M6L2\_Assignment

Xueyi Fan

June 30, 2016

# Assignemnt

1. Go to the UC Irvine Machine Learning Repository and find a dataset for supervised classification. Every student MUST use a different dataset so you MUST get approved for which you can going to use. This can be the same dataset you used for the unsupervised clustering as long as the data has some labeled data.
2. Generate a Decision Tree with your data. You can use any method/package you wish. Answer the following questions:

* Does the size of the data set make a difference?
* Do the rules make sense? If so why did the algorithm generate good rules? If not, why not?
* Does scaling, normalization or leaving the data unscaled make a difference?

library("ggplot2")  
library("C50")  
library("gmodels")  
library("rpart")  
library("RColorBrewer")  
library("tree")  
library("party")

## Loading required package: grid

## Loading required package: mvtnorm

## Loading required package: modeltools

## Loading required package: stats4

## Loading required package: strucchange

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

## Loading required package: sandwich

# Answer:

## 1. Loading the dataset

Here, I choose the [Breast Cancer Wisconsin (Diagnostic) data set](https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/)

data\_url <- 'https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.data'  
  
data <- read.table(url(data\_url), sep = ',')  
  
names(data) <- c('ID number', 'Diagnosis','radius\_mean','texure\_mean','perimeter\_mean','area\_mean','smoothness\_mean','compactness\_mean','concavity\_mean','concave\_points\_mean','symmetry\_mean','fractal\_dimension\_mean', 'radius\_SE','texure\_SE','perimeter\_SE','area\_SE','smoothness\_SE','compactness\_SE','concavity\_SE','concave\_points\_SE','symmetry\_SE','fractal\_dimension\_SE','radius\_worst','texure\_worst','perimeter\_worst','area\_worst','smoothness\_worst','compactness\_worst','concavity\_worst','concave\_points\_worst','symmetry\_worst','fractal\_dimension\_worst')   
  
head(data)

## ID number Diagnosis radius\_mean texure\_mean perimeter\_mean area\_mean  
## 1 842302 M 17.99 10.38 122.80 1001.0  
## 2 842517 M 20.57 17.77 132.90 1326.0  
## 3 84300903 M 19.69 21.25 130.00 1203.0  
## 4 84348301 M 11.42 20.38 77.58 386.1  
## 5 84358402 M 20.29 14.34 135.10 1297.0  
## 6 843786 M 12.45 15.70 82.57 477.1  
## smoothness\_mean compactness\_mean concavity\_mean concave\_points\_mean  
## 1 0.11840 0.27760 0.3001 0.14710  
## 2 0.08474 0.07864 0.0869 0.07017  
## 3 0.10960 0.15990 0.1974 0.12790  
## 4 0.14250 0.28390 0.2414 0.10520  
## 5 0.10030 0.13280 0.1980 0.10430  
## 6 0.12780 0.17000 0.1578 0.08089  
## symmetry\_mean fractal\_dimension\_mean radius\_SE texure\_SE perimeter\_SE  
## 1 0.2419 0.07871 1.0950 0.9053 8.589  
## 2 0.1812 0.05667 0.5435 0.7339 3.398  
## 3 0.2069 0.05999 0.7456 0.7869 4.585  
## 4 0.2597 0.09744 0.4956 1.1560 3.445  
## 5 0.1809 0.05883 0.7572 0.7813 5.438  
## 6 0.2087 0.07613 0.3345 0.8902 2.217  
## area\_SE smoothness\_SE compactness\_SE concavity\_SE concave\_points\_SE  
## 1 153.40 0.006399 0.04904 0.05373 0.01587  
## 2 74.08 0.005225 0.01308 0.01860 0.01340  
## 3 94.03 0.006150 0.04006 0.03832 0.02058  
## 4 27.23 0.009110 0.07458 0.05661 0.01867  
## 5 94.44 0.011490 0.02461 0.05688 0.01885  
## 6 27.19 0.007510 0.03345 0.03672 0.01137  
## symmetry\_SE fractal\_dimension\_SE radius\_worst texure\_worst  
## 1 0.03003 0.006193 25.38 17.33  
## 2 0.01389 0.003532 24.99 23.41  
## 3 0.02250 0.004571 23.57 25.53  
## 4 0.05963 0.009208 14.91 26.50  
## 5 0.01756 0.005115 22.54 16.67  
## 6 0.02165 0.005082 15.47 23.75  
## perimeter\_worst area\_worst smoothness\_worst compactness\_worst  
## 1 184.60 2019.0 0.1622 0.6656  
## 2 158.80 1956.0 0.1238 0.1866  
## 3 152.50 1709.0 0.1444 0.4245  
## 4 98.87 567.7 0.2098 0.8663  
## 5 152.20 1575.0 0.1374 0.2050  
## 6 103.40 741.6 0.1791 0.5249  
## concavity\_worst concave\_points\_worst symmetry\_worst  
## 1 0.7119 0.2654 0.4601  
## 2 0.2416 0.1860 0.2750  
## 3 0.4504 0.2430 0.3613  
## 4 0.6869 0.2575 0.6638  
## 5 0.4000 0.1625 0.2364  
## 6 0.5355 0.1741 0.3985  
## fractal\_dimension\_worst  
## 1 0.11890  
## 2 0.08902  
## 3 0.08758  
## 4 0.17300  
## 5 0.07678  
## 6 0.12440

str(data)

