-03
45	3	3.325E-01	6.221E-03	3.956E-03	3.365E-01	6.220E-03
46	2	3.751E-01	6.393E-03	4.248E-03	3.801E-01	3.186E-03
47	1	3.808E-01	6.412E-03	2.782E-04	3.805E-01	4.832E-04

0-SE tree based on mean is marked with * and has 12 terminal nodes
0-SE tree based on median is marked with + and has 17 terminal nodes
Selected-SE tree based on mean using naive SE is marked with **
Selected-SE tree based on mean using bootstrap SE is marked with --
Selected-SE tree based on median and bootstrap SE is marked with ++
** tree same as -- tree
* tree same as ++ tree

Following tree is based on mean CV with naive SE estimate (**).

Structure of final tree. Each terminal node is marked with a T.

Node cost is node misclassification cost divided by number of training cases

Node label	Total cases	Train cases	Predicted class	Node cost	Split variables	Interacting variable
1	5735	5735	NoRHC	3.808E-01	cat1	
2	1683	1683	RHC	4.599E-01	meanbp1	
4	1117	1117	RHC	3.796E-01	pafi1	
8T	655	655	RHC	3.038E-01	resp1	
9	462	462	RHC	4.870E-01	ninsclas	
18T	244	244	RHC	3.730E-01	bili1	
19T	218	218	NoRHC	3.853E-01	card	
5T	566	566	NoRHC	3.816E-01	alb1	
3	4052	4052	NoRHC	3.147E-01	pafi1	
6	1292	1292	NoRHC	4.837E-01	resp	
12	581	581	RHC	4.200E-01	dnr1	
24	515	515	RHC	3.903E-01	cat1	
48T	438	438	RHC	3.447E-01	meanbp1	
49T	77	77	NoRHC	3.506E-01	-	
25T	66	66	NoRHC	3.485E-01	-	
13	711	711	NoRHC	4.051E-01	seps	
26T	110	110	RHC	3.636E-01	-	
27T	601	601	NoRHC	3.627E-01	aps1	
7T	2760	2760	NoRHC	2.355E-01	aps1	

Number of terminal nodes of final tree: 10
Total number of nodes of final tree: 19
Second best split variable (based on curvature test) at root node is aps1

Classification tree:
For categorical variable splits, values not in training data go to the right

Node 1: cat1 = "CHF", "MOSF w/Sepsis"
  Node 2: meanbp1 <= 68.500000 or NA

```
    Node 4: pafi1 <= 266.15625
      Node 8: RHC
    Node 4: pafi1 > 266.15625 or NA
      Node 9: ninsclas = "No insurance", "Private", "Private & Medicare"
        Node 18: RHC
      Node 9: ninsclas /= "No insurance", "Private", "Private & Medicare"
        Node 19: NoRHC
  Node 2: meanbp1 > 68.500000
    Node 5: NoRHC
 Node 1: cat1 /= "CHF", "MOSF w/Sepsis"
  Node 3: pafi1 <= 142.35938
    Node 6: resp = "No"
      Node 12: dnr1 = "No"
        Node 24: cat1 = "ARF", "Lung Cancer", "MOSF w/Malignancy"
          Node 48: RHC
        Node 24: cat1 /= "ARF", "Lung Cancer", "MOSF w/Malignancy"
          Node 49: NoRHC
      Node 12: dnr1 /= "No"
        Node 25: NoRHC
    Node 6: resp /= "No"
      Node 13: seps = "Yes"
        Node 26: RHC
      Node 13: seps /= "Yes"
        Node 27: NoRHC
  Node 3: pafi1 > 142.35938 or NA
    Node 7: NoRHC

 ************************************************************

 Predictor means below are means of cases with no missing values.

Node 1: Intermediate node
 A case goes into Node 2 if cat1 = "CHF", "MOSF w/Sepsis"
 cat1 mode = "ARF"
 Class      Number   Posterior
 NoRHC        3551  0.6192E+00
 RHC          2184  0.3808E+00
 Number of training cases misclassified = 2184
 Predicted class is NoRHC
 ---------------------------
Node 2: Intermediate node
 A case goes into Node 4 if meanbp1 <= 68.500000 or NA
 meanbp1 mean = 72.674985
 Class      Number   Posterior
```

NoRHC 774 0.4599E+00
RHC 909 0.5401E+00
Number of training cases misclassified = 774
Predicted class is RHC

Node 4: Intermediate node
A case goes into Node 8 if pafi1 <= 266.15625
pafi1 mean = 241.37331
Class Number Posterior
NoRHC 424 0.3796E+00
RHC 693 0.6204E+00
Number of training cases misclassified = 424
Predicted class is RHC

