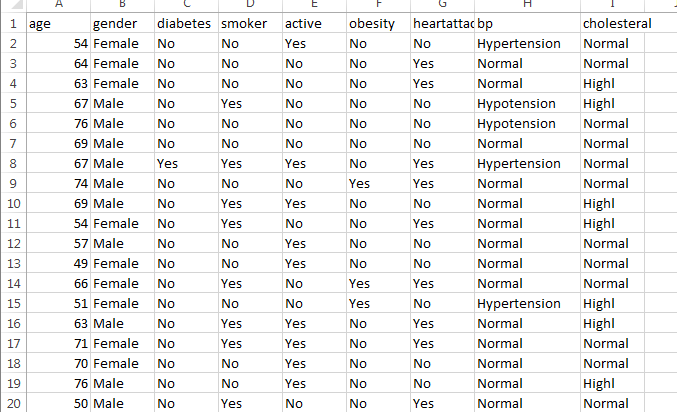
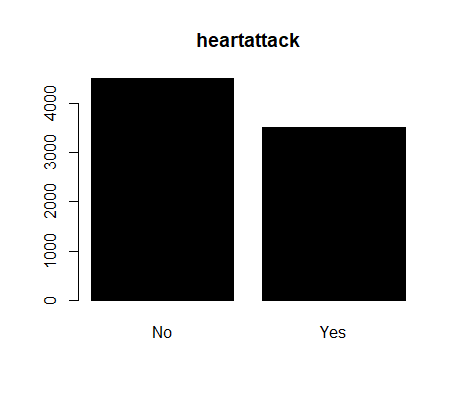
Exercise 6

This patients’ (heart attack diagnosis) dataset (Patient\_data.xlsx) was retrieved from the Internet. But I forgot about its source. We should give the provider/publisher credits for their efforts (if you find its source, please let me know).

The file contains 7998 records. The following screenshot shows you what the dataset actually contains.



1. (50 points) Explore the data set, then use C5.0 to model this classification problem (no partition at this step).
2. ackages
3. > dim(Patient\_Data)
4. [1] 7998 9
5. >
6. > barplot(table(Patient\_Data$heartattack\_s),main="heartattack", col="black")



boxplot(Patient\_Data$age ~ Patient\_Data$heartattack\_s,main="Age",ylab="Age",xlab="Heart disease")



missingd= sum(is.na(Patient\_Data))

>

> Patient\_Data <- na.omit(Patient\_Data)

library(caret)

> set.seed(266)

> install.packages('C50')

> library(C50)

> summary(Patient\_Data)

