report_addition_Switalska_Kohut.rmd

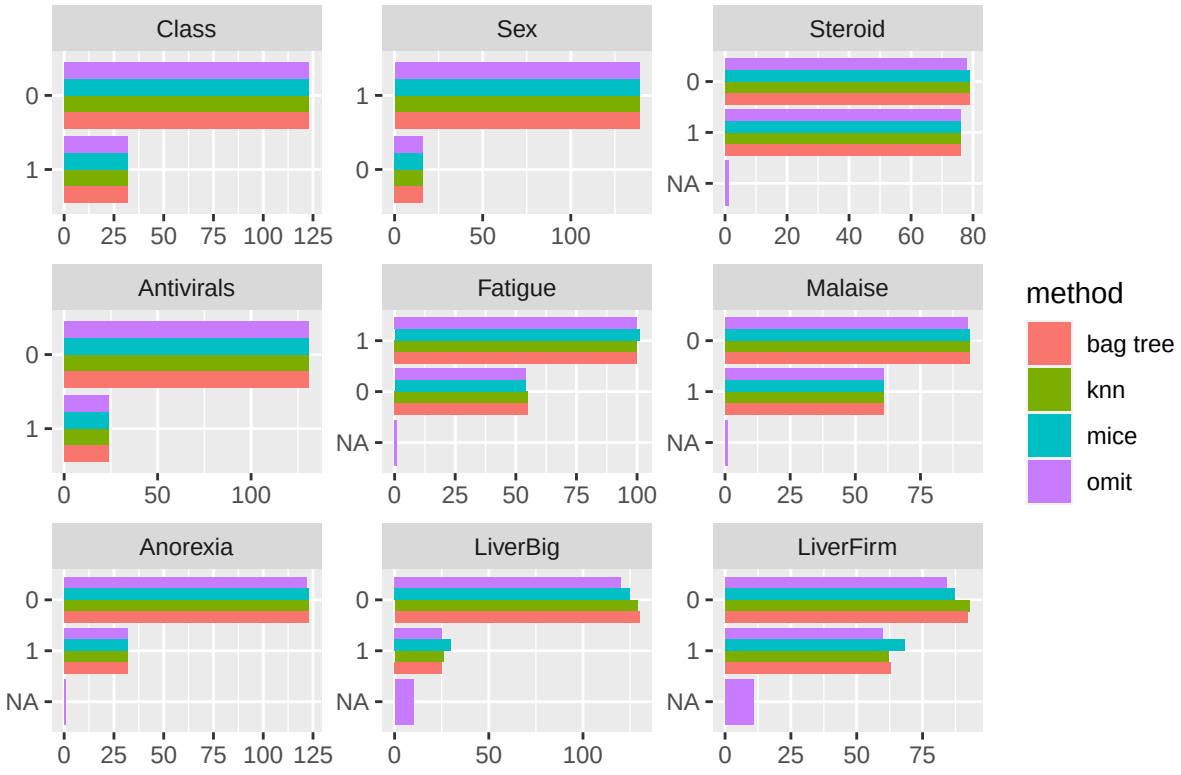
2023-12-08

3. EDA

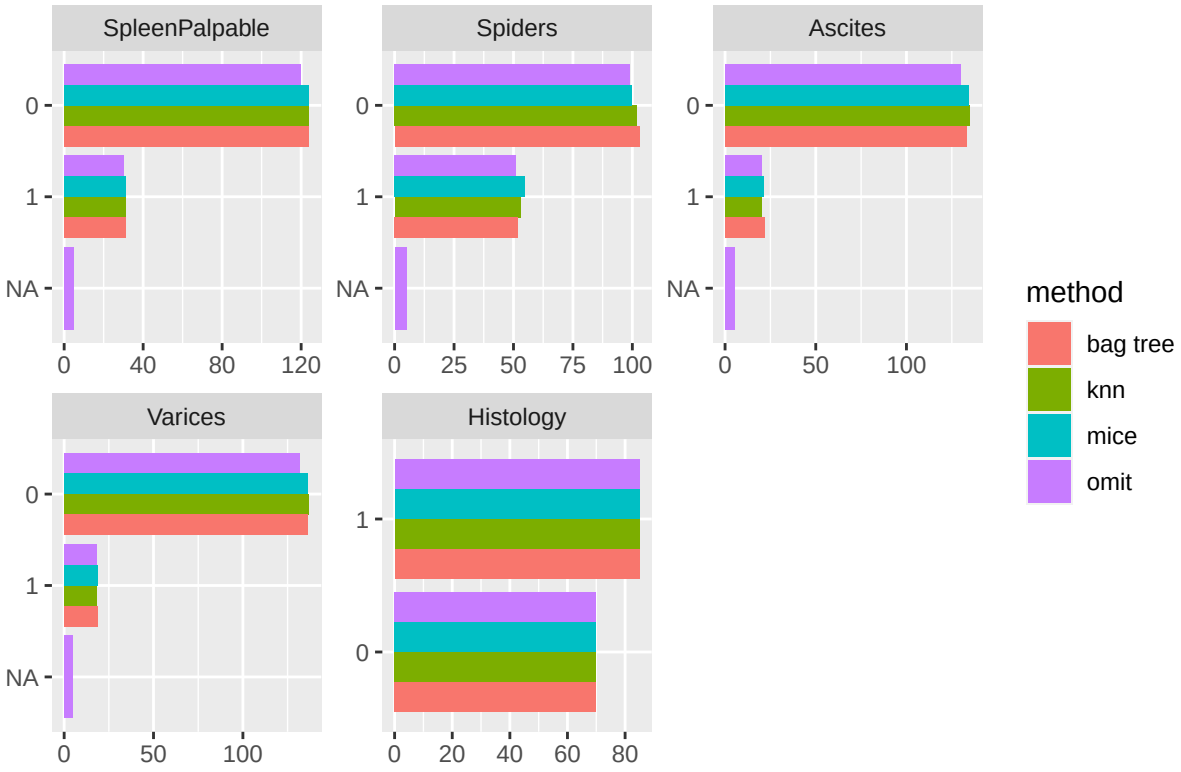
3.2. Barplots

We can see that bar plots for different methods are very similar.

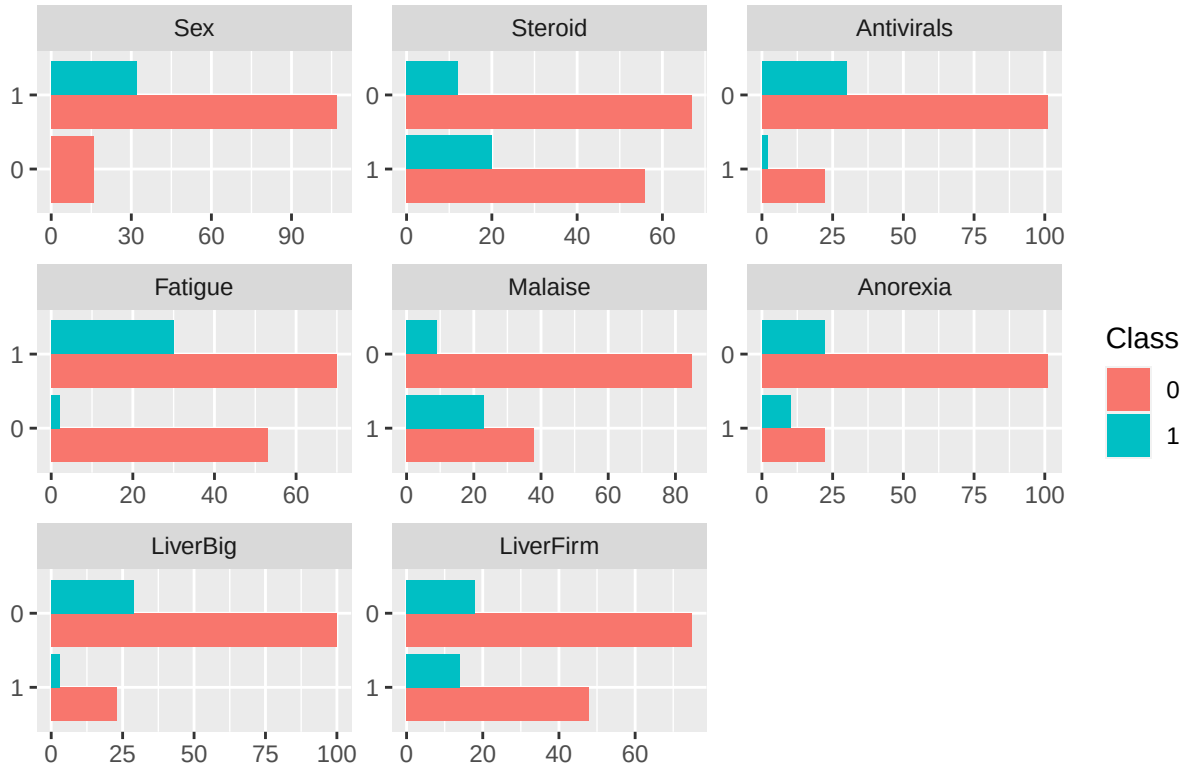
```
plot_bar(df_all, by = "method", by_position = "dodge")
```

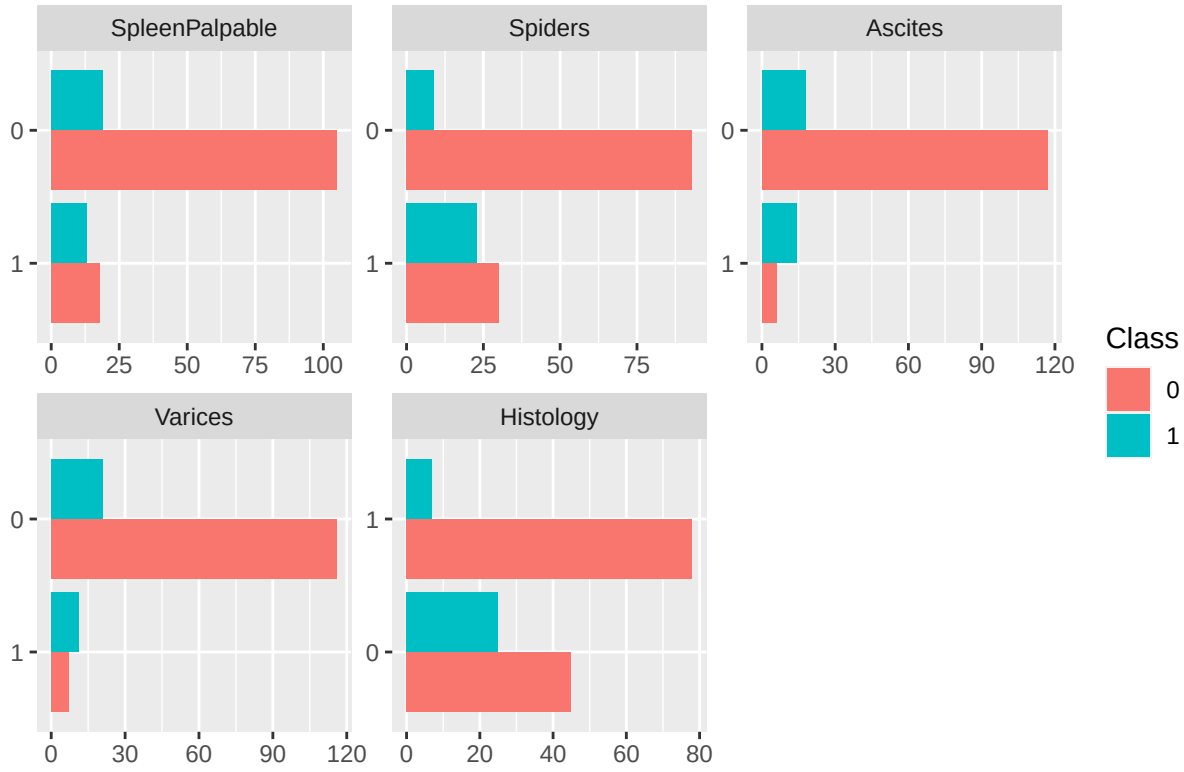


Page 1

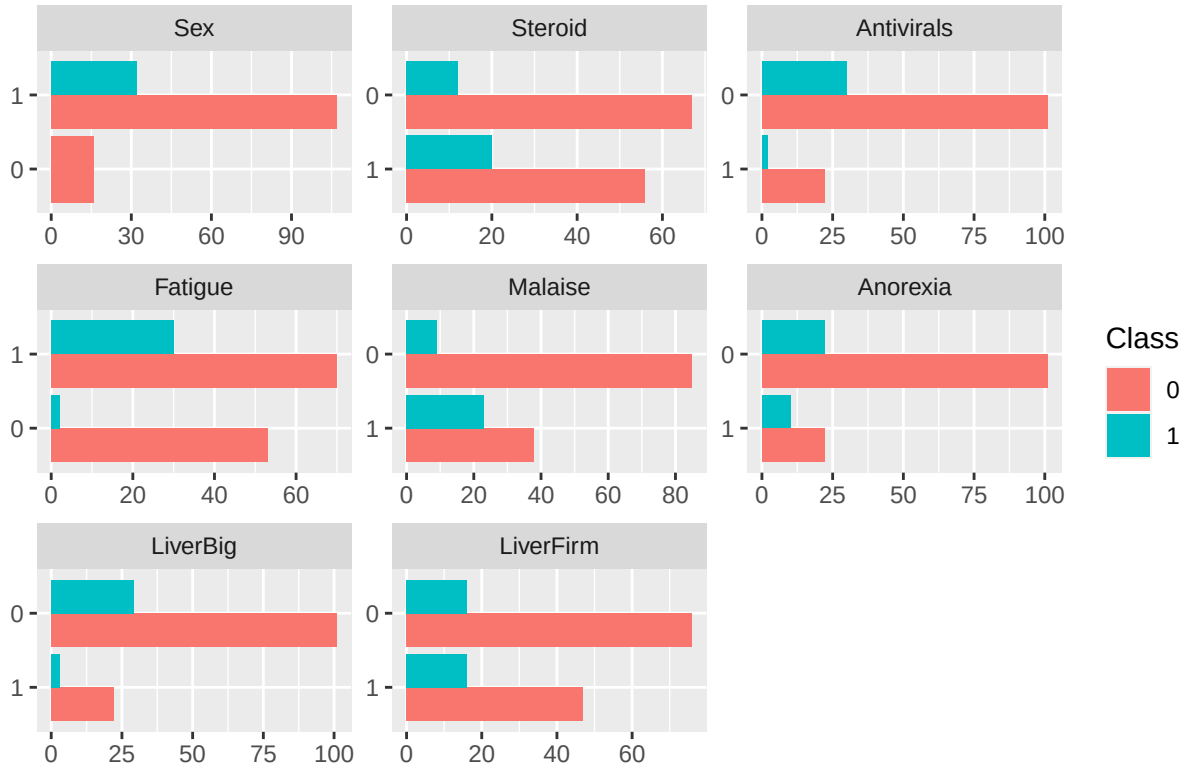



```
df1[, categorical] <-  lapply(df1[, categorical], as.factor)
plot_bar(df1[-21], by = "Class", by_position = "dodge")
```

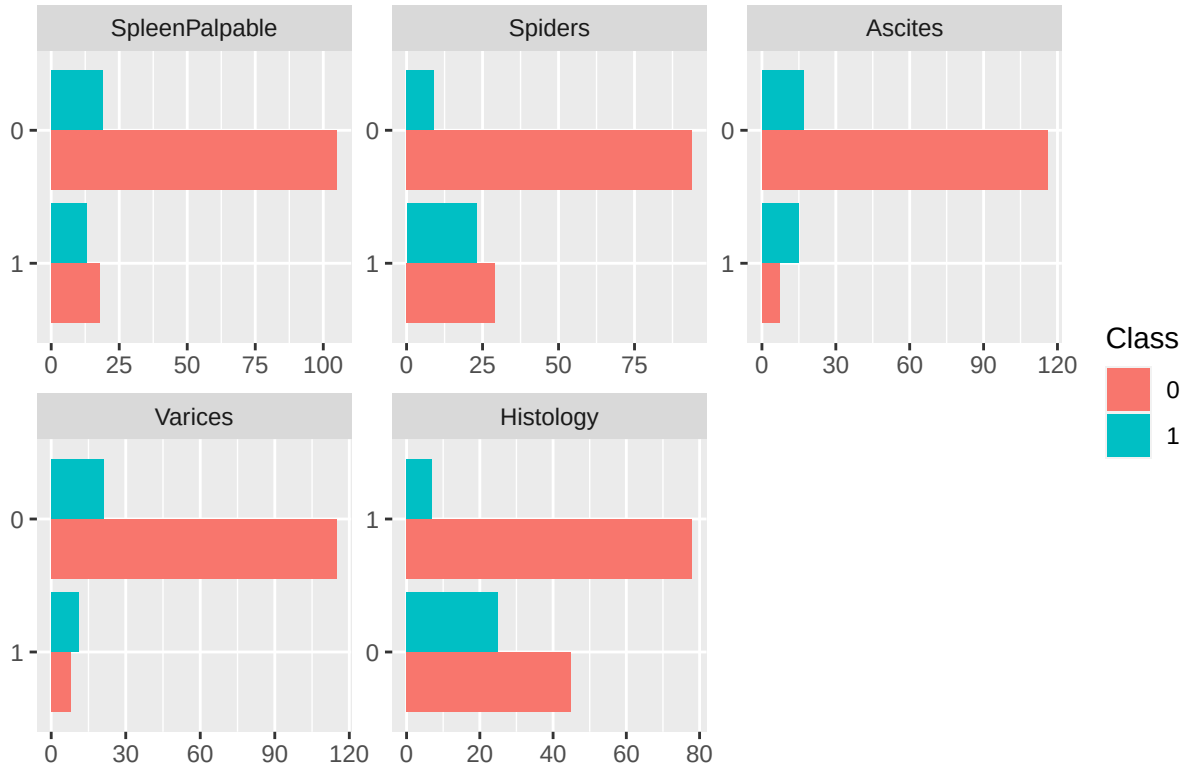

Sex
Steroid
Antivirals
Fatigue
Malaise
Anorexia
LiverBig
LiverFirm
Class
0
1



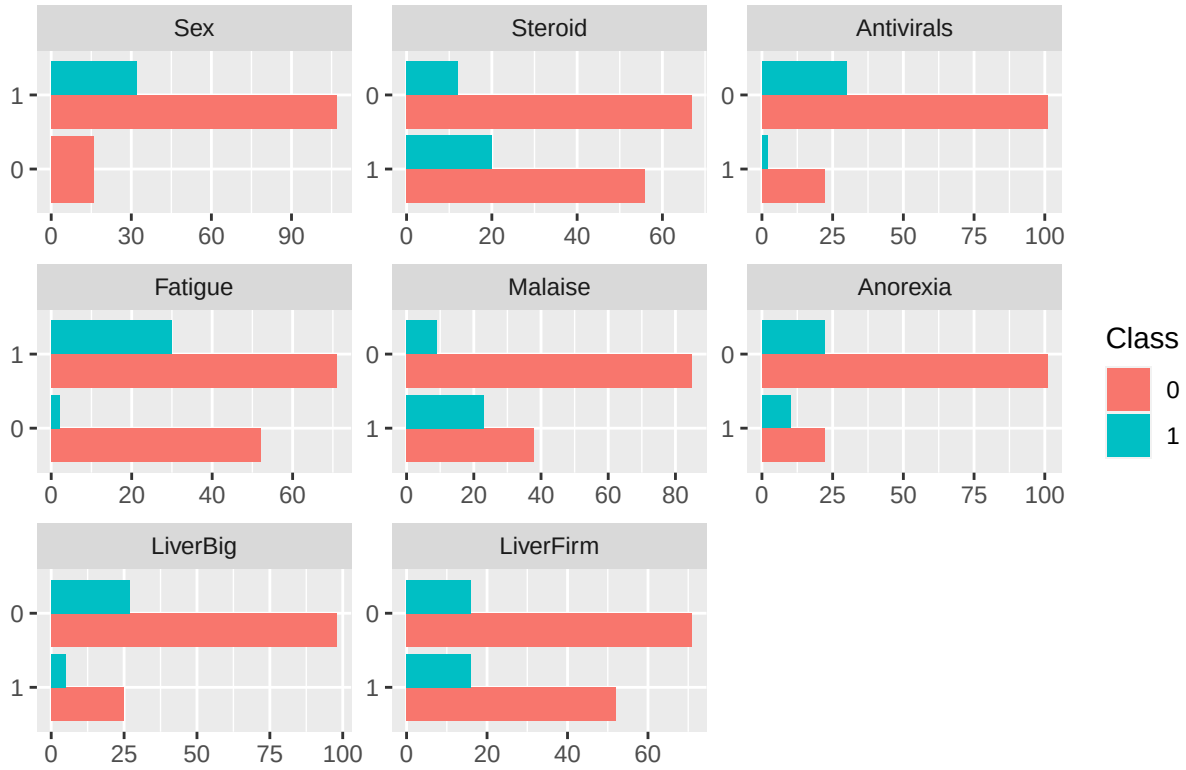

```
df2[, categorical] <-  lapply(df2[, categorical], as.factor)
plot_bar(df2[-21], by = "Class", by_position = "dodge")
```


Sex
Steroid
Antivirals
Fatigue
Malaise
Anorexia
LiverBig
LiverFirm
Class
0
1

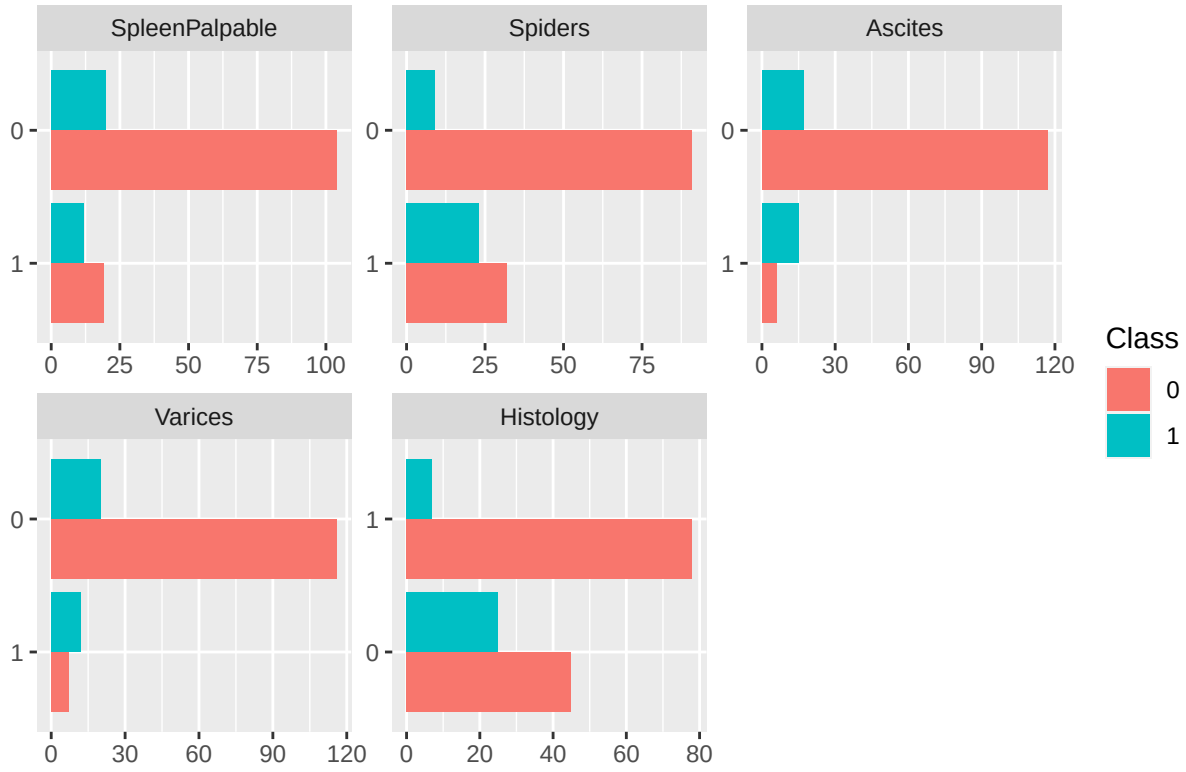
Page 1




```
df5[, categorical] <-  lapply(df5[, categorical], as.factor)
plot_bar(df5[-21], by = "Class", by_position = "dodge")
```


Sex
Steroid
Antivirals
Fatigue
Malaise
Anorexia
LiverBig
LiverFirm
Class
0
1

Page 1

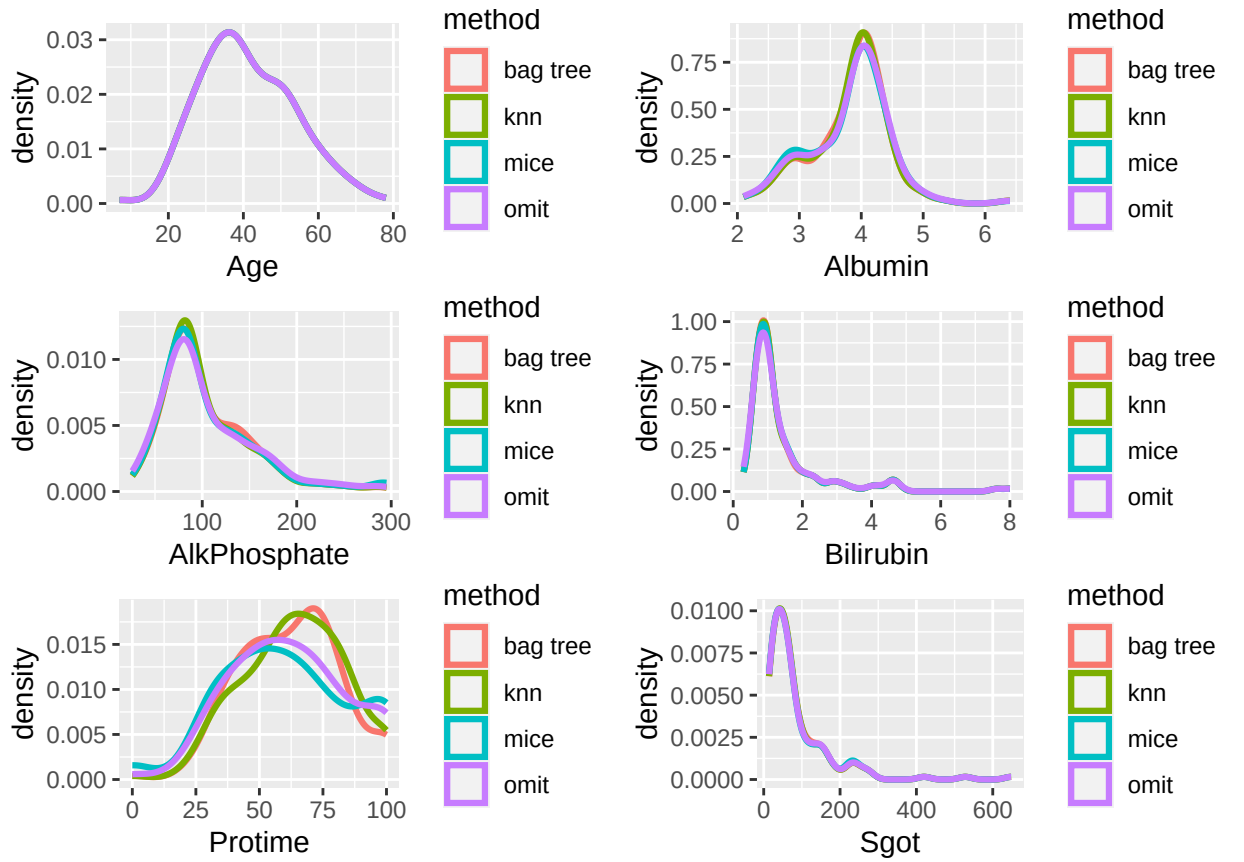


3.3. Histograms

We can see that density plots for different methods are very similar. Only for Protime results are a bit different.