## 'data.frame': 569 obs. of 32 variables:  
## $ ID number : int 842302 842517 84300903 84348301 84358402 843786 844359 84458202 844981 84501001 ...  
## $ Diagnosis : Factor w/ 2 levels "B","M": 2 2 2 2 2 2 2 2 2 2 ...  
## $ radius\_mean : num 18 20.6 19.7 11.4 20.3 ...  
## $ texure\_mean : num 10.4 17.8 21.2 20.4 14.3 ...  
## $ perimeter\_mean : num 122.8 132.9 130 77.6 135.1 ...  
## $ area\_mean : num 1001 1326 1203 386 1297 ...  
## $ smoothness\_mean : num 0.1184 0.0847 0.1096 0.1425 0.1003 ...  
## $ compactness\_mean : num 0.2776 0.0786 0.1599 0.2839 0.1328 ...  
## $ concavity\_mean : num 0.3001 0.0869 0.1974 0.2414 0.198 ...  
## $ concave\_points\_mean : num 0.1471 0.0702 0.1279 0.1052 0.1043 ...  
## $ symmetry\_mean : num 0.242 0.181 0.207 0.26 0.181 ...  
## $ fractal\_dimension\_mean : num 0.0787 0.0567 0.06 0.0974 0.0588 ...  
## $ radius\_SE : num 1.095 0.543 0.746 0.496 0.757 ...  
## $ texure\_SE : num 0.905 0.734 0.787 1.156 0.781 ...  
## $ perimeter\_SE : num 8.59 3.4 4.58 3.44 5.44 ...  
## $ area\_SE : num 153.4 74.1 94 27.2 94.4 ...  
## $ smoothness\_SE : num 0.0064 0.00522 0.00615 0.00911 0.01149 ...  
## $ compactness\_SE : num 0.049 0.0131 0.0401 0.0746 0.0246 ...  
## $ concavity\_SE : num 0.0537 0.0186 0.0383 0.0566 0.0569 ...  
## $ concave\_points\_SE : num 0.0159 0.0134 0.0206 0.0187 0.0188 ...  
## $ symmetry\_SE : num 0.03 0.0139 0.0225 0.0596 0.0176 ...  
## $ fractal\_dimension\_SE : num 0.00619 0.00353 0.00457 0.00921 0.00511 ...  
## $ radius\_worst : num 25.4 25 23.6 14.9 22.5 ...  
## $ texure\_worst : num 17.3 23.4 25.5 26.5 16.7 ...  
## $ perimeter\_worst : num 184.6 158.8 152.5 98.9 152.2 ...  
## $ area\_worst : num 2019 1956 1709 568 1575 ...  
## $ smoothness\_worst : num 0.162 0.124 0.144 0.21 0.137 ...  
## $ compactness\_worst : num 0.666 0.187 0.424 0.866 0.205 ...  
## $ concavity\_worst : num 0.712 0.242 0.45 0.687 0.4 ...  
## $ concave\_points\_worst : num 0.265 0.186 0.243 0.258 0.163 ...  
## $ symmetry\_worst : num 0.46 0.275 0.361 0.664 0.236 ...  
## $ fractal\_dimension\_worst: num 0.1189 0.089 0.0876 0.173 0.0768 ...

summary(data)

## ID number Diagnosis radius\_mean texure\_mean   
## Min. : 8670 B:357 Min. : 6.981 Min. : 9.71   
## 1st Qu.: 869218 M:212 1st Qu.:11.700 1st Qu.:16.17   
## Median : 906024 Median :13.370 Median :18.84   
## Mean : 30371831 Mean :14.127 Mean :19.29   
## 3rd Qu.: 8813129 3rd Qu.:15.780 3rd Qu.:21.80   
## Max. :911320502 Max. :28.110 Max. :39.28   
## perimeter\_mean area\_mean smoothness\_mean compactness\_mean   
## Min. : 43.79 Min. : 143.5 Min. :0.05263 Min. :0.01938   
## 1st Qu.: 75.17 1st Qu.: 420.3 1st Qu.:0.08637 1st Qu.:0.06492   
## Median : 86.24 Median : 551.1 Median :0.09587 Median :0.09263   
## Mean : 91.97 Mean : 654.9 Mean :0.09636 Mean :0.10434   
## 3rd Qu.:104.10 3rd Qu.: 782.7 3rd Qu.:0.10530 3rd Qu.:0.13040   
## Max. :188.50 Max. :2501.0 Max. :0.16340 Max. :0.34540   
## concavity\_mean concave\_points\_mean symmetry\_mean   
## Min. :0.00000 Min. :0.00000 Min. :0.1060   
## 1st Qu.:0.02956 1st Qu.:0.02031 1st Qu.:0.1619   
## Median :0.06154 Median :0.03350 Median :0.1792   
## Mean :0.08880 Mean :0.04892 Mean :0.1812   
## 3rd Qu.:0.13070 3rd Qu.:0.07400 3rd Qu.:0.1957   
## Max. :0.42680 Max. :0.20120 Max. :0.3040   
## fractal\_dimension\_mean radius\_SE texure\_SE perimeter\_SE   
## Min. :0.04996 Min. :0.1115 Min. :0.3602 Min. : 0.757   
## 1st Qu.:0.05770 1st Qu.:0.2324 1st Qu.:0.8339 1st Qu.: 1.606   
## Median :0.06154 Median :0.3242 Median :1.1080 Median : 2.287   
## Mean :0.06280 Mean :0.4052 Mean :1.2169 Mean : 2.866   
## 3rd Qu.:0.06612 3rd Qu.:0.4789 3rd Qu.:1.4740 3rd Qu.: 3.357   
## Max. :0.09744 Max. :2.8730 Max. :4.8850 Max. :21.980   
## area\_SE smoothness\_SE compactness\_SE concavity\_SE   
## Min. : 6.802 Min. :0.001713 Min. :0.002252 Min. :0.00000   
## 1st Qu.: 17.850 1st Qu.:0.005169 1st Qu.:0.013080 1st Qu.:0.01509   
## Median : 24.530 Median :0.006380 Median :0.020450 Median :0.02589   
## Mean : 40.337 Mean :0.007041 Mean :0.025478 Mean :0.03189   
## 3rd Qu.: 45.190 3rd Qu.:0.008146 3rd Qu.:0.032450 3rd Qu.:0.04205   
## Max. :542.200 Max. :0.031130 Max. :0.135400 Max. :0.39600   
## concave\_points\_SE symmetry\_SE fractal\_dimension\_SE  
## Min. :0.000000 Min. :0.007882 Min. :0.0008948   
## 1st Qu.:0.007638 1st Qu.:0.015160 1st Qu.:0.0022480   
## Median :0.010930 Median :0.018730 Median :0.0031870   
## Mean :0.011796 Mean :0.020542 Mean :0.0037949   
## 3rd Qu.:0.014710 3rd Qu.:0.023480 3rd Qu.:0.0045580   
## Max. :0.052790 Max. :0.078950 Max. :0.0298400   
## radius\_worst texure\_worst perimeter\_worst area\_worst   
## Min. : 7.93 Min. :12.02 Min. : 50.41 Min. : 185.2   
## 1st Qu.:13.01 1st Qu.:21.08 1st Qu.: 84.11 1st Qu.: 515.3   
## Median :14.97 Median :25.41 Median : 97.66 Median : 686.5   
## Mean :16.27 Mean :25.68 Mean :107.26 Mean : 880.6   
## 3rd Qu.:18.79 3rd Qu.:29.72 3rd Qu.:125.40 3rd Qu.:1084.0   
## Max. :36.04 Max. :49.54 Max. :251.20 Max. :4254.0   
## smoothness\_worst compactness\_worst concavity\_worst concave\_points\_worst  
## Min. :0.07117 Min. :0.02729 Min. :0.0000 Min. :0.00000   
## 1st Qu.:0.11660 1st Qu.:0.14720 1st Qu.:0.1145 1st Qu.:0.06493   
## Median :0.13130 Median :0.21190 Median :0.2267 Median :0.09993   
## Mean :0.13237 Mean :0.25427 Mean :0.2722 Mean :0.11461   
## 3rd Qu.:0.14600 3rd Qu.:0.33910 3rd Qu.:0.3829 3rd Qu.:0.16140   
## Max. :0.22260 Max. :1.05800 Max. :1.2520 Max. :0.29100   
## symmetry\_worst fractal\_dimension\_worst  
## Min. :0.1565 Min. :0.05504   
## 1st Qu.:0.2504 1st Qu.:0.07146   
## Median :0.2822 Median :0.08004   
## Mean :0.2901 Mean :0.08395   
## 3rd Qu.:0.3179 3rd Qu.:0.09208   
## Max. :0.6638 Max. :0.20750