age gender diabetes

Min. :45.00 Length:7998 Length:7998

1st Qu.:55.00 Class :character Class :character

Median :61.00 Mode :character Mode :character

Mean :61.85

3rd Qu.:68.00

Max. :89.00

smoker active obesity

Length:7998 Length:7998 Length:7998

Class :character Class :character Class :character

Mode :character Mode :character Mode :character

heartattack\_s bp cholesteral

No :4491 Hypertension:2011 Highl :3064

Yes:3507 Hypotension : 985 Normal:4934

Normal :5002

|  |
| --- |
| > build\_tree<-C5.0(heartattack\_s~.,data=Patient\_Data)  > summary(build\_tree)  Call:  C5.0.formula(formula = heartattack\_s ~ ., data = Patient\_Data)  C5.0 [Release 2.07 GPL Edition] Mon Oct 15 03:43:48 2018  -------------------------------  Class specified by attribute `outcome'  Read 7998 cases (9 attributes) from undefined.data  Decision tree:  diabetes = Yes: Yes (542/97)  diabetes = No:  :...smoker = Yes:  :...age <= 59:  : :...bp = Hypertension: Yes (158/43)  : : bp in {Hypotension,Normal}:  : : :...active = Yes:  : : :...obesity = No: No (262/75)  : : : obesity = Yes: Yes (42/17)  : : active = No:  : : :...cholesteral = Highl: Yes (79/27)  : : cholesteral = Normal:  : : :...gender = Female: Yes (60/26)  : : gender = Male: No (64/26)  : age > 59:  : :...obesity = Yes: Yes (229/26)  : obesity = No:  : :...cholesteral = Highl: Yes (272/54)  : cholesteral = Normal:  : :...active = Yes:  : :...age <= 66: No (90/34)  : : age > 66: Yes (104/39)  : active = No:  : :...age > 60: Yes (154/36)  : age <= 60:  : :...bp = Hypertension: Yes (3)  : bp in {Hypotension,Normal}: No (9/1)  smoker = No:  :...age <= 60:  :...bp in {Hypotension,Normal}: No (2202/445)  : bp = Hypertension:  : :...obesity = Yes:  : :...age <= 51: No (45/16)  : : age > 51: Yes (121/44)  : obesity = No:  : :...cholesteral = Normal: No (293/64)  : cholesteral = Highl:  : :...active = Yes: No (76/29)  : active = No:  : :...age <= 53: No (51/19)  : age > 53: Yes (65/24)  age > 60:  :...active = Yes:  :...bp = Hypertension:  : :...obesity = Yes: Yes (68/21)  : : obesity = No:  : : :...cholesteral = Highl: Yes (78/32)  : : cholesteral = Normal: No (138/57)  : bp in {Hypotension,Normal}:  : :...obesity = No: No (1083/285)  : obesity = Yes:  : :...cholesteral = Normal:  : :...age <= 73: No (104/31)  : : age > 73: Yes (26/10)  : cholesteral = Highl:  : :...bp = Hypotension: Yes (16/2)  : bp = Normal:  : :...gender = Female: Yes (35/12)  : gender = Male:  : :...age <= 68: No (15/2)  : age > 68: Yes (6)  active = No:  :...obesity = Yes: Yes (380/93)  obesity = No:  :...bp = Hypertension: Yes (302/100)  bp in {Hypotension,Normal}:  :...cholesteral = Highl: Yes (295/124)  cholesteral = Normal:  :...age <= 72: No (383/136)  age > 72:  :...bp = Hypotension: Yes (26/8)  bp = Normal:  :...gender = Female: Yes (62/28)  gender = Male: No (60/27)  Evaluation on training data (7998 cases):  Decision Tree  ----------------  Size Errors  38 2110(26.4%) <<  (a) (b) <-classified as  ---- ----  3628 863 (a): class No  1247 2260 (b): class Yes  Attribute usage:  100.00% diabetes  93.22% age  93.22% smoker  77.86% bp  61.18% obesity  51.71% active  32.06% cholesteral  3.78% gender  Time: 0.0 secs |
|  |
| |  | | --- | | > | |

1. (30 points) Comparing with a CART model, is there any difference?

There are about 38 nodes for C50, node error rate =26.4%

For CART there are 25 nodes, node error =43.85%

> library(rpart)

> carttree<-rpart(heartattack\_s~.,method="class",data=Patient\_Data)

> printcp(carttree)

Classification tree:

rpart(formula = heartattack\_s ~ ., data = Patient\_Data, method = "class")

Variables actually used in tree construction:

[1] active age bp diabetes smoker

Root node error: 3507/7998 = 0.43848

n= 7998

CP nsplit rel error xerror xstd

1 0.111491 0 1.00000 1.00000 0.012654

2 0.085258 1 0.88851 0.89050 0.012441

3 0.043342 2 0.80325 0.81123 0.012208

4 0.018249 3 0.75991 0.76048 0.012022

5 0.014542 4 0.74166 0.74194 0.011947

6 0.010000 7 0.69803 0.71029 0.011809

> plotcp(carttree)

> summary(carttree)

Call:

rpart(formula = heartattack\_s ~ ., data = Patient\_Data, method = "class")

n= 7998

CP nsplit rel error xerror xstd

1 0.11149130 0 1.0000000 1.0000000 0.01265357

2 0.08525806 1 0.8885087 0.8905047 0.01244076

3 0.04334189 2 0.8032506 0.8112347 0.01220801

4 0.01824922 3 0.7599088 0.7604790 0.01202235

5 0.01454234 4 0.7416595 0.7419447 0.01194712

6 0.01000000 7 0.6980325 0.7102937 0.01180913

Variable importance

age smoker active diabetes bp obesity

33 24 18 16 6 2

Node number 1: 7998 observations, complexity param=0.1114913

predicted class=No expected loss=0.4384846 P(node) =1

class counts: 4491 3507

probabilities: 0.562 0.438

left son=2 (3793 obs) right son=3 (4205 obs)