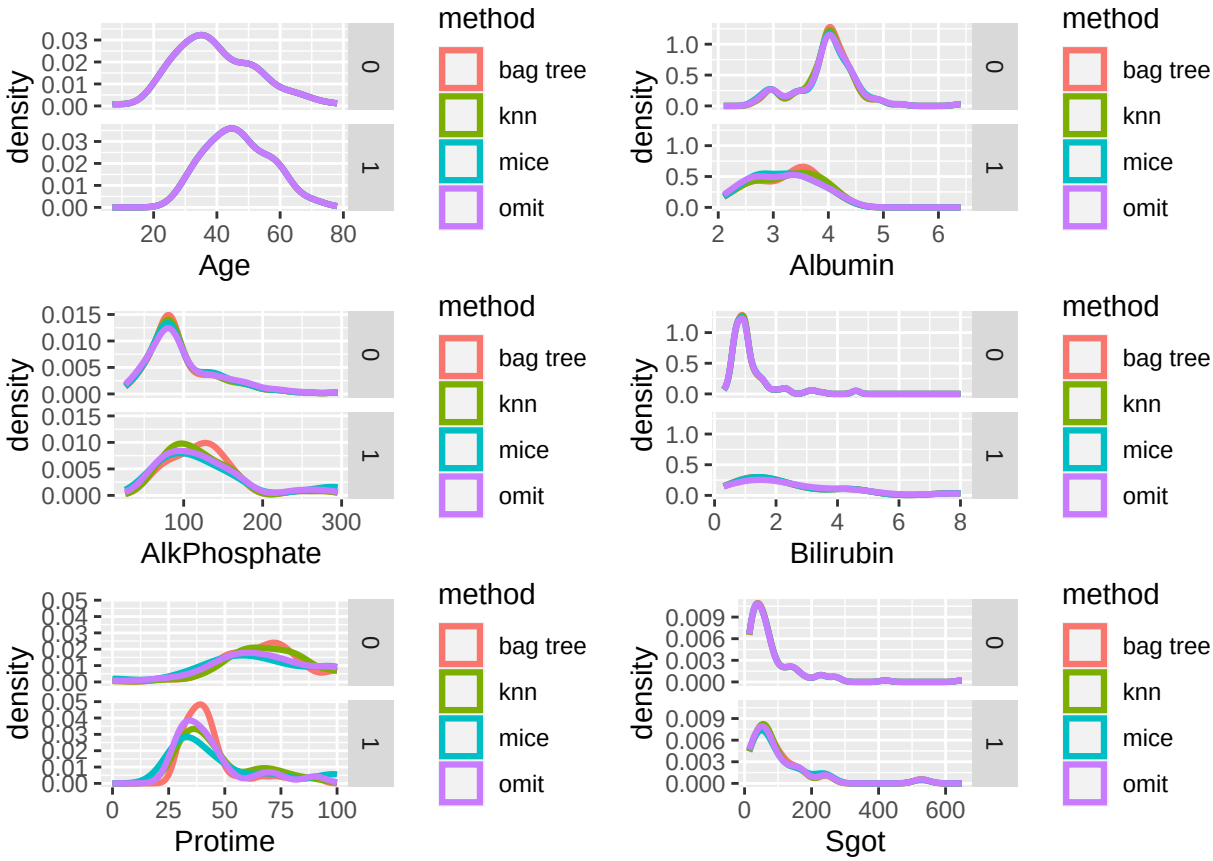
```
p7 <- ggplot(df_all, aes(x = Age, color = method)) +
  scale_fill_brewer(palette = "Set2") +
  geom_density(size = 1.1)

p8 <- ggplot(df_all, aes(x = Albumin, color = method)) +
  scale_fill_brewer(palette = "Set2") +
  geom_density(size = 1.1)

p9 <- ggplot(df_all, aes(x = AlkPhosphate, color = method)) +
  scale_fill_brewer(palette = "Set2") +
  geom_density(size = 1.1)

p10 <- ggplot(df_all, aes(x = Bilirubin, color = method)) +
  scale_fill_brewer(palette = "Set2") +
  geom_density(size = 1.1)

p11 <- ggplot(df_all, aes(x = Protime, color = method)) +
  scale_fill_brewer(palette = "Set2") +
  geom_density(size = 1.1)

p12 <- ggplot(df_all, aes(x = Sgot, color = method)) +
```

```
  scale_fill_brewer(palette = "Set2") +
  geom_density(size = 1.1)

ggarrange(p7, p8, p9, p10, p11, p12, ncol = 2, nrow = 3)
```




```
p13 <- ggplot(df_all, aes(x = Age, color = method)) +
  scale_fill_brewer(palette = "Set2") +
  geom_density(size = 1.1) +
  facet_grid(Class ~ .)

p14 <- ggplot(df_all, aes(x = Albumin, color = method)) +
  scale_fill_brewer(palette = "Set2") +
  geom_density(size = 1.1) +
  facet_grid(Class ~ .)

p15 <- ggplot(df_all, aes(x = AlkPhosphate, color = method)) +
  scale_fill_brewer(palette = "Set2") +
  geom_density(size = 1.1) +
  facet_grid(Class ~ .)

p16 <- ggplot(df_all, aes(x = Bilirubin, color = method)) +
  scale_fill_brewer(palette = "Set2") +
  geom_density(size = 1.1) +
  facet_grid(Class ~ .)

p17 <- ggplot(df_all, aes(x = Protime, color = method)) +
```

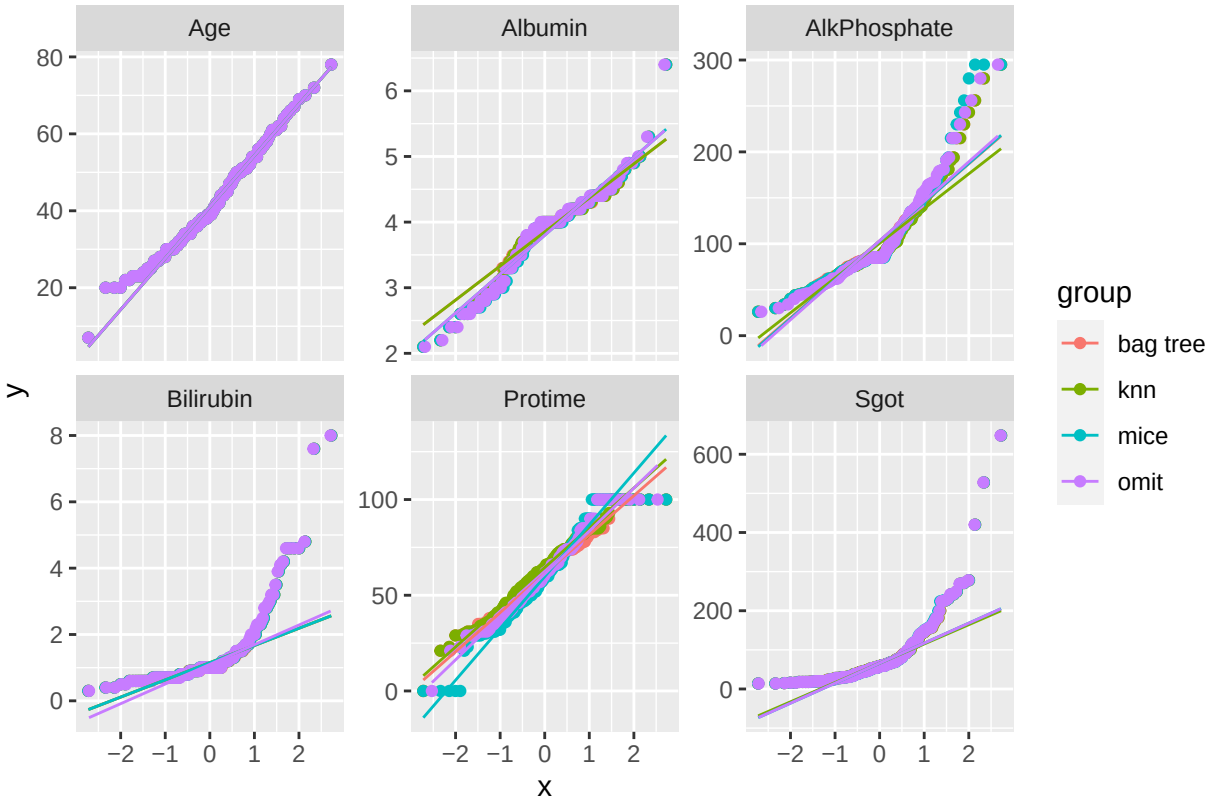
```
  scale_fill_brewer(palette = "Set2") +
  geom_density(size = 1.1) +
  facet_grid(Class ~ .)

p18 <- ggplot(df_all, aes(x = Sgot, color = method)) +
  scale_fill_brewer(palette = "Set2") +
  geom_density(size = 1.1) +
  facet_grid(Class ~ .)

ggarrange(p13, p14, p15, p16, p17, p18, ncol = 2, nrow = 3)
```



3.4. Q-Q plot

We can see that Q-Q plots for different methods are very similar. Only for Protime results are a bit different.

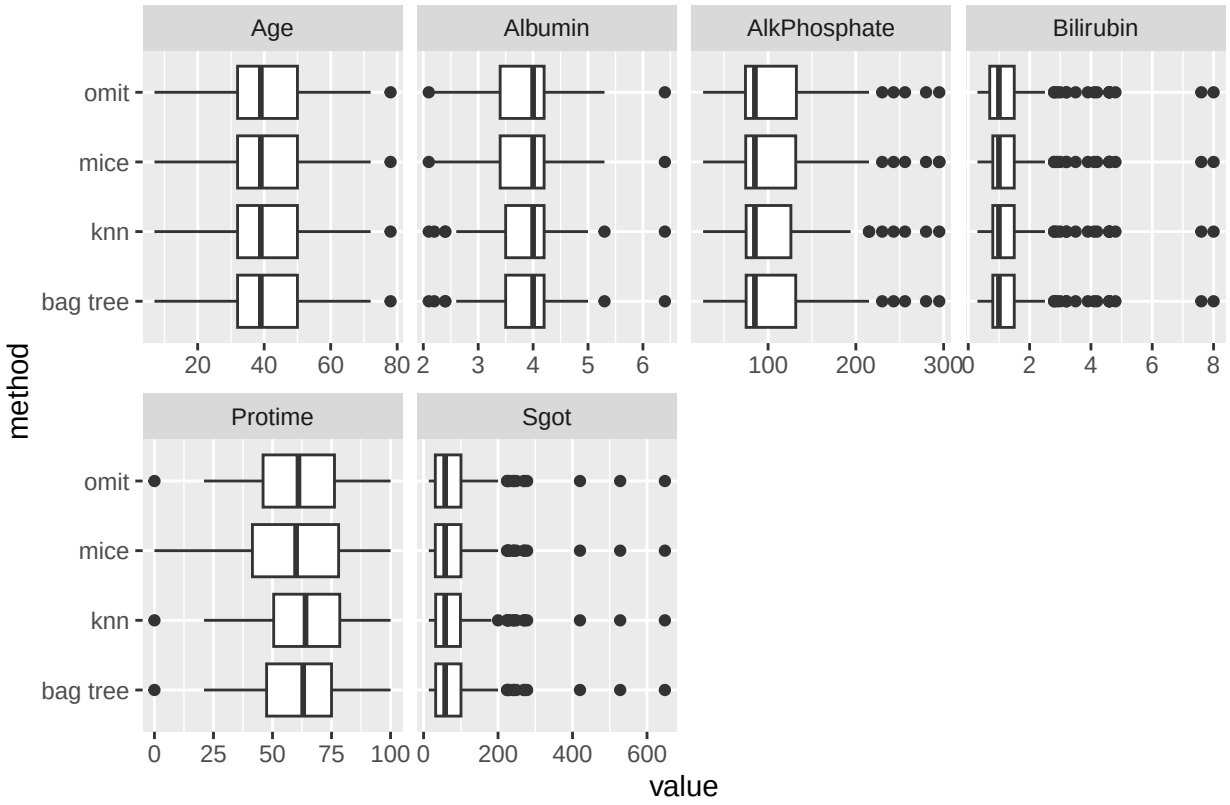
```
plot_qq(df_all, by = "method")
```



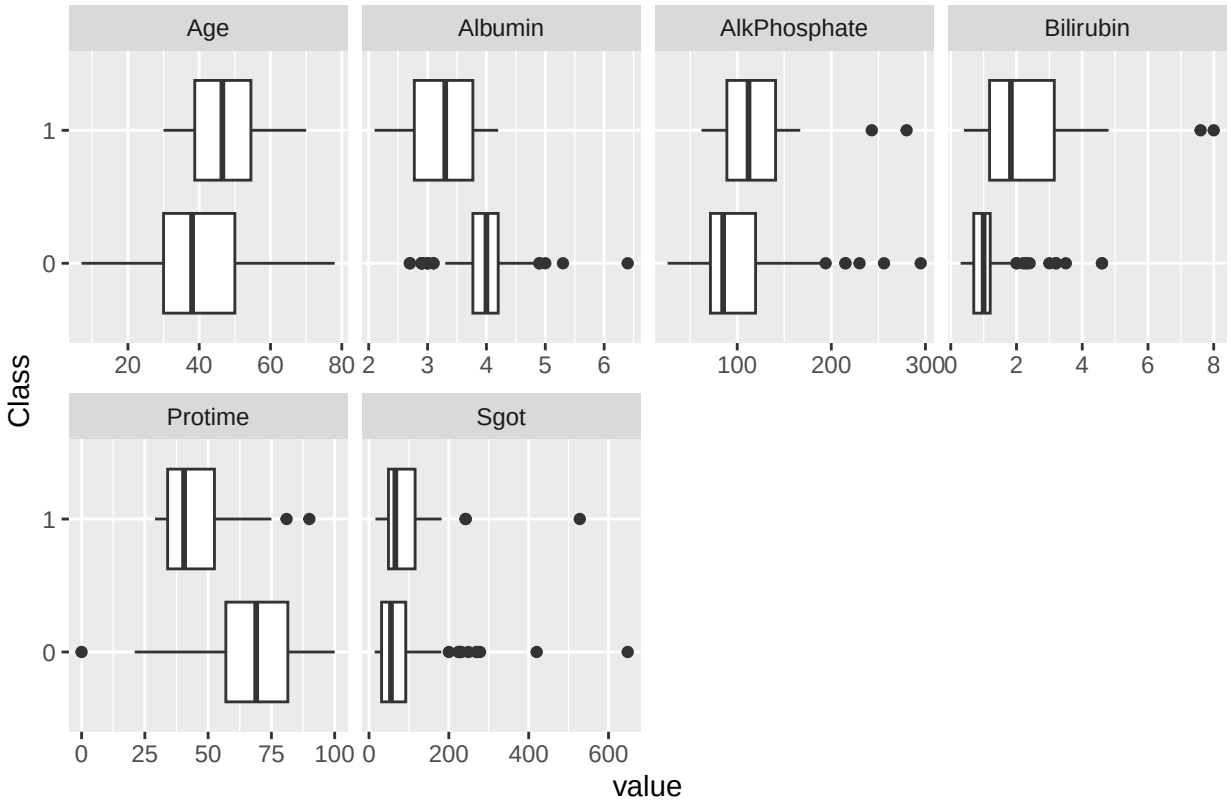
3.5. Boxplots

We can see that bar plots for different methods are very similar.

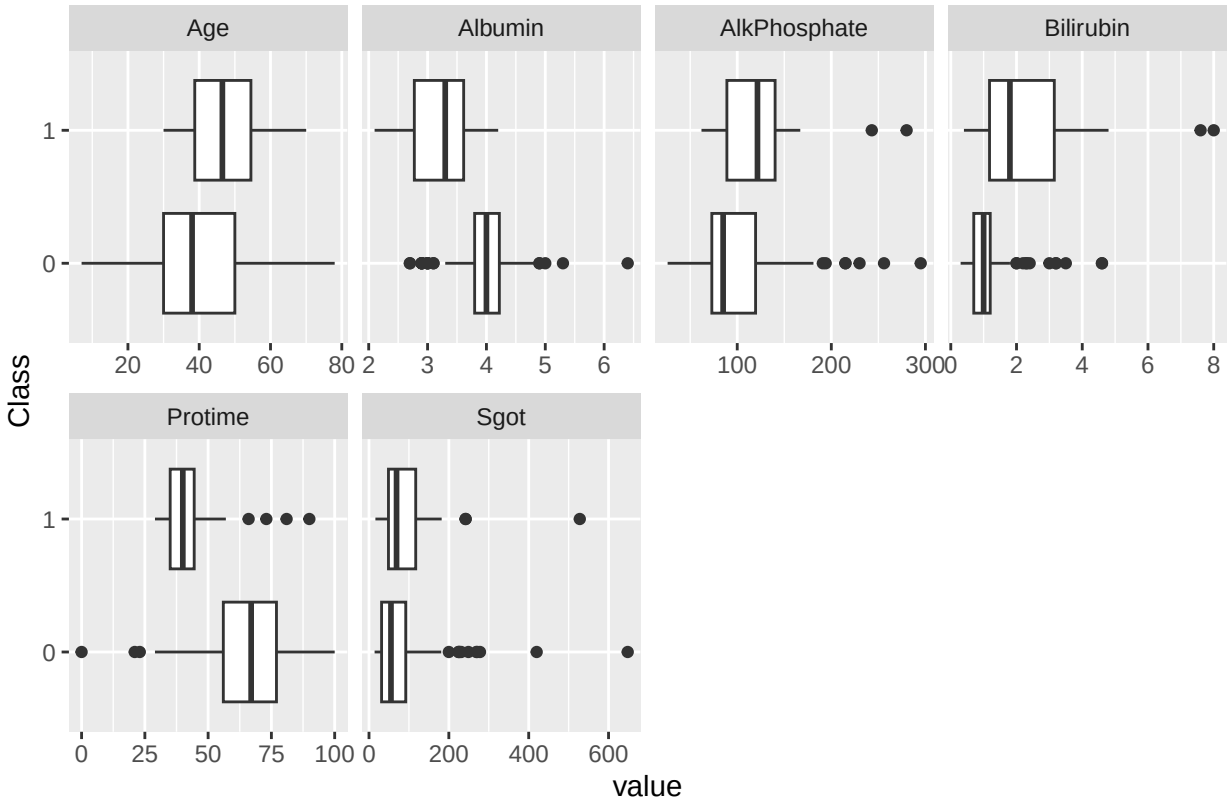
```
plot_boxplot(df_all, by = "method")
```



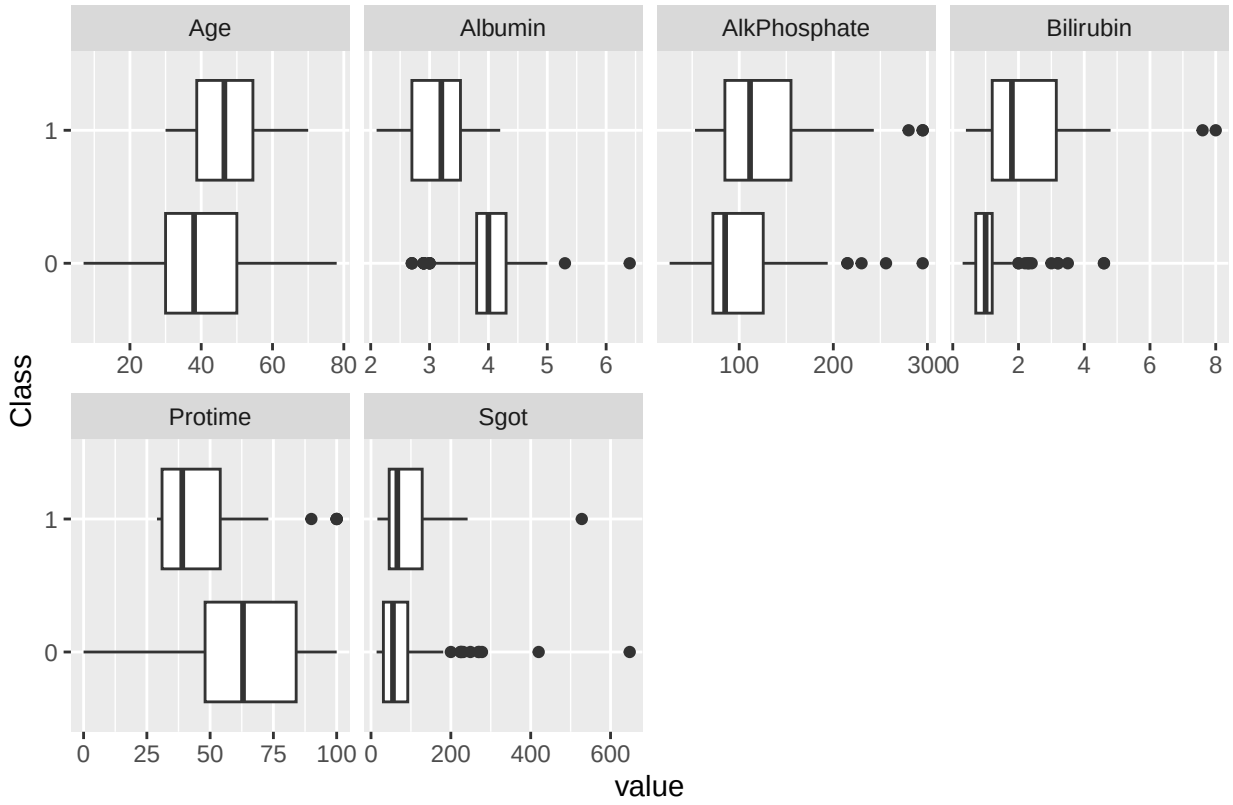

```
plot_boxplot(df1, by = "Class")
```




```
plot_boxplot(df2, by = "Class")
```



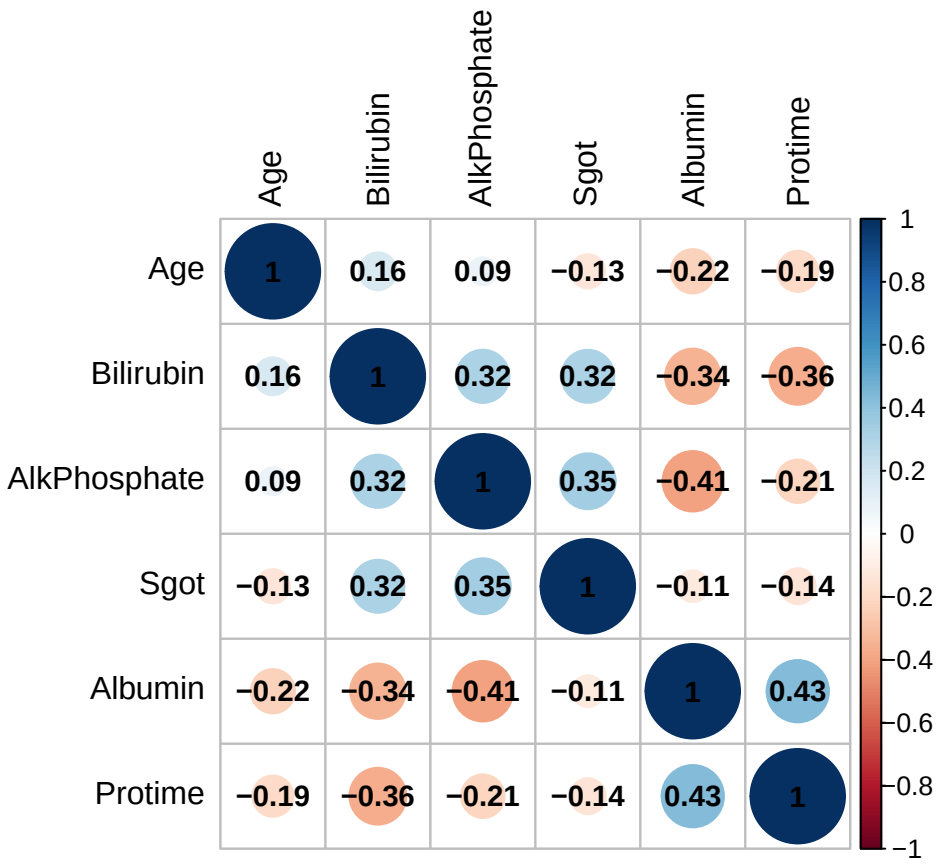

```
plot_boxplot(df5, by = "Class")
```



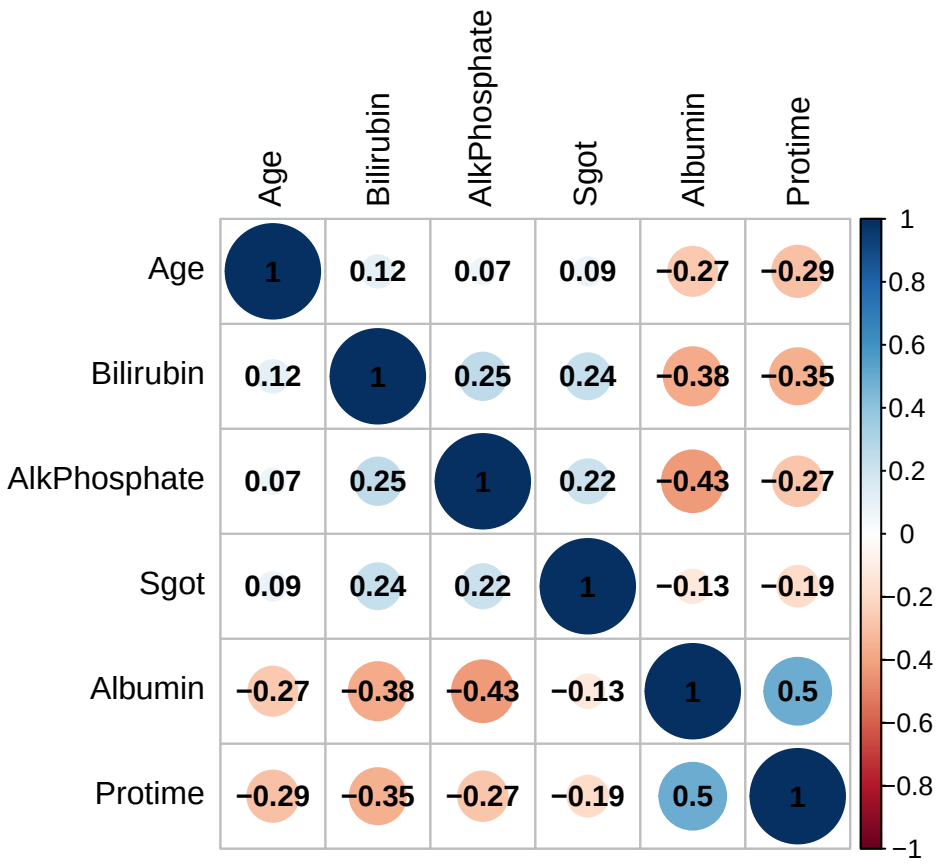
3.6. Correlation

We can see that correlation matrices for different methods are very similar. We also added a matrix for initial data with omitting objects with missing values.

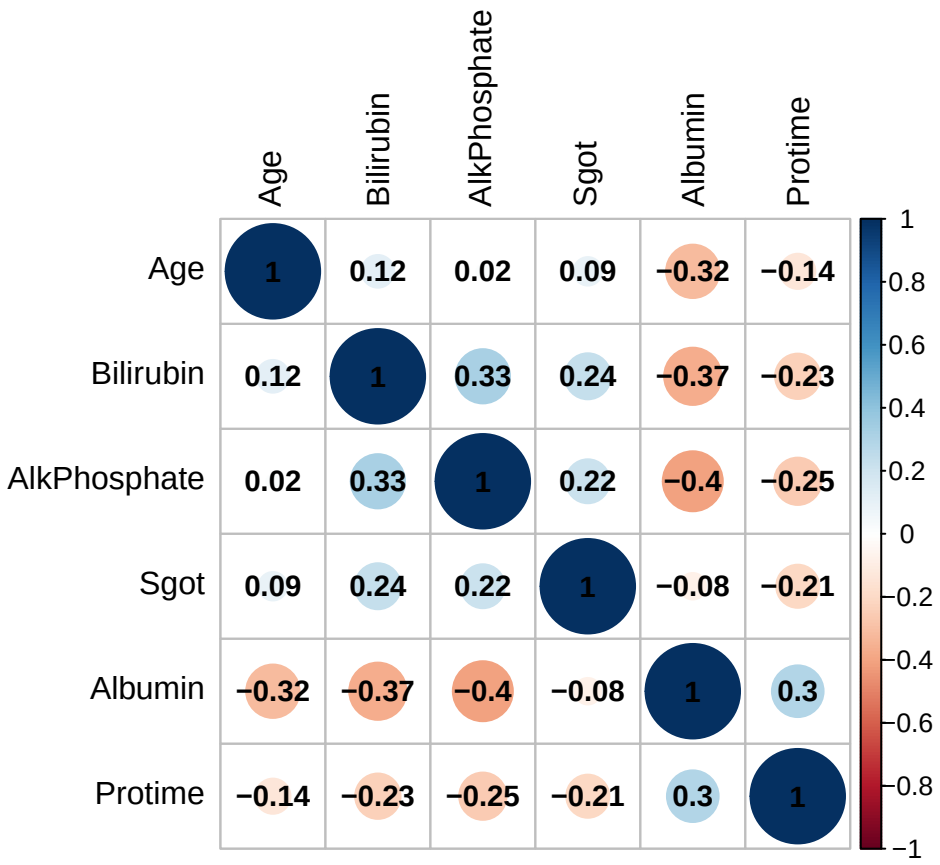
```
df_omit <- na.omit(df)
cor_matrix1 <- cor(df_omit[, sapply(df_omit, is.numeric)], method = "pearson")
cor_matrix2 <- cor(df2[, sapply(df2, is.numeric)], method = "pearson")
cor_matrix3 <- cor(df5[, sapply(df5, is.numeric)], method = "pearson")

corrplot(cor_matrix1, tl.col = "black", addCoef.col = 1, number.cex = 0.9)
```




```
corrplot(cor_matrix2, tl.col = "black", addCoef.col = 1, number.cex = 0.9)
```




```
corrplot(cor_matrix3, tl.col = "black", addCoef.col = 1, number.cex = 0.9)
```