#shuffle the data   
set.seed(123)  
cancer\_data <- data[order(runif(nrow(data))),]  
cancer\_data[,2]

## [1] B M B B B B B B B M B B B B B B M M B M B B B B B B B M B B B B B M B  
## [36] B M B B B B M B B B B B B B B B B M M B B B B B B B B M B M B B B B B  
## [71] M B B M B B B B M M M B M B B B B M B B B M B B B B B B M B B M B B M  
## [106] M M B B B M B B M B B B B B M M B M M B B B B M B B B M B M B B B B M  
## [141] B B B M B B M B B M B B M B B M B B B M B B B M M M M B B B B B B M B  
## [176] B B B M B M B B M M M B M B B B M M B B M M M B B B M M M B B M B B M  
## [211] B B B M B M M B B B B B B B M B B B B B B B M B B B M B B B M M M M B  
## [246] M M M M M M B M B M B B B M M B B B B B B M B B M M M M M B B B M M B  
## [281] B M B M B B B B M B M B B B B M B B B B B B M M M M B M B B M B M B M  
## [316] B B B B B B M B B B B M B M B B B M M M B M B M B B M B M M M B M M B  
## [351] M M M B M B M B B B B B B B B M M B B B B M M M M M B M M B B B B B B  
## [386] B B M M M M M M B B M M B B M M B B B B M B M M B M M B M B B B M B B  
## [421] B B B B M B B M B B B B M B M B M M M B B M M M B M B M M M M B M M M  
## [456] B M B B B B B M B B B M B B M B M B M B B B B M B B M B B M B B B B B  
## [491] B M B M M M M B B M B M M B B B B M B M M M B B M B M B B B M B M M M  
## [526] M B B B M B M B B M B M M B B M B M M B B B M B M B B M B M M B B B M  
## [561] B B M B B B M M B  
## Levels: B M

# 2. Generate a Decision Tree with my data

First of all, I will normalize my data

normalize <- function(x){  
 return((x-min(x))/(max(x)-min(x)))  
}  
  
cancer.normalized <- as.data.frame(lapply(cancer\_data[,c(3:32)], normalize))  
head(cancer.normalized)

## radius\_mean texure\_mean perimeter\_mean area\_mean smoothness\_mean  
## 1 0.32746462 0.2336828 0.31221063 0.19338282 0.1412837  
## 2 0.32273179 0.2056138 0.32229977 0.18689290 0.4339623  
## 3 0.20393772 0.1126141 0.19653099 0.10371156 0.4126569  
## 4 0.09555587 0.1586067 0.08686338 0.04360551 0.1572628  
## 5 0.25931185 0.3442678 0.25319605 0.13904560 0.2878036  
## 6 0.21056368 0.2570172 0.20641283 0.10795334 0.5106978  
## compactness\_mean concavity\_mean concave\_points\_mean symmetry\_mean  
## 1 0.10370529 0.052108716 0.06655070 0.3803030  
## 2 0.33316974 0.182497657 0.25193837 0.3040404  
## 3 0.17391571 0.076499531 0.13692843 0.3580808  
## 4 0.03613275 0.008624649 0.01725646 0.3676768  
## 5 0.25157966 0.160028116 0.15402584 0.3641414  
## 6 0.23151954 0.047586692 0.09249503 0.2954545  
## fractal\_dimension\_mean radius\_SE texure\_SE perimeter\_SE area\_SE  
## 1 0.1137321 0.01593337 0.04773692 0.02992037 0.013500237  
## 2 0.3306655 0.06054680 0.05752740 0.05654243 0.030907848  
## 3 0.2683235 0.02857143 0.03737182 0.01917731 0.011893956  
## 4 0.3868997 0.02284990 0.20880481 0.01823493 0.005579027  
## 5 0.2639006 0.09150824 0.15023868 0.08585026 0.040190662  
## 6 0.3298231 0.06257469 0.34207037 0.05560006 0.027228342  
## smoothness\_SE compactness\_SE concavity\_SE concave\_points\_SE symmetry\_SE  
## 1 0.05422035 0.08184877 0.025010101 0.09153249 0.07426690  
## 2 0.10211782 0.13817707 0.044419192 0.17438909 0.06075871  
## 3 0.16177720 0.06735362 0.032373737 0.16762644 0.12717397  
## 4 0.19294965 0.01975997 0.009295455 0.06577003 0.26915067  
## 5 0.16374885 0.23611320 0.118989899 0.26577003 0.27463837  
## 6 0.16830404 0.11339262 0.029242424 0.18033719 0.21019305  
## fractal\_dimension\_SE radius\_worst texure\_worst perimeter\_worst  
## 1 0.04146456 0.25649235 0.2606610 0.25295084  
## 2 0.07722178 0.30736393 0.2356077 0.29827183  
## 3 0.06640825 0.15617218 0.1108742 0.13875193  
## 4 0.04346835 0.06293134 0.2145522 0.05224364  
## 5 0.12976936 0.19672714 0.2945096 0.18785796  
## 6 0.09090281 0.17253646 0.3928571 0.16061557  
## area\_worst smoothness\_worst compactness\_worst concavity\_worst  
## 1 0.13116890 0.1497061 0.16814623 0.11054313  
## 2 0.15414864 0.4617975 0.31716972 0.22196486  
## 3 0.06692391 0.4267979 0.11973300 0.10183706  
## 4 0.02465100 0.1812058 0.02428423 0.01175719  
## 5 0.08997739 0.2234036 0.18289335 0.15207668  
## 6 0.07810657 0.4631183 0.14738384 0.06156550  
## concave\_points\_worst symmetry\_worst fractal\_dimension\_worst  
## 1 0.21381443 0.2195939 0.1439066  
## 2 0.47525773 0.2018529 0.3145743  
## 3 0.30068729 0.2302385 0.1651581  
## 4 0.04773196 0.2810960 0.1508592  
## 5 0.26202749 0.2207767 0.1482356  
## 6 0.22000000 0.2008673 0.1700118

summary(cancer.normalized)