Primary splits:

age < 60.5 to the left, improve=206.8729, (0 missing)

smoker splits as LR, improve=197.3318, (0 missing)

active splits as RL, improve=188.0524, (0 missing)

bp splits as RLL, improve=180.1255, (0 missing)

diabetes splits as LR, improve=170.1680, (0 missing)

Node number 2: 3793 observations, complexity param=0.01454234

predicted class=No expected loss=0.3187451 P(node) =0.4742436

class counts: 2584 1209

probabilities: 0.681 0.319

left son=4 (3022 obs) right son=5 (771 obs)

Primary splits:

smoker splits as LR, improve=83.60781, (0 missing)

bp splits as RLL, improve=71.29711, (0 missing)

diabetes splits as LR, improve=71.17877, (0 missing)

active splits as RL, improve=59.03512, (0 missing)

obesity splits as LR, improve=51.13655, (0 missing)

Node number 3: 4205 observations, complexity param=0.08525806

predicted class=Yes expected loss=0.4535077 P(node) =0.5257564

class counts: 1907 2298

probabilities: 0.454 0.546

left son=6 (2089 obs) right son=7 (2116 obs)

Primary splits:

active splits as RL, improve=115.71960, (0 missing)

smoker splits as LR, improve=110.93340, (0 missing)

bp splits as RLL, improve=104.50390, (0 missing)

obesity splits as LR, improve= 99.09606, (0 missing)

diabetes splits as LR, improve= 85.30530, (0 missing)

Surrogate splits:

bp splits as RLL, agree=0.555, adj=0.105, (0 split)

obesity splits as LR, agree=0.549, adj=0.092, (0 split)

age < 65.5 to the left, agree=0.521, adj=0.036, (0 split)

diabetes splits as LR, agree=0.512, adj=0.019, (0 split)

gender splits as RL, agree=0.509, adj=0.011, (0 split)

Node number 4: 3022 observations, complexity param=0.01454234

predicted class=No expected loss=0.2657181 P(node) =0.3778445

class counts: 2219 803

probabilities: 0.734 0.266

left son=8 (2853 obs) right son=9 (169 obs)

Primary splits:

diabetes splits as LR, improve=56.42853, (0 missing)

bp splits as RLL, improve=46.25869, (0 missing)

active splits as RL, improve=40.62085, (0 missing)

cholesteral splits as RL, improve=31.72896, (0 missing)

obesity splits as LR, improve=31.02963, (0 missing)

Node number 5: 771 observations, complexity param=0.01454234

predicted class=Yes expected loss=0.4734112 P(node) =0.0963991

class counts: 365 406

probabilities: 0.473 0.527

left son=10 (579 obs) right son=11 (192 obs)

Primary splits:

bp splits as RLL, improve=26.72597, (0 missing)

active splits as RL, improve=22.07201, (0 missing)

obesity splits as LR, improve=18.57453, (0 missing)

cholesteral splits as RL, improve=14.88015, (0 missing)

diabetes splits as LR, improve=12.89033, (0 missing)

Node number 6: 2089 observations, complexity param=0.04334189

predicted class=No expected loss=0.4284347 P(node) =0.2611903

class counts: 1194 895

probabilities: 0.572 0.428

left son=12 (1669 obs) right son=13 (420 obs)

Primary splits:

smoker splits as LR, improve=67.04172, (0 missing)

diabetes splits as LR, improve=45.94836, (0 missing)

bp splits as RLL, improve=41.62284, (0 missing)

obesity splits as LR, improve=35.40730, (0 missing)

cholesteral splits as RL, improve=33.77798, (0 missing)

Node number 7: 2116 observations

predicted class=Yes expected loss=0.3369565 P(node) =0.2645661

class counts: 713 1403

probabilities: 0.337 0.663

Node number 8: 2853 observations

predicted class=No expected loss=0.2422012 P(node) =0.3567142

class counts: 2162 691

probabilities: 0.758 0.242

Node number 9: 169 observations

predicted class=Yes expected loss=0.3372781 P(node) =0.02113028

class counts: 57 112

probabilities: 0.337 0.663

Node number 10: 579 observations

predicted class=No expected loss=0.4507772 P(node) =0.0723931

class counts: 318 261

probabilities: 0.549 0.451

Node number 11: 192 observations

predicted class=Yes expected loss=0.2447917 P(node) =0.024006

class counts: 47 145

probabilities: 0.245 0.755

Node number 12: 1669 observations, complexity param=0.01824922

predicted class=No expected loss=0.3648892 P(node) =0.2086772

class counts: 1060 609

probabilities: 0.635 0.365

left son=24 (1569 obs) right son=25 (100 obs)