4. Classification

4.6. Logistic regression (LR)

We can see that summary for different methods are similar.

```
summary(model.logit1)
```

```
##
## Call:
## glm(formula = Class ~ . - Class, family = binomial(link = "logit"),
##     data = train.balanced1)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.030e+01  1.561e+03  -0.013  0.98963
## Age             1.152e-01  4.866e-02   2.367  0.01794 *
## Sex             2.046e+01  1.561e+03   0.013  0.98954
## Steroid         3.580e+00  1.369e+00   2.615  0.00892 **
## Antivirals     -9.730e-01  1.688e+00  -0.577  0.56426
## Fatigue         1.686e+00  1.590e+00   1.060  0.28895
## Malaise         1.142e+00  1.154e+00   0.989  0.32244
## Anorexia       -3.738e+00  1.182e+00  -3.163  0.00156 **
## LiverBig       -1.809e+00  1.596e+00  -1.134  0.25684
## LiverFirm      -5.856e-01  1.114e+00  -0.526  0.59918
## SpleenPalpable  3.086e-01  1.094e+00   0.282  0.77779
```

```
## Spiders         2.605e+00  1.153e+00   2.259  0.02388 *
## Ascites         7.618e-01  1.770e+00   0.430  0.66697
## Varices        -4.583e-01  1.534e+00  -0.299  0.76509
## Bilirubin       1.254e+00  4.536e-01   2.764  0.00571 **
## AlkPhosphate   -2.814e-03  7.666e-03  -0.367  0.71360
## Sgot           -2.227e-03  6.193e-03  -0.360  0.71914
## Albumin        -2.050e+00  1.123e+00  -1.826  0.06784 .
## Protime        -8.527e-02  3.647e-02  -2.338  0.01940 *
## Histology       1.289e+00  9.392e-01   1.372  0.17003
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 257.851  on 185  degrees of freedom
## Residual deviance:  68.755  on 166  degrees of freedom
## AIC: 108.76
##
## Number of Fisher Scoring iterations: 17
```

```
summary(model.logit2)
```

```
##
## Call:
## glm(formula = Class ~ . - Class, family = binomial(link = "logit"),
##     data = train.balanced2)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.419e+01  1.684e+03  -0.014  0.98854
## Age             8.595e-02  4.121e-02   2.086  0.03700 *
## Sex             1.902e+01  1.684e+03   0.011  0.99099
## Steroid         2.484e+00  1.067e+00   2.328  0.01991 *
## Antivirals     -7.505e-01  1.333e+00  -0.563  0.57341
## Fatigue         3.549e-01  1.289e+00   0.275  0.78310
## Malaise         1.064e+00  1.082e+00   0.983  0.32570
## Anorexia       -2.753e+00  1.000e+00  -2.753  0.00591 **
## LiverBig       -1.894e+00  1.127e+00  -1.680  0.09286 .
## LiverFirm       3.556e-01  9.451e-01   0.376  0.70678
## SpleenPalpable  9.053e-01  1.033e+00   0.876  0.38087
## Spiders         1.112e+00  9.134e-01   1.217  0.22363
## Ascites         2.711e+00  1.330e+00   2.038  0.04156 *
## Varices        -5.952e-02  1.166e+00  -0.051  0.95928
## Bilirubin       1.131e+00  4.387e-01   2.578  0.00995 **
## AlkPhosphate    9.912e-03  7.006e-03   1.415  0.15713
## Sgot           -1.912e-04  4.243e-03  -0.045  0.96406
## Albumin        -5.053e-01  8.137e-01  -0.621  0.53456
## Protime        -5.367e-02  2.320e-02  -2.313  0.02073 *
## Histology      -3.479e-02  7.806e-01  -0.045  0.96445
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 257.851  on 185  degrees of freedom
```

```
## Residual deviance:  79.775  on 166  degrees of freedom
## AIC: 119.77
##
## Number of Fisher Scoring iterations: 17
```

```
summary(model.logit3)
```

```
##
## Call:
## glm(formula = Class ~ . - Class, family = binomial(link = "logit"),
##     data = train.balanced3)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.857e+01  1.560e+03  -0.012   0.9905
## Age             4.959e-02  4.608e-02   1.076   0.2819
## Sex             1.911e+01  1.560e+03   0.012   0.9902
## Steroid         1.585e+00  1.117e+00   1.419   0.1558
## Antivirals      1.857e+00  1.429e+00   1.299   0.1939
## Fatigue         1.431e+00  1.684e+00   0.850   0.3955
## Malaise         1.811e+00  1.115e+00   1.624   0.1043
## Anorexia       -3.516e+00  1.409e+00  -2.496   0.0125 *
## LiverBig        6.857e-01  1.247e+00   0.550   0.5822
## LiverFirm      -1.942e+00  1.583e+00  -1.227   0.2198
## SpleenPalpable  9.144e-01  1.103e+00   0.829   0.4069
## Spiders         1.827e+00  1.001e+00   1.825   0.0681 .
## Ascites         2.102e+00  1.327e+00   1.584   0.1131
## Varices         4.188e-01  1.241e+00   0.337   0.7358
## Bilirubin       1.418e+00  5.844e-01   2.427   0.0152 *
## AlkPhosphate   -6.290e-03  9.502e-03  -0.662   0.5080
## Sgot            6.720e-03  4.844e-03   1.387   0.1653
## Albumin        -2.696e+00  1.082e+00  -2.491   0.0127 *
## Protime         1.468e-02  2.790e-02   0.526   0.5988
## Histology      -4.628e-01  9.276e-01  -0.499   0.6178
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 257.851  on 185  degrees of freedom
## Residual deviance:  71.801  on 166  degrees of freedom
## AIC: 111.8
##
## Number of Fisher Scoring iterations: 17
```

```
summary(model.logit4)
```

```
##
## Call:
## glm(formula = Class ~ . - Class - Age - Anorexia - Bilirubin,
##     family = binomial(link = "logit"), data = train.balanced1)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.397e+01  1.805e+03  -0.008   0.9938
```

```
## Sex             1.865e+01  1.805e+03   0.010   0.9918
## Steroid         1.425e+00  7.365e-01   1.935   0.0530 .
## Antivirals      4.030e-02  1.017e+00   0.040   0.9684
## Fatigue         2.026e+00  1.291e+00   1.569   0.1166
## Malaise         8.756e-01  8.106e-01   1.080   0.2801
## LiverBig       -7.743e-01  9.587e-01  -0.808   0.4193
## LiverFirm      -6.358e-01  7.415e-01  -0.857   0.3912
## SpleenPalpable  1.275e+00  8.195e-01   1.555   0.1199
## Spiders         1.076e+00  6.875e-01   1.565   0.1175
## Ascites         1.146e+00  1.118e+00   1.026   0.3050
## Varices        -4.592e-01  1.070e+00  -0.429   0.6678
## AlkPhosphate    4.040e-04  5.868e-03   0.069   0.9451
## Sgot           -2.156e-03  3.223e-03  -0.669   0.5035
## Albumin        -1.669e+00  7.459e-01  -2.238   0.0253 *
## Protime        -4.044e-02  1.913e-02  -2.114   0.0346 *
## Histology       9.870e-03  7.066e-01   0.014   0.9889
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 257.85  on 185  degrees of freedom
## Residual deviance: 101.14  on 169  degrees of freedom
## AIC: 135.14
##
## Number of Fisher Scoring iterations: 17
```

```
summary(model.logit5)
```

```
##
## Call:
## glm(formula = Class ~ . - Class - Age - Anorexia - Bilirubin,
##     family = binomial(link = "logit"), data = train.balanced2)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.357e+01  1.738e+03  -0.008  0.99377
## Sex             1.871e+01  1.738e+03   0.011  0.99141
## Steroid         1.182e+00  7.307e-01   1.618  0.10575
## Antivirals     -5.719e-01  1.105e+00  -0.518  0.60464
## Fatigue         9.768e-01  1.124e+00   0.869  0.38473
## Malaise         8.048e-01  8.142e-01   0.988  0.32294
## LiverBig       -1.136e+00  8.647e-01  -1.314  0.18898
## LiverFirm       4.764e-01  7.315e-01   0.651  0.51491
## SpleenPalpable  9.117e-01  8.014e-01   1.138  0.25524
## Spiders         6.444e-01  6.896e-01   0.934  0.35006
## Ascites         1.438e+00  1.063e+00   1.353  0.17616
## Varices        -5.385e-01  9.528e-01  -0.565  0.57192
## AlkPhosphate    5.640e-03  5.762e-03   0.979  0.32766
## Sgot           -1.116e-03  3.206e-03  -0.348  0.72768
## Albumin        -1.430e+00  7.018e-01  -2.037  0.04161 *
## Protime        -5.708e-02  1.942e-02  -2.939  0.00329 **
## Histology      -6.452e-01  6.428e-01  -1.004  0.31546
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 257.85  on 185  degrees of freedom
## Residual deviance: 102.09  on 169  degrees of freedom
## AIC: 136.09
##
## Number of Fisher Scoring iterations: 17
```

```
summary(model.logit6)
```

```
##
## Call:
## glm(formula = Class ~ . - Class - Age - Anorexia - Bilirubin,
##     family = binomial(link = "logit"), data = train.balanced3)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.015e+01  1.839e+03  -0.006  0.99560
## Sex             1.777e+01  1.839e+03   0.010  0.99229
## Steroid         8.657e-01  8.475e-01   1.022  0.30699
## Antivirals      9.319e-01  1.135e+00   0.821  0.41148
## Fatigue         1.079e+00  1.445e+00   0.747  0.45496
## Malaise         1.376e+00  9.495e-01   1.449  0.14725
## LiverBig        6.535e-01  9.561e-01   0.684  0.49428
## LiverFirm      -5.168e-01  9.734e-01  -0.531  0.59543
## SpleenPalpable  1.808e+00  7.930e-01   2.279  0.02264 *
## Spiders         1.222e+00  7.889e-01   1.549  0.12150
## Ascites         9.997e-01  1.015e+00   0.985  0.32458
## Varices         5.655e-01  1.037e+00   0.545  0.58548
## AlkPhosphate   -1.740e-02  8.130e-03  -2.140  0.03235 *
## Sgot            4.508e-03  3.689e-03   1.222  0.22176
## Albumin        -2.464e+00  7.798e-01  -3.159  0.00158 **
## Protime        -2.208e-02  1.816e-02  -1.216  0.22399
## Histology      -7.392e-01  7.447e-01  -0.993  0.32090
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 257.851  on 185  degrees of freedom
## Residual deviance:  88.159  on 169  degrees of freedom
## AIC: 122.16
##
## Number of Fisher Scoring iterations: 17
```

```
summary(model.logit7)
```

```
##
## Call:
## glm(formula = Class ~ . - Class - Age - Steroid - AlkPhosphate -
##     Sgot, family = binomial(link = "logit"), data = train.balanced1)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)     -13.75786 1709.38618  -0.008 0.993578
## Sex              18.83603 1709.38375   0.011 0.991208
## Antivirals        0.12920    1.17947   0.110 0.912773
## Fatigue           1.76994    1.24643   1.420 0.155607
## Malaise           1.95626    0.81782   2.392 0.016755 *
## Anorexia         -2.78879    0.83408  -3.344 0.000827 ***
## LiverBig         -0.75707    1.13804  -0.665 0.505896
## LiverFirm        -1.28415    0.88930  -1.444 0.148738
## SpleenPalpable    1.39971    0.86055   1.627 0.103837
## Spiders           1.62689    0.82982   1.961 0.049933 *
## Ascites           0.19876    1.25332   0.159 0.873993
## Varices          -1.27272    1.10776  -1.149 0.250594
## Bilirubin         0.82404    0.31897   2.583 0.009782 **
## Albumin          -1.84133    0.82090  -2.243 0.024893 *
## Protime          -0.03951    0.02197  -1.798 0.072149 .
## Histology         0.30671    0.77655   0.395 0.692874
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 257.851  on 185  degrees of freedom
## Residual deviance:  84.748  on 170  degrees of freedom
## AIC: 116.75
##
## Number of Fisher Scoring iterations: 17
```

```
summary(model.logit8)
```

```
##
## Call:
## glm(formula = Class ~ . - Class - Age - Steroid - AlkPhosphate -
##     Sgot, family = binomial(link = "logit"), data = train.balanced2)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -13.18976 1746.02601  -0.008  0.99397
## Sex              18.34547 1746.02399   0.011  0.99162
## Antivirals       -0.19893    1.11282  -0.179  0.85813
## Fatigue           0.43351    1.02379   0.423  0.67198
## Malaise           1.95153    0.82374   2.369  0.01783 *
## Anorexia         -2.25367    0.81783  -2.756  0.00586 **
## LiverBig         -1.29737    0.94044  -1.380  0.16773
## LiverFirm         0.32007    0.77771   0.412  0.68067
## SpleenPalpable    1.41388    0.80998   1.746  0.08089 .
## Spiders           0.69000    0.78797   0.876  0.38121
## Ascites           0.71461    1.05508   0.677  0.49821
## Varices          -1.10167    0.89232  -1.235  0.21698
## Bilirubin         0.87276    0.34418   2.536  0.01122 *
## Albumin          -1.47049    0.71813  -2.048  0.04059 *
## Protime          -0.04765    0.01868  -2.550  0.01076 *
## Histology        -0.52041    0.68098  -0.764  0.44474
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 257.851  on 185  degrees of freedom
## Residual deviance:  91.399  on 170  degrees of freedom
## AIC: 123.4
##
## Number of Fisher Scoring iterations: 17
```

```
summary(model.logit9)
```

```
##
## Call:
## glm(formula = Class ~ . - Class - Age - Steroid - AlkPhosphate -
##     Sgot, family = binomial(link = "logit"), data = train.balanced3)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.401e+01  1.580e+03  -0.009  0.99293
## Sex             1.889e+01  1.580e+03   0.012  0.99046
## Antivirals      1.402e+00  1.247e+00   1.125  0.26071
## Fatigue         1.065e+00  1.422e+00   0.749  0.45388
## Malaise         2.389e+00  9.812e-01   2.435  0.01491 *
## Anorexia       -2.497e+00  9.274e-01  -2.693  0.00708 **
## LiverBig        6.493e-01  1.094e+00   0.594  0.55272
## LiverFirm      -1.461e+00  1.096e+00  -1.332  0.18273
## SpleenPalpable  1.510e+00  9.491e-01   1.591  0.11152
## Spiders         1.319e+00  8.945e-01   1.474  0.14042
## Ascites         1.124e+00  1.053e+00   1.067  0.28580
## Varices        -1.307e+00  9.719e-01  -1.345  0.17877
## Bilirubin       1.296e+00  4.412e-01   2.938  0.00330 **
## Albumin        -2.712e+00  9.037e-01  -3.001  0.00269 **
## Protime         7.681e-03  1.903e-02   0.404  0.68648
## Histology      -1.059e+00  7.851e-01  -1.349  0.17743
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 257.851  on 185  degrees of freedom
## Residual deviance:  78.127  on 170  degrees of freedom
## AIC: 110.13
##
## Number of Fisher Scoring iterations: 17
```

We can see confusion matrices for all models. The results are pretty similar.

```
confusionMatrix(table(pred.labels1, real.labels), positive = "1")
```

```
## Confusion Matrix and Statistics
##
##             real.labels
## pred.labels1  0  1
##            0 28  3
##            1  2  5
##
##                Accuracy : 0.8684
```

```
##                  95% CI : (0.7191, 0.9559)
##     No Information Rate : 0.7895
##     P-Value [Acc > NIR] : 0.1596
##
##                   Kappa : 0.5852
##
##  Mcnemar's Test P-Value : 1.0000
##
##             Sensitivity : 0.6250
##             Specificity : 0.9333
##          Pos Pred Value : 0.7143
##          Neg Pred Value : 0.9032
##              Prevalence : 0.2105
##          Detection Rate : 0.1316
##    Detection Prevalence : 0.1842
##       Balanced Accuracy : 0.7792
##
##        'Positive' Class : 1
##
```

```
confusionMatrix(table(pred.labels2, real.labels), positive = "1")
```

```
## Confusion Matrix and Statistics
##
##             real.labels
## pred.labels2  0  1
##            0 28  2
##            1  2  6
##
##                Accuracy : 0.8947
##                  95% CI : (0.752, 0.9706)
##     No Information Rate : 0.7895
##     P-Value [Acc > NIR] : 0.07462
##
##                   Kappa : 0.6833
##
##  Mcnemar's Test P-Value : 1.00000
##
##             Sensitivity : 0.7500
##             Specificity : 0.9333
##          Pos Pred Value : 0.7500
##          Neg Pred Value : 0.9333
##              Prevalence : 0.2105
##          Detection Rate : 0.1579
##    Detection Prevalence : 0.2105
##       Balanced Accuracy : 0.8417
##
##        'Positive' Class : 1
##
```

```
confusionMatrix(table(pred.labels3, real.labels), positive = "1")
```

```
## Confusion Matrix and Statistics
##
##             real.labels
```

```
## pred.labels3  0  1
##            0 28  2
##            1  2  6
##
##                Accuracy : 0.8947
##                  95% CI : (0.752, 0.9706)
##     No Information Rate : 0.7895
##     P-Value [Acc > NIR] : 0.07462
##
##                   Kappa : 0.6833
##
##  Mcnemar's Test P-Value : 1.00000
##
##             Sensitivity : 0.7500
##             Specificity : 0.9333
##          Pos Pred Value : 0.7500
##          Neg Pred Value : 0.9333
##              Prevalence : 0.2105
##          Detection Rate : 0.1579
##    Detection Prevalence : 0.2105
##       Balanced Accuracy : 0.8417
##
##        'Positive' Class : 1
##
```

```
confusionMatrix(table(pred.labels4, real.labels), positive = "1")
```

```
## Confusion Matrix and Statistics
##
##             real.labels
## pred.labels4  0  1
##            0 29  2
##            1  1  6
##
##                Accuracy : 0.9211
##                  95% CI : (0.7862, 0.9834)
##     No Information Rate : 0.7895
##     P-Value [Acc > NIR] : 0.02776
##
##                   Kappa : 0.7511
##
##  Mcnemar's Test P-Value : 1.00000
##
##             Sensitivity : 0.7500
##             Specificity : 0.9667
##          Pos Pred Value : 0.8571
##          Neg Pred Value : 0.9355
##              Prevalence : 0.2105
##          Detection Rate : 0.1579
##    Detection Prevalence : 0.1842
##       Balanced Accuracy : 0.8583
##
##        'Positive' Class : 1
##
```

```r
confusionMatrix(table(pred.labels5, real.labels), positive = "1")
```

```
## Confusion Matrix and Statistics
##
##             real.labels
## pred.labels5  0  1
##            0 28  2
##            1  2  6
##
##                Accuracy : 0.8947
##                  95% CI : (0.752, 0.9706)
##     No Information Rate : 0.7895
##     P-Value [Acc > NIR] : 0.07462
##
##                   Kappa : 0.6833
##
##  Mcnemar's Test P-Value : 1.00000
##
##             Sensitivity : 0.7500
##             Specificity : 0.9333
##          Pos Pred Value : 0.7500
##          Neg Pred Value : 0.9333
##              Prevalence : 0.2105
##          Detection Rate : 0.1579
##    Detection Prevalence : 0.2105
##       Balanced Accuracy : 0.8417
##
##        'Positive' Class : 1
##
```

```r
confusionMatrix(table(pred.labels6, real.labels), positive = "1")
```

```
## Confusion Matrix and Statistics
##
##             real.labels
## pred.labels6  0  1
##            0 28  2
##            1  2  6
##
##                Accuracy : 0.8947
##                  95% CI : (0.752, 0.9706)
##     No Information Rate : 0.7895
##     P-Value [Acc > NIR] : 0.07462
##
##                   Kappa : 0.6833
##
##  Mcnemar's Test P-Value : 1.00000
##
##             Sensitivity : 0.7500
##             Specificity : 0.9333
##          Pos Pred Value : 0.7500
##          Neg Pred Value : 0.9333
##              Prevalence : 0.2105
##          Detection Rate : 0.1579
```

```
##    Detection Prevalence : 0.2105
##       Balanced Accuracy : 0.8417
##
##        'Positive' Class : 1
##
```

```
confusionMatrix(table(pred.labels7, real.labels), positive = "1")
```

```
## Confusion Matrix and Statistics
##
##             real.labels
## pred.labels7  0  1
##            0 27  2
##            1  3  6
##
##                Accuracy : 0.8684
##                  95% CI : (0.7191, 0.9559)
##     No Information Rate : 0.7895
##     P-Value [Acc > NIR] : 0.1596
##
##                   Kappa : 0.6215
##
##  Mcnemar's Test P-Value : 1.0000
##
##             Sensitivity : 0.7500
##             Specificity : 0.9000
##          Pos Pred Value : 0.6667
##          Neg Pred Value : 0.9310
##              Prevalence : 0.2105
##          Detection Rate : 0.1579
##    Detection Prevalence : 0.2368
##       Balanced Accuracy : 0.8250
##
##        'Positive' Class : 1
##
```

```
confusionMatrix(table(pred.labels8, real.labels), positive = "1")
```

```
## Confusion Matrix and Statistics
##
##             real.labels
## pred.labels8  0  1
##            0 27  2
##            1  3  6
##
##                Accuracy : 0.8684
##                  95% CI : (0.7191, 0.9559)
##     No Information Rate : 0.7895
##     P-Value [Acc > NIR] : 0.1596
##
##                   Kappa : 0.6215
##
##  Mcnemar's Test P-Value : 1.0000
##
##             Sensitivity : 0.7500
```

```
##             Specificity : 0.9000
##          Pos Pred Value : 0.6667
##          Neg Pred Value : 0.9310
##              Prevalence : 0.2105
##          Detection Rate : 0.1579
##    Detection Prevalence : 0.2368
##       Balanced Accuracy : 0.8250
##
##        'Positive' Class : 1
##
```

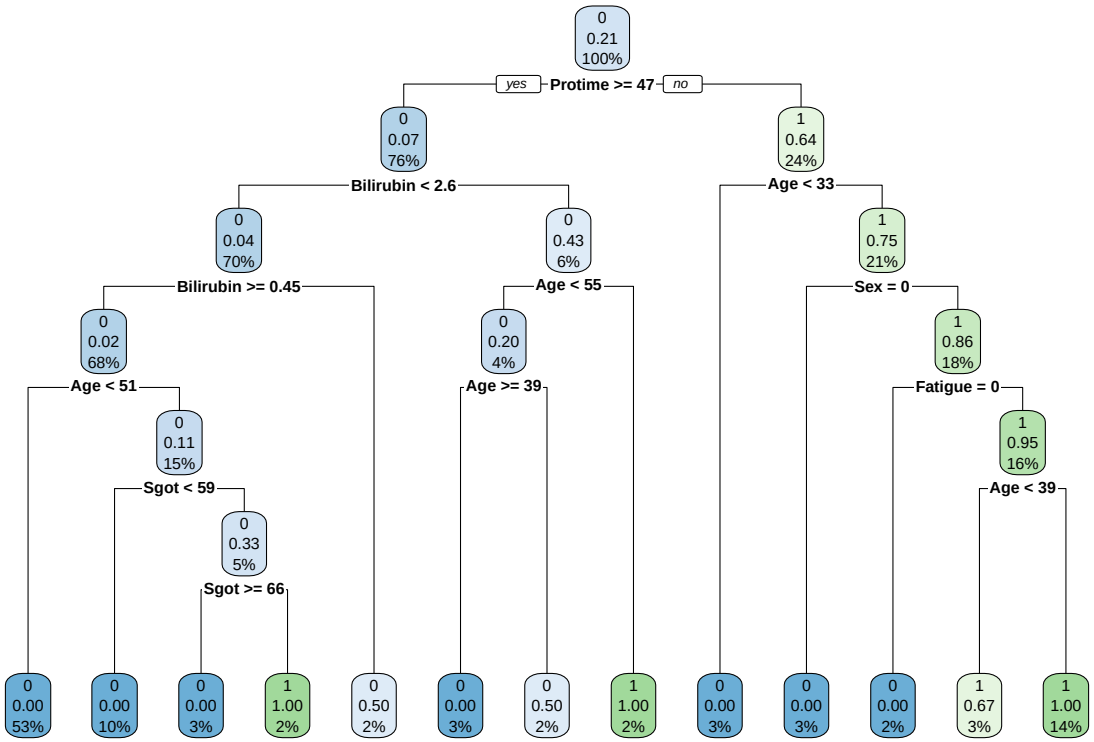
```
confusionMatrix(table(pred.labels9, real.labels), positive = "1")
```

```
## Confusion Matrix and Statistics
##
##             real.labels
## pred.labels9  0  1
##            0 29  2
##            1  1  6
##
##                Accuracy : 0.9211
##                  95% CI : (0.7862, 0.9834)
##     No Information Rate : 0.7895
##     P-Value [Acc > NIR] : 0.02776
##
##                   Kappa : 0.7511
##
##  Mcnemar's Test P-Value : 1.00000
##
##             Sensitivity : 0.7500
##             Specificity : 0.9667
##          Pos Pred Value : 0.8571
##          Neg Pred Value : 0.9355
##              Prevalence : 0.2105
##          Detection Rate : 0.1579
##    Detection Prevalence : 0.1842
##       Balanced Accuracy : 0.8583
##
##        'Positive' Class : 1
##
```

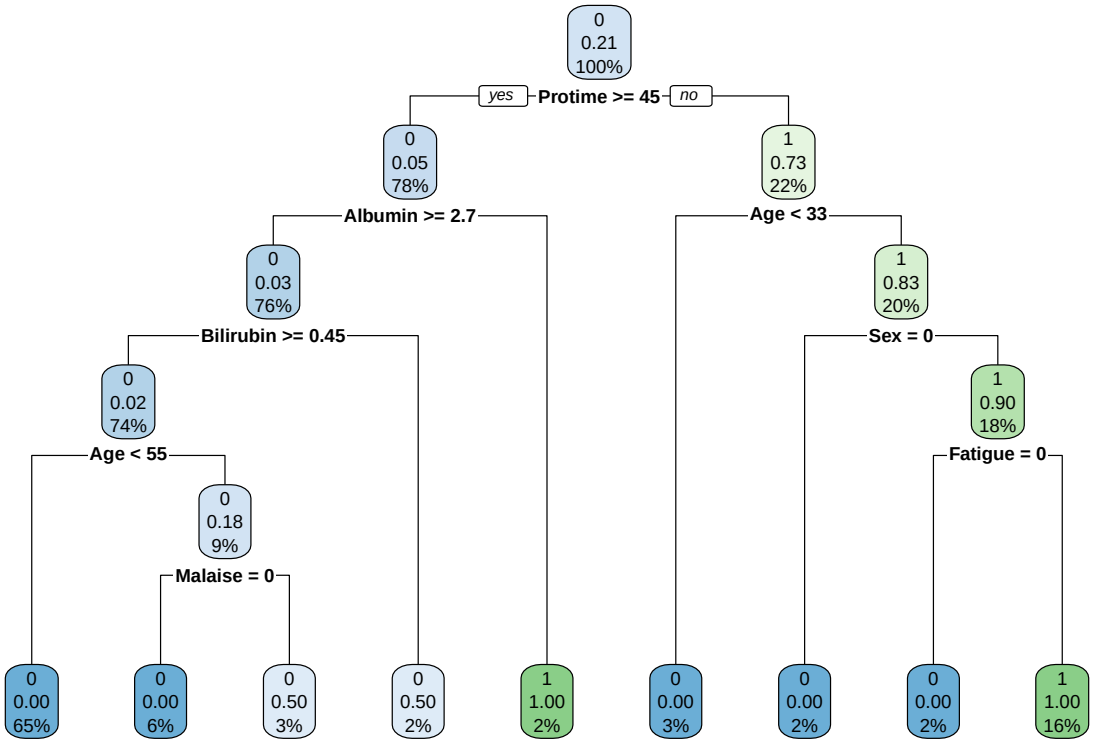
4.7. Random tree

We can see that full trees for different models and different datasets (for different imputation methods) are pretty different.