## radius\_mean texure\_mean perimeter\_mean area\_mean   
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.2233 1st Qu.:0.2185 1st Qu.:0.2168 1st Qu.:0.1174   
## Median :0.3024 Median :0.3088 Median :0.2933 Median :0.1729   
## Mean :0.3382 Mean :0.3240 Mean :0.3329 Mean :0.2169   
## 3rd Qu.:0.4164 3rd Qu.:0.4089 3rd Qu.:0.4168 3rd Qu.:0.2711   
## Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :1.0000   
## smoothness\_mean compactness\_mean concavity\_mean concave\_points\_mean  
## Min. :0.0000 Min. :0.0000 Min. :0.00000 Min. :0.0000   
## 1st Qu.:0.3046 1st Qu.:0.1397 1st Qu.:0.06926 1st Qu.:0.1009   
## Median :0.3904 Median :0.2247 Median :0.14419 Median :0.1665   
## Mean :0.3948 Mean :0.2606 Mean :0.20806 Mean :0.2431   
## 3rd Qu.:0.4755 3rd Qu.:0.3405 3rd Qu.:0.30623 3rd Qu.:0.3678   
## Max. :1.0000 Max. :1.0000 Max. :1.00000 Max. :1.0000   
## symmetry\_mean fractal\_dimension\_mean radius\_SE   
## Min. :0.0000 Min. :0.0000 Min. :0.00000   
## 1st Qu.:0.2823 1st Qu.:0.1630 1st Qu.:0.04378   
## Median :0.3697 Median :0.2439 Median :0.07702   
## Mean :0.3796 Mean :0.2704 Mean :0.10635   
## 3rd Qu.:0.4530 3rd Qu.:0.3404 3rd Qu.:0.13304   
## Max. :1.0000 Max. :1.0000 Max. :1.00000   
## texure\_SE perimeter\_SE area\_SE smoothness\_SE   
## Min. :0.0000 Min. :0.00000 Min. :0.00000 Min. :0.0000   
## 1st Qu.:0.1047 1st Qu.:0.04000 1st Qu.:0.02064 1st Qu.:0.1175   
## Median :0.1653 Median :0.07209 Median :0.03311 Median :0.1586   
## Mean :0.1893 Mean :0.09938 Mean :0.06264 Mean :0.1811   
## 3rd Qu.:0.2462 3rd Qu.:0.12251 3rd Qu.:0.07170 3rd Qu.:0.2187   
## Max. :1.0000 Max. :1.00000 Max. :1.00000 Max. :1.0000   
## compactness\_SE concavity\_SE concave\_points\_SE symmetry\_SE   
## Min. :0.00000 Min. :0.00000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.08132 1st Qu.:0.03811 1st Qu.:0.1447 1st Qu.:0.1024   
## Median :0.13667 Median :0.06538 Median :0.2070 Median :0.1526   
## Mean :0.17444 Mean :0.08054 Mean :0.2235 Mean :0.1781   
## 3rd Qu.:0.22680 3rd Qu.:0.10619 3rd Qu.:0.2787 3rd Qu.:0.2195   
## Max. :1.00000 Max. :1.00000 Max. :1.0000 Max. :1.0000   
## fractal\_dimension\_SE radius\_worst texure\_worst perimeter\_worst   
## Min. :0.00000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.04675 1st Qu.:0.1807 1st Qu.:0.2415 1st Qu.:0.1678   
## Median :0.07919 Median :0.2504 Median :0.3569 Median :0.2353   
## Mean :0.10019 Mean :0.2967 Mean :0.3640 Mean :0.2831   
## 3rd Qu.:0.12656 3rd Qu.:0.3863 3rd Qu.:0.4717 3rd Qu.:0.3735   
## Max. :1.00000 Max. :1.0000 Max. :1.0000 Max. :1.0000   
## area\_worst smoothness\_worst compactness\_worst concavity\_worst   
## Min. :0.00000 Min. :0.0000 Min. :0.0000 Min. :0.00000   
## 1st Qu.:0.08113 1st Qu.:0.3000 1st Qu.:0.1163 1st Qu.:0.09145   
## Median :0.12321 Median :0.3971 Median :0.1791 Median :0.18107   
## Mean :0.17091 Mean :0.4041 Mean :0.2202 Mean :0.21740   
## 3rd Qu.:0.22090 3rd Qu.:0.4942 3rd Qu.:0.3025 3rd Qu.:0.30583   
## Max. :1.00000 Max. :1.0000 Max. :1.0000 Max. :1.00000   
## concave\_points\_worst symmetry\_worst fractal\_dimension\_worst  
## Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.2231 1st Qu.:0.1851 1st Qu.:0.1077   
## Median :0.3434 Median :0.2478 Median :0.1640   
## Mean :0.3938 Mean :0.2633 Mean :0.1896   
## 3rd Qu.:0.5546 3rd Qu.:0.3182 3rd Qu.:0.2429   
## Max. :1.0000 Max. :1.0000 Max. :1.0000

#split the data frames  
cancer.normalized.train <- cancer.normalized[1:500,]  
cancer.normalized.test <- cancer.normalized[501:569,]  
cancer.normalized.train.target<- cancer\_data[1:500,2]  
cancer.normalized.test.target <- cancer\_data[501:569,2]  
  
#check the proportion of class variable  
  
prop.table(table(cancer\_data[,2]))

##   
## B M   
## 0.6274165 0.3725835

Then, I will use the method C5.0, rpart, tree, ctree to generate the decision trees.

#use C5.0  
cancer.c5.0 <- C5.0(cancer.normalized.train, cancer.normalized.train.target)  
cancer.c5.0

##   
## Call:  
## C5.0.default(x = cancer.normalized.train, y  
## = cancer.normalized.train.target)  
##   
## Classification Tree  
## Number of samples: 500   
## Number of predictors: 30   
##   
## Tree size: 9   
##   
## Non-standard options: attempt to group attributes

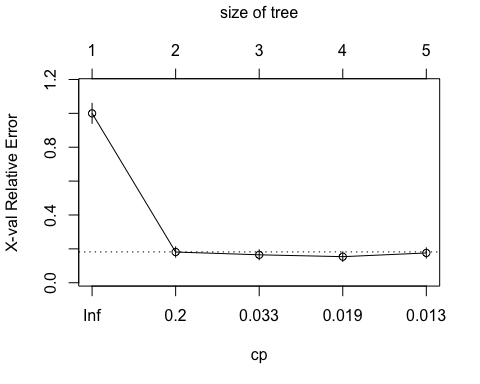
summary(cancer.c5.0)