Primary splits:

diabetes splits as LR, improve=44.06540, (0 missing)

bp splits as RLL, improve=38.34859, (0 missing)

obesity splits as LR, improve=26.38097, (0 missing)

cholesteral splits as RL, improve=22.65515, (0 missing)

age < 71.5 to the left, improve=16.19905, (0 missing)

Node number 13: 420 observations

predicted class=Yes expected loss=0.3190476 P(node) =0.05251313

class counts: 134 286

probabilities: 0.319 0.681

Node number 24: 1569 observations

predicted class=No expected loss=0.3358827 P(node) =0.196174

class counts: 1042 527

probabilities: 0.664 0.336

Node number 25: 100 observations

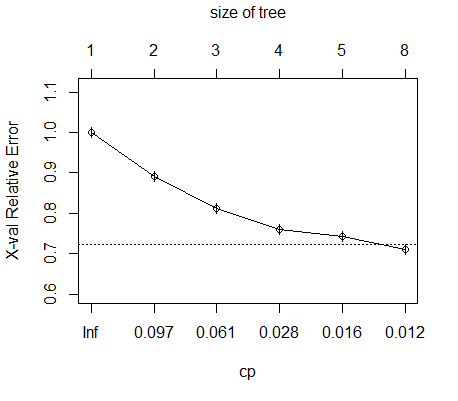
predicted class=Yes expected loss=0.18 P(node) =0.01250313