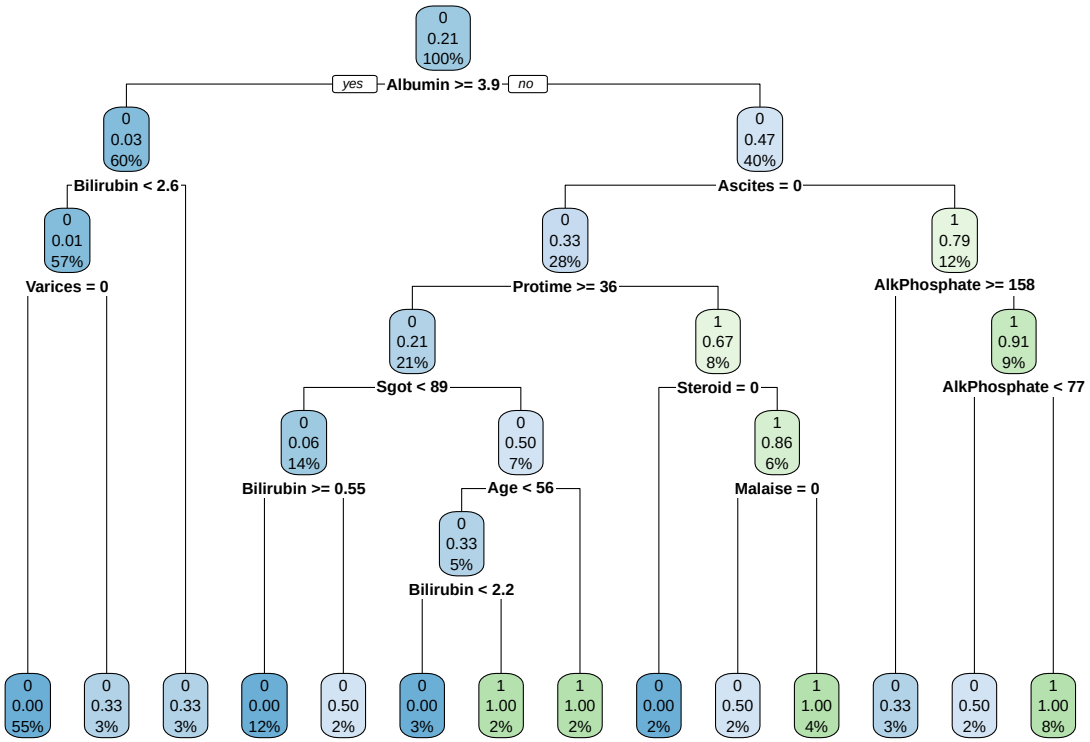
```
rpart.plot(full.tree1)
```



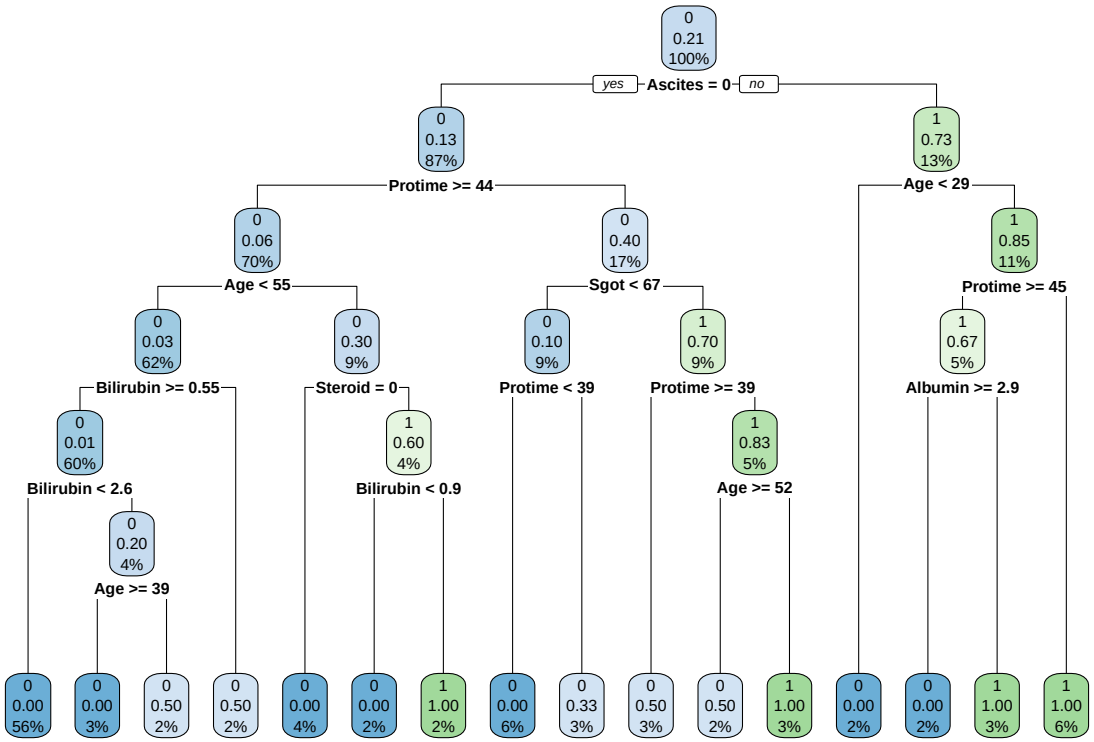

```
rpart.plot(full.tree2)
```



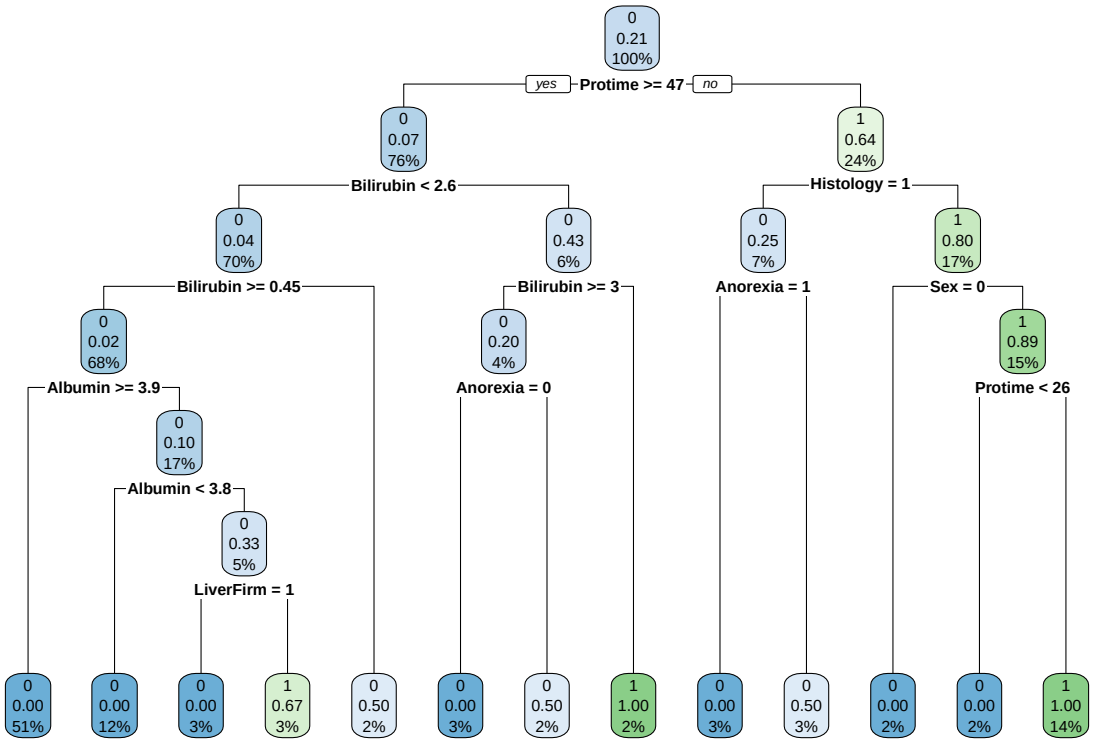

```
rpart.plot(full.tree3)
```



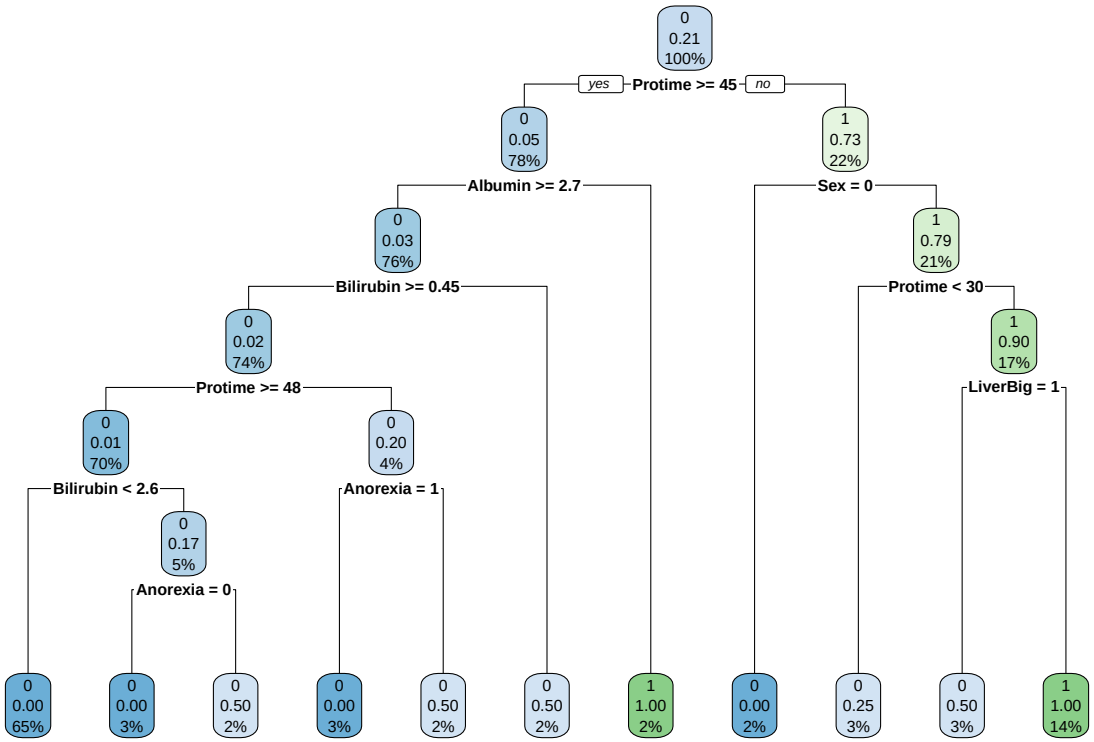

```
rpart.plot(full.tree4)
```



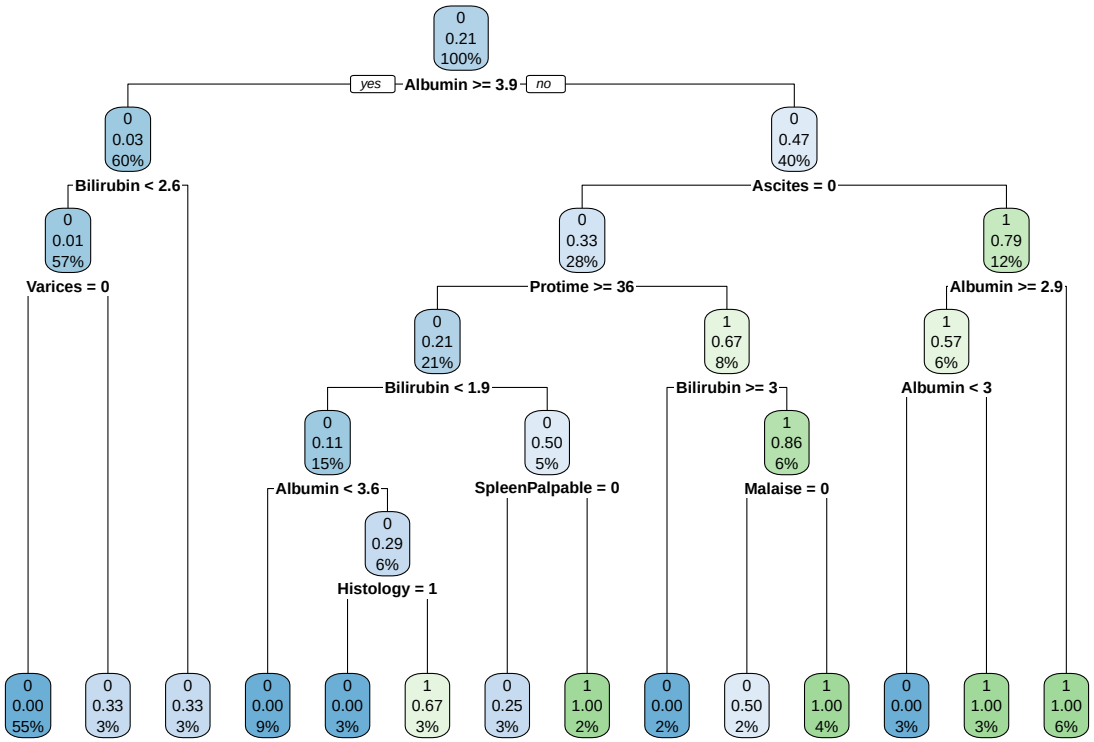

```
rpart.plot(full.tree5)
```



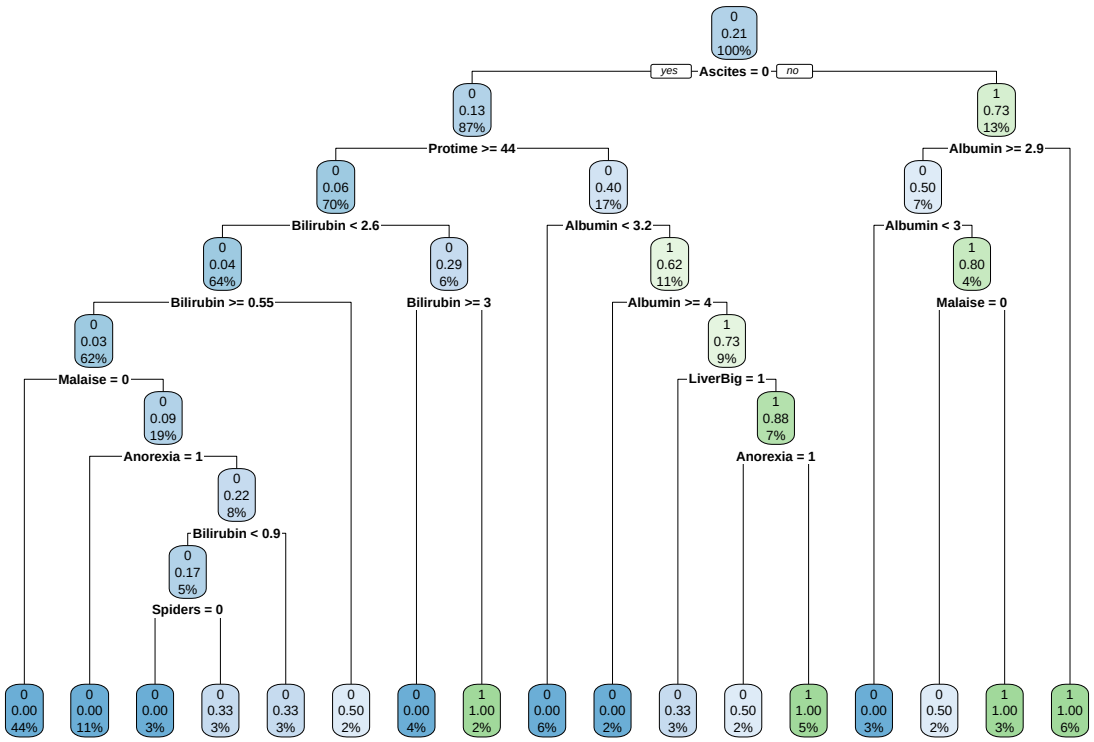

```
rpart.plot(full.tree6)
```




```
rpart.plot(full.tree7)
```

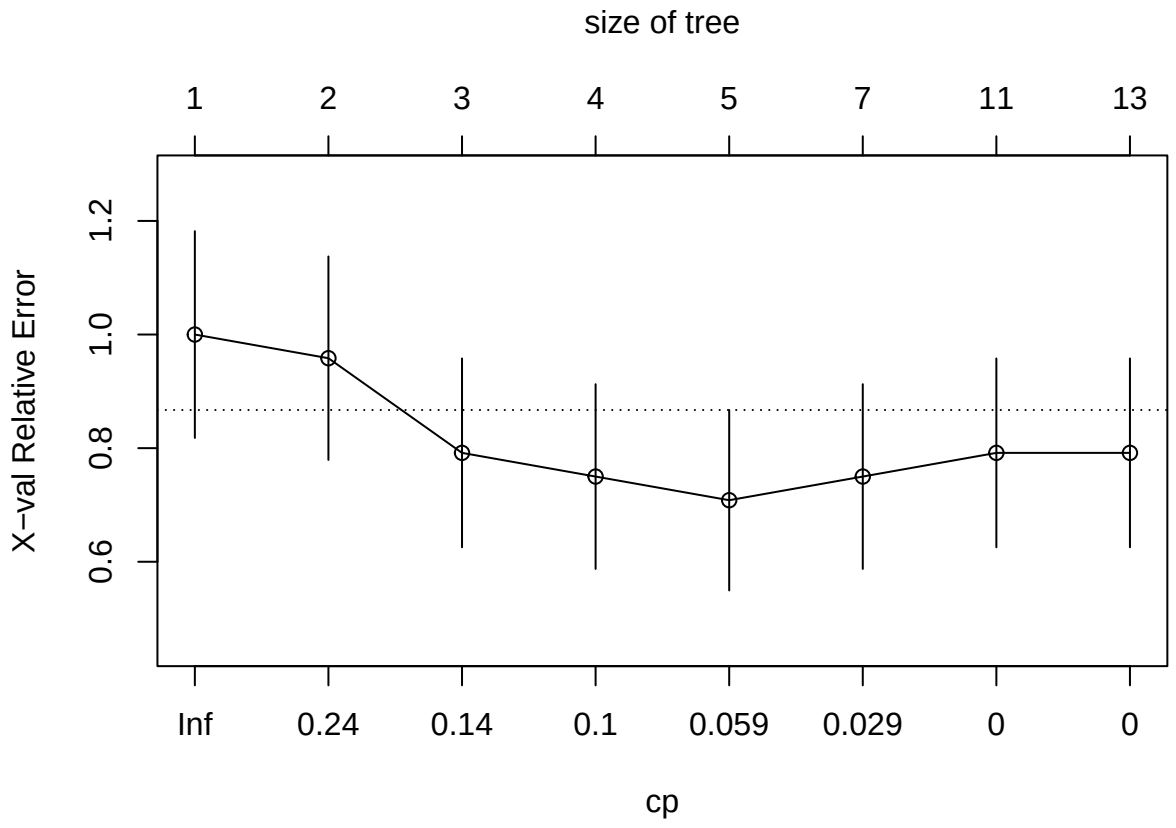



```
rpart.plot(full.tree8)
```



We can see plots and information about misclassification error.

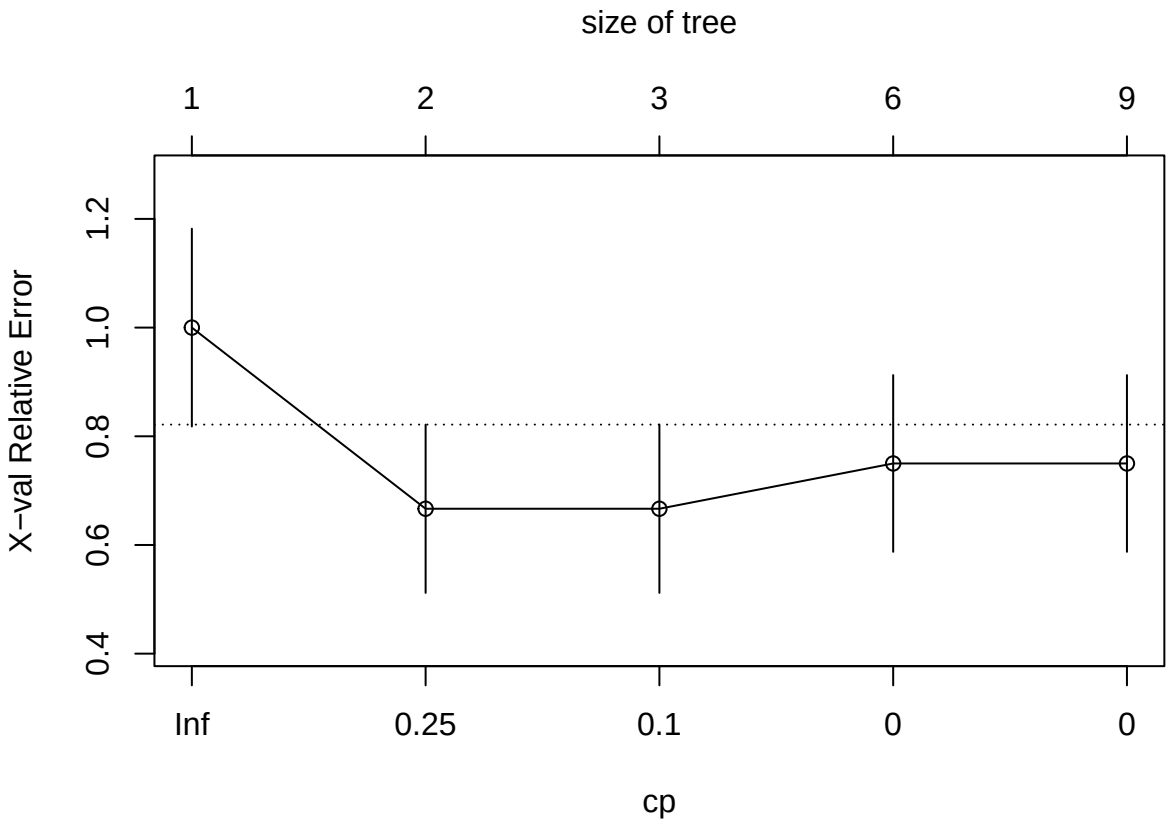
```
plotcp(full.tree1)
```




```
printcp(full.tree1)
```

```
##
## Classification tree:
## rpart(formula = mod1, data = train1, control = rpart.control(cp = -1,
##     minsplit = 5))
##
## Variables actually used in tree construction:
## [1] Age       Bilirubin Fatigue   Protime   Sex       Sgot
##
## Root node error: 24/117 = 0.20513
##
## n= 117
##
##          CP nsplit rel error  xerror    xstd
## 1  0.333333      0   1.00000 1.00000 0.18199
## 2  0.166667      1   0.66667 0.95833 0.17911
## 3  0.125000      2   0.50000 0.79167 0.16622
## 4  0.083333      3   0.37500 0.75000 0.16261
## 5  0.041667      4   0.29167 0.70833 0.15883
## 6  0.020833      6   0.20833 0.75000 0.16261
## 7  0.000000     10   0.12500 0.79167 0.16622
## 8 -1.000000     12   0.12500 0.79167 0.16622
```

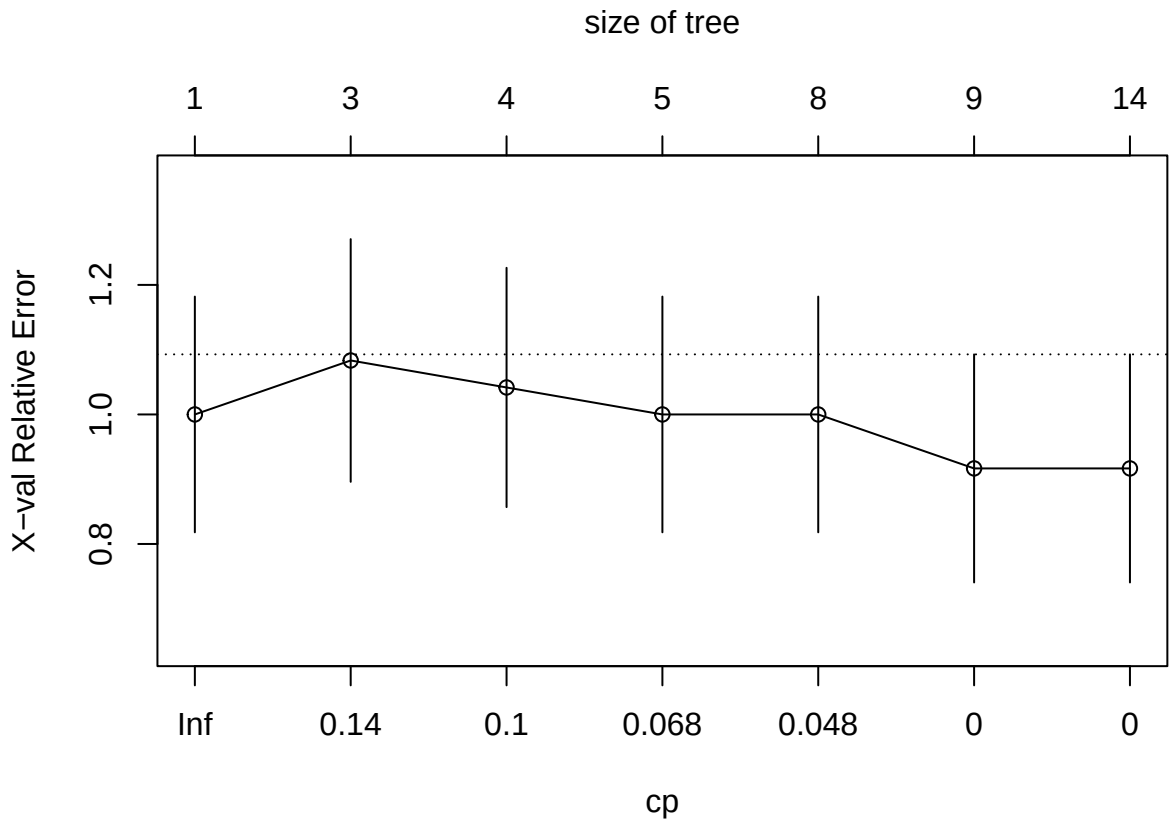
```
plotcp(full.tree2)
```




```
printcp(full.tree2)
```

```
##
## Classification tree:
## rpart(formula = mod1, data = train2, control = rpart.control(cp = -1,
##     minsplit = 5))
##
## Variables actually used in tree construction:
## [1] Age       Albumin   Bilirubin Fatigue   Malaise   Protime   Sex
##
## Root node error: 24/117 = 0.20513
##
## n= 117
##
##          CP nsplit rel error  xerror    xstd
## 1  0.500000      0     1.000 1.00000 0.18199
## 2  0.125000      1     0.500 0.66667 0.15485
## 3  0.083333      2     0.375 0.66667 0.15485
## 4  0.000000      5     0.125 0.75000 0.16261
## 5 -1.000000      8     0.125 0.75000 0.16261
```

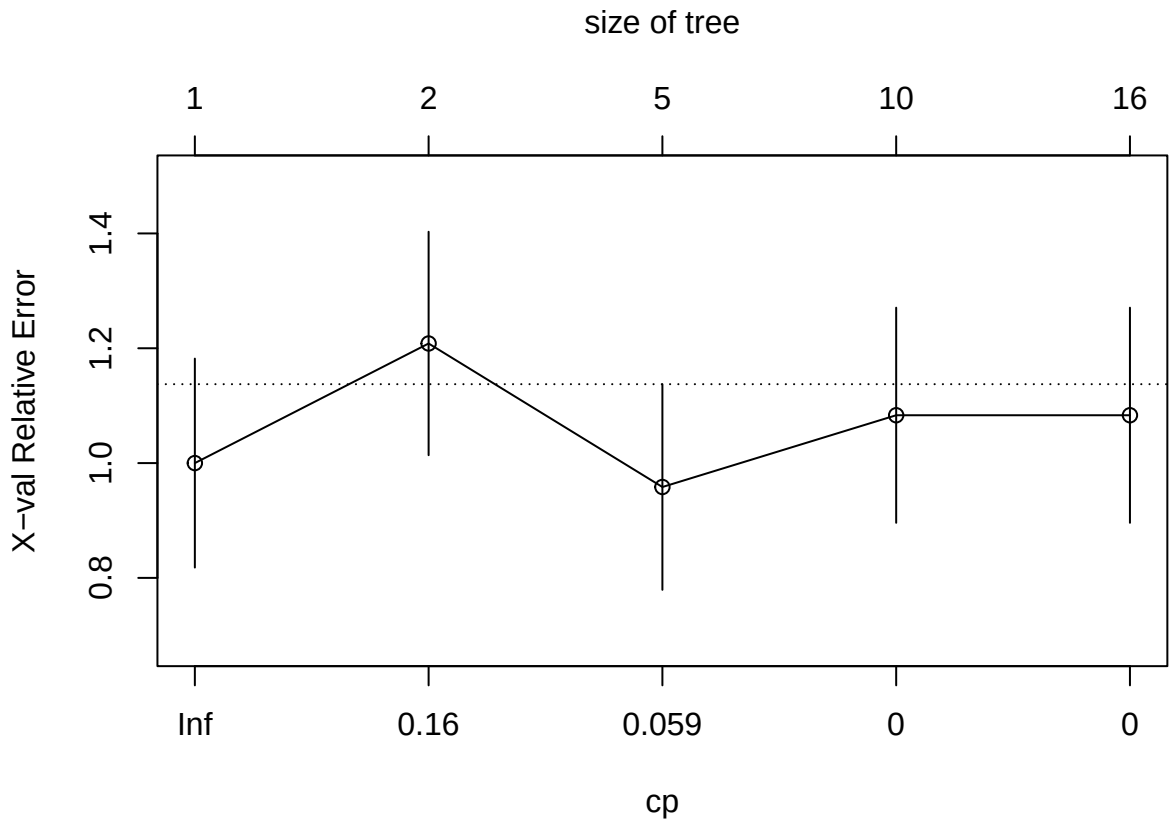
```
plotcp(full.tree3)
```




```
printcp(full.tree3)
```

```
##
## Classification tree:
## rpart(formula = mod1, data = train3, control = rpart.control(cp = -1,
##     minsplit = 5))
##
## Variables actually used in tree construction:
##  [1] Age          Albumin      AlkPhosphate Ascites      Bilirubin
##  [6] Malaise      Protime      Sgot         Steroid      Varices
##
## Root node error: 24/117 = 0.20513
##
## n= 117
##
##          CP nsplit rel error  xerror    xstd
## 1  0.166667      0   1.00000 1.00000 0.18199
## 2  0.125000      2   0.66667 1.08333 0.18737
## 3  0.083333      3   0.54167 1.04167 0.18474
## 4  0.055556      4   0.45833 1.00000 0.18199
## 5  0.041667      7   0.29167 1.00000 0.18199
## 6  0.000000      8   0.25000 0.91667 0.17610
## 7 -1.000000     13   0.25000 0.91667 0.17610
```

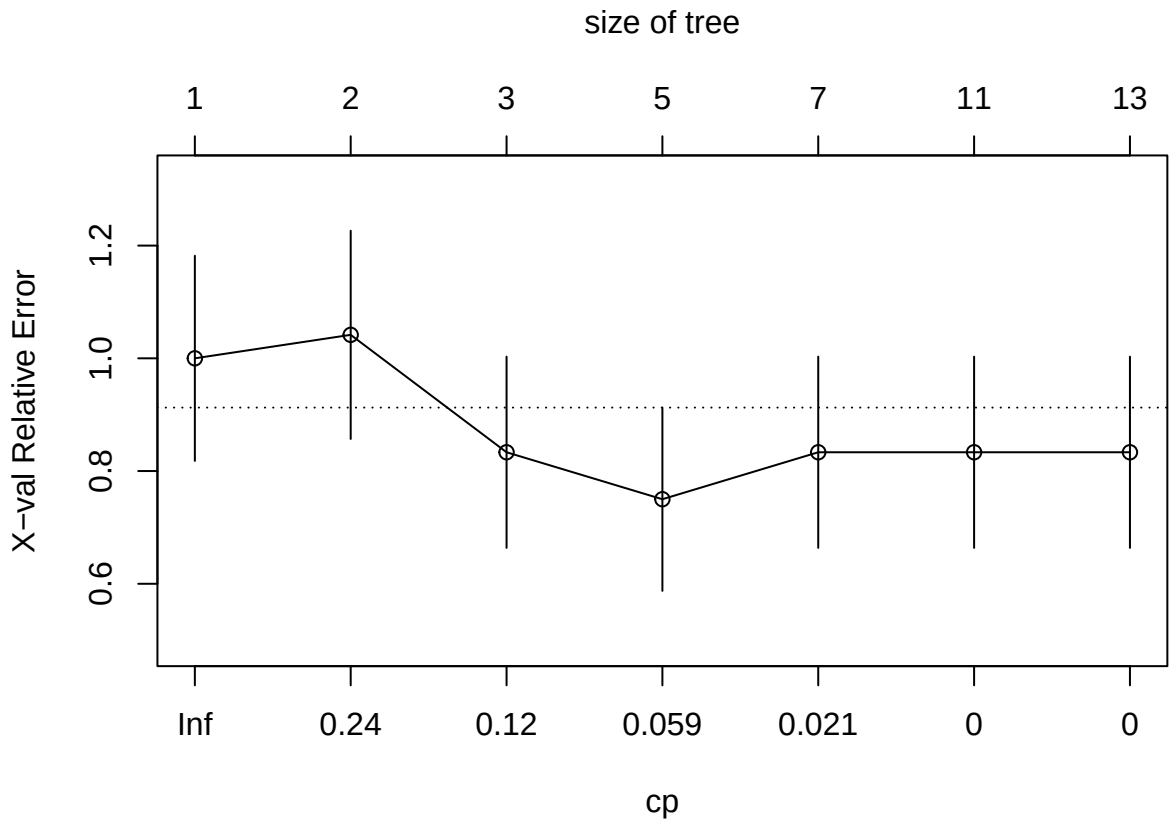
```
plotcp(full.tree4)
```




```
printcp(full.tree4)
```

```
##
## Classification tree:
## rpart(formula = mod1, data = train0, control = rpart.control(cp = -1,
##     minsplit = 5))
##
## Variables actually used in tree construction:
## [1] Age       Albumin   Ascites   Bilirubin Protime   Sgot      Steroid
##
## Root node error: 24/117 = 0.20513
##
## n= 117
##
##          CP nsplit rel error  xerror    xstd
## 1  0.291667      0   1.00000 1.00000 0.18199
## 2  0.083333      1   0.70833 1.20833 0.19460
## 3  0.041667      4   0.45833 0.95833 0.17911
## 4  0.000000      9   0.25000 1.08333 0.18737
## 5 -1.000000     15   0.25000 1.08333 0.18737
```

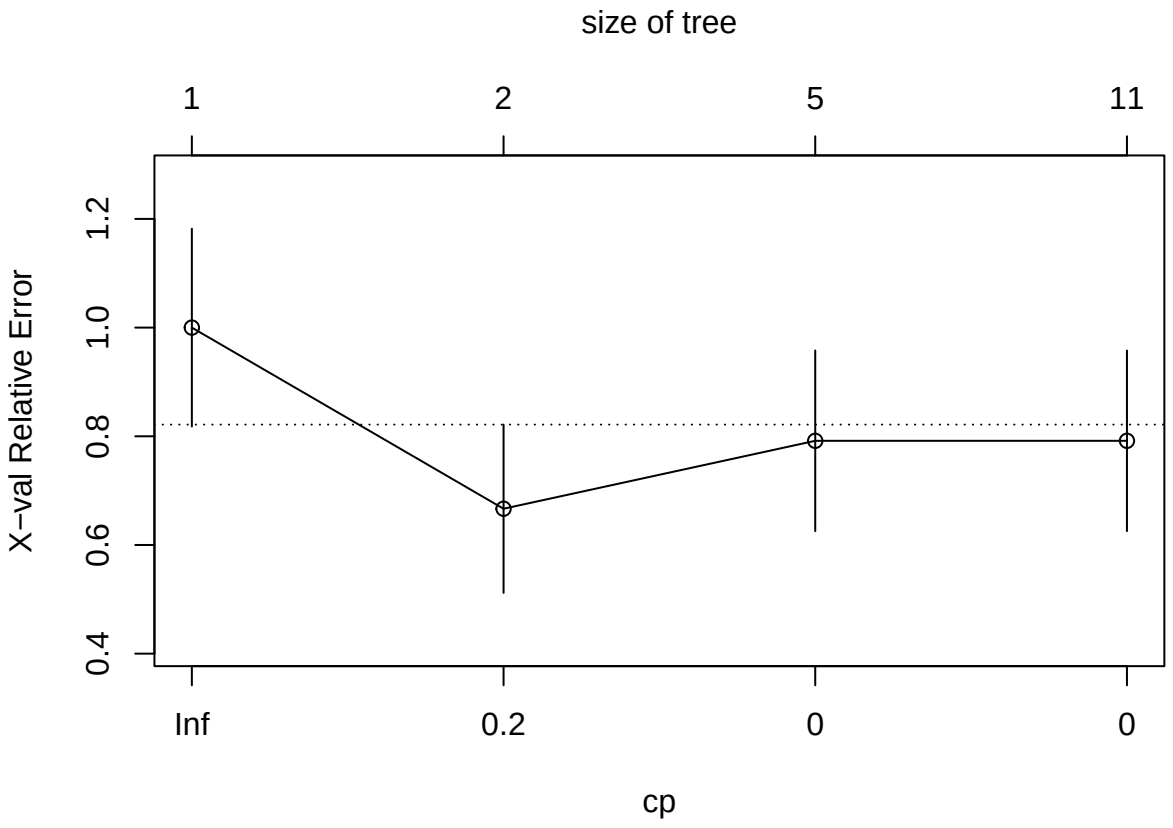
```
plotcp(full.tree5)
```




```
printcp(full.tree5)
```

```
##
## Classification tree:
## rpart(formula = mod2, data = train1, control = rpart.control(cp = -1,
##     minsplit = 5))
##
## Variables actually used in tree construction:
## [1] Albumin   Anorexia  Bilirubin Histology LiverFirm Protime   Sex
##
## Root node error: 24/117 = 0.20513
##
## n= 117
##
##          CP nsplit rel error  xerror    xstd
## 1  0.333333      0   1.00000 1.00000 0.18199
## 2  0.166667      1   0.66667 1.04167 0.18474
## 3  0.083333      2   0.50000 0.83333 0.16967
## 4  0.041667      4   0.33333 0.75000 0.16261
## 5  0.010417      6   0.25000 0.83333 0.16967
## 6  0.000000     10   0.20833 0.83333 0.16967
## 7 -1.000000     12   0.20833 0.83333 0.16967
```

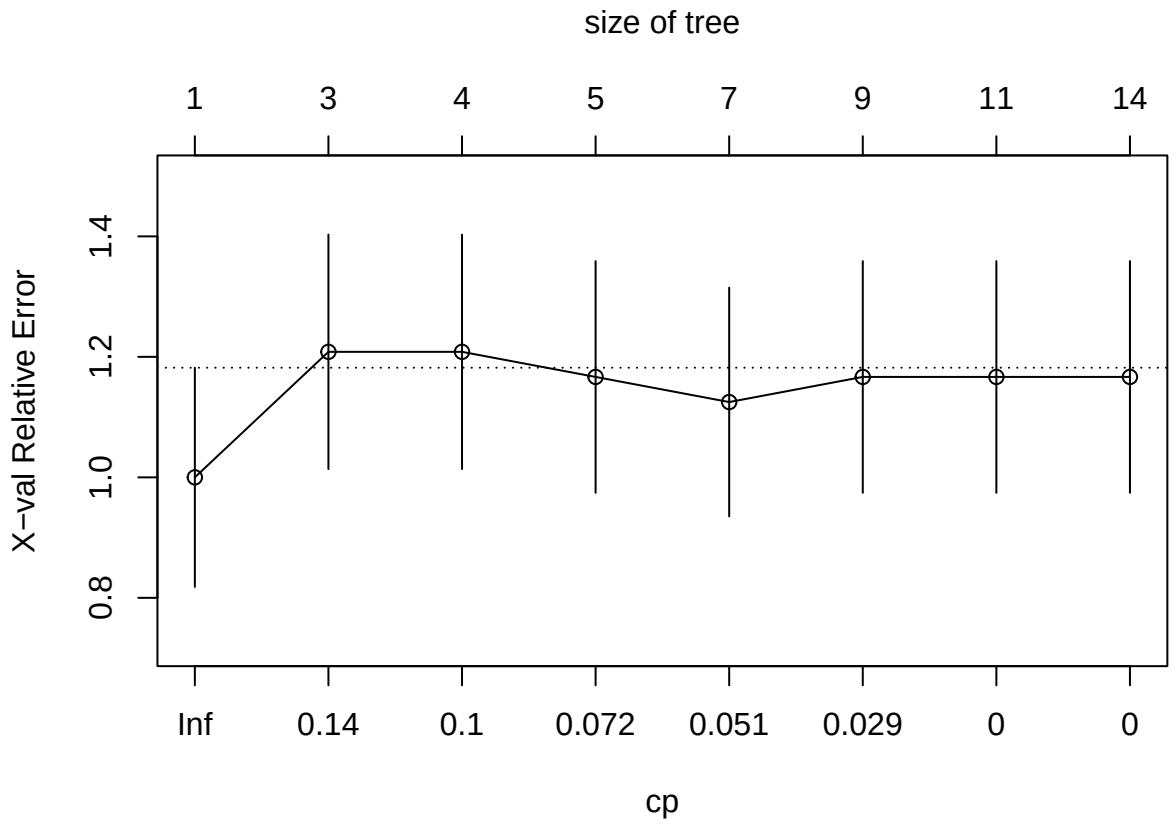
```
plotcp(full.tree6)
```




```
printcp(full.tree6)
```

```
##
## Classification tree:
## rpart(formula = mod2, data = train2, control = rpart.control(cp = -1,
##     minsplit = 5))
##
## Variables actually used in tree construction:
## [1] Albumin   Anorexia  Bilirubin LiverBig  Protime   Sex
##
## Root node error: 24/117 = 0.20513
##
## n= 117
##
##          CP nsplit rel error  xerror    xstd
## 1  0.500000      0      1.00 1.00000 0.18199
## 2  0.083333      1      0.50 0.66667 0.15485
## 3  0.000000      4      0.25 0.79167 0.16622
## 4 -1.000000     10      0.25 0.79167 0.16622
```