##   
## Call:  
## C5.0.default(x = cancer.normalized.train, y  
## = cancer.normalized.train.target)  
##   
##   
## C5.0 [Release 2.07 GPL Edition] Thu Jun 30 19:43:09 2016  
## -------------------------------  
##   
## Class specified by attribute `outcome'  
##   
## Read 500 cases (31 attributes) from undefined.data  
##   
## Decision tree:  
##   
## radius\_worst > 0.3144788:  
## :...texure\_mean > 0.1782212: M (156/2)  
## : texure\_mean <= 0.1782212:  
## : :...concavity\_mean <= 0.3202905: B (6)  
## : concavity\_mean > 0.3202905: M (3)  
## radius\_worst <= 0.3144788:  
## :...concave\_points\_worst <= 0.466323: B (298/4)  
## concave\_points\_worst > 0.466323:  
## :...texure\_worst > 0.4123134: M (14)  
## texure\_worst <= 0.4123134:  
## :...concave\_points\_worst > 0.6147766: M (2)  
## concave\_points\_worst <= 0.6147766:  
## :...area\_worst <= 0.1311689: B (10)  
## area\_worst > 0.1311689:  
## :...radius\_mean <= 0.3411898: M (4)  
## radius\_mean > 0.3411898: B (7/1)  
##   
##   
## Evaluation on training data (500 cases):  
##   
## Decision Tree   
## ----------------   
## Size Errors   
##   
## 9 7( 1.4%) <<  
##   
##   
## (a) (b) <-classified as  
## ---- ----  
## 316 2 (a): class B  
## 5 177 (b): class M  
##   
##   
## Attribute usage:  
##   
## 100.00% radius\_worst  
## 67.00% concave\_points\_worst  
## 33.00% texure\_mean  
## 7.40% texure\_worst  
## 4.20% area\_worst  
## 2.20% radius\_mean  
## 1.80% concavity\_mean  
##   
##   
## Time: 0.0 secs

#use rpart  
cancer.rpart <- rpart(cancer\_data[1:500,2]~., data=cancer.normalized.train, method="class")  
printcp(cancer.rpart)

##   
## Classification tree:  
## rpart(formula = cancer\_data[1:500, 2] ~ ., data = cancer.normalized.train,   
## method = "class")  
##   
## Variables actually used in tree construction:  
## [1] concave\_points\_worst radius\_worst texure\_mean   
##   
## Root node error: 182/500 = 0.364  
##   
## n= 500   
##   
## CP nsplit rel error xerror xstd  
## 1 0.818681 0 1.000000 1.00000 0.059114  
## 2 0.049451 1 0.181319 0.18132 0.030504  
## 3 0.021978 2 0.131868 0.16484 0.029178  
## 4 0.016484 3 0.109890 0.15385 0.028248  
## 5 0.010000 4 0.093407 0.17582 0.030071

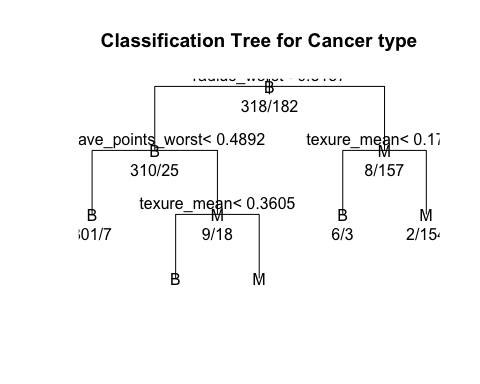
plotcp(cancer.rpart)



summary(cancer.rpart)