class counts: 18 82

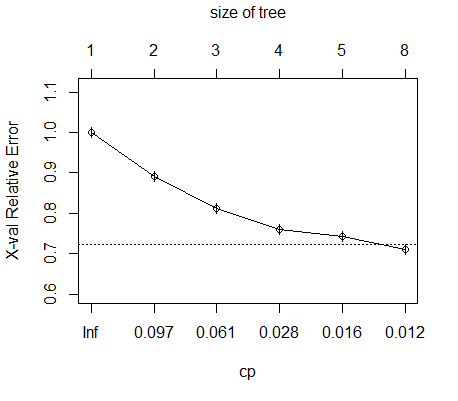
probabilities: 0.180 0.820

|  |
| --- |
| prunTree<- prune(carttree, cp=Patient\_Data$cptable[which.min(carttree$cptable[,"xerror"]),"CP"])  > summary(prunTree)  Call:  rpart(formula = heartattack\_s ~ ., data = Patient\_Data, method = "class")  n= 7998  CP nsplit rel error xerror xstd  1 0.11149130 0 1.0000000 1.0000000 0.01265357  2 0.08525806 1 0.8885087 0.8905047 0.01244076  3 0.04334189 2 0.8032506 0.8112347 0.01220801  4 0.01824922 3 0.7599088 0.7604790 0.01202235  5 0.01454234 4 0.7416595 0.7419447 0.01194712  6 0.01000000 7 0.6980325 0.7102937 0.01180913  Variable importance  age smoker active diabetes bp obesity  33 24 18 16 6 2  Node number 1: 7998 observations, complexity param=0.1114913  predicted class=No expected loss=0.4384846 P(node) =1  class counts: 4491 3507  probabilities: 0.562 0.438  left son=2 (3793 obs) right son=3 (4205 obs)  Primary splits:  age < 60.5 to the left, improve=206.8729, (0 missing)  smoker splits as LR, improve=197.3318, (0 missing)  active splits as RL, improve=188.0524, (0 missing)  bp splits as RLL, improve=180.1255, (0 missing)  diabetes splits as LR, improve=170.1680, (0 missing)  Node number 2: 3793 observations, complexity param=0.01454234  predicted class=No expected loss=0.3187451 P(node) =0.4742436  class counts: 2584 1209  probabilities: 0.681 0.319  left son=4 (3022 obs) right son=5 (771 obs)  Primary splits:  smoker splits as LR, improve=83.60781, (0 missing)  bp splits as RLL, improve=71.29711, (0 missing)  diabetes splits as LR, improve=71.17877, (0 missing)  active splits as RL, improve=59.03512, (0 missing)  obesity splits as LR, improve=51.13655, (0 missing)  Node number 3: 4205 observations, complexity param=0.08525806  predicted class=Yes expected loss=0.4535077 P(node) =0.5257564  class counts: 1907 2298  probabilities: 0.454 0.546  left son=6 (2089 obs) right son=7 (2116 obs)  Primary splits:  active splits as RL, improve=115.71960, (0 missing)  smoker splits as LR, improve=110.93340, (0 missing)  bp splits as RLL, improve=104.50390, (0 missing)  obesity splits as LR, improve= 99.09606, (0 missing)  diabetes splits as LR, improve= 85.30530, (0 missing)  Surrogate splits:  bp splits as RLL, agree=0.555, adj=0.105, (0 split)  obesity splits as LR, agree=0.549, adj=0.092, (0 split)  age < 65.5 to the left, agree=0.521, adj=0.036, (0 split)  diabetes splits as LR, agree=0.512, adj=0.019, (0 split)  gender splits as RL, agree=0.509, adj=0.011, (0 split)  Node number 4: 3022 observations, complexity param=0.01454234  predicted class=No expected loss=0.2657181 P(node) =0.3778445  class counts: 2219 803  probabilities: 0.734 0.266  left son=8 (2853 obs) right son=9 (169 obs)  Primary splits:  diabetes splits as LR, improve=56.42853, (0 missing)  bp splits as RLL, improve=46.25869, (0 missing)  active splits as RL, improve=40.62085, (0 missing)  cholesteral splits as RL, improve=31.72896, (0 missing)  obesity splits as LR, improve=31.02963, (0 missing)  Node number 5: 771 observations, complexity param=0.01454234  predicted class=Yes expected loss=0.4734112 P(node) =0.0963991  class