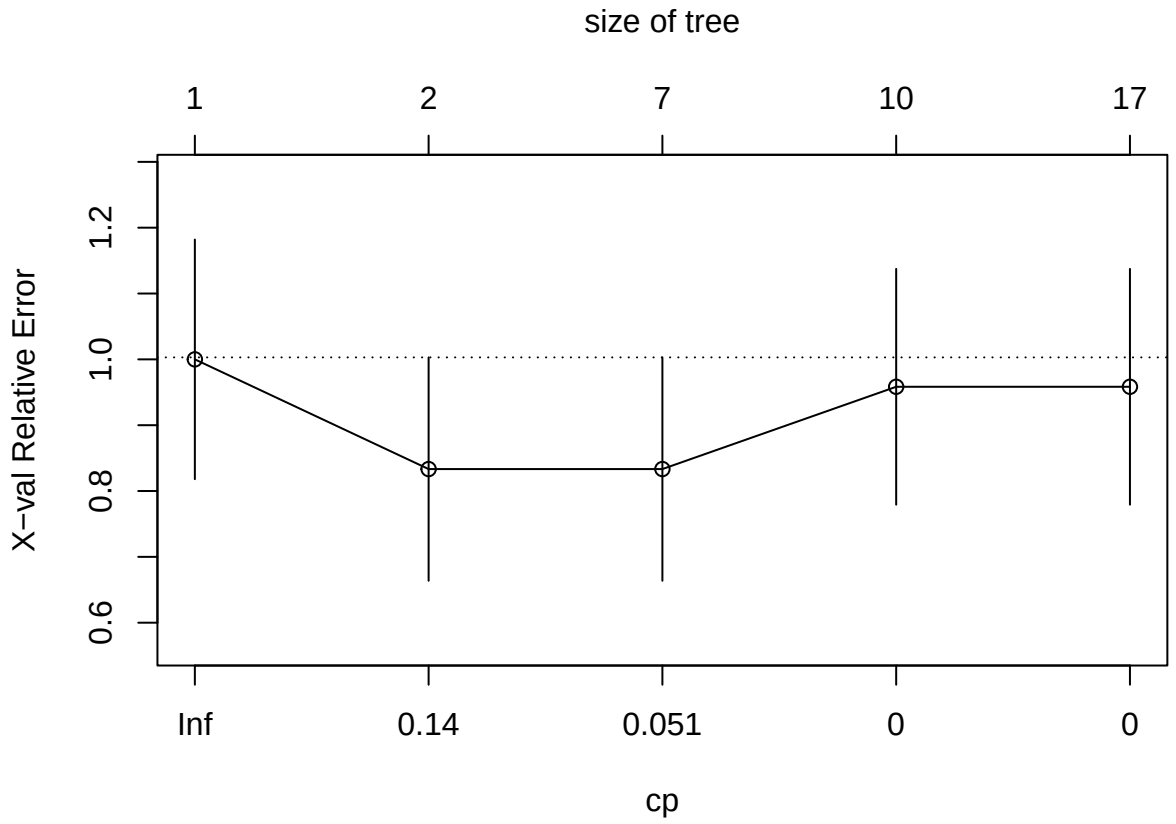
```
plotcp(full.tree7)
```




```
printcp(full.tree7)
```

```
##
## Classification tree:
## rpart(formula = mod2, data = train3, control = rpart.control(cp = -1,
##     minsplit = 5))
##
## Variables actually used in tree construction:
## [1] Albumin        Ascites        Bilirubin      Histology      Malaise
## [6] Protime        SpleenPalpable Varices
##
## Root node error: 24/117 = 0.20513
##
## n= 117
##
##          CP nsplit rel error xerror    xstd
## 1  0.166667      0   1.00000 1.0000 0.18199
## 2  0.125000      2   0.66667 1.2083 0.19460
## 3  0.083333      3   0.54167 1.2083 0.19460
## 4  0.062500      4   0.45833 1.1667 0.19230
## 5  0.041667      6   0.33333 1.1250 0.18989
## 6  0.020833      8   0.25000 1.1667 0.19230
## 7  0.000000     10   0.20833 1.1667 0.19230
## 8 -1.000000     13   0.20833 1.1667 0.19230
```

```
plotcp(full.tree8)
```




```
printcp(full.tree8)
```

```
##
## Classification tree:
## rpart(formula = mod2, data = train0, control = rpart.control(cp = -1,
##     minsplit = 5))
##
## Variables actually used in tree construction:
## [1] Albumin   Anorexia  Ascites   Bilirubin LiverBig  Malaise   Protime
## [8] Spiders
##
## Root node error: 24/117 = 0.20513
##
## n= 117
##
##          CP nsplit rel error  xerror    xstd
## 1  0.291667      0   1.00000 1.00000 0.18199
## 2  0.062500      1   0.70833 0.83333 0.16967
## 3  0.041667      6   0.37500 0.83333 0.16967
## 4  0.000000      9   0.25000 0.95833 0.17911
## 5 -1.000000     16   0.25000 0.95833 0.17911
```

We can see basic information about all pruned trees. There are both similar and different trees.

```
print(full.tree1.pruned)
```

```
## n= 117
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
## 1) root 117 24 0 (0.79487179 0.20512821)
##   2) Protime>=46.5 89  6 0 (0.93258427 0.06741573) *
##   3) Protime< 46.5 28 10 1 (0.35714286 0.64285714)
##     6) Age< 32.5 4  0 0 (1.00000000 0.00000000) *
##     7) Age>=32.5 24  6 1 (0.25000000 0.75000000) *
```

```
summary(full.tree1.pruned)
```

```
## Call:
## rpart(formula = mod1, data = train1, control = rpart.control(cp = -1,
##     minsplit = 5))
##   n= 117
##
##          CP nsplit rel error    xerror      xstd
## 1 0.3333333      0 1.0000000 1.0000000 0.1819881
## 2 0.1666667      1 0.6666667 0.9583333 0.1791116
## 3 0.1250000      2 0.5000000 0.7916667 0.1662209
##
## Variable importance
##      Protime      Albumin      Ascites          Age AlkPhosphate    Bilirubin
##           43           18           14           12            3            3
##      Varices     Anorexia
##            3            3
##
## Node number 1: 117 observations,    complexity param=0.3333333
##   predicted class=0  expected loss=0.2051282  P(node) =1
##     class counts:    93    24
##    probabilities: 0.795 0.205
##   left son=2 (89 obs) right son=3 (28 obs)
##   Primary splits:
##       Protime   < 46.5 to the right, improve=14.105690, (0 missing)
##       Albumin   < 3.88 to the right, improve= 9.633249, (0 missing)
##       Ascites   splits as  LR,       improve= 9.600905, (0 missing)
##       Bilirubin < 1.65 to the left,  improve= 6.304281, (0 missing)
##       Malaise   splits as  LR,       improve= 5.254887, (0 missing)
##   Surrogate splits:
##       Albumin      < 3.45 to the right, agree=0.846, adj=0.357, (0 split)
##       Ascites      splits as  LR,       agree=0.838, adj=0.321, (0 split)
##       Varices      splits as  LR,       agree=0.778, adj=0.071, (0 split)
##       Bilirubin    < 3.1  to the left,  agree=0.778, adj=0.071, (0 split)
##       AlkPhosphate < 159  to the left,  agree=0.778, adj=0.071, (0 split)
##
## Node number 2: 89 observations
##   predicted class=0  expected loss=0.06741573  P(node) =0.7606838
##     class counts:    83     6
##    probabilities: 0.933 0.067
##
```

```
## Node number 3: 28 observations,    complexity param=0.1666667
##   predicted class=1  expected loss=0.3571429  P(node) =0.2393162
##     class counts:    10    18
##    probabilities: 0.357 0.643
##   left son=6 (4 obs) right son=7 (24 obs)
##   Primary splits:
##       Age       < 32.5 to the left,  improve=3.857143, (0 missing)
##       Histology splits as  RL,       improve=3.457143, (0 missing)
##       Sex       splits as  LR,       improve=2.777143, (0 missing)
##       Albumin   < 3.9  to the right, improve=2.777143, (0 missing)
##       Anorexia  splits as  RL,       improve=2.380952, (0 missing)
##   Surrogate splits:
##       Anorexia splits as  RL,       agree=0.893, adj=0.25, (0 split)
##       Albumin  < 3.9  to the right, agree=0.893, adj=0.25, (0 split)
##
## Node number 6: 4 observations
##   predicted class=0  expected loss=0  P(node) =0.03418803
##     class counts:     4     0
##    probabilities: 1.000 0.000
##
## Node number 7: 24 observations
##   predicted class=1  expected loss=0.25  P(node) =0.2051282
##     class counts:     6    18
##    probabilities: 0.250 0.750
```

```
print(full.tree2.pruned)
```

```
## n= 117
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 117 24 0 (0.79487179 0.20512821)
##    2) Protime>=45 91  5 0 (0.94505495 0.05494505)
##      4) Albumin>=2.65 89  3 0 (0.96629213 0.03370787) *
##      5) Albumin< 2.65 2  0 1 (0.00000000 1.00000000) *
##    3) Protime< 45 26  7 1 (0.26923077 0.73076923)
##      6) Age< 32.5 3  0 0 (1.00000000 0.00000000) *
##      7) Age>=32.5 23  4 1 (0.17391304 0.82608696)
##       14) Sex=0 2  0 0 (1.00000000 0.00000000) *
##       15) Sex=1 21  2 1 (0.09523810 0.90476190)
##         30) Fatigue=0 2  0 0 (1.00000000 0.00000000) *
##         31) Fatigue=1 19  0 1 (0.00000000 1.00000000) *
```

```
summary(full.tree2.pruned)
```

```
## Call:
## rpart(formula = mod1, data = train2, control = rpart.control(cp = -1,
##     minsplit = 5))
##   n= 117
##
##           CP nsplit rel error    xerror      xstd
## 1 0.50000000      0     1.000 1.0000000 0.1819881
## 2 0.12500000      1     0.500 0.6666667 0.1548519
## 3 0.08333333      2     0.375 0.6666667 0.1548519
```

```
## 4 0.08300000      5     0.125 0.7500000 0.1626109
##
## Variable importance
##      Protime      Albumin          Age      Fatigue      Ascites          Sex
##           41           19            8            8            8            7
## AlkPhosphate    Bilirubin
##            6            3
##
## Node number 1: 117 observations,    complexity param=0.5
##   predicted class=0  expected loss=0.2051282  P(node) =1
##     class counts:    93    24
##    probabilities: 0.795 0.205
##   left son=2 (91 obs) right son=3 (26 obs)
##   Primary splits:
##       Protime   < 45       to the right, improve=18.472530, (0 missing)
##       Albumin   < 3.85     to the right, improve=11.775090, (0 missing)
##       Ascites   splits as  LR,          improve= 9.600905, (0 missing)
##       Bilirubin < 1.057866 to the left,  improve= 5.553846, (0 missing)
##       Malaise   splits as  LR,          improve= 5.254887, (0 missing)
##   Surrogate splits:
##       Albumin      < 3.618841 to the right, agree=0.838, adj=0.269, (0 split)
##       Ascites      splits as  LR,          agree=0.821, adj=0.192, (0 split)
##       Bilirubin    < 3.35     to the left,  agree=0.795, adj=0.077, (0 split)
##       AlkPhosphate < 159      to the left,  agree=0.795, adj=0.077, (0 split)
##
## Node number 2: 91 observations,    complexity param=0.08333333
##   predicted class=0  expected loss=0.05494505  P(node) =0.7777778
##     class counts:    86     5
##    probabilities: 0.945 0.055
##   left son=4 (89 obs) right son=5 (2 obs)
##   Primary splits:
##       Albumin   < 2.65     to the right, improve=3.6527970, (0 missing)
##       Ascites   splits as  LR,          improve=1.2598520, (0 missing)
##       Protime   < 47.5     to the right, improve=0.9956475, (0 missing)
##       Age       < 55       to the left,  improve=0.9377289, (0 missing)
##       Bilirubin < 1.65     to the left,  improve=0.8401598, (0 missing)
##
## Node number 3: 26 observations,    complexity param=0.125
##   predicted class=1  expected loss=0.2692308  P(node) =0.2222222
##     class counts:     7    19
##    probabilities: 0.269 0.731
##   left son=6 (3 obs) right son=7 (23 obs)
##   Primary splits:
##       Age          < 32.5     to the left,  improve=3.622074, (0 missing)
##       AlkPhosphate < 157.5    to the right, improve=2.925214, (0 missing)
##       Sex          splits as  LR,          improve=2.314103, (0 missing)
##       Fatigue      splits as  LR,          improve=2.314103, (0 missing)
##       Albumin      < 3.95     to the right, improve=2.314103, (0 missing)
##
## Node number 4: 89 observations
##   predicted class=0  expected loss=0.03370787  P(node) =0.7606838
##     class counts:    86     3
##    probabilities: 0.966 0.034
##
```

```
## Node number 5: 2 observations
##   predicted class=1  expected loss=0  P(node) =0.01709402
##     class counts:     0     2
##    probabilities: 0.000 1.000
##
## Node number 6: 3 observations
##   predicted class=0  expected loss=0  P(node) =0.02564103
##     class counts:     3     0
##    probabilities: 1.000 0.000
##
## Node number 7: 23 observations,    complexity param=0.08333333
##   predicted class=1  expected loss=0.173913  P(node) =0.1965812
##     class counts:     4    19
##    probabilities: 0.174 0.826
##   left son=14 (2 obs) right son=15 (21 obs)
##   Primary splits:
##       Sex          splits as  LR,           improve=2.989648, (0 missing)
##       Fatigue      splits as  LR,           improve=2.989648, (0 missing)
##       Malaise      splits as  LR,           improve=2.164251, (0 missing)
##       AlkPhosphate < 159.5    to the right, improve=1.726343, (0 missing)
##       Bilirubin    < 1.057866 to the left,  improve=1.305124, (0 missing)
##   Surrogate splits:
##       AlkPhosphate < 171      to the right, agree=0.957, adj=0.5, (0 split)
##
## Node number 14: 2 observations
##   predicted class=0  expected loss=0  P(node) =0.01709402
##     class counts:     2     0
##    probabilities: 1.000 0.000
##
## Node number 15: 21 observations,    complexity param=0.08333333
##   predicted class=1  expected loss=0.0952381  P(node) =0.1794872
##     class counts:     2    19
##    probabilities: 0.095 0.905
##   left son=30 (2 obs) right son=31 (19 obs)
##   Primary splits:
##       Fatigue   splits as  LR,           improve=3.6190480, (0 missing)
##       Bilirubin < 1.057866 to the left,  improve=0.9523810, (0 missing)
##       Sgot      < 49.5     to the left,  improve=0.9523810, (0 missing)
##       Malaise   splits as  LR,           improve=0.7619048, (0 missing)
##       Spiders   splits as  LR,           improve=0.7619048, (0 missing)
##
## Node number 30: 2 observations
##   predicted class=0  expected loss=0  P(node) =0.01709402
##     class counts:     2     0
##    probabilities: 1.000 0.000
##
## Node number 31: 19 observations
##   predicted class=1  expected loss=0  P(node) =0.1623932
##     class counts:     0    19
##    probabilities: 0.000 1.000
```

```
print(full.tree3.pruned)
```

```
## n= 117
##
```

```
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 117 24 0 (0.79487179 0.20512821)
##    2) Albumin>=3.85 70  2 0 (0.97142857 0.02857143) *
##    3) Albumin< 3.85 47 22 0 (0.53191489 0.46808511)
##      6) Ascites=0 33 11 0 (0.66666667 0.33333333)
##       12) Protime>=35.5 24  5 0 (0.79166667 0.20833333) *
##       13) Protime< 35.5 9  3 1 (0.33333333 0.66666667)
##         26) Steroid=0 2  0 0 (1.00000000 0.00000000) *
##         27) Steroid=1 7  1 1 (0.14285714 0.85714286) *
##      7) Ascites=1 14  3 1 (0.21428571 0.78571429) *
```

```
summary(full.tree3.pruned)
```

```
## Call:
## rpart(formula = mod1, data = train3, control = rpart.control(cp = -1,
##     minsplit = 5))
##   n= 117
##
##           CP nsplit rel error   xerror      xstd
## 1 0.16666667      0 1.0000000 1.000000 0.1819881
## 2 0.12500000      2 0.6666667 1.083333 0.1873714
## 3 0.08333333      3 0.5416667 1.041667 0.1847395
## 4 0.08300000      4 0.4583333 1.000000 0.1819881
##
## Variable importance
##      Albumin      Protime      Ascites    Bilirubin          Age AlkPhosphate
##           28           16           16           15            8            8
##      Steroid    LiverFirm         Sgot
##            5            1            1
##
## Node number 1: 117 observations,    complexity param=0.1666667
##   predicted class=0  expected loss=0.2051282  P(node) =1
##     class counts:    93    24
##    probabilities: 0.795 0.205
##   left son=2 (70 obs) right son=3 (47 obs)
##   Primary splits:
##       Albumin   < 3.85  to the right, improve=10.863880, (0 missing)
##       Ascites   splits as  LR,        improve= 9.600905, (0 missing)
##       Protime   < 44.5  to the right, improve= 8.835664, (0 missing)
##       Bilirubin < 1.65  to the left,  improve= 6.304281, (0 missing)
##       Malaise   splits as  LR,        improve= 5.254887, (0 missing)
##   Surrogate splits:
##       Protime      < 44.5  to the right, agree=0.761, adj=0.404, (0 split)
##       Bilirubin    < 1.35  to the left,  agree=0.718, adj=0.298, (0 split)
##       AlkPhosphate < 102.5 to the left,  agree=0.718, adj=0.298, (0 split)
##       Ascites      splits as  LR,        agree=0.709, adj=0.277, (0 split)
##       Age          < 44.5  to the left,  agree=0.701, adj=0.255, (0 split)
##
## Node number 2: 70 observations
##   predicted class=0  expected loss=0.02857143  P(node) =0.5982906
##     class counts:    68     2
##    probabilities: 0.971 0.029
##
```

```
## Node number 3: 47 observations,    complexity param=0.1666667
##   predicted class=0  expected loss=0.4680851  P(node) =0.4017094
##     class counts:    25    22
##    probabilities: 0.532 0.468
##   left son=6 (33 obs) right son=7 (14 obs)
##   Primary splits:
##       Ascites   splits as  LR,        improve=4.023303, (0 missing)
##       Protime   < 35.5  to the right, improve=3.855868, (0 missing)
##       Bilirubin < 3.7   to the left,  improve=3.166160, (0 missing)
##       Albumin   < 2.65  to the right, improve=3.166160, (0 missing)
##       Malaise   splits as  LR,        improve=2.525134, (0 missing)
##   Surrogate splits:
##       Albumin   < 2.65  to the right, agree=0.809, adj=0.357, (0 split)
##       Bilirubin < 3.7   to the left,  agree=0.766, adj=0.214, (0 split)
##       Sgot      < 19.5  to the right, agree=0.745, adj=0.143, (0 split)
##
## Node number 6: 33 observations,    complexity param=0.125
##   predicted class=0  expected loss=0.3333333  P(node) =0.2820513
##     class counts:    22    11
##    probabilities: 0.667 0.333
##   left son=12 (24 obs) right son=13 (9 obs)
##   Primary splits:
##       Protime      < 35.5  to the right, improve=2.750000, (0 missing)
##       Sgot         < 66.5  to the left,  improve=2.200000, (0 missing)
##       AlkPhosphate < 72.5  to the right, improve=1.580460, (0 missing)
##       Malaise      splits as  LR,        improve=1.350877, (0 missing)
##       Sex          splits as  LR,        improve=1.011494, (0 missing)
##   Surrogate splits:
##       Age          < 59.5  to the left,  agree=0.818, adj=0.333, (0 split)
##       LiverFirm    splits as  RL,        agree=0.788, adj=0.222, (0 split)
##       AlkPhosphate < 63.5  to the right, agree=0.758, adj=0.111, (0 split)
##
## Node number 7: 14 observations
##   predicted class=1  expected loss=0.2142857  P(node) =0.1196581
##     class counts:     3    11
##    probabilities: 0.214 0.786
##
## Node number 12: 24 observations
##   predicted class=0  expected loss=0.2083333  P(node) =0.2051282
##     class counts:    19     5
##    probabilities: 0.792 0.208
##
## Node number 13: 9 observations,    complexity param=0.08333333
##   predicted class=1  expected loss=0.3333333  P(node) =0.07692308
##     class counts:     3     6
##    probabilities: 0.333 0.667
##   left son=26 (2 obs) right son=27 (7 obs)
##   Primary splits:
##       Steroid      splits as  LR,        improve=2.285714, (0 missing)
##       Bilirubin    < 3     to the right, improve=2.285714, (0 missing)
##       AlkPhosphate < 102   to the right, improve=1.600000, (0 missing)
##       Malaise      splits as  LR,        improve=1.000000, (0 missing)
##       Sgot         < 57    to the left,  improve=1.000000, (0 missing)
##   Surrogate splits:
```

```
##       Bilirubin < 3     to the right, agree=1, adj=1, (0 split)
##
## Node number 26: 2 observations
##   predicted class=0  expected loss=0  P(node) =0.01709402
##     class counts:     2     0
##    probabilities: 1.000 0.000
##
## Node number 27: 7 observations
##   predicted class=1  expected loss=0.1428571  P(node) =0.05982906
##     class counts:     1     6
##    probabilities: 0.143 0.857
```

```
print(full.tree4.pruned)
```

```
## n= 117
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 117 24 0 (0.79487179 0.20512821)
##    2) Ascites=0 102 13 0 (0.87254902 0.12745098)
##      4) Protime>=43.5 82  5 0 (0.93902439 0.06097561)
##        8) Age< 55 72  2 0 (0.97222222 0.02777778) *
##        9) Age>=55 10  3 0 (0.70000000 0.30000000)
##         18) Steroid=0 5  0 0 (1.00000000 0.00000000) *
##         19) Steroid=1 5  2 1 (0.40000000 0.60000000)
##           38) Bilirubin< 0.9 2  0 0 (1.00000000 0.00000000) *
##           39) Bilirubin>=0.9 2  0 1 (0.00000000 1.00000000) *
##      5) Protime< 43.5 20  8 0 (0.60000000 0.40000000)
##       10) Sgot< 66.5 10  1 0 (0.90000000 0.10000000) *
##       11) Sgot>=66.5 10  3 1 (0.30000000 0.70000000) *
##    3) Ascites=1 15  4 1 (0.26666667 0.73333333)
##      6) Age< 29 2  0 0 (1.00000000 0.00000000) *
##      7) Age>=29 13  2 1 (0.15384615 0.84615385)
##       14) Protime>=44.5 6  2 1 (0.33333333 0.66666667)
##         28) Albumin>=2.85 2  0 0 (1.00000000 0.00000000) *
##         29) Albumin< 2.85 4  0 1 (0.00000000 1.00000000) *
##       15) Protime< 44.5 7  0 1 (0.00000000 1.00000000) *
```

```
summary(full.tree4.pruned)
```

```
## Call:
## rpart(formula = mod1, data = train0, control = rpart.control(cp = -1,
##     minsplit = 5))
##   n= 117
##
##           CP nsplit rel error    xerror      xstd
## 1 0.29166667      0 1.0000000 1.0000000 0.1819881
## 2 0.08333333      1 0.7083333 1.2083333 0.1945970
## 3 0.04166667      4 0.4583333 0.9583333 0.1791116
## 4 0.04100000      9 0.2500000 1.0833333 0.1873714
##
## Variable importance
##        Ascites           Sgot            Age        Albumin        Protime
##             21             17             13             10             10
```

```
##      Bilirubin        Steroid        Malaise        Spiders   AlkPhosphate
##              5              4              4              4              3
##            Sex     Antivirals        Fatigue       Anorexia      Histology
##              2              2              2              1              1
## SpleenPalpable
##              1
##
## Node number 1: 117 observations,    complexity param=0.2916667
##   predicted class=0  expected loss=0.2051282  P(node) =1
##     class counts:    93    24
##    probabilities: 0.795 0.205
##   left son=2 (102 obs) right son=3 (15 obs)
##   Primary splits:
##       Ascites   splits as  LR,        improve=9.442212, (3 missing)
##       Albumin   < 3.85  to the right, improve=8.759234, (15 missing)
##       Protime   < 46.5  to the right, improve=8.727273, (51 missing)
##       Bilirubin < 1.65  to the left,  improve=5.578130, (6 missing)
##       Malaise   splits as  LR,        improve=5.184493, (1 missing)
##
## Node number 2: 102 observations,    complexity param=0.08333333
##   predicted class=0  expected loss=0.127451  P(node) =0.8717949
##     class counts:    89    13
##    probabilities: 0.873 0.127
##   left son=4 (82 obs) right son=5 (20 obs)
##   Primary splits:
##       Protime   < 43.5  to the right, improve=3.234583, (49 missing)
##       Albumin   < 3.85  to the right, improve=2.191699, (15 missing)
##       Spiders   splits as  LR,        improve=2.125971, (3 missing)
##       Malaise   splits as  LR,        improve=1.895783, (1 missing)
##       Bilirubin < 0.55  to the right, improve=1.887769, (6 missing)
##   Surrogate splits:
##       Albumin        < 3.45  to the right, agree=0.830, adj=0.182, (35 split)
##       SpleenPalpable splits as  LR,        agree=0.811, adj=0.091, (13 split)
##       Bilirubin      < 2.25  to the left,  agree=0.811, adj=0.091, (0 split)
##       AlkPhosphate   < 156.5 to the left,  agree=0.811, adj=0.091, (0 split)
##
## Node number 3: 15 observations,    complexity param=0.08333333
##   predicted class=1  expected loss=0.2666667  P(node) =0.1282051
##     class counts:     4    11
##    probabilities: 0.267 0.733
##   left son=6 (2 obs) right son=7 (13 obs)
##   Primary splits:
##       Age          < 29    to the left,  improve=2.482051, (0 missing)
##       AlkPhosphate < 173   to the right, improve=2.265734, (2 missing)
##       Albumin      < 2.85  to the right, improve=1.866667, (0 missing)
##       Sgot         < 84.5  to the right, improve=1.666667, (0 missing)
##       Protime      < 44.5  to the right, improve=1.388462, (2 missing)
##   Surrogate splits:
##       Sgot < 116   to the right, agree=0.933, adj=0.5, (0 split)
##
## Node number 4: 82 observations,    complexity param=0.04166667
##   predicted class=0  expected loss=0.06097561  P(node) =0.7008547
##     class counts:    77     5
##    probabilities: 0.939 0.061
```

```
##   left son=8 (72 obs) right son=9 (10 obs)
##   Primary splits:
##       Age          < 55    to the left,  improve=1.3013550, (0 missing)
##       AlkPhosphate < 229   to the left,  improve=0.8477823, (18 missing)
##       Bilirubin    < 2.55  to the left,  improve=0.8452768, (4 missing)
##       Steroid      splits as  LR,        improve=0.8112875, (1 missing)
##       Malaise      splits as  LR,        improve=0.6498489, (1 missing)
##
## Node number 5: 20 observations,    complexity param=0.08333333
##   predicted class=0  expected loss=0.4  P(node) =0.1709402
##     class counts:    12     8
##    probabilities: 0.600 0.400
##   left son=10 (10 obs) right son=11 (10 obs)
##   Primary splits:
##       Sgot       < 66.5  to the left,  improve=3.042105, (1 missing)
##       Albumin    < 3.2   to the left,  improve=2.700000, (4 missing)
##       Fatigue    splits as  LR,        improve=2.133333, (0 missing)
##       Age        < 37.5  to the left,  improve=1.600000, (0 missing)
##       Antivirals splits as  LR,        improve=1.600000, (0 missing)
##   Surrogate splits:
##       Spiders    splits as  LR,        agree=0.684, adj=0.333, (1 split)
##       Albumin    < 3.2   to the left,  agree=0.684, adj=0.333, (0 split)
##       Antivirals splits as  LR,        agree=0.632, adj=0.222, (0 split)
##       Fatigue    splits as  LR,        agree=0.632, adj=0.222, (0 split)
##       Malaise    splits as  LR,        agree=0.632, adj=0.222, (0 split)
##
## Node number 6: 2 observations
##   predicted class=0  expected loss=0  P(node) =0.01709402
##     class counts:     2     0
##    probabilities: 1.000 0.000
##
## Node number 7: 13 observations,    complexity param=0.04166667
##   predicted class=1  expected loss=0.1538462  P(node) =0.1111111
##     class counts:     2    11
##    probabilities: 0.154 0.846
##   left son=14 (6 obs) right son=15 (7 obs)
##   Primary splits:
##       Protime   < 44.5  to the right, improve=1.2727270, (2 missing)
##       Albumin   < 2.85  to the right, improve=0.7179487, (0 missing)
##       Bilirubin < 0.85  to the left,  improve=0.5664336, (0 missing)
##       Histology splits as  RL,        improve=0.5664336, (0 missing)
##       Sgot      < 50.5  to the right, improve=0.5274725, (0 missing)
##   Surrogate splits:
##       Anorexia  splits as  RL,        agree=0.818, adj=0.50, (2 split)
##       Histology splits as  RL,        agree=0.818, adj=0.50, (0 split)
##       Age       < 55.5  to the right, agree=0.727, adj=0.25, (0 split)
##       Albumin   < 2.95  to the left,  agree=0.727, adj=0.25, (0 split)
##
## Node number 8: 72 observations
##   predicted class=0  expected loss=0.02777778  P(node) =0.6153846
##     class counts:    70     2
##    probabilities: 0.972 0.028
##
## Node number 9: 10 observations,    complexity param=0.04166667
```

```
##   predicted class=0  expected loss=0.3  P(node) =0.08547009
##     class counts:     7     3
##    probabilities: 0.700 0.300
##   left son=18 (5 obs) right son=19 (5 obs)
##   Primary splits:
##       Steroid  splits as  LR,        improve=1.800000, (0 missing)
##       Malaise  splits as  LR,        improve=1.800000, (0 missing)
##       Age      < 64    to the right, improve=1.200000, (0 missing)
##       Sgot     < 56.5  to the left,  improve=1.111111, (1 missing)
##       LiverBig splits as  LR,        improve=0.750000, (2 missing)
##   Surrogate splits:
##       Age          < 61.5  to the right, agree=0.9, adj=0.8, (0 split)
##       Malaise      splits as  LR,        agree=0.8, adj=0.6, (0 split)
##       AlkPhosphate < 93    to the right, agree=0.8, adj=0.6, (0 split)
##       Sex          splits as  LR,        agree=0.7, adj=0.4, (0 split)
##       Spiders      splits as  LR,        agree=0.7, adj=0.4, (0 split)
##
## Node number 10: 10 observations
##   predicted class=0  expected loss=0.1  P(node) =0.08547009
##     class counts:     9     1
##    probabilities: 0.900 0.100
##
## Node number 11: 10 observations
##   predicted class=1  expected loss=0.3  P(node) =0.08547009
##     class counts:     3     7
##    probabilities: 0.300 0.700
##
## Node number 14: 6 observations,    complexity param=0.04166667
##   predicted class=1  expected loss=0.3333333  P(node) =0.05128205
##     class counts:     2     4
##    probabilities: 0.333 0.667
##   left son=28 (2 obs) right son=29 (4 obs)
##   Primary splits:
##       Albumin < 2.85  to the right, improve=2.6666670, (0 missing)
##       Sgot    < 50.5  to the right, improve=1.3333330, (0 missing)
##       Varices splits as  LR,        improve=0.6666667, (0 missing)
##       Age     < 41.5  to the left,  improve=0.1666667, (0 missing)
##       Steroid splits as  RL,        improve=0.1666667, (0 missing)
##   Surrogate splits:
##       Sgot < 50.5  to the right, agree=0.833, adj=0.5, (0 split)
##
## Node number 15: 7 observations
##   predicted class=1  expected loss=0  P(node) =0.05982906
##     class counts:     0     7
##    probabilities: 0.000 1.000
##
## Node number 18: 5 observations
##   predicted class=0  expected loss=0  P(node) =0.04273504
##     class counts:     5     0
##    probabilities: 1.000 0.000
##
## Node number 19: 5 observations,    complexity param=0.04166667
##   predicted class=1  expected loss=0.4  P(node) =0.04273504
##     class counts:     2     3
```

```
##    probabilities: 0.400 0.600
##   left son=38 (2 obs) right son=39 (2 obs), 1 observation remains
##   Primary splits:
##       Bilirubin < 0.9   to the left,  improve=2.00000000, (1 missing)
##       Sgot      < 40    to the left,  improve=2.00000000, (1 missing)
##       Age       < 58.5  to the left,  improve=0.06666667, (0 missing)
##       Spiders   splits as  LR,        improve=0.06666667, (0 missing)
##       Histology splits as  RL,        improve=0.06666667, (0 missing)
##   Surrogate splits:
##       Sgot < 40    to the left,  agree=1, adj=1, (0 split)
##
## Node number 28: 2 observations
##   predicted class=0  expected loss=0  P(node) =0.01709402
##     class counts:     2     0
##    probabilities: 1.000 0.000
##
## Node number 29: 4 observations
##   predicted class=1  expected loss=0  P(node) =0.03418803
##     class counts:     0     4
##    probabilities: 0.000 1.000
##
## Node number 38: 2 observations
##   predicted class=0  expected loss=0  P(node) =0.01709402
##     class counts:     2     0
##    probabilities: 1.000 0.000
##
## Node number 39: 2 observations
##   predicted class=1  expected loss=0  P(node) =0.01709402
##     class counts:     0     2
##    probabilities: 0.000 1.000
```

```
print(full.tree5.pruned)
```

```
## n= 117
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 117 24 0 (0.79487179 0.20512821)
##    2) Protime>=46.5 89  6 0 (0.93258427 0.06741573) *
##    3) Protime< 46.5 28 10 1 (0.35714286 0.64285714)
##      6) Histology=1 8  2 0 (0.75000000 0.25000000) *
##      7) Histology=0 20  4 1 (0.20000000 0.80000000)
##       14) Sex=0 2  0 0 (1.00000000 0.00000000) *
##       15) Sex=1 18  2 1 (0.11111111 0.88888889)
##         30) Protime< 26 2  0 0 (1.00000000 0.00000000) *
##         31) Protime>=26 16  0 1 (0.00000000 1.00000000) *
```

```
summary(full.tree5.pruned)
```

```
## Call:
## rpart(formula = mod2, data = train1, control = rpart.control(cp = -1,
##     minsplit = 5))
##   n= 117
##
```

```
##           CP nsplit rel error    xerror      xstd
## 1 0.33333333      0 1.0000000 1.0000000 0.1819881
## 2 0.16666667      1 0.6666667 1.0416667 0.1847395
## 3 0.08333333      2 0.5000000 0.8333333 0.1696667
## 4 0.08300000      4 0.3333333 0.7500000 0.1626109
##
## Variable importance
##   Protime   Albumin   Ascites Histology       Sex Bilirubin   Varices  LiverBig
##        48        15        12         9         8         3         3         2
##  Anorexia
##         1
##
## Node number 1: 117 observations,    complexity param=0.3333333
##   predicted class=0  expected loss=0.2051282  P(node) =1
##     class counts:    93    24
##    probabilities: 0.795 0.205
##   left son=2 (89 obs) right son=3 (28 obs)
##   Primary splits:
##       Protime   < 46.5 to the right, improve=14.105690, (0 missing)
##       Albumin   < 3.88 to the right, improve= 9.633249, (0 missing)
##       Ascites   splits as  LR,       improve= 9.600905, (0 missing)
##       Bilirubin < 1.65 to the left,  improve= 6.304281, (0 missing)
##       Malaise   splits as  LR,       improve= 5.254887, (0 missing)
##   Surrogate splits:
##       Albumin   < 3.45 to the right, agree=0.846, adj=0.357, (0 split)
##       Ascites   splits as  LR,       agree=0.838, adj=0.321, (0 split)
##       Varices   splits as  LR,       agree=0.778, adj=0.071, (0 split)
##       Bilirubin < 3.1  to the left,  agree=0.778, adj=0.071, (0 split)
##
## Node number 2: 89 observations
##   predicted class=0  expected loss=0.06741573  P(node) =0.7606838
##     class counts:    83     6
##    probabilities: 0.933 0.067
##
## Node number 3: 28 observations,    complexity param=0.1666667
##   predicted class=1  expected loss=0.3571429  P(node) =0.2393162
##     class counts:    10    18
##    probabilities: 0.357 0.643
##   left son=6 (8 obs) right son=7 (20 obs)
##   Primary splits:
##       Histology splits as  RL,       improve=3.457143, (0 missing)
##       Sex       splits as  LR,       improve=2.777143, (0 missing)
##       Albumin   < 3.9  to the right, improve=2.777143, (0 missing)
##       Anorexia  splits as  RL,       improve=2.380952, (0 missing)
##       Fatigue   splits as  LR,       improve=1.780220, (0 missing)
##   Surrogate splits:
##       LiverBig splits as  RL,       agree=0.786, adj=0.250, (0 split)
##       Anorexia splits as  RL,       agree=0.750, adj=0.125, (0 split)
##       Albumin  < 3.75 to the right, agree=0.750, adj=0.125, (0 split)
##       Protime  < 45.5 to the right, agree=0.750, adj=0.125, (0 split)
##
## Node number 6: 8 observations
##   predicted class=0  expected loss=0.25  P(node) =0.06837607
##     class counts:     6     2
```

```
##    probabilities: 0.750 0.250
##
## Node number 7: 20 observations,    complexity param=0.08333333
##   predicted class=1  expected loss=0.2  P(node) =0.1709402
##     class counts:     4    16
##    probabilities: 0.200 0.800
##   left son=14 (2 obs) right son=15 (18 obs)
##   Primary splits:
##       Sex       splits as  LR,       improve=2.844444, (0 missing)
##       Protime   < 26   to the left,  improve=2.844444, (0 missing)
##       Malaise   splits as  LR,       improve=1.542857, (0 missing)
##       Albumin   < 3.66 to the right, improve=1.537255, (0 missing)
##       Bilirubin < 1.1  to the left,  improve=1.125275, (0 missing)
##
## Node number 14: 2 observations
##   predicted class=0  expected loss=0  P(node) =0.01709402
##     class counts:     2     0
##    probabilities: 1.000 0.000
##
## Node number 15: 18 observations,    complexity param=0.08333333
##   predicted class=1  expected loss=0.1111111  P(node) =0.1538462
##     class counts:     2    16
##    probabilities: 0.111 0.889
##   left son=30 (2 obs) right son=31 (16 obs)
##   Primary splits:
##       Protime   < 26   to the left,  improve=3.5555560, (0 missing)
##       Bilirubin < 1.1  to the left,  improve=0.8888889, (0 missing)
##       Albumin   < 3.71 to the right, improve=0.6805556, (0 missing)
##       Anorexia  splits as  RL,       improve=0.3555556, (0 missing)
##       LiverFirm splits as  LR,       improve=0.2828283, (0 missing)
##
## Node number 30: 2 observations
##   predicted class=0  expected loss=0  P(node) =0.01709402
##     class counts:     2     0
##    probabilities: 1.000 0.000
##
## Node number 31: 16 observations
##   predicted class=1  expected loss=0  P(node) =0.1367521
##     class counts:     0    16
##    probabilities: 0.000 1.000
```

```
print(full.tree6.pruned)
```

```
## n= 117
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 117 24 0 (0.79487179 0.20512821)
##    2) Protime>=45 91  5 0 (0.94505495 0.05494505)
##      4) Albumin>=2.65 89  3 0 (0.96629213 0.03370787) *
##      5) Albumin< 2.65 2  0 1 (0.00000000 1.00000000) *
##    3) Protime< 45 26  7 1 (0.26923077 0.73076923)
##      6) Sex=0 2  0 0 (1.00000000 0.00000000) *
##      7) Sex=1 24  5 1 (0.20833333 0.79166667)
```

```
##       14) Protime< 29.5 4  1 0 (0.75000000 0.25000000) *
##       15) Protime>=29.5 20  2 1 (0.10000000 0.90000000) *
```

```
summary(full.tree6.pruned)
```

```
## Call:
## rpart(formula = mod2, data = train2, control = rpart.control(cp = -1,
##     minsplit = 5))
##   n= 117
##
##           CP nsplit rel error    xerror      xstd
## 1 0.50000000      0      1.00 1.0000000 0.1819881
## 2 0.08333333      1      0.50 0.6666667 0.1548519
## 3 0.08300000      4      0.25 0.7916667 0.1662209
##
## Variable importance
##   Protime   Albumin   Ascites       Sex Bilirubin
##        57        23        10         6         4
##
## Node number 1: 117 observations,    complexity param=0.5
##   predicted class=0  expected loss=0.2051282  P(node) =1
##     class counts:    93    24
##    probabilities: 0.795 0.205
##   left son=2 (91 obs) right son=3 (26 obs)
##   Primary splits:
##       Protime   < 45       to the right, improve=18.472530, (0 missing)
##       Albumin   < 3.85     to the right, improve=11.775090, (0 missing)
##       Ascites   splits as  LR,          improve= 9.600905, (0 missing)
##       Bilirubin < 1.057866 to the left,  improve= 5.553846, (0 missing)
##       Malaise   splits as  LR,          improve= 5.254887, (0 missing)
##   Surrogate splits:
##       Albumin   < 3.618841 to the right, agree=0.838, adj=0.269, (0 split)
##       Ascites   splits as  LR,          agree=0.821, adj=0.192, (0 split)
##       Bilirubin < 3.35     to the left,  agree=0.795, adj=0.077, (0 split)
##
## Node number 2: 91 observations,    complexity param=0.08333333
##   predicted class=0  expected loss=0.05494505  P(node) =0.7777778
##     class counts:    86     5
##    probabilities: 0.945 0.055
##   left son=4 (89 obs) right son=5 (2 obs)
##   Primary splits:
##       Albumin   < 2.65     to the right, improve=3.6527970, (0 missing)
##       Ascites   splits as  LR,          improve=1.2598520, (0 missing)
##       Protime   < 47.5     to the right, improve=0.9956475, (0 missing)
##       Bilirubin < 1.65     to the left,  improve=0.8401598, (0 missing)
##       Malaise   splits as  LR,          improve=0.5862558, (0 missing)
##
## Node number 3: 26 observations,    complexity param=0.08333333
##   predicted class=1  expected loss=0.2692308  P(node) =0.2222222
##     class counts:     7    19
##    probabilities: 0.269 0.731
##   left son=6 (2 obs) right son=7 (24 obs)
##   Primary splits:
##       Sex      splits as  LR,          improve=2.314103, (0 missing)
##       Fatigue  splits as  LR,          improve=2.314103, (0 missing)
```

```
##       Albumin  < 3.95     to the right, improve=2.314103, (0 missing)
##       Protime  < 26       to the left,  improve=2.314103, (0 missing)
##       LiverBig splits as  RL,           improve=1.354579, (0 missing)
##
## Node number 4: 89 observations
##   predicted class=0  expected loss=0.03370787  P(node) =0.7606838
##     class counts:    86     3
##    probabilities: 0.966 0.034
##
## Node number 5: 2 observations
##   predicted class=1  expected loss=0  P(node) =0.01709402
##     class counts:     0     2
##    probabilities: 0.000 1.000
##
## Node number 6: 2 observations
##   predicted class=0  expected loss=0  P(node) =0.01709402
##     class counts:     2     0
##    probabilities: 1.000 0.000
##
## Node number 7: 24 observations,    complexity param=0.08333333
##   predicted class=1  expected loss=0.2083333  P(node) =0.2051282
##     class counts:     5    19
##    probabilities: 0.208 0.792
##   left son=14 (4 obs) right son=15 (20 obs)
##   Primary splits:
##       Protime   < 29.5     to the left,  improve=2.8166670, (0 missing)
##       Fatigue   splits as  LR,           improve=2.7348480, (0 missing)
##       Albumin   < 3.95     to the right, improve=2.7348480, (0 missing)
##       Anorexia  splits as  RL,           improve=1.3611110, (0 missing)
##       Bilirubin < 1.258523 to the left,  improve=0.9796037, (0 missing)
##
## Node number 14: 4 observations
##   predicted class=0  expected loss=0.25  P(node) =0.03418803
##     class counts:     3     1
##    probabilities: 0.750 0.250
##
## Node number 15: 20 observations
##   predicted class=1  expected loss=0.1  P(node) =0.1709402
##     class counts:     2    18
##    probabilities: 0.100 0.900
```

```
print(full.tree7.pruned)
```

```
## n= 117
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
## 1) root 117 24 0 (0.79487179 0.20512821)
##   2) Albumin>=3.85 70  2 0 (0.97142857 0.02857143) *
##   3) Albumin< 3.85 47 22 0 (0.53191489 0.46808511)
##     6) Ascites=0 33 11 0 (0.66666667 0.33333333) *
##     7) Ascites=1 14  3 1 (0.21428571 0.78571429) *
```

```r
summary(full.tree7.pruned)
```

```
## Call:
## rpart(formula = mod2, data = train3, control = rpart.control(cp = -1,
##     minsplit = 5))
##   n= 117
##
##          CP nsplit rel error   xerror      xstd
## 1 0.1666667      0 1.0000000 1.000000 0.1819881
## 2 0.1660000      2 0.6666667 1.208333 0.1945970
##
## Variable importance
##   Albumin   Ascites   Protime Bilirubin   Spiders   Malaise
##        38        22        14        13         9         4
##
## Node number 1: 117 observations,    complexity param=0.1666667
##   predicted class=0  expected loss=0.2051282  P(node) =1
##     class counts:    93    24
##    probabilities: 0.795 0.205
##   left son=2 (70 obs) right son=3 (47 obs)
##   Primary splits:
##       Albumin   < 3.85 to the right, improve=10.863880, (0 missing)
##       Ascites   splits as  LR,       improve= 9.600905, (0 missing)
##       Protime   < 44.5 to the right, improve= 8.835664, (0 missing)
##       Bilirubin < 1.65 to the left,  improve= 6.304281, (0 missing)
##       Malaise   splits as  LR,       improve= 5.254887, (0 missing)
##   Surrogate splits:
##       Protime   < 44.5 to the right, agree=0.761, adj=0.404, (0 split)
##       Bilirubin < 1.35 to the left,  agree=0.718, adj=0.298, (0 split)
##       Ascites   splits as  LR,       agree=0.709, adj=0.277, (0 split)
##       Spiders   splits as  LR,       agree=0.701, adj=0.255, (0 split)
##       Malaise   splits as  LR,       agree=0.650, adj=0.128, (0 split)
##
## Node number 2: 70 observations
##   predicted class=0  expected loss=0.02857143  P(node) =0.5982906
##     class counts:    68     2
##    probabilities: 0.971 0.029
##
## Node number 3: 47 observations,    complexity param=0.1666667
##   predicted class=0  expected loss=0.4680851  P(node) =0.4017094
##     class counts:    25    22
##    probabilities: 0.532 0.468
##   left son=6 (33 obs) right son=7 (14 obs)
##   Primary splits:
##       Ascites   splits as  LR,       improve=4.023303, (0 missing)
##       Protime   < 35.5 to the right, improve=3.855868, (0 missing)
##       Bilirubin < 3.7  to the left,  improve=3.166160, (0 missing)
##       Albumin   < 2.65 to the right, improve=3.166160, (0 missing)
##       Malaise   splits as  LR,       improve=2.525134, (0 missing)
##   Surrogate splits:
##       Albumin   < 2.65 to the right, agree=0.809, adj=0.357, (0 split)
##       Bilirubin < 3.7  to the left,  agree=0.766, adj=0.214, (0 split)
##
## Node number 6: 33 observations
```

```
##   predicted class=0  expected loss=0.3333333  P(node) =0.2820513
##     class counts:    22    11
##    probabilities: 0.667 0.333
##
## Node number 7: 14 observations
##   predicted class=1  expected loss=0.2142857  P(node) =0.1196581
##     class counts:     3    11
##    probabilities: 0.214 0.786
```

```
print(full.tree8.pruned)
```

```
## n= 117
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 117 24 0 (0.79487179 0.20512821)
##    2) Ascites=0 102 13 0 (0.87254902 0.12745098)
##      4) Protime>=43.5 82  5 0 (0.93902439 0.06097561) *
##      5) Protime< 43.5 20  8 0 (0.60000000 0.40000000)
##       10) Albumin< 3.2 7  0 0 (1.00000000 0.00000000) *
##       11) Albumin>=3.2 13  5 1 (0.38461538 0.61538462)
##         22) Albumin>=3.95 2  0 0 (1.00000000 0.00000000) *
##         23) Albumin< 3.95 11  3 1 (0.27272727 0.72727273) *
##    3) Ascites=1 15  4 1 (0.26666667 0.73333333)
##      6) Albumin>=2.85 8  4 0 (0.50000000 0.50000000)
##       12) Albumin< 2.95 3  0 0 (1.00000000 0.00000000) *
##       13) Albumin>=2.95 5  1 1 (0.20000000 0.80000000) *
##      7) Albumin< 2.85 7  0 1 (0.00000000 1.00000000) *
```

```
summary(full.tree8.pruned)
```

```
## Call:
## rpart(formula = mod2, data = train0, control = rpart.control(cp = -1,
##     minsplit = 5))
##   n= 117
##
##          CP nsplit rel error    xerror      xstd
## 1 0.2916667      0 1.0000000 1.0000000 0.1819881
## 2 0.0625000      1 0.7083333 0.8333333 0.1696667
## 3 0.0620000      6 0.3750000 0.8333333 0.1696667
##
## Variable importance
##        Ascites        Albumin        Protime      Bilirubin        Varices
##             31             31             13              7              3
##        Fatigue        Spiders SpleenPalpable      Histology      LiverFirm
##              3              3              3              3              2
##        Malaise       Anorexia
##              1              1
##
## Node number 1: 117 observations,    complexity param=0.2916667
##   predicted class=0  expected loss=0.2051282  P(node) =1
##     class counts:    93    24
##    probabilities: 0.795 0.205
##   left son=2 (102 obs) right son=3 (15 obs)
```

```
##   Primary splits:
##       Ascites   splits as  LR,       improve=9.442212, (3 missing)
##       Albumin   < 3.85 to the right, improve=8.759234, (15 missing)
##       Protime   < 46.5 to the right, improve=8.727273, (51 missing)
##       Bilirubin < 1.65 to the left,  improve=5.578130, (6 missing)
##       Malaise   splits as  LR,       improve=5.184493, (1 missing)
##
## Node number 2: 102 observations,    complexity param=0.0625
##   predicted class=0  expected loss=0.127451  P(node) =0.8717949
##     class counts:    89    13
##    probabilities: 0.873 0.127
##   left son=4 (82 obs) right son=5 (20 obs)
##   Primary splits:
##       Protime   < 43.5 to the right, improve=3.234583, (49 missing)
##       Albumin   < 3.85 to the right, improve=2.191699, (15 missing)
##       Spiders   splits as  LR,       improve=2.125971, (3 missing)
##       Malaise   splits as  LR,       improve=1.895783, (1 missing)
##       Bilirubin < 0.55 to the right, improve=1.887769, (6 missing)
##   Surrogate splits:
##       Albumin        < 3.45 to the right, agree=0.830, adj=0.182, (35 split)
##       SpleenPalpable splits as  LR,       agree=0.811, adj=0.091, (13 split)
##       Bilirubin      < 2.25 to the left,  agree=0.811, adj=0.091, (0 split)
##
## Node number 3: 15 observations,    complexity param=0.0625
##   predicted class=1  expected loss=0.2666667  P(node) =0.1282051
##     class counts:     4    11
##    probabilities: 0.267 0.733
##   left son=6 (8 obs) right son=7 (7 obs)
##   Primary splits:
##       Albumin        < 2.85 to the right, improve=1.8666670, (0 missing)
##       Protime        < 44.5 to the right, improve=1.3884620, (2 missing)
##       Bilirubin      < 2.4  to the left,  improve=1.0666670, (0 missing)
##       Anorexia       splits as  RL,       improve=0.2666667, (0 missing)
##       SpleenPalpable splits as  RL,       improve=0.2666667, (0 missing)
##   Surrogate splits:
##       Varices        splits as  LR,       agree=0.800, adj=0.571, (0 split)
##       LiverFirm      splits as  LR,       agree=0.667, adj=0.286, (0 split)
##       SpleenPalpable splits as  RL,       agree=0.667, adj=0.286, (0 split)
##       Bilirubin      < 1.5  to the left,  agree=0.667, adj=0.286, (0 split)
##       Anorexia       splits as  LR,       agree=0.600, adj=0.143, (0 split)
##
## Node number 4: 82 observations
##   predicted class=0  expected loss=0.06097561  P(node) =0.7008547
##     class counts:    77     5
##    probabilities: 0.939 0.061
##
## Node number 5: 20 observations,    complexity param=0.0625
##   predicted class=0  expected loss=0.4  P(node) =0.1709402
##     class counts:    12     8
##    probabilities: 0.600 0.400
##   left son=10 (7 obs) right son=11 (13 obs)
##   Primary splits:
##       Albumin    < 3.2  to the left,  improve=2.700000, (4 missing)
##       Fatigue    splits as  LR,       improve=2.133333, (0 missing)
```

```
##       Antivirals splits as  LR,       improve=1.600000, (0 missing)
##       Malaise    splits as  LR,       improve=1.350000, (0 missing)
##       Bilirubin  < 1.85 to the left,  improve=1.341270, (2 missing)
##   Surrogate splits:
##       Fatigue   splits as  LR,       agree=0.750, adj=0.333, (4 split)
##       Spiders   splits as  LR,       agree=0.750, adj=0.333, (0 split)
##       Malaise   splits as  LR,       agree=0.688, adj=0.167, (0 split)
##       Bilirubin < 0.8  to the left,  agree=0.688, adj=0.167, (0 split)
##
## Node number 6: 8 observations,    complexity param=0.0625
##   predicted class=0  expected loss=0.5  P(node) =0.06837607
##     class counts:     4     4
##    probabilities: 0.500 0.500
##   left son=12 (3 obs) right son=13 (5 obs)
##   Primary splits:
##       Albumin   < 2.95 to the left,  improve=2.400000, (0 missing)
##       Protime   < 44.5 to the right, improve=2.400000, (0 missing)
##       Anorexia  splits as  RL,       improve=1.333333, (0 missing)
##       Bilirubin < 2.95 to the left,  improve=1.333333, (0 missing)
##       LiverFirm splits as  RL,       improve=1.028571, (1 missing)
##   Surrogate splits:
##       Bilirubin < 1.1  to the left,  agree=0.75, adj=0.333, (0 split)
##       Protime   < 44.5 to the right, agree=0.75, adj=0.333, (0 split)
##       Histology splits as  RL,       agree=0.75, adj=0.333, (0 split)
##
## Node number 7: 7 observations
##   predicted class=1  expected loss=0  P(node) =0.05982906
##     class counts:     0     7
##    probabilities: 0.000 1.000
##
## Node number 10: 7 observations
##   predicted class=0  expected loss=0  P(node) =0.05982906
##     class counts:     7     0
##    probabilities: 1.000 0.000
##
## Node number 11: 13 observations,    complexity param=0.0625
##   predicted class=1  expected loss=0.3846154  P(node) =0.1111111
##     class counts:     5     8
##    probabilities: 0.385 0.615
##   left son=22 (2 obs) right son=23 (11 obs)
##   Primary splits:
##       Albumin    < 3.95 to the right, improve=1.8000000, (3 missing)
##       LiverBig   splits as  RL,       improve=1.5427350, (0 missing)
##       Bilirubin  < 1.3  to the left,  improve=1.0242420, (2 missing)
##       Antivirals splits as  LR,       improve=0.6993007, (0 missing)
##       Anorexia   splits as  RL,       improve=0.6205128, (0 missing)
##
## Node number 12: 3 observations
##   predicted class=0  expected loss=0  P(node) =0.02564103
##     class counts:     3     0
##    probabilities: 1.000 0.000
##
## Node number 13: 5 observations
##   predicted class=1  expected loss=0.2  P(node) =0.04273504
```

```
##     class counts:     1     4
##    probabilities: 0.200 0.800
##
## Node number 22: 2 observations
##   predicted class=0  expected loss=0  P(node) =0.01709402
##     class counts:     2     0
##    probabilities: 1.000 0.000
##
## Node number 23: 11 observations
##   predicted class=1  expected loss=0.2727273  P(node) =0.09401709
##     class counts:     3     8
##    probabilities: 0.273 0.727
```