## Call:  
## rpart(formula = cancer\_data[1:500, 2] ~ ., data = cancer.normalized.train,   
## method = "class")  
## n= 500   
##   
## CP nsplit rel error xerror xstd  
## 1 0.81868132 0 1.00000000 1.0000000 0.05911434  
## 2 0.04945055 1 0.18131868 0.1813187 0.03050416  
## 3 0.02197802 2 0.13186813 0.1648352 0.02917784  
## 4 0.01648352 3 0.10989011 0.1538462 0.02824839  
## 5 0.01000000 4 0.09340659 0.1758242 0.03007056  
##   
## Variable importance  
## radius\_worst area\_worst perimeter\_worst   
## 16 16 16   
## area\_mean radius\_mean perimeter\_mean   
## 15 14 14   
## concave\_points\_worst texure\_mean concavity\_worst   
## 2 1 1   
## compactness\_worst texure\_worst concavity\_mean   
## 1 1 1   
## concave\_points\_mean symmetry\_worst   
## 1 1   
##   
## Node number 1: 500 observations, complexity param=0.8186813  
## predicted class=B expected loss=0.364 P(node) =1  
## class counts: 318 182  
## probabilities: 0.636 0.364   
## left son=2 (335 obs) right son=3 (165 obs)  
## Primary splits:  
## radius\_worst < 0.3157239 to the left, improve=170.0111, (0 missing)  
## area\_worst < 0.1718811 to the left, improve=165.0636, (0 missing)  
## concave\_points\_worst < 0.4891753 to the left, improve=163.1507, (0 missing)  
## perimeter\_worst < 0.3042482 to the left, improve=162.9821, (0 missing)  
## concave\_points\_mean < 0.2557406 to the left, improve=157.7055, (0 missing)  
## Surrogate splits:  
## area\_worst < 0.1678628 to the left, agree=0.994, adj=0.982, (0 split)  
## perimeter\_worst < 0.3052443 to the left, agree=0.982, adj=0.945, (0 split)  
## area\_mean < 0.2362036 to the left, agree=0.964, adj=0.891, (0 split)  
## radius\_mean < 0.3816555 to the left, agree=0.962, adj=0.885, (0 split)  
## perimeter\_mean < 0.3635892 to the left, agree=0.956, adj=0.867, (0 split)  
##   
## Node number 2: 335 observations, complexity param=0.04945055  
## predicted class=B expected loss=0.07462687 P(node) =0.67  
## class counts: 310 25  
## probabilities: 0.925 0.075   
## left son=4 (308 obs) right son=5 (27 obs)  
## Primary splits:  
## concave\_points\_worst < 0.4891753 to the left, improve=20.58684, (0 missing)  
## concave\_points\_mean < 0.2479374 to the left, improve=13.83297, (0 missing)  
## compactness\_worst < 0.346955 to the left, improve=12.72930, (0 missing)  
## symmetry\_worst < 0.3954268 to the left, improve=12.49617, (0 missing)  
## concavity\_worst < 0.2951278 to the left, improve=12.32023, (0 missing)  
## Surrogate splits:  
## concavity\_worst < 0.3500799 to the left, agree=0.955, adj=0.444, (0 split)  
## compactness\_worst < 0.346955 to the left, agree=0.952, adj=0.407, (0 split)  
## concavity\_mean < 0.3662137 to the left, agree=0.946, adj=0.333, (0 split)  
## concave\_points\_mean < 0.3285288 to the left, agree=0.946, adj=0.333, (0 split)  
## symmetry\_worst < 0.3954268 to the left, agree=0.940, adj=0.259, (0 split)  
##   
## Node number 3: 165 observations, complexity param=0.01648352  
## predicted class=M expected loss=0.04848485 P(node) =0.33  
## class counts: 8 157  
## probabilities: 0.048 0.952   
## left son=6 (9 obs) right son=7 (156 obs)  
## Primary splits:  
## texure\_mean < 0.1785594 to the left, improve=7.275524, (0 missing)  
## concavity\_mean < 0.1690253 to the left, improve=6.753654, (0 missing)  
## concavity\_worst < 0.1741613 to the left, improve=5.588892, (0 missing)  
## concavity\_SE < 0.04690657 to the left, improve=5.276530, (0 missing)  
## concave\_points\_mean < 0.2491054 to the left, improve=5.002462, (0 missing)  
## Surrogate splits:  
## texure\_worst < 0.171242 to the left, agree=0.976, adj=0.556, (0 split)  
## texure\_SE < 0.02085175 to the left, agree=0.958, adj=0.222, (0 split)  
## concavity\_mean < 0.10888 to the left, agree=0.952, adj=0.111, (0 split)  
## symmetry\_mean < 0.1734848 to the left, agree=0.952, adj=0.111, (0 split)  
##   
## Node number 4: 308 observations  
## predicted class=B expected loss=0.02272727 P(node) =0.616  
## class counts: 301 7  
## probabilities: 0.977 0.023   
##   
## Node number 5: 27 observations, complexity param=0.02197802  
## predicted class=M expected loss=0.3333333 P(node) =0.054  
## class counts: 9 18  
## probabilities: 0.333 0.667   
## left son=10 (14 obs) right son=11 (13 obs)  
## Primary splits:  
## texure\_mean < 0.3605005 to the left, improve=5.571429, (0 missing)  
## texure\_worst < 0.4145789 to the left, improve=5.571429, (0 missing)  
## smoothness\_worst < 0.4740144 to the left, improve=5.333333, (0 missing)  
## area\_worst < 0.1290307 to the left, improve=4.270588, (0 missing)  
## perimeter\_worst < 0.2344738 to the left, improve=3.947368, (0 missing)  
## Surrogate splits:  
## texure\_worst < 0.4473614 to the left, agree=0.889, adj=0.769, (0 split)  
## smoothness\_worst < 0.5156178 to the left, agree=0.815, adj=0.615, (0 split)  
## perimeter\_worst < 0.2716271 to the left, agree=0.741, adj=0.462, (0 split)  
## smoothness\_mean < 0.5165659 to the left, agree=0.704, adj=0.385, (0 split)  
## area\_SE < 0.04210513 to the left, agree=0.704, adj=0.385, (0 split)  
##   
## Node number 6: 9 observations  
## predicted class=B expected loss=0.3333333 P(node) =0.018  
## class counts: 6 3  
## probabilities: 0.667 0.333   
##   
## Node number 7: 156 observations  
## predicted class=M expected loss=0.01282051 P(node) =0.312  
## class counts: 2 154  
## probabilities: 0.013 0.987   
##   
## Node number 10: 14 observations  
## predicted class=B expected loss=0.3571429 P(node) =0.028  
## class counts: 9 5  
## probabilities: 0.643 0.357   
##   
## Node number 11: 13 observations  
## predicted class=M expected loss=0 P(node) =0.026  
## class counts: 0 13  
## probabilities: 0.000 1.000

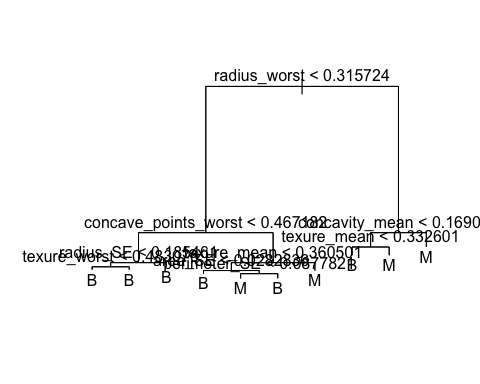
#plot tree  
plot(cancer.rpart, uniform = T, main = "Classification Tree for Cancer type")  
text(cancer.rpart, use.n = T, all= T)



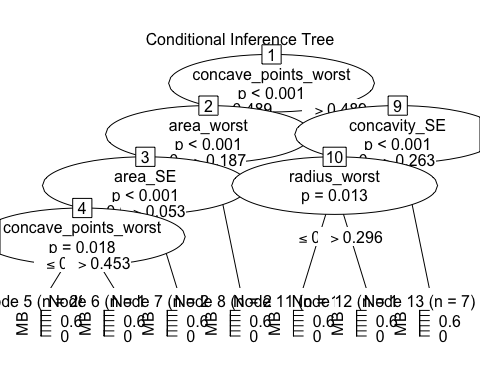
#use tree  
cancer.tree <- tree(cancer\_data[1:500,2]~., data=cancer.normalized.train)  
summary(cancer.tree)

##   
## Classification tree:  
## tree(formula = cancer\_data[1:500, 2] ~ ., data = cancer.normalized.train)  
## Variables actually used in tree construction:  
## [1] "radius\_worst" "concave\_points\_worst" "radius\_SE"   
## [4] "texure\_worst" "texure\_mean" "area\_SE"   
## [7] "perimeter\_SE" "concavity\_mean"   
## Number of terminal nodes: 10   
## Residual mean deviance: 0.05824 = 28.54 / 490   
## Misclassification error rate: 0.01 = 5 / 500

plot(cancer.tree)  
text(cancer.tree)



#use ctree  
cancer.ctree <- ctree(cancer\_data[1:500,2]~., data=cancer.normalized.train)  
plot(cancer.ctree, main = "Conditional Inference Tree")



# Answer the following questions:

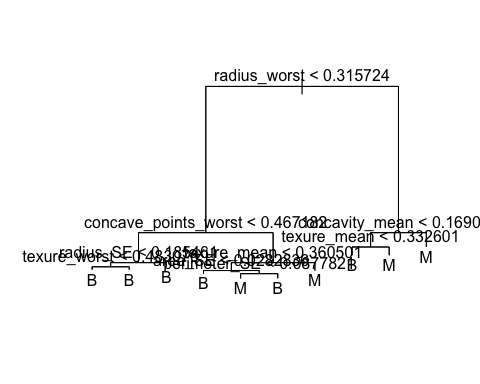
I choose "tree" as my method. And the original size of my data is 569.