Node 8: Terminal node
Class Number Posterior
NoRHC 199 0.3038E+00
RHC 456 0.6962E+00
Number of training cases misclassified = 199
Predicted class is RHC

Node 9: Intermediate node
A case goes into Node 18 if ninsclas = "No insurance", "Private",
"Private & Medicare"
ninsclas mode = "Private"
Class Number Posterior
NoRHC 225 0.4870E+00
RHC 237 0.5130E+00
Number of training cases misclassified = 225
Predicted class is RHC

Node 18: Terminal node
Class Number Posterior
NoRHC 91 0.3730E+00
RHC 153 0.6270E+00
Number of training cases misclassified = 91
Predicted class is RHC

Node 19: Terminal node
Class Number Posterior
NoRHC 134 0.6147E+00
RHC 84 0.3853E+00
Number of training cases misclassified = 84
Predicted class is NoRHC

Node 5: Terminal node
Class Number Posterior
NoRHC 350 0.6184E+00
RHC 216 0.3816E+00
Number of training cases misclassified = 216
Predicted class is NoRHC

Node 3: Intermediate node
A case goes into Node 6 if pafi1 <= 142.35938
pafi1 mean = 211.08630
Class Number Posterior
NoRHC 2777 0.6853E+00
RHC 1275 0.3147E+00
Number of training cases misclassified = 1275
Predicted class is NoRHC

Node 6: Intermediate node
A case goes into Node 12 if resp = "No"
resp mode = "Yes"
Class Number Posterior
NoRHC 667 0.5163E+00
RHC 625 0.4837E+00
Number of training cases misclassified = 625
Predicted class is NoRHC

Node 12: Intermediate node
A case goes into Node 24 if dnr1 = "No"
dnr1 mode = "No"
Class Number Posterior
NoRHC 244 0.4200E+00
RHC 337 0.5800E+00
Number of training cases misclassified = 244
Predicted class is RHC

Node 24: Intermediate node
A case goes into Node 48 if cat1 = "ARF", "Lung Cancer", "MOSF w/Malignancy"
cat1 mode = "ARF"
Class Number Posterior
NoRHC 201 0.3903E+00
RHC 314 0.6097E+00
Number of training cases misclassified = 201
Predicted class is RHC

Node 48: Terminal node
Class Number Posterior
NoRHC 151 0.3447E+00
RHC 287 0.6553E+00
Number of training cases misclassified = 151
Predicted class is RHC

Node 49: Terminal node
Class Number Posterior
NoRHC 50 0.6494E+00
RHC 27 0.3506E+00
Number of training cases misclassified = 27
Predicted class is NoRHC

Node 25: Terminal node
Class Number Posterior
NoRHC 43 0.6515E+00
RHC 23 0.3485E+00
Number of training cases misclassified = 23
Predicted class is NoRHC

Node 13: Intermediate node
A case goes into Node 26 if seps = "Yes"
seps mode = "No"
Class Number Posterior
NoRHC 423 0.5949E+00
RHC 288 0.4051E+00
Number of training cases misclassified = 288
Predicted class is NoRHC

Node 26: Terminal node
Class Number Posterior
NoRHC 40 0.3636E+00
RHC 70 0.6364E+00
Number of training cases misclassified = 40
Predicted class is RHC

Node 27: Terminal node
Class Number Posterior
NoRHC 383 0.6373E+00
RHC 218 0.3627E+00
Number of training cases misclassified = 218
Predicted class is NoRHC

Node 7: Terminal node
Class Number Posterior
NoRHC 2110 0.7645E+00
RHC 650 0.2355E+00
Number of training cases misclassified = 650
Predicted class is NoRHC

Classification matrix for training sample:

Predicted class	True class NoRHC	RHC
NoRHC	3070	1218
RHC	481	966
Total	3551	2184

Number of cases used for tree construction: 5735
Number misclassified: 1699
Resubstitution estimate of mean misclassification cost: 0.29625109

Observed and fitted values are stored in classpred.txt
LaTeX code for tree is in class.tex
R code is stored in class.r
Elapsed time in seconds: 13.965