4.8. Bagging (bootstrap aggregating)

We can see basic information about all random forests. The error is pretty similar.

```
print(btree1)
```

```
##
## Bagging classification trees with 150 bootstrap replications
##
## Call: bagging.data.frame(formula = mod1, data = train1, nbagg = 150,
##     coob = TRUE, minsplit = 2, cp = 0)
##
## Out-of-bag estimate of misclassification error:  0.1538
```

```
print(btree2)
```

```
##
## Bagging classification trees with 150 bootstrap replications
##
## Call: bagging.data.frame(formula = mod1, data = train2, nbagg = 150,
##     coob = TRUE, minsplit = 2, cp = 0)
##
## Out-of-bag estimate of misclassification error:  0.1368
```

```
print(btree3)
```

```
##
## Bagging classification trees with 150 bootstrap replications
##
## Call: bagging.data.frame(formula = mod1, data = train3, nbagg = 150,
##     coob = TRUE, minsplit = 2, cp = 0)
##
## Out-of-bag estimate of misclassification error:  0.1966
```

```
print(btree4)
```

```
##
## Bagging classification trees with 150 bootstrap replications
##
## Call: bagging.data.frame(formula = mod1, data = train0, nbagg = 150,
##     coob = TRUE, minsplit = 2, cp = 0)
##
## Out-of-bag estimate of misclassification error:  0.2167
```

```
print(btree5)
```

```
##
## Bagging classification trees with 150 bootstrap replications
##
## Call: bagging.data.frame(formula = mod2, data = train1, nbagg = 150,
##     coob = TRUE, minsplit = 2, cp = 0)
##
## Out-of-bag estimate of misclassification error:  0.1709
```

```
print(btree6)
```

```
##
## Bagging classification trees with 150 bootstrap replications
##
## Call: bagging.data.frame(formula = mod2, data = train2, nbagg = 150,
##     coob = TRUE, minsplit = 2, cp = 0)
##
## Out-of-bag estimate of misclassification error:  0.1197
```

```
print(btree7)
```

```
##
## Bagging classification trees with 150 bootstrap replications
##
## Call: bagging.data.frame(formula = mod2, data = train3, nbagg = 150,
##     coob = TRUE, minsplit = 2, cp = 0)
##
## Out-of-bag estimate of misclassification error:  0.2051
```

```
print(btree8)
```

```
##
## Bagging classification trees with 150 bootstrap replications
##
## Call: bagging.data.frame(formula = mod2, data = train0, nbagg = 150,
##     coob = TRUE, minsplit = 2, cp = 0)
##
## Out-of-bag estimate of misclassification error:  0.2167
```

4.9. Boosting

We can see basic information about all random forests. The error is pretty similar.

```
print(boost1)
```

```
## Call:
## ada(mod1, data = train1, iter = 10)
##
## Loss: exponential Method: discrete   Iteration: 10
##
## Final Confusion Matrix for Data:
##           Final Prediction
## True value  0  1
##          0 84  9
##          1  6 18
##
```

```
## Train Error: 0.128
##
## Out-Of-Bag Error:  0.137  iteration= 6
##
## Additional Estimates of number of iterations:
##
## train.err1 train.kap1
##          9          9
```

```
print(boost2)
```

```
## Call:
## ada(mod1, data = train2, iter = 10)
##
## Loss: exponential Method: discrete   Iteration: 10
##
## Final Confusion Matrix for Data:
##           Final Prediction
## True value  0  1
##          0 87  6
##          1  4 20
##
## Train Error: 0.085
##
## Out-Of-Bag Error:  0.103  iteration= 7
##
## Additional Estimates of number of iterations:
##
## train.err1 train.kap1
##          8          8
```

```
print(boost3)
```

```
## Call:
## ada(mod1, data = train3, iter = 10)
##
## Loss: exponential Method: discrete   Iteration: 10
##
## Final Confusion Matrix for Data:
##           Final Prediction
## True value  0  1
##          0 88  5
##          1  9 15
##
## Train Error: 0.12
##
## Out-Of-Bag Error:  0.12  iteration= 9
##
## Additional Estimates of number of iterations:
##
## train.err1 train.kap1
##          5          5
```

```
print(boost4)
```

```
## Call:
```

```
## ada(mod1, data = train0, iter = 10)
##
## Loss: exponential Method: discrete   Iteration: 10
##
## Final Confusion Matrix for Data:
##           Final Prediction
## True value  0  1
##          0 90  3
##          1 10 14
##
## Train Error: 0.111
##
## Out-Of-Bag Error:  0.12  iteration= 7
##
## Additional Estimates of number of iterations:
##
## train.err1 train.kap1
##          9          6
```

```
print(boost5)
```

```
## Call:
## ada(mod2, data = train1, iter = 10)
##
## Loss: exponential Method: discrete   Iteration: 10
##
## Final Confusion Matrix for Data:
##           Final Prediction
## True value  0  1
##          0 86  7
##          1  6 18
##
## Train Error: 0.111
##
## Out-Of-Bag Error:  0.103  iteration= 9
##
## Additional Estimates of number of iterations:
##
## train.err1 train.kap1
##          7          7
```

```
print(boost6)
```

```
## Call:
## ada(mod2, data = train2, iter = 10)
##
## Loss: exponential Method: discrete   Iteration: 10
##
## Final Confusion Matrix for Data:
##           Final Prediction
## True value  0  1
##          0 88  5
##          1  5 19
##
## Train Error: 0.085
```

```
##
## Out-Of-Bag Error:  0.094  iteration= 8
##
## Additional Estimates of number of iterations:
##
## train.err1 train.kap1
##          8          8
```

```
print(boost7)
```

```
## Call:
## ada(mod2, data = train3, iter = 10)
##
## Loss: exponential Method: discrete   Iteration: 10
##
## Final Confusion Matrix for Data:
##           Final Prediction
## True value  0  1
##          0 84  9
##          1  7 17
##
## Train Error: 0.137
##
## Out-Of-Bag Error:  0.128  iteration= 9
##
## Additional Estimates of number of iterations:
##
## train.err1 train.kap1
##          4          4
```

```
print(boost8)
```

```
## Call:
## ada(mod2, data = train0, iter = 10)
##
## Loss: exponential Method: discrete   Iteration: 10
##
## Final Confusion Matrix for Data:
##           Final Prediction
## True value  0  1
##          0 88  5
##          1 11 13
##
## Train Error: 0.137
##
## Out-Of-Bag Error:  0.154  iteration= 8
##
## Additional Estimates of number of iterations:
##
## train.err1 train.kap1
##          1          9
```

4.10. Random forest

We can see basic information about all random forests. The error is pretty similar.

```
print(rf1)
```

```
##
## Call:
##  randomForest(formula = mod1, data = train1, ntree = 500, mtry = p1,      importance = TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 19
##
##         OOB estimate of  error rate: 17.09%
## Confusion matrix:
##    0  1 class.error
## 0 82 11   0.1182796
## 1  9 15   0.3750000
```

```
print(rf2)
```

```
##
## Call:
##  randomForest(formula = mod1, data = train2, ntree = 500, mtry = p1,      importance = TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 19
##
##         OOB estimate of  error rate: 12.82%
## Confusion matrix:
##    0  1 class.error
## 0 84  9  0.09677419
## 1  6 18  0.25000000
```

```
print(rf3)
```

```
##
## Call:
##  randomForest(formula = mod1, data = train3, ntree = 500, mtry = p1,      importance = TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 19
##
##         OOB estimate of  error rate: 18.8%
## Confusion matrix:
##    0  1 class.error
## 0 83 10   0.1075269
## 1 12 12   0.5000000
```

```
print(rf4)
```

```
##
## Call:
##  randomForest(formula = mod1, data = train1, ntree = 500, mtry = sqrt(p1),      importance = TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 14.53%
## Confusion matrix:
```

```
##    0  1 class.error
## 0 88  5  0.05376344
## 1 12 12  0.50000000
```

```
print(rf5)
```

```
##
## Call:
##  randomForest(formula = mod1, data = train2, ntree = 500, mtry = sqrt(p1),      importance = TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 11.97%
## Confusion matrix:
##    0  1 class.error
## 0 87  6  0.06451613
## 1  8 16  0.33333333
```

```
print(rf6)
```

```
##
## Call:
##  randomForest(formula = mod1, data = train3, ntree = 500, mtry = sqrt(p1),      importance = TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 17.95%
## Confusion matrix:
##    0  1 class.error
## 0 84  9  0.09677419
## 1 12 12  0.50000000
```

```
print(rf7)
```

```
##
## Call:
##  randomForest(formula = mod2, data = train1, ntree = 500, mtry = p2,      importance = TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 15
##
##         OOB estimate of  error rate: 14.53%
## Confusion matrix:
##    0  1 class.error
## 0 85  8  0.08602151
## 1  9 15  0.37500000
```

```
print(rf8)
```

```
##
## Call:
##  randomForest(formula = mod2, data = train2, ntree = 500, mtry = p2,      importance = TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 15
```

```
##
##         OOB estimate of  error rate: 11.11%
## Confusion matrix:
##    0  1 class.error
## 0 85  8  0.08602151
## 1  5 19  0.20833333
```

```
print(rf9)
```

```
##
## Call:
##  randomForest(formula = mod2, data = train3, ntree = 500, mtry = p2,      importance = TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 15
##
##         OOB estimate of  error rate: 18.8%
## Confusion matrix:
##    0  1 class.error
## 0 82 11   0.1182796
## 1 11 13   0.4583333
```

```
print(rf10)
```

```
##
## Call:
##  randomForest(formula = mod2, data = train1, ntree = 500, mtry = sqrt(p2),      importance = TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 13.68%
## Confusion matrix:
##    0  1 class.error
## 0 87  6  0.06451613
## 1 10 14  0.41666667
```

```
print(rf11)
```

```
##
## Call:
##  randomForest(formula = mod2, data = train2, ntree = 500, mtry = sqrt(p2),      importance = TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 11.11%
## Confusion matrix:
##    0  1 class.error
## 0 87  6  0.06451613
## 1  7 17  0.29166667
```