## Does the size of the data set make a difference?

#use the orginal size of data  
cancer.tree.whole <- tree(cancer\_data[1:500,2]~., data=cancer.normalized.train)  
summary(cancer.tree.whole)

##   
## Classification tree:  
## tree(formula = cancer\_data[1:500, 2] ~ ., data = cancer.normalized.train)  
## Variables actually used in tree construction:  
## [1] "radius\_worst" "concave\_points\_worst" "radius\_SE"   
## [4] "texure\_worst" "texure\_mean" "area\_SE"   
## [7] "perimeter\_SE" "concavity\_mean"   
## Number of terminal nodes: 10   
## Residual mean deviance: 0.05824 = 28.54 / 490   
## Misclassification error rate: 0.01 = 5 / 500

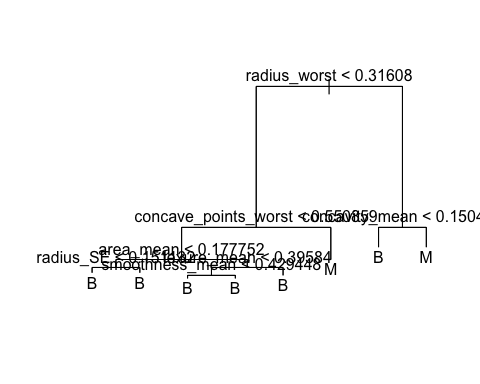
plot(cancer.tree.whole)  
text(cancer.tree.whole)



#use half of the data  
cancer.half.train <- cancer.normalized[1:250,]  
cancer.half.test <- cancer.normalized[501:569,]  
cancer.half.train.target<- cancer\_data[1:250,2]  
cancer.half.test.target <- cancer\_data[501:569,2]  
  
cancer.tree.half<- tree(cancer.half.train.target~., data= cancer.half.train)  
summary(cancer.tree.half)

##   
## Classification tree:  
## tree(formula = cancer.half.train.target ~ ., data = cancer.half.train)  
## Variables actually used in tree construction:  
## [1] "radius\_worst" "concave\_points\_worst" "area\_mean"   
## [4] "radius\_SE" "texure\_mean" "smoothness\_mean"   
## [7] "concavity\_mean"   
## Number of terminal nodes: 8   
## Residual mean deviance: 0.1309 = 31.68 / 242   
## Misclassification error rate: 0.036 = 9 / 250

plot(cancer.tree.half)  
text(cancer.tree.half)



#evaluate the performances of the models of the two datasets  
  
#whole data set  
cancer.whole.pred <- predict(cancer.tree.whole, cancer.normalized.test, type="class")  
CrossTable(cancer.normalized.test.target, cancer.whole.pred, prop.chisq = F, prop.c = F, prop.r = F, dnn = c('actual type', 'predicted type'))

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 69   
##   
##   
## | predicted type   
## actual type | B | M | Row Total |   
## -------------|-----------|-----------|-----------|  
## B | 38 | 1 | 39 |   
## | 0.551 | 0.014 | |   
## -------------|-----------|-----------|-----------|  
## M | 4 | 26 | 30 |   
## | 0.058 | 0.377 | |   
## -------------|-----------|-----------|-----------|  
## Column Total | 42 | 27 | 69 |   
## -------------|-----------|-----------|-----------|  
##   
##

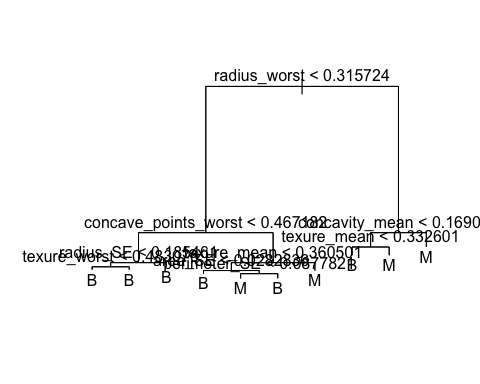
#half-size data set   
cancer.half.pred <- predict(cancer.tree.half, cancer.half.test, type= "class")  
CrossTable(cancer.half.test.target, cancer.half.pred, prop.chisq = F, prop.c = F, prop.r = F, dnn = c('actual type', 'predicted type'))

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 69   
##   
##   
## | predicted type   
## actual type | B | M | Row Total |   
## -------------|-----------|-----------|-----------|  
## B | 36 | 3 | 39 |   
## | 0.522 | 0.043 | |   
## -------------|-----------|-----------|-----------|  
## M | 3 | 27 | 30 |   
## | 0.043 | 0.391 | |   
## -------------|-----------|-----------|-----------|  
## Column Total | 39 | 30 | 69 |   
## -------------|-----------|-----------|-----------|  
##   
##

I used different size of data to generate the models and used the same data to test the performances of these models. When I used whole data set, the misclassification is 5, while I used half size of data, the misclassification is 6. That means the size of data set makes a difference.

## Do the rules make sense? If so why did the algorithm generate good rules? If not, why not?

#use the whole data set and "tree" as the method  
plot(cancer.tree.whole)  
text(cancer.tree.whole)



According to my decision tree, there are 10 rules. For example, if the largest radius (radius\_worst) is less than 0.3157, largest concave point is less than 0.467, the standard error of radius is less than 0.185, and the largest texure is less than 0.48, then the cancer is begign; If the largest radius is greater than 0.31, and the mean of concavity is greater than 0.169, the cancer is malignant.

The rules make sense. For example, if the tumor is small, it is highly possible that this tumor is benign and if the tumor is large, it is highly possible that this tumor is malignant. A decision tree is split the data set into subsets based on an attribute value test. and in general, doctor make a dignostic of the cancer is also depend on the attributes of the tumor or tissue for example the radius, the standard deviation of gray-scale values (texture). So the algorithm generate good rules.

## Does scaling, normalization or leaving the data unscaled make a difference?

Here, I use the whole data set and "tree" as the method

#scaling data  
cancer.scaled <- as.data.frame(lapply(cancer\_data[,c(3:32)], scale))  
head(cancer.scaled)