counts: 365 406  probabilities: 0.473 0.527  left son=10 (579 obs) right son=11 (192 obs)  Primary splits:  bp splits as RLL, improve=26.72597, (0 missing)  active splits as RL, improve=22.07201, (0 missing)  obesity splits as LR, improve=18.57453, (0 missing)  cholesteral splits as RL, improve=14.88015, (0 missing)  diabetes splits as LR, improve=12.89033, (0 missing)  Node number 6: 2089 observations, complexity param=0.04334189  predicted class=No expected loss=0.4284347 P(node) =0.2611903  class counts: 1194 895  probabilities: 0.572 0.428  left son=12 (1669 obs) right son=13 (420 obs)  Primary splits:  smoker splits as LR, improve=67.04172, (0 missing)  diabetes splits as LR, improve=45.94836, (0 missing)  bp splits as RLL, improve=41.62284, (0 missing)  obesity splits as LR, improve=35.40730, (0 missing)  cholesteral splits as RL, improve=33.77798, (0 missing)  Node number 7: 2116 observations  predicted class=Yes expected loss=0.3369565 P(node) =0.2645661  class counts: 713 1403  probabilities: 0.337 0.663  Node number 8: 2853 observations  predicted class=No expected loss=0.2422012 P(node) =0.3567142  class counts: 2162 691  probabilities: 0.758 0.242  Node number 9: 169 observations  predicted class=Yes expected loss=0.3372781 P(node) =0.02113028  class counts: 57 112  probabilities: 0.337 0.663  Node number 10: 579 observations  predicted class=No expected loss=0.4507772 P(node) =0.0723931  class counts: 318 261  probabilities: 0.549 0.451  Node number 11: 192 observations  predicted class=Yes expected loss=0.2447917 P(node) =0.024006  class counts: 47 145  probabilities: 0.245 0.755  Node number 12: 1669 observations, complexity param=0.01824922  predicted class=No expected loss=0.3648892 P(node) =0.2086772  class counts: 1060 609  probabilities: 0.635 0.365  left son=24 (1569 obs) right son=25 (100 obs)  Primary splits:  diabetes splits as LR, improve=44.06540, (0 missing)  bp splits as RLL, improve=38.34859, (0 missing)  obesity splits as LR, improve=26.38097, (0 missing)  cholesteral splits as RL, improve=22.65515, (0 missing)  age < 71.5 to the left, improve=16.19905, (0 missing)  Node number 13: 420 observations  predicted class=Yes expected loss=0.3190476 P(node) =0.05251313  class counts: 134 286  probabilities: 0.319 0.681  Node number 24: 1569 observations  predicted class=No expected loss=0.3358827 P(node) =0.196174  class counts: 1042 527  probabilities: 0.664 0.336  Node number 25: 100 observations  predicted class=Yes expected loss=0.18 P(node) =0.01250313  class counts: 18 82  probabilities: 0.180 0.820  > plot(prunTree)  > plotcp(prunTree)  > printcp(prunTree)  Classification tree:  rpart(formula = heartattack\_s ~ ., data = Patient\_Data, method = "class")  Variables actually used in tree construction:  [1] active age bp diabetes smoker  Root node error: 3507/7998 = 0.43848  n= 7998  CP nsplit rel error xerror xstd  1 0.111491 0 1.00000 1.00000 0.012654  2 0.085258 1 0.88851 0.89050 0.012441  3 0.043342 2 0.80325 0.81123 0.012208  4 0.018249 3 0.75991 0.76048 0.012022  5 0.014542 4 0.74166 0.74194 0.011947  6 0.010000 7 0.69803 0.71029 0.011809 |
|  |
| |  | | --- | | > | |