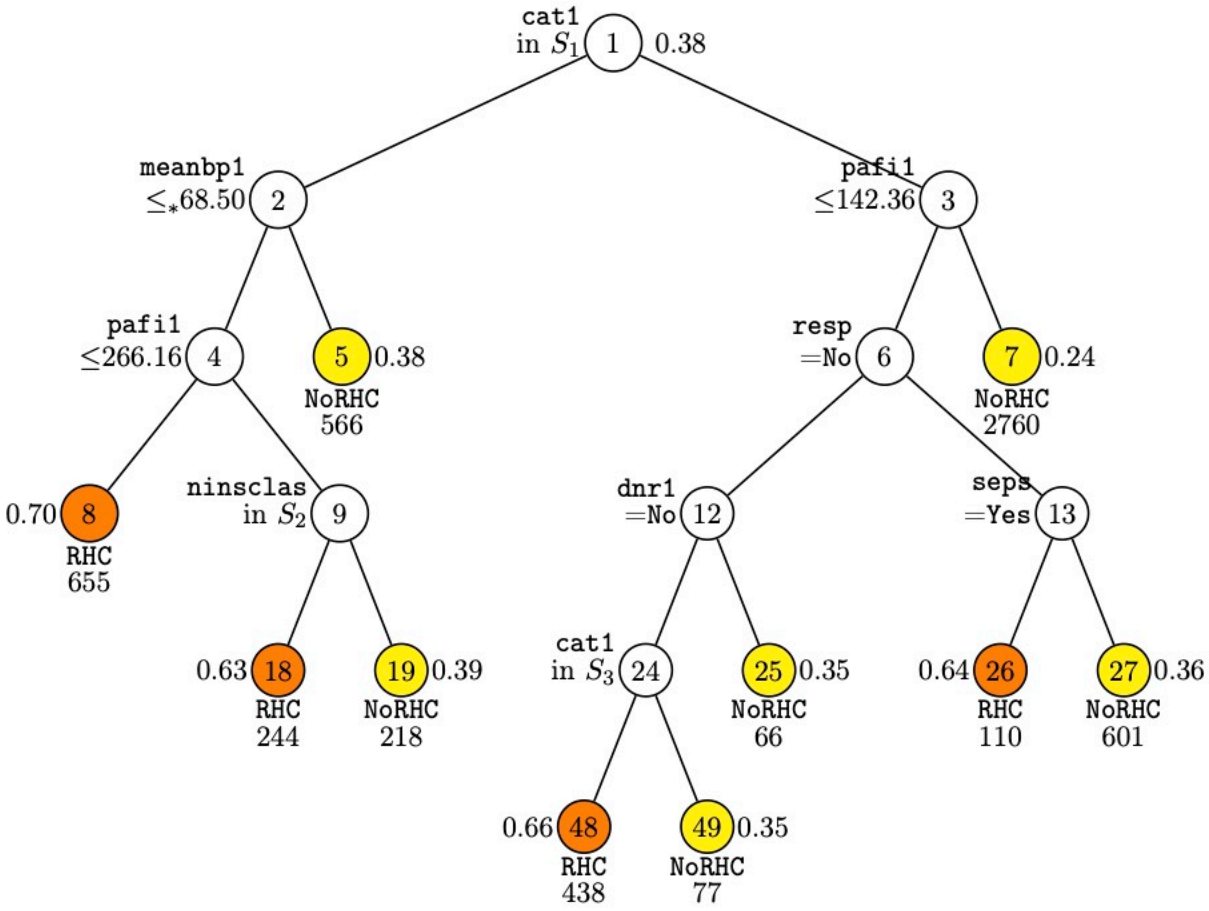


Figure 1: GUIDE v.36.2 0.50-SE classification tree for predicting `swang1` using estimated priors and unit misclassification costs. Tree constructed with 5735 observations. Maximum number of split levels is 15 and minimum node sample size is 57. At each split, an observation goes to the left branch if and only if the condition is satisfied. Symbol '$\leq_*$' stands for '$\leq$ or missing'. Set S_1 = {`CHF`, `MOSF w/Sepsis`}. Set S_2 = {`No insurance`, `Private`, `Private & Medicare`}. Set S_3 = {`ARF`, `Lung Cancer`, `MOSF w/Malignancy`}. Predicted classes and sample sizes printed below terminal nodes; class sample proportion for `swang1` = `RHC` beside nodes. Second best split variable at root node is `aps1`.

Homework #1 Extra Writing

Purpose:

The goal of this homework was to get used to the guide, and create a tree diagram with the program.

Analysis:

Summary table:

Total #cases	#cases w/ miss. D	#missing ord. vals	#X-var	#N-var	#F-var	#S-var
5735	0	3443	13	0	0	20

#P-var	#M-var	#B-var	#C-var	#I-var
0	0	0	30	0

This table shows what is in a classout.txt. According to this file, there are total of 5735 records of patients. Among those 5735, there were 3443 datasets that contained one or more missing values.