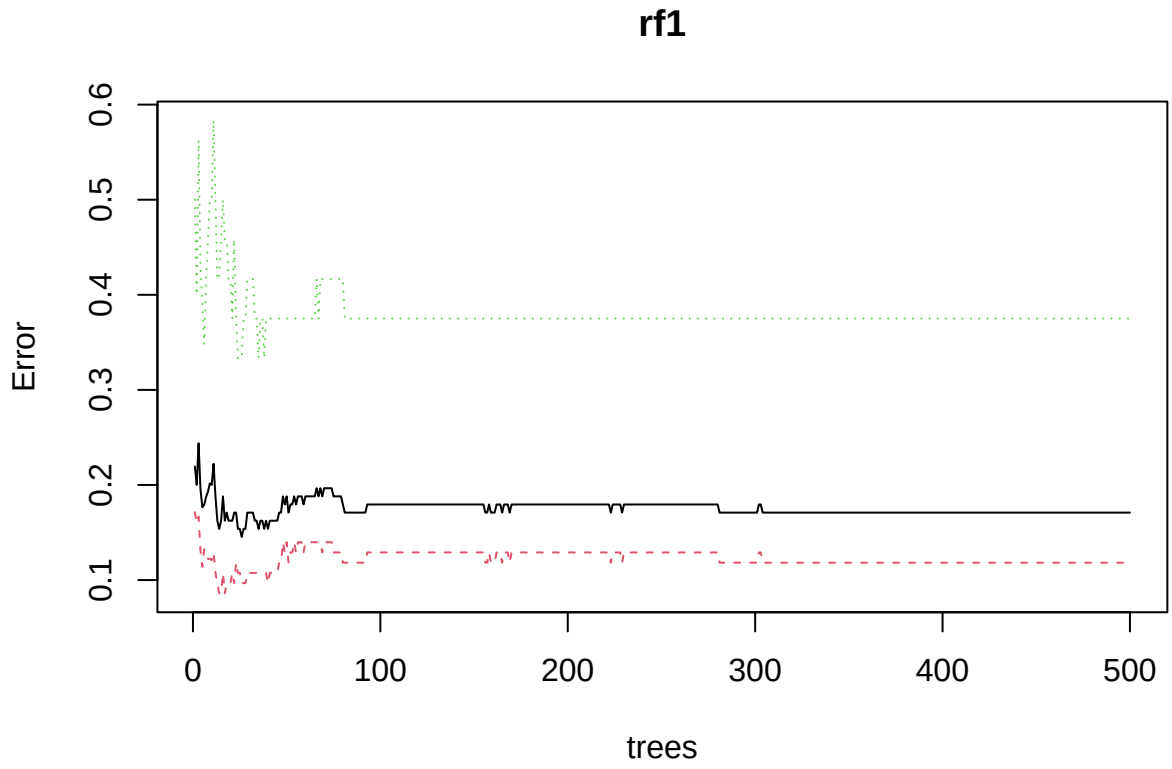
```
print(rf12)
```

```
##
## Call:
##  randomForest(formula = mod2, data = train3, ntree = 500, mtry = sqrt(p2),      importance = TRUE)
```

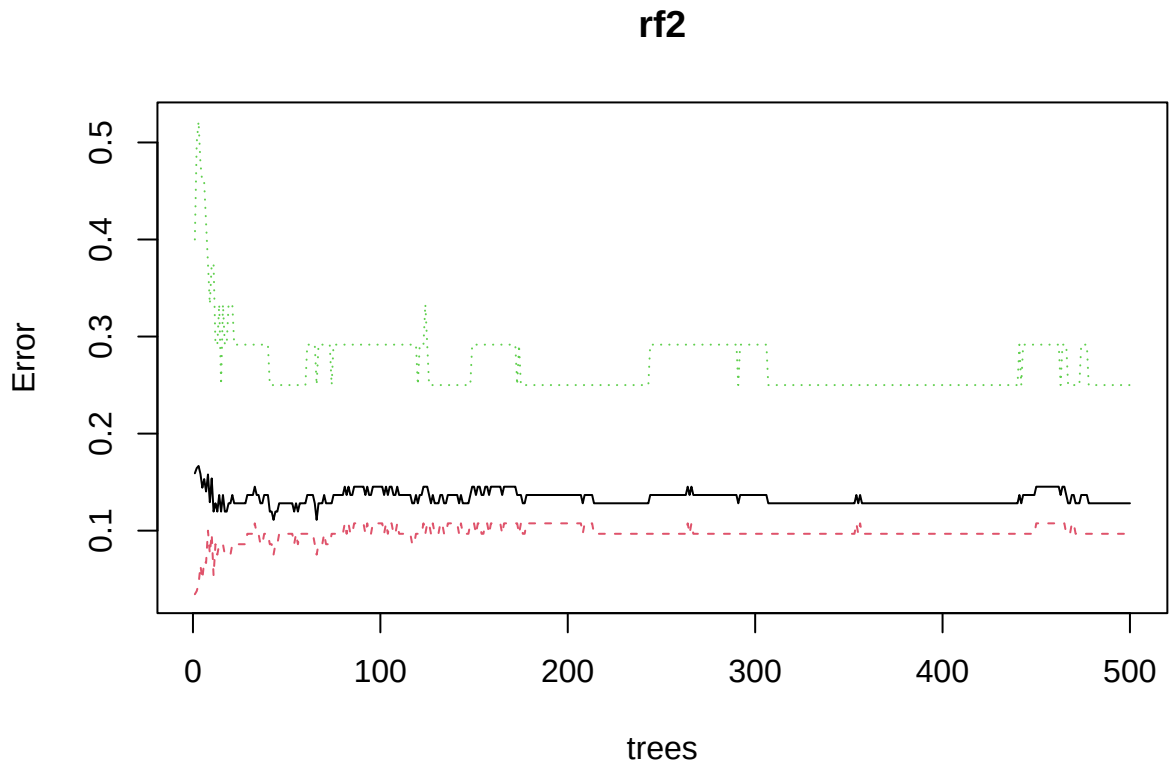
```
##                 Type of random forest: classification
##                       Number of trees: 500
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 15.38%
## Confusion matrix:
##    0  1 class.error
## 0 87  6  0.06451613
## 1 12 12  0.50000000
```

We can see classification error plot for all random forest models. There are both similar and different results

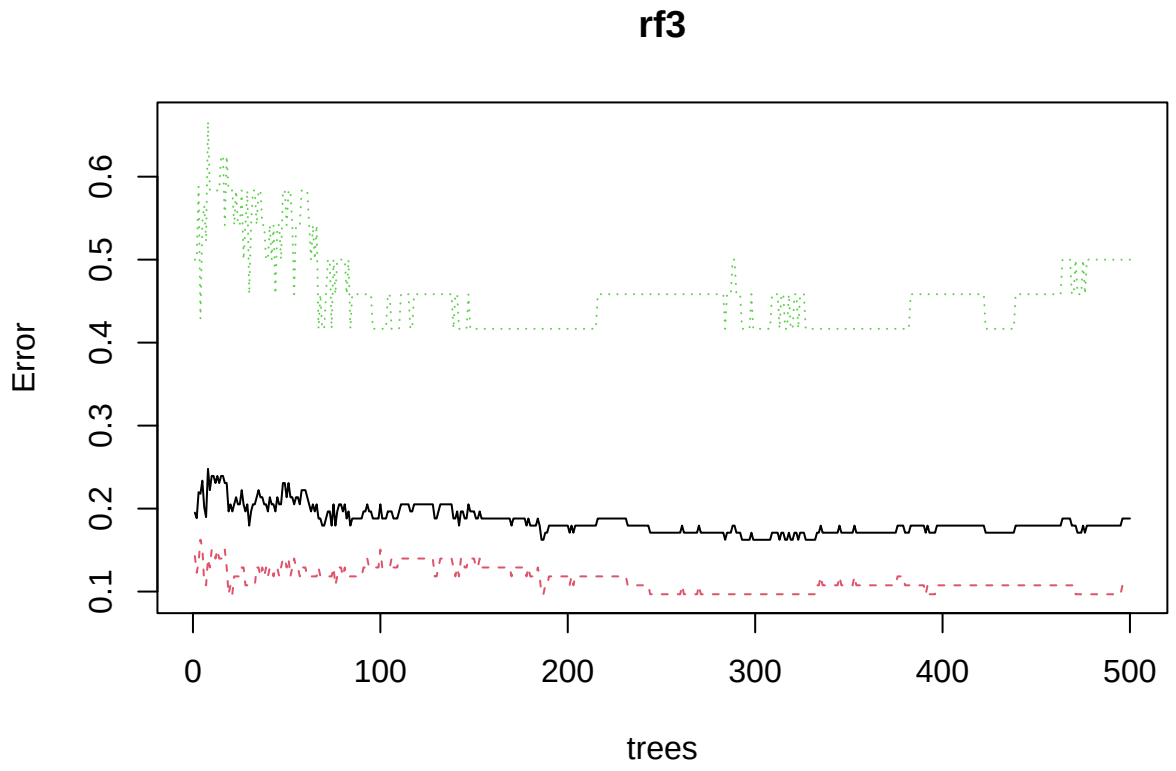
```
plot(rf1)
```



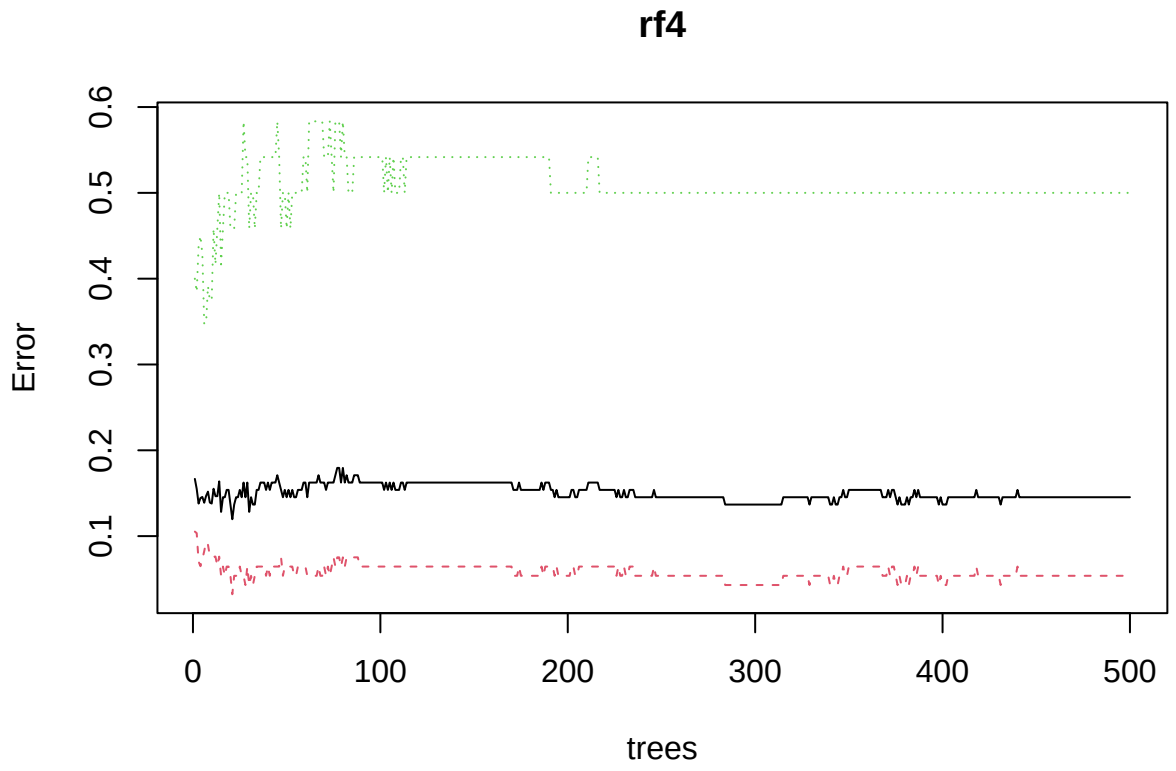

```
plot(rf2)
```



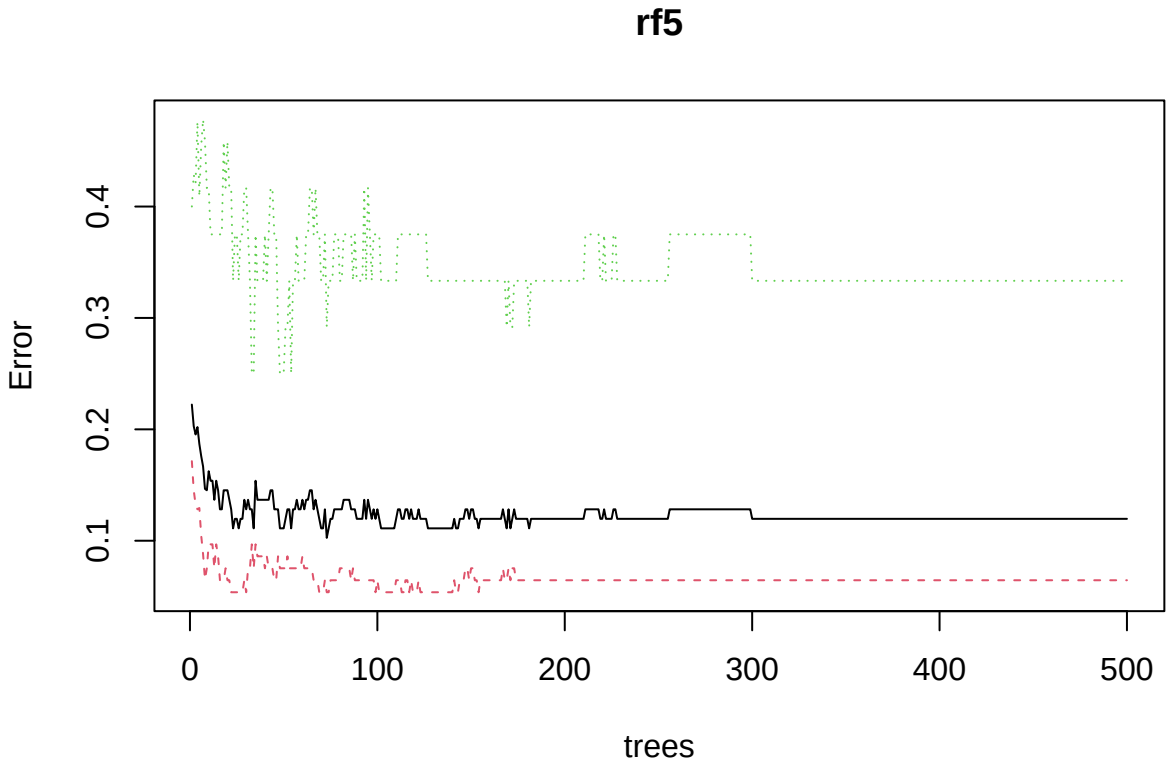

```
plot(rf3)
```



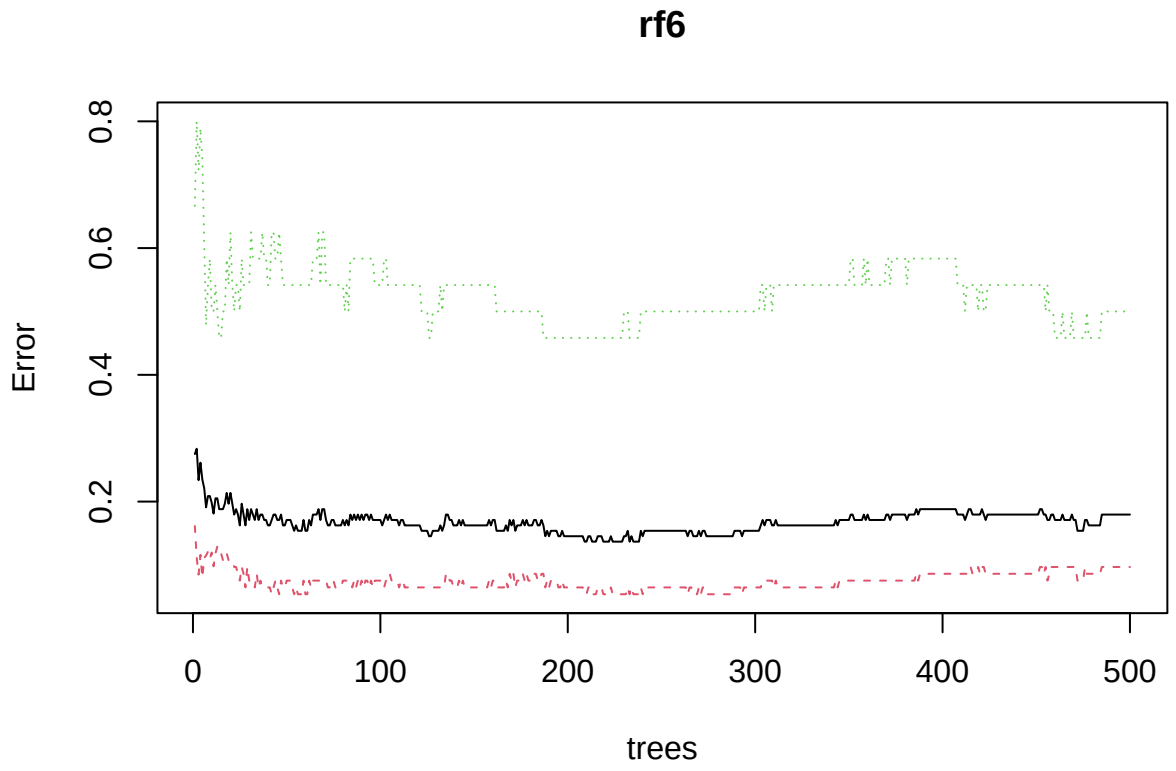

```
plot(rf4)
```



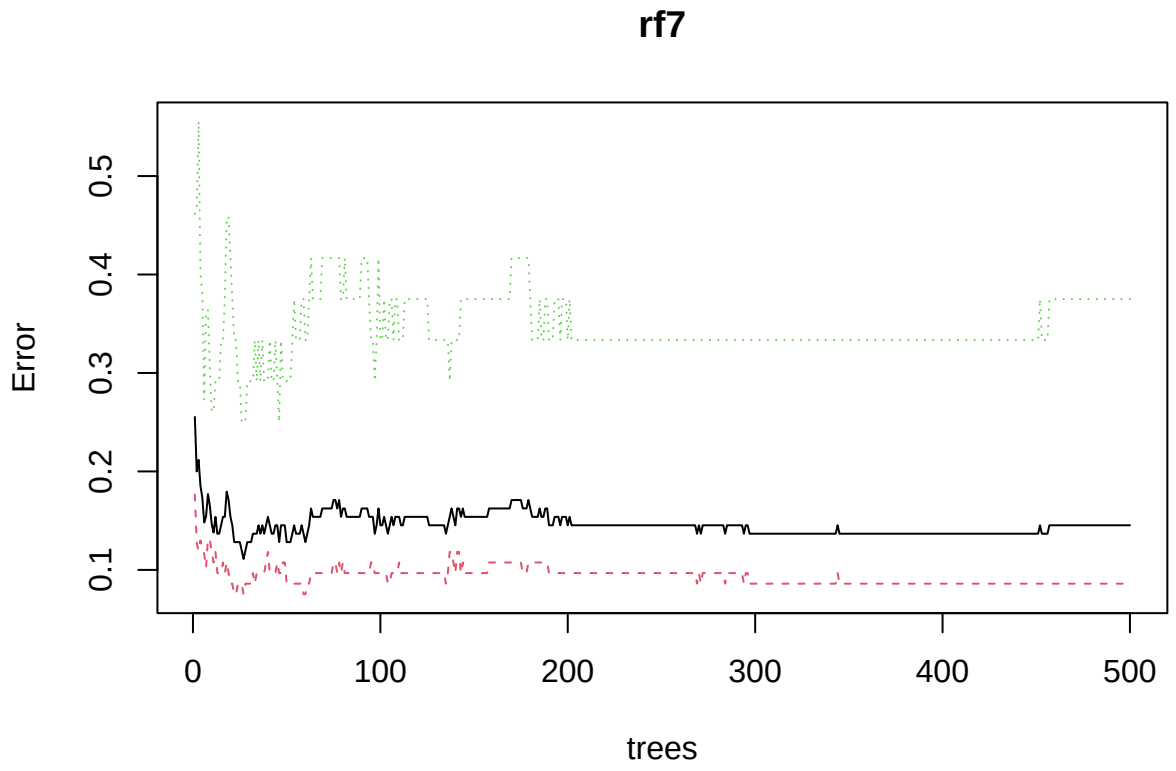

```
plot(rf5)
```



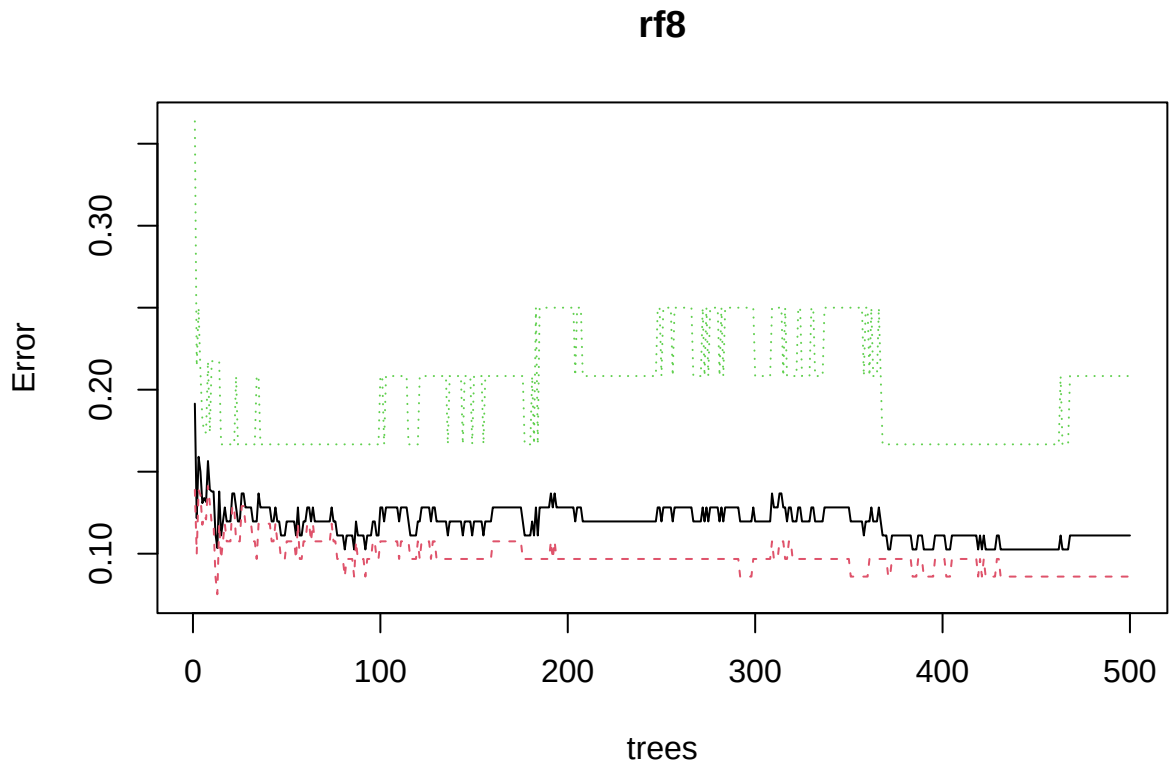

```
plot(rf6)
```



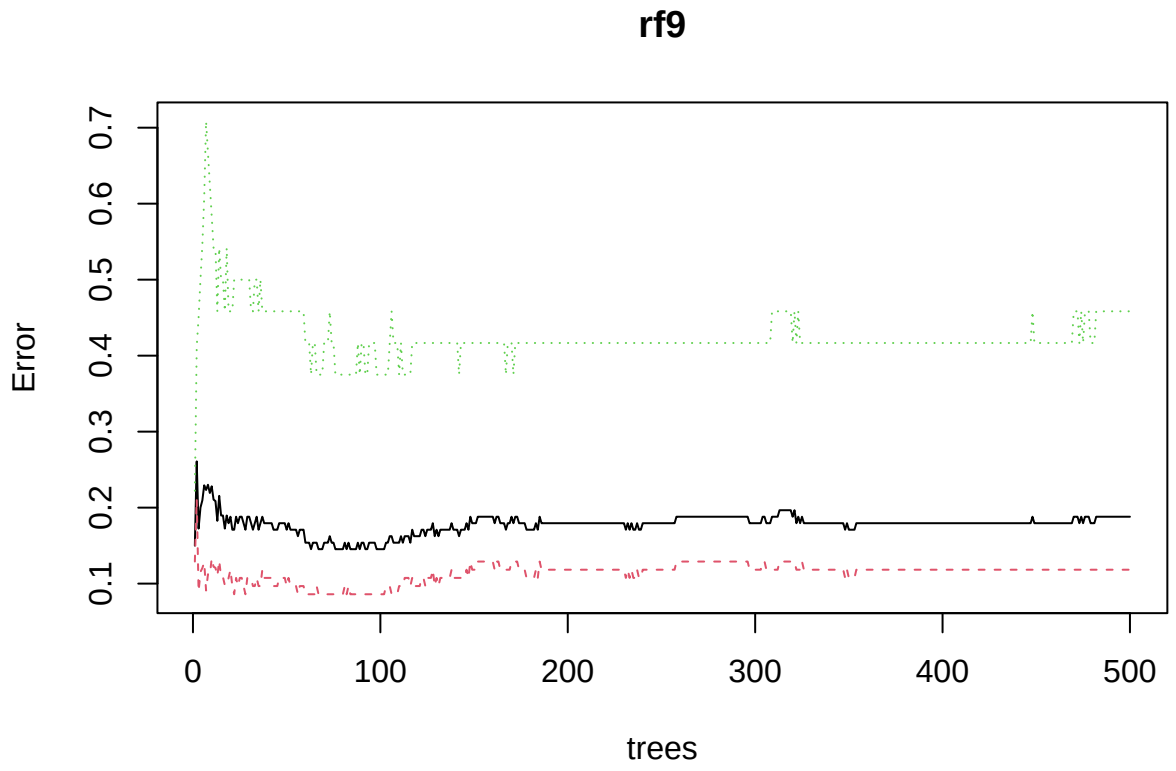

```
plot(rf7)
```



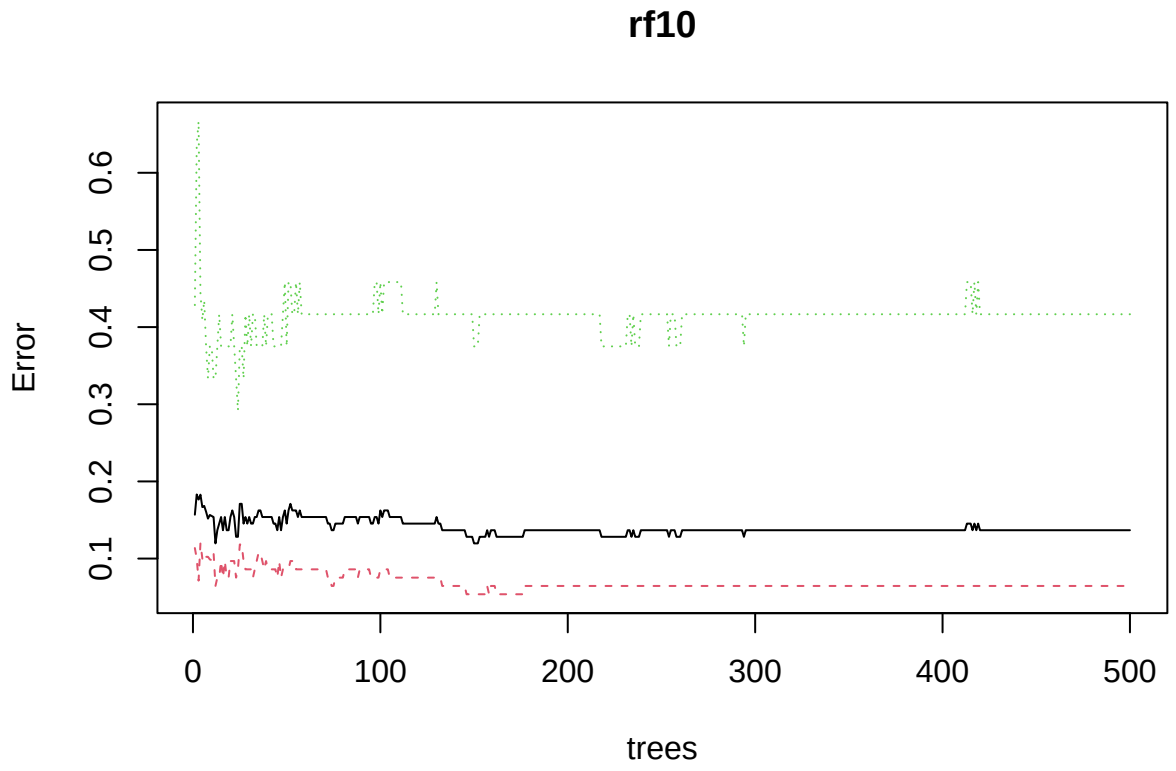

```
plot(rf8)
```



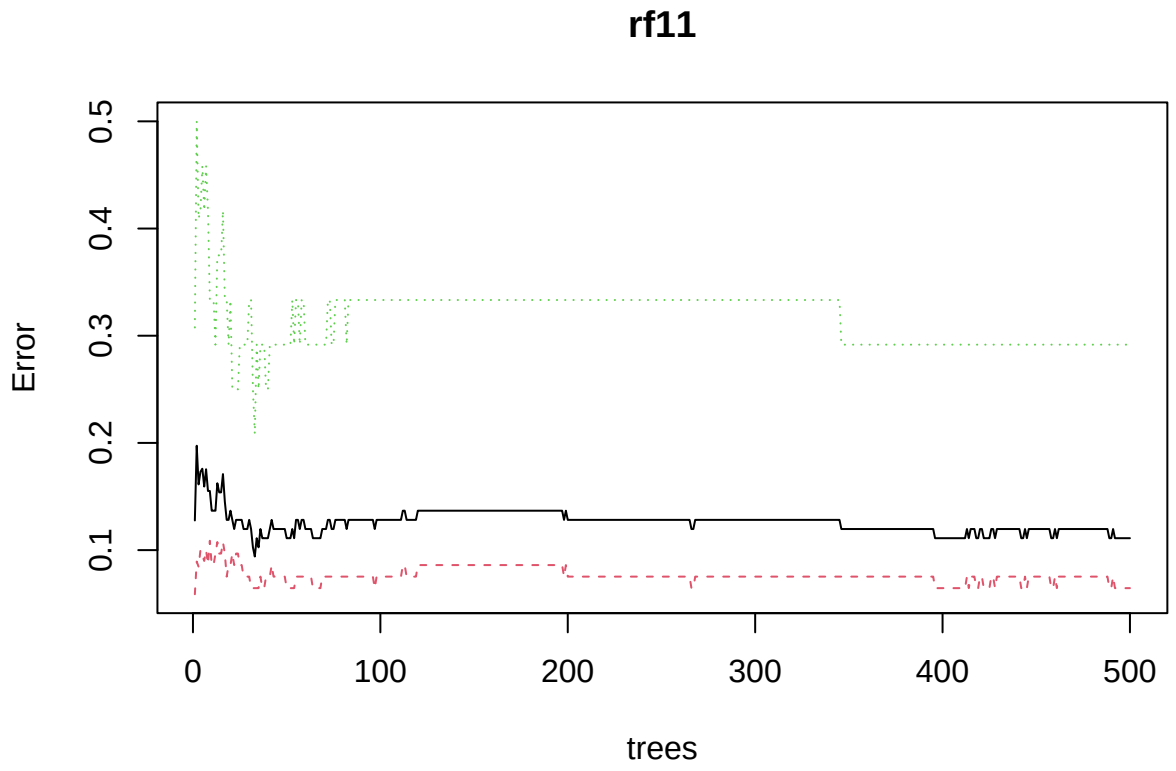

```
plot(rf9)
```



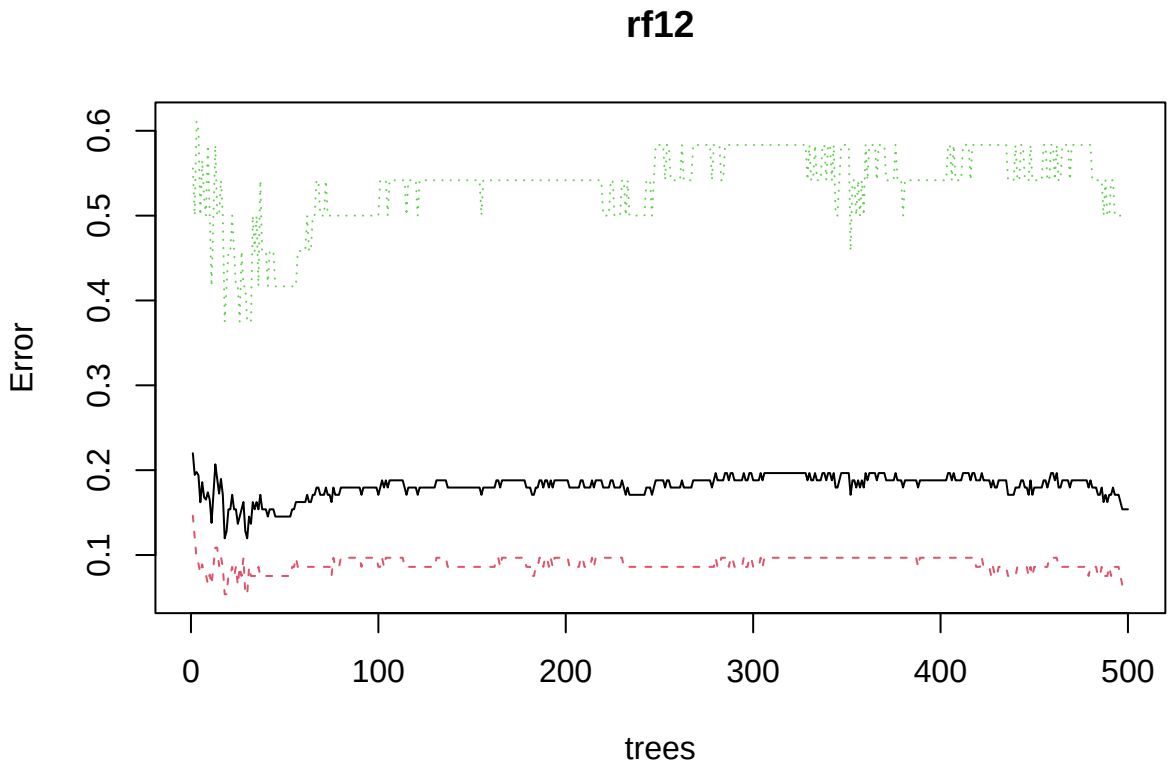

```
plot(rf10)
```




```
plot(rf11)
```