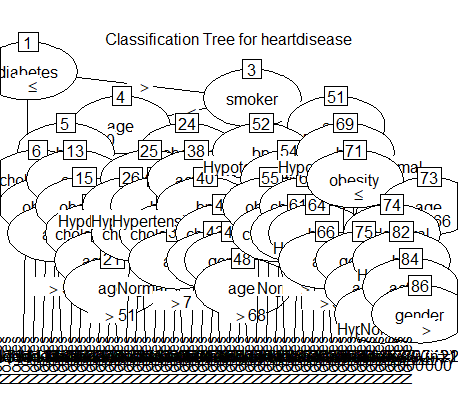
Original Tree cross validation results:



Prune Tree cross validation results



1. (20 points) Using partition ratio 80:20 to rerun you C5.0 model. Do you have the similar accuracy for the test data? If not, how will you improve your model?
2. > library(caret)
3. > set.seed(266)
4. > > SelectedData\_sampling\_vector <- createDataPartition(Patient\_Data$heartattack\_s, p=0.80, list = FALSE)
5. > data\_train<-Patient\_Data[SelectedData\_sampling\_vector,]
6. > data\_test<-Patient\_Data[-SelectedData\_sampling\_vector,]
7. > library(C50)
8. > tree\_C51<-C5.0(heartattack\_s ~ ., data=data\_train)
9. >
10. > summary(tree\_C51)
11. Call:
12. C5.0.formula(formula = heartattack\_s ~ ., data = data\_train)
13. C5.0 [Release 2.07 GPL Edition] Mon Oct 15 05:18:48 2018
14. -------------------------------
15. Class specified by attribute `outcome'
16. Read 6399 cases (9 attributes) from undefined.data
17. Decision tree:
18. diabetes = Yes: Yes (433/83)
19. diabetes = No:
20. :...smoker = No:
21. :...age <= 60:
22. : :...bp in {Hypotension,Normal}:
23. : : :...cholesteral = Normal: No (1093/164)
24. : : : cholesteral = Highl:
25. : : : :...obesity = No: No (523/126)
26. : : : obesity = Yes:
27. : : : :...age <= 56: No (95/29)
28. : : : age > 56: Yes (44/15)
29. : : bp = Hypertension:
30. : : :...age <= 49: No (83/13)
31. : : age > 49:
32. : : :...obesity = Yes: Yes (111/45)
33. : : obesity = No:
34. : : :...cholesteral = Normal: No (189/45)
35. : : cholesteral = Highl:
36. : : :...active = No: Yes (79/34)
37. : : active = Yes:
38. : : :...age <= 51: Yes (8/1)
39. : : age > 51: No (43/15)
40. : age > 60:
41. : :...obesity = No:
42. : :...active = Yes:
43. : : :...bp in {Hypotension,Normal}: No (857/225)
44. : : : bp = Hypertension:
45. : : : :...cholesteral = Highl: Yes (57/23)
46. : : : cholesteral = Normal: No (107/43)
47. : : active = No:
48. : : :...bp = Hypertension: Yes (238/78)
49. : : bp in {Hypotension,Normal}:
50. : : :...cholesteral = Highl: Yes (237/98)
51. : : cholesteral = Normal:
52. : : :...age <= 75: No (352/127)
53. : : age > 75: Yes (73/32)
54. : obesity = Yes:
55. : :...active = No: Yes (313/77)
56. : active = Yes:
57. : :...bp in {Hypertension,Hypotension}: Yes (95/31)
58. : bp = Normal:
59. : :...cholesteral = Normal:
60. : :...age <= 76: No (75/21)
61. : : age > 76: Yes (9/2)
62. : cholesteral = Highl:
63. : :...gender = Female: Yes (29/10)
64. : gender = Male:
65. : :...age <= 68: No (13/2)
66. : age > 68: Yes (6)
67. smoker = Yes:
68. :...age <= 59:
69. :...bp = Hypertension: Yes (132/35)
70. : bp in {Hypotension,Normal}:
71. : :...active = Yes:
72. : :...obesity = No: No (201/56)
73. : : obesity = Yes:
74. : : :...cholesteral = Highl: Yes (11/2)
75. : : cholesteral = Normal: No (22/9)
76. : active = No:
77. : :...cholesteral = Normal:
78. : :...gender = Female: Yes (49/21)
79. : : gender = Male: No (53/22)
80. : cholesteral = Highl:
81. : :...bp = Hypotension: Yes (16/3)
82. : bp = Normal:
83. : :...age <= 52: No (18/6)
84. : age > 52: Yes (32/8)
85. age > 59:
86. :...bp = Hypertension: Yes (167/16)
87. bp in {Hypotension,Normal}:
88. :...obesity = Yes: Yes (121/19)
89. obesity = No:
90. :...age > 66: Yes (227/52)
91. age <= 66:
92. :...cholesteral = Highl:
93. :...gender = Female: Yes (52/13)
94. : gender = Male:
95. : :...active = No: No (12/4)
96. : active = Yes:
97. : :...bp = Hypotension: No (7/2)
98. : bp = Normal: Yes (17/4)
99. cholesteral = Normal:
100. :...bp = Hypotension: Yes (11/4)
101. bp = Normal:
102. :...active = Yes: No (54/19)
103. active = No:
104. :...gender = Female: Yes (14/4)
105. gender = Male: No (21/7)
106. Evaluation on training data (6399 cases):
107. Decision Tree
108. ----------------
109. Size Errors
110. 45 1645(25.7%) <<
111. (a) (b) <-classified as
112. ---- ----
113. 2883 710 (a): class No
114. 935 1871 (b): class Yes
115. Attribute usage:
116. 100.00% diabetes
117. 93.23% age
118. 93.23% smoker
119. 88.34% bp
120. 67.56% obesity
121. 53.46% cholesteral
122. 48.73% active
123. 4.27% gender
124. Time: 0.0 secs
125. > plot(tree\_C51,main="Classification Tree for heartdisease")
126. > train\_predictions <- predict(tree\_C51, data\_train)
127. > mean(train\_predictions==data\_train$heartattack\_s)
128. [1] 0.7429286
129. > test\_predictions<- predict(tree\_C51, data\_test)
130. > mean(test\_predictions==data\_test$heartattack\_s)
131. [1] 0.7123202
132. > table(predicted = train\_predictions, actual = data\_train$heartattack\_s)
133. actual
134. predicted No Yes
135. No 2883 935
136. Yes 710 1871
137. > table(predicted = test\_predictions, actual = data\_test$heartattack\_s)
138. actual
139. predicted No Yes
140. No 717 279
141. Yes 181 422



> rpt<- rpart(heartattack\_s~ .,data=data\_train,method='class')

> printcp(rpt)

Classification tree:

rpart(formula = heartattack\_s ~ ., data = data\_train, method = "class")

Variables actually used in tree construction:

[1] active age bp cholesteral

[5] diabetes smoker

Root node error: 2806/6399 = 0.43851

n= 6399

CP nsplit rel error xerror xstd

1 0.115467 0 1.00000 1.00000 0.014146

2 0.063435 1 0.88453 0.89665 0.013925

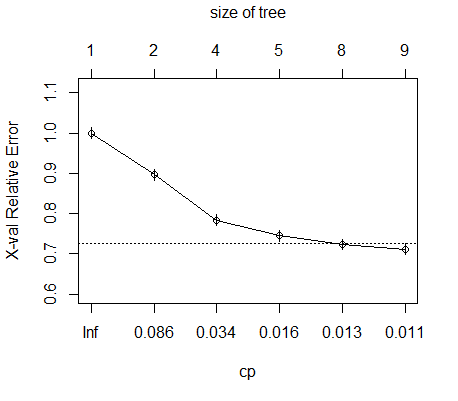
3 0.017819 3 0.75766 0.78297 0.013536

4 0.014849 4 0.73984 0.74519 0.013371

5 0.011761 7 0.69530 0.72345 0.013268

6 0.010000 8 0.68354 0.71169 0.013209

> plotcp(rpt)



> rptpred <- predict(rpt,type='class')

> mean(rptpred == data\_train$heartattack\_s)

[1] 0.7002657

rpt1<- rpart(heartattack\_s~ .,data=data\_test,method='class')

> printcp(rpt1)

Classification tree:

rpart(formula = heartattack\_s ~ ., data = data\_test, method = "class")

Variables actually used in tree construction:

[1] active age cholesteral diabetes

[5] obesity smoker

Root node error: 701/1599 = 0.4384

n= 1599

CP nsplit rel error xerror xstd

1 0.115549 0 1.00000 1.00000 0.028304

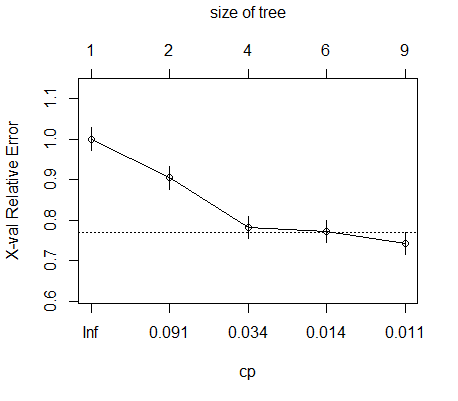
2 0.072040 1 0.88445 0.90442 0.027904

3 0.016405 3 0.74037 0.78317 0.027086

4 0.011888 5 0.70756 0.77175 0.026990

5 0.010000 8 0.67190 0.74322 0.026735

> plotcp(rpt1)



rptpred1 <- predict(rpt1,type='class')

> mean(rptpred1 == data\_test$heartattack\_s)

[1] 0.7054409

table(predicted = rptpred, actual = data\_train$heartattack\_s)

actual

predicted No Yes

No 2739 1064

Yes 854 1742

> table(predicted = rptpred1, actual = data\_test$heartattack\_s)

actual

predicted No Yes

No 739 312

Yes 159 389

C50:

1. train\_predictions <- predict(tree\_C51, data\_train)
2. > mean(train\_predictions==data\_train$heartattack\_s)
3. [1] 0.7429286
4. > test\_predictions<- predict(tree\_C51, data\_test)
5. > mean(test\_predictions==data\_test$heartattack\_s)
6. [1] 0.7123202
7. > table(predicted = train\_predictions, actual = data\_train$heartattack\_s)
8. actual
9. predicted No Yes
10. No 2883 935
11. Yes 710 1871
12. > table(predicted = test\_predictions, actual = data\_test$heartattack\_s)
13. actual
14. predicted No Yes
15. No 717 279
16. Yes 181 422

CART:

rptpred <- predict(rpt,type='class')

> mean(rptpred == data\_train$heartattack\_s)

[1] 0.7002657

rptpred1 <- predict(rpt1,type='class')

> mean(rptpred1 == data\_test$heartattack\_s)

[1] 0.7054409

table(predicted = rptpred, actual = data\_train$heartattack\_s)

actual

predicted No Yes

No 2739 1064

Yes 854 1742

> table(predicted = rptpred1, actual = data\_test$heartattack\_s)

actual

predicted No Yes

No 739 312

Yes 159 389

C50 has better accuracy values than CART, but they are not similar.

We can use the following 4 questions to narrow down to the best split we want for CART in order to match the similar results C50 yields.

* “For every node (including the root node), how should we choose the input feature to split on and, given this feature, what is the value of the split point?”
* “How do we decide whether a node should become a leaf node or if we should make another split point?”
* “How deep should our tree be allowed to become?”
* “Once we arrive at a leaf node, what value should we predict?”

**Appendix** (The complete version of your solution scripts in R)