Some trivial analysis:

I have found some data analysis through this classout.txt.

```
Minimum node sample size: 57
```

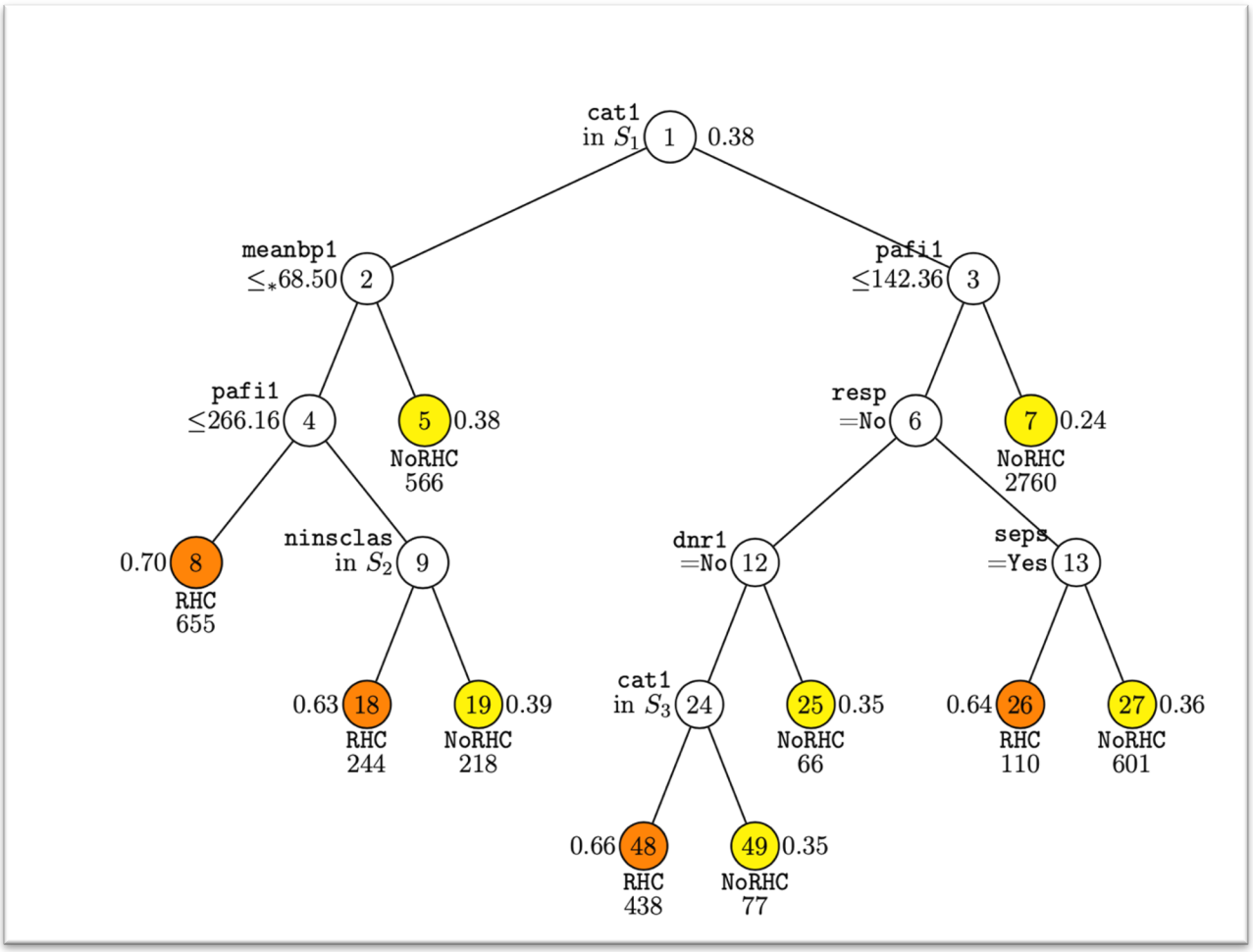
This line shows that the smallest node size in this data is 57.

```
Size and CV mean cost and SE of subtrees:
```

Tree	#Tnodes	Mean Cost	SE(Mean)	BSE(Mean)	Median Cost	BSE(Median)
1	72	3.259E-01	6.189E-03	5.481E-03	3.235E-01	5.839E-03
2	71	3.259E-01	6.189E-03	5.481E-03	3.235E-01	5.839E-03
:						
38	19	3.179E-01	6.149E-03	5.334E-03	3.191E-01	5.314E-03
39+	17	3.172E-01	6.145E-03	3.358E-03	3.139E-01	3.477E-03
40++	12	3.167E-01	6.143E-03	2.739E-03	3.147E-01	2.121E-03

This analysis shows that if we were to create the largest tree diagram, it will contain 72 possible nodes.

Classification tree:



This is a final classification tree diagram. If the person's mean blood pressure Day1 is less than or equal to OR the value does not exist, you would move to down left (to 4). * after the less than equal to sign means that the value missing. If there are no * value next to the scale function, you would want to go right down since that means that the value is not satisfied.

Therefore, if you want to predict the person with a person who has 60 mean blood pressure in day 1 and has 200 PaO2, they have 0.70% of benefiting from cardiac function by right heart catherization (RHC).