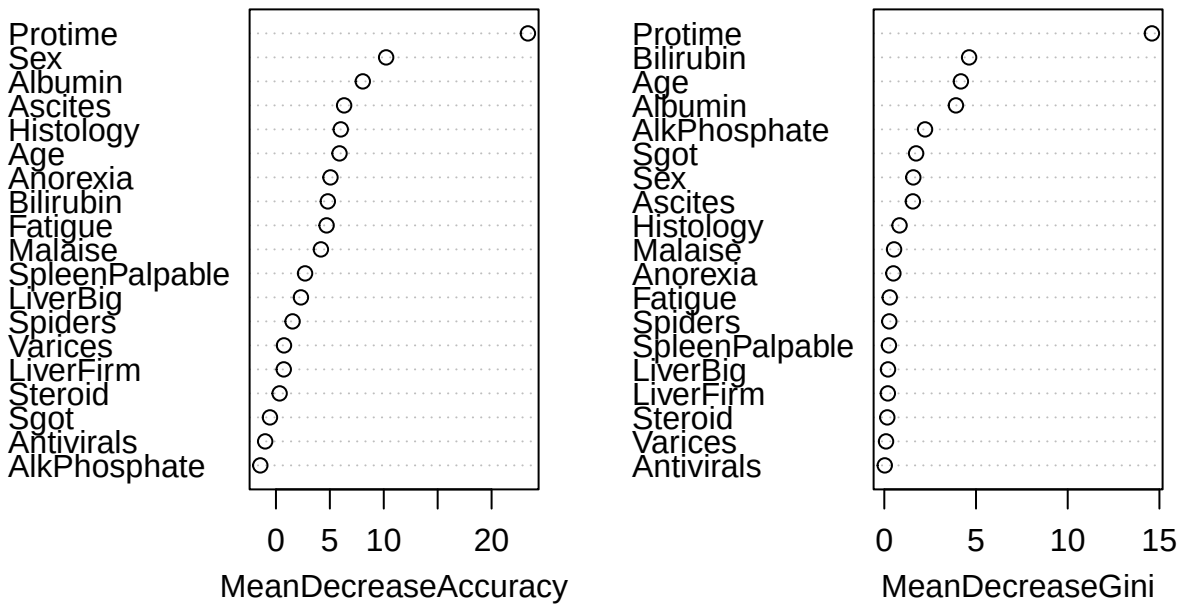



```
plot(rf12)
```

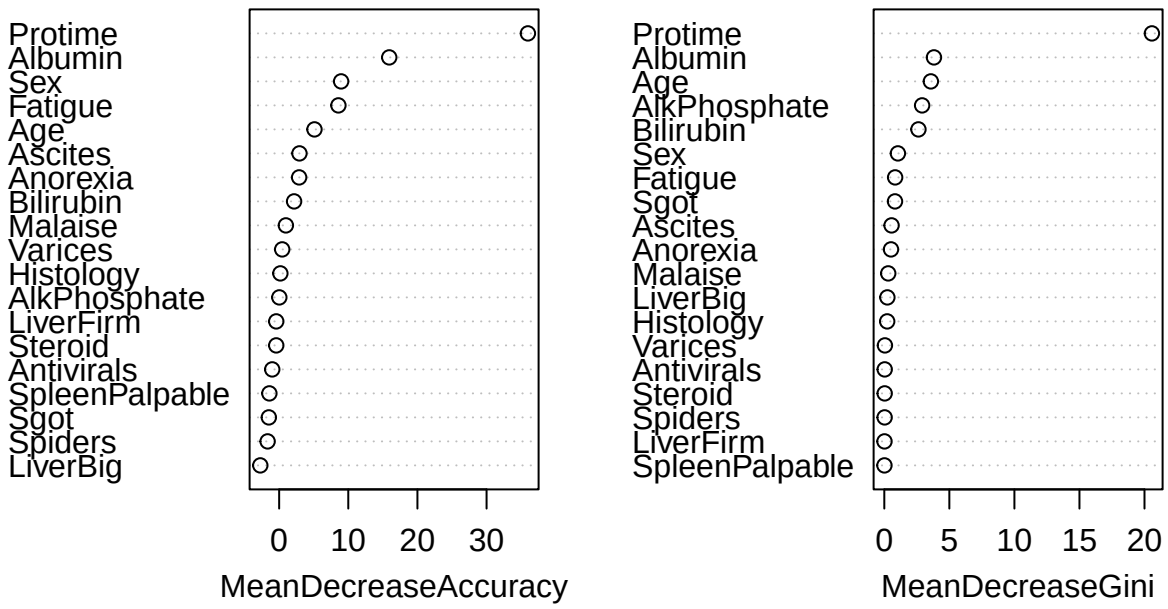


We can see variable importance ranking for other random forest models. The results are different, but for each model, Protime is in the first place.

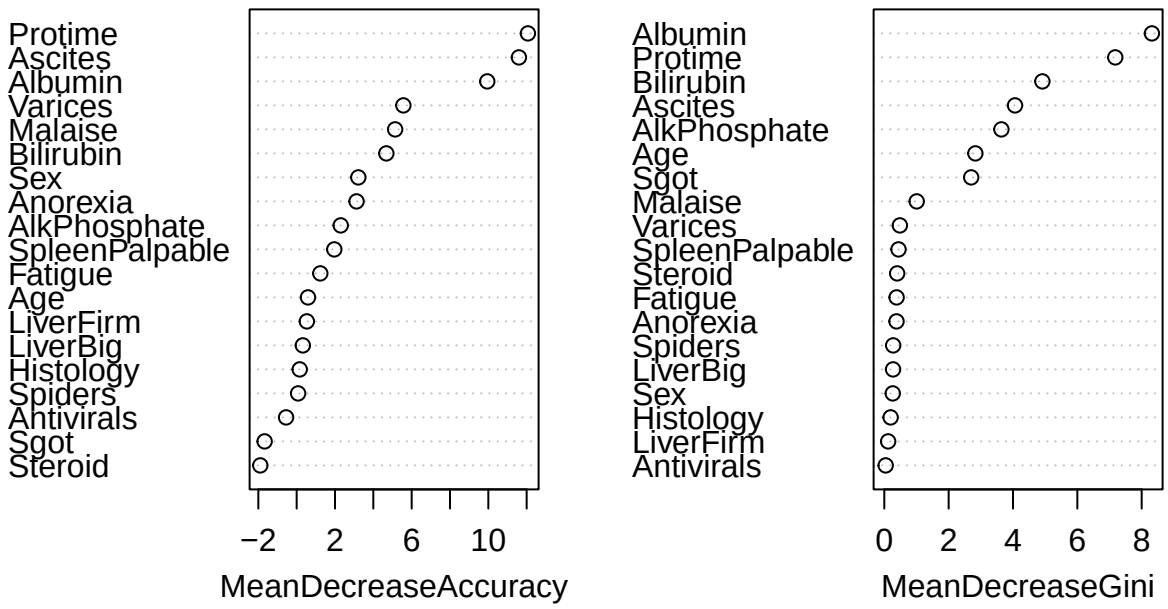
```
# Variable importance ranking
varImpPlot(rf1, main = "Variable Importance Plot")
```



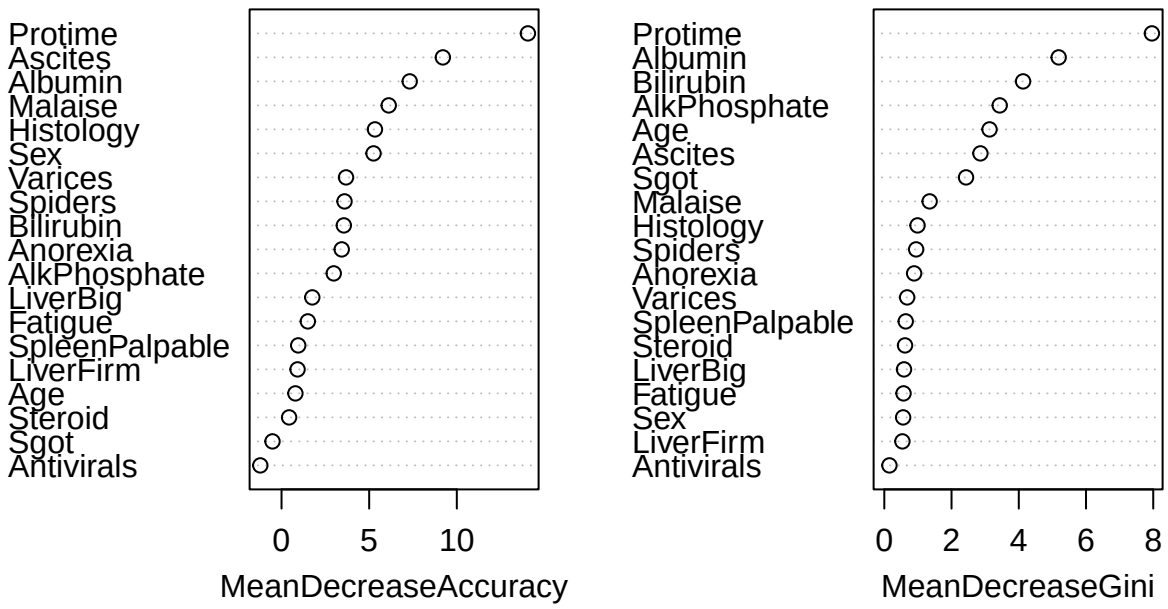

```
varImpPlot(rf2, main = "Variable Importance Plot")
```



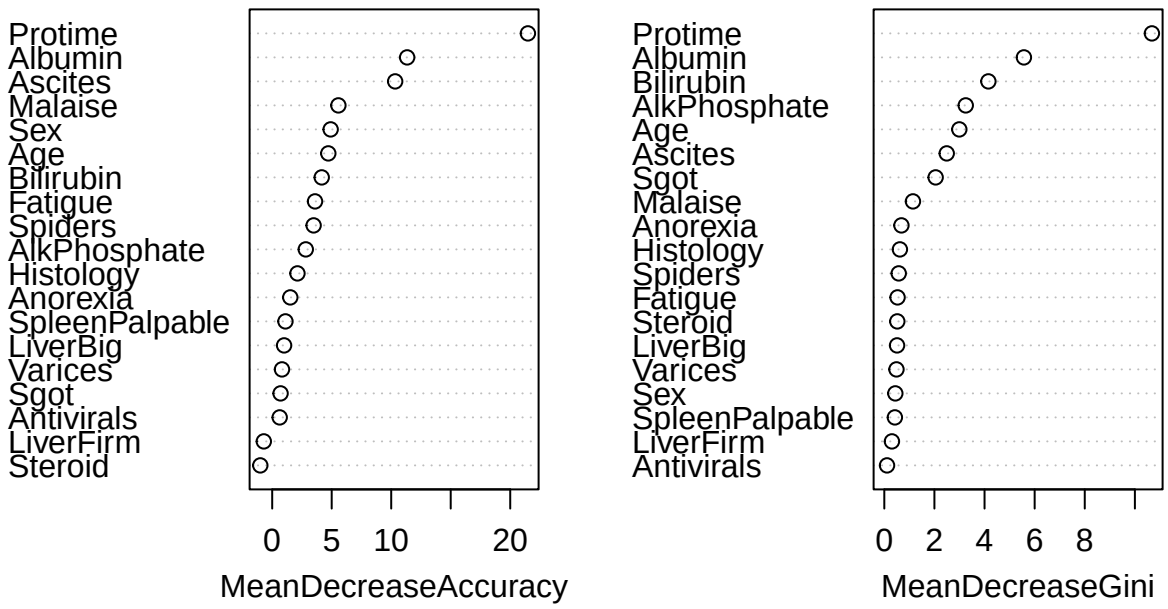

```
varImpPlot(rf3, main = "Variable Importance Plot")
```



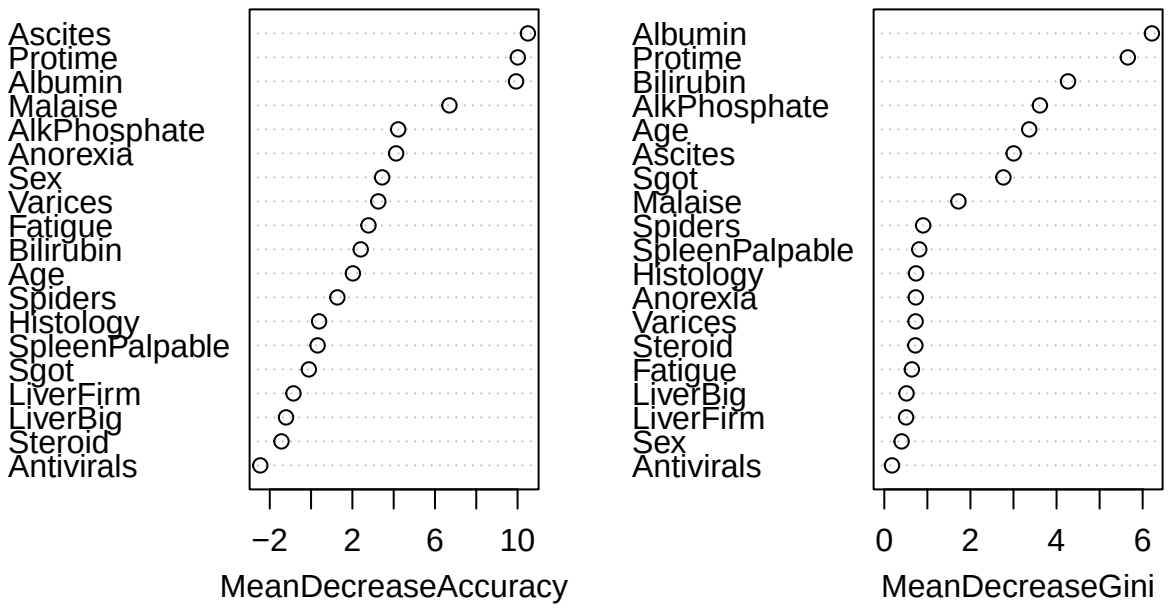

```
varImpPlot(rf4, main = "Variable Importance Plot")
```



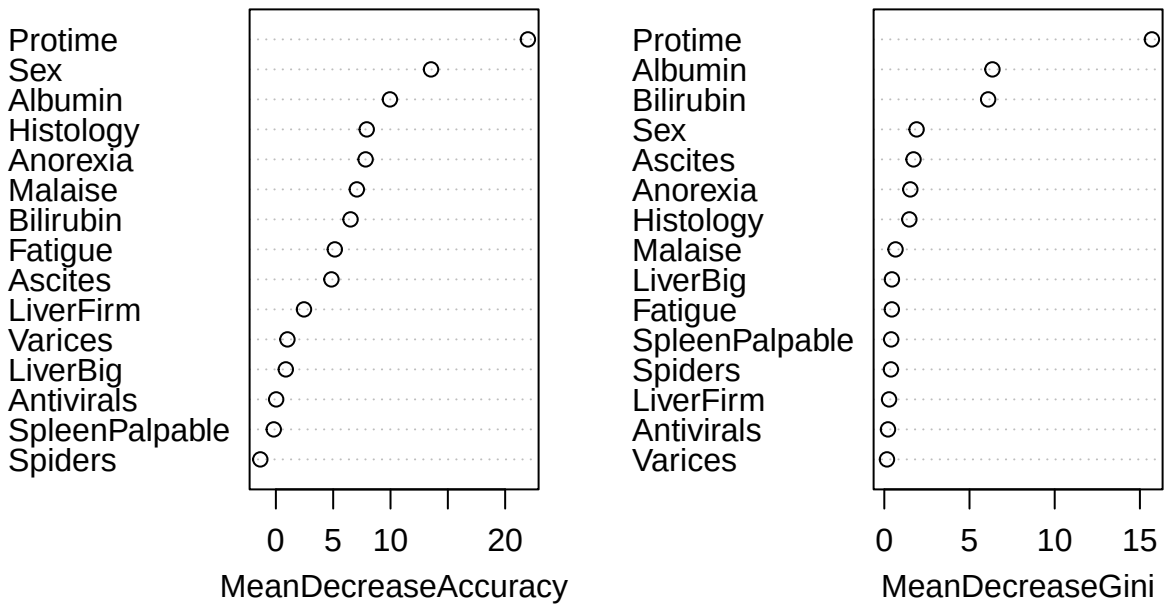

```
varImpPlot(rf5, main = "Variable Importance Plot")
```



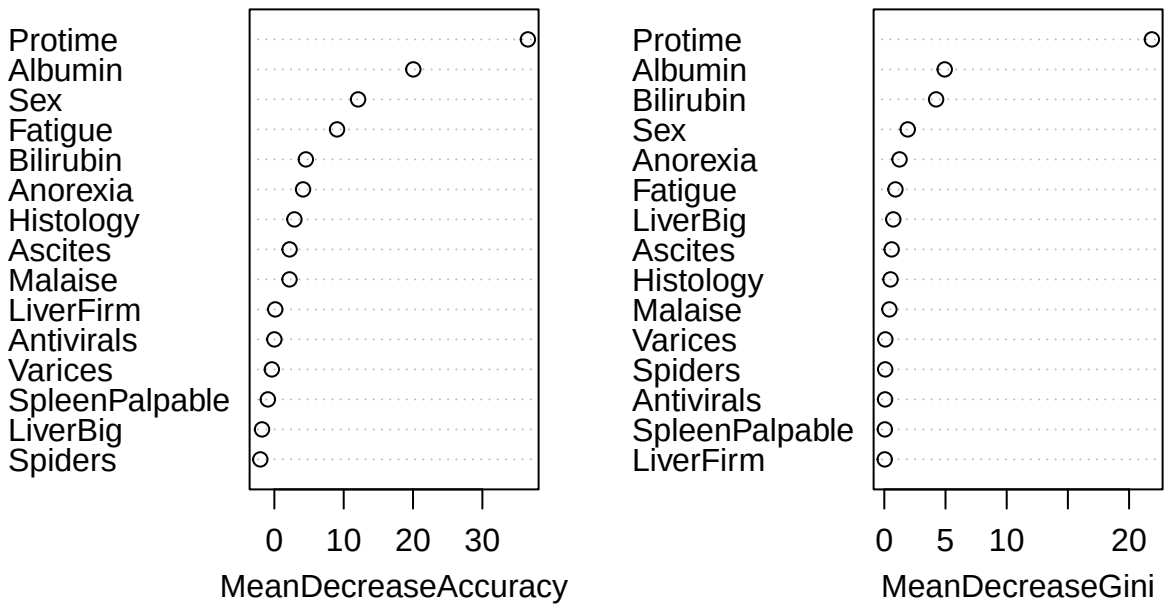

```r
varImpPlot(rf6, main = "Variable Importance Plot")
```



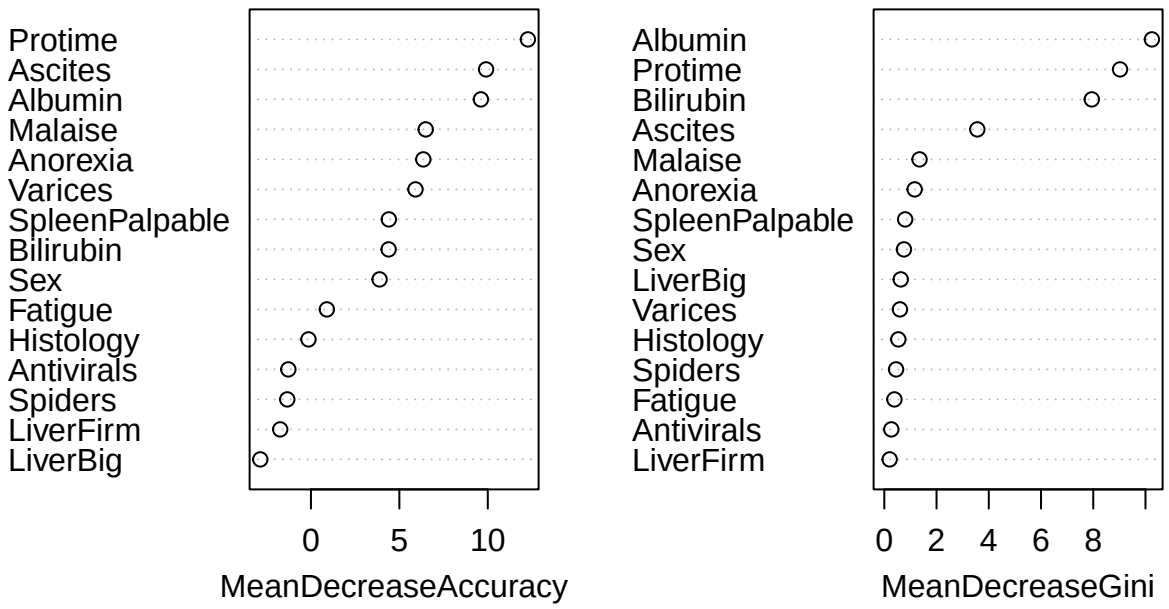

```r
varImpPlot(rf7, main = "Variable Importance Plot")
```



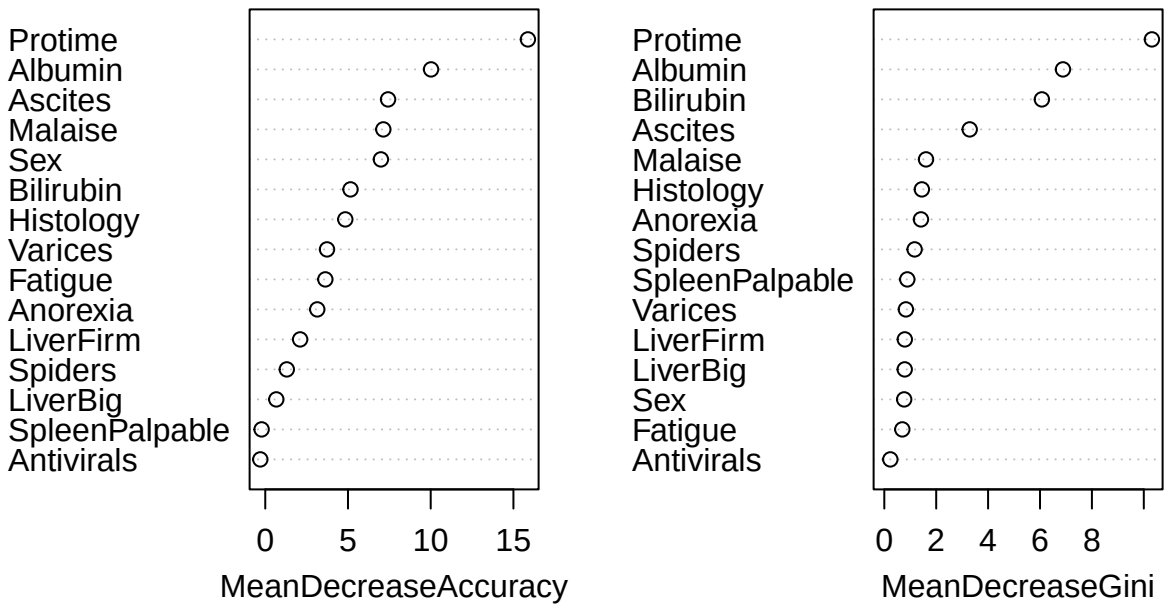

```
varImpPlot(rf8, main = "Variable Importance Plot")
```



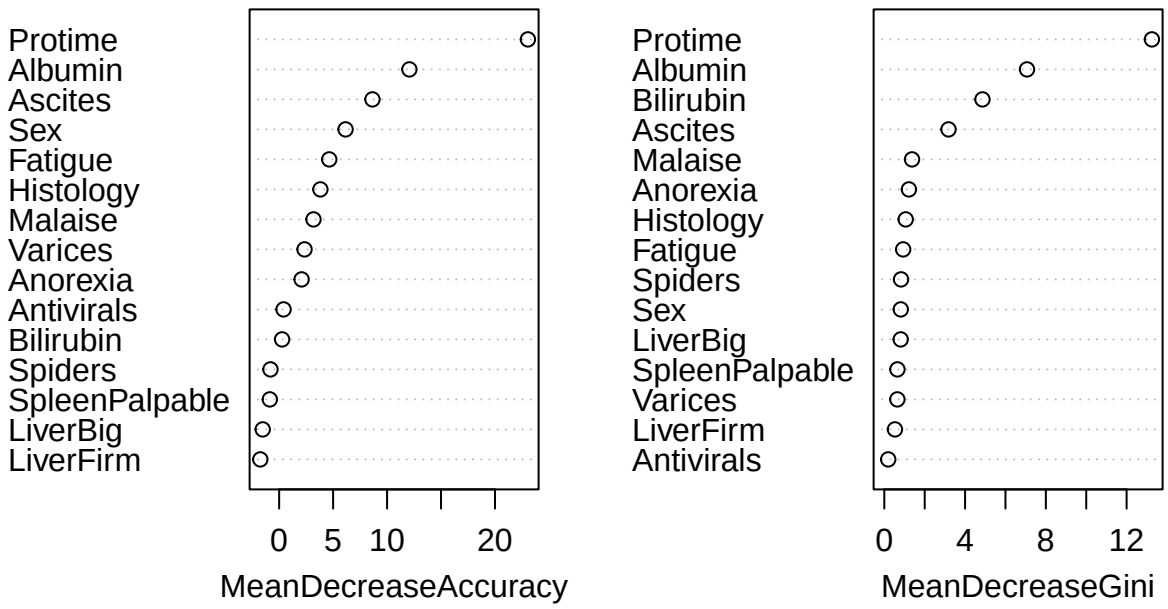

```r
varImpPlot(rf9, main = "Variable Importance Plot")
```



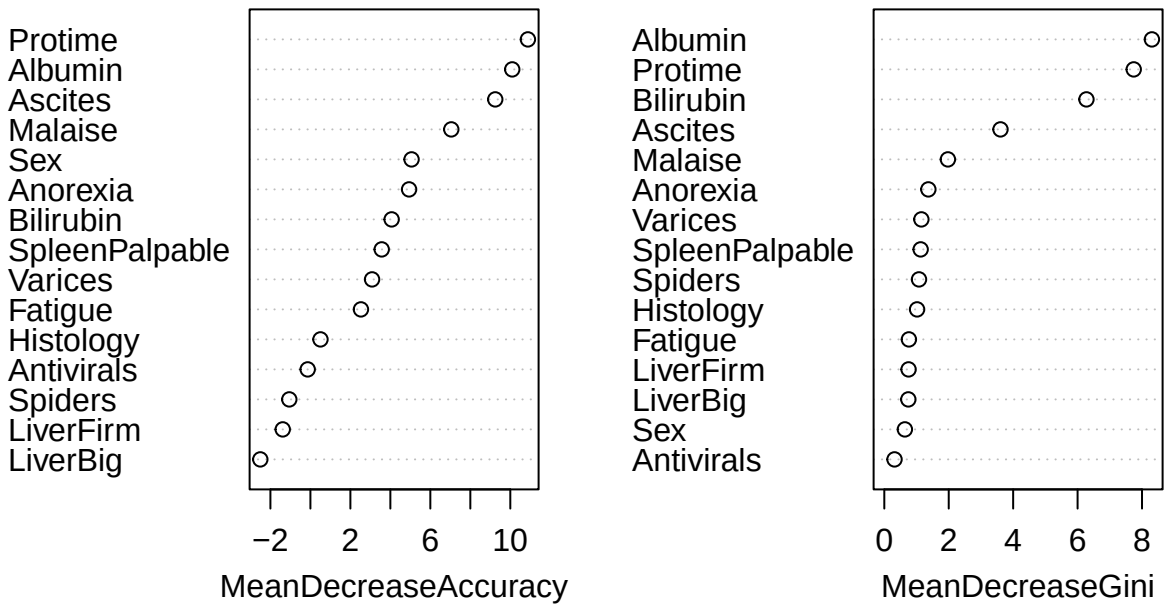

```r
varImpPlot(rf10, main = "Variable Importance Plot")
```




```r
varImpPlot(rf11, main = "Variable Importance Plot")
```




```
varImpPlot(rf12, main = "Variable Importance Plot")
```


Variable Importance Plot
Protime
Albumin
Ascites
Malaise
Sex
Anorexia
Bilirubin
SpleenPalpable
Varices
Fatigue
Histology
Antivirals
Spiders
LiverFirm
LiverBig
−2
2
6
10
MeanDecreaseAccuracy
Albumin
Protime
Bilirubin
Ascites
Malaise
Anorexia
Varices
SpleenPalpable
Spiders
Histology
Fatigue
LiverFirm
LiverBig
Sex
Antivirals
0
2
4
6
8
MeanDecreaseGini