## radius\_mean texure\_mean perimeter\_mean area\_mean smoothness\_mean  
## 1 -0.06449733 -0.6206990 -0.12342219 -0.1576780 -1.9965888  
## 2 -0.09287378 -0.8136758 -0.06333736 -0.2011545 0.3085665  
## 3 -0.80512271 -1.4530566 -0.81233996 -0.7583927 0.1407637  
## 4 -1.45494345 -1.1368537 -1.46545377 -1.1610477 -1.8707367  
## 5 -0.47311823 0.1395830 -0.47487725 -0.5216872 -0.8425891  
## 6 -0.76539568 -0.4602725 -0.75348976 -0.7299767 0.9129410  
## compactness\_mean concavity\_mean concave\_points\_mean symmetry\_mean  
## 1 -0.9685346 -0.8349156 -0.91563250 0.005038873  
## 2 0.4479792 -0.1368457 0.04563722 -0.545768927  
## 3 -0.5351166 -0.7043333 -0.55071080 -0.155461413  
## 4 -1.3856687 -1.0677185 -1.17123232 -0.086154471  
## 5 -0.0556870 -0.2571421 -0.46205751 -0.111688608  
## 6 -0.1795207 -0.8591254 -0.78110628 -0.607780402  
## fractal\_dimension\_mean radius\_SE texure\_SE perimeter\_SE area\_SE  
## 1 -1.05343168 -0.9003267 -1.1613438 -0.7290629 -0.5782919  
## 2 0.40541687 -0.4560629 -1.0810390 -0.4496165 -0.3734162  
## 3 -0.01382505 -0.7744760 -1.2463617 -0.8418307 -0.5971967  
## 4 0.78358440 -0.8314514 0.1597876 -0.8517226 -0.6715191  
## 5 -0.04356856 -0.1477468 -0.3205908 -0.1419782 -0.2641638  
## 6 0.39975143 -0.4358691 1.2528752 -0.4595084 -0.4167215  
## smoothness\_SE compactness\_SE concavity\_SE concave\_points\_SE symmetry\_SE  
## 1 -1.2432828 -0.6884083 -0.7284726 -1.1286572 -0.8930519  
## 2 -0.7740100 -0.2696052 -0.4738517 -0.4197759 -1.0091851  
## 3 -0.1895006 -0.7961803 -0.6318717 -0.4776339 -0.4381969  
## 4 0.1159097 -1.1500409 -0.9346273 -1.3490685 0.7824111  
## 5 -0.1701835 0.4585537 0.5044144 0.3620356 0.8295903  
## 6 -0.1255543 -0.4538786 -0.6729502 -0.3688869 0.2755382  
## fractal\_dimension\_SE radius\_worst texure\_worst perimeter\_worst  
## 1 -0.6424257 -0.23362991 -0.6308267 -0.18037959  
## 2 -0.2512797 0.06223777 -0.7837653 0.09043326  
## 3 -0.3695683 -0.81708926 -1.5452042 -0.86276843  
## 4 -0.6205064 -1.35937542 -0.9122988 -1.37969360  
## 5 0.3235348 -0.58122272 -0.4241969 -0.56933825  
## 6 -0.1016238 -0.72191504 0.1761685 -0.73212355  
## area\_worst smoothness\_worst compactness\_worst concavity\_worst  
## 1 -0.2839750 -1.6874505 -0.3410845 -0.64128913  
## 2 -0.1197546 0.3824125 0.6351671 0.02737705  
## 3 -0.7430894 0.1502865 -0.6582392 -0.69353616  
## 4 -1.0451845 -1.4785371 -1.2835233 -1.23412521  
## 5 -0.5783421 -1.1986720 -0.2444763 -0.39203722  
## 6 -0.6631747 0.3911720 -0.4770988 -0.93521465  
## concave\_points\_worst symmetry\_worst fractal\_dimension\_worst  
## 1 -0.7969627 -0.3584367 -0.3856771  
## 2 0.3604584 -0.5039090 1.0549748  
## 3 -0.4123727 -0.2711534 -0.2062877  
## 4 -1.5322172 0.1458671 -0.3269880  
## 5 -0.5835213 -0.3487386 -0.3491348  
## 6 -0.7695789 -0.5119908 -0.1653160

summary(cancer.scaled)

## radius\_mean texure\_mean perimeter\_mean area\_mean   
## Min. :-2.0279 Min. :-2.2273 Min. :-1.9828 Min. :-1.4532   
## 1st Qu.:-0.6888 1st Qu.:-0.7253 1st Qu.:-0.6913 1st Qu.:-0.6666   
## Median :-0.2149 Median :-0.1045 Median :-0.2358 Median :-0.2949   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.4690 3rd Qu.: 0.5837 3rd Qu.: 0.4992 3rd Qu.: 0.3632   
## Max. : 3.9678 Max. : 4.6478 Max. : 3.9726 Max. : 5.2459   
## smoothness\_mean compactness\_mean concavity\_mean   
## Min. :-3.10935 Min. :-1.6087 Min. :-1.1139   
## 1st Qu.:-0.71034 1st Qu.:-0.7464 1st Qu.:-0.7431   
## Median :-0.03486 Median :-0.2217 Median :-0.3419   
## Mean : 0.00000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.63564 3rd Qu.: 0.4934 3rd Qu.: 0.5256   
## Max. : 4.76672 Max. : 4.5644 Max. : 4.2399   
## concave\_points\_mean symmetry\_mean fractal\_dimension\_mean  
## Min. :-1.2607 Min. :-2.74171 Min. :-1.8183   
## 1st Qu.:-0.7373 1st Qu.:-0.70262 1st Qu.:-0.7220   
## Median :-0.3974 Median :-0.07156 Median :-0.1781   
## Mean : 0.0000 Mean : 0.00000 Mean : 0.0000   
## 3rd Qu.: 0.6464 3rd Qu.: 0.53031 3rd Qu.: 0.4706   
## Max. : 3.9245 Max. : 4.48081 Max. : 4.9066   
## radius\_SE texure\_SE perimeter\_SE area\_SE   
## Min. :-1.0590 Min. :-1.5529 Min. :-1.0431 Min. :-0.7372   
## 1st Qu.:-0.6230 1st Qu.:-0.6942 1st Qu.:-0.6232 1st Qu.:-0.4943   
## Median :-0.2920 Median :-0.1973 Median :-0.2864 Median :-0.3475   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.2659 3rd Qu.: 0.4661 3rd Qu.: 0.2428 3rd Qu.: 0.1067   
## Max. : 8.8991 Max. : 6.6494 Max. : 9.4537 Max. :11.0321   
## smoothness\_SE compactness\_SE concavity\_SE concave\_points\_SE  
## Min. :-1.7745 Min. :-1.2970 Min. :-1.0566 Min. :-1.9118   
## 1st Qu.:-0.6235 1st Qu.:-0.6923 1st Qu.:-0.5567 1st Qu.:-0.6739   
## Median :-0.2201 Median :-0.2808 Median :-0.1989 Median :-0.1404   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.3680 3rd Qu.: 0.3893 3rd Qu.: 0.3365 3rd Qu.: 0.4722   
## Max. : 8.0229 Max. : 6.1381 Max. :12.0621 Max. : 6.6438   
## symmetry\_SE fractal\_dimension\_SE radius\_worst   
## Min. :-1.5315 Min. :-1.0960 Min. :-1.7254   
## 1st Qu.:-0.6511 1st Qu.:-0.5846 1st Qu.:-0.6743   
## Median :-0.2192 Median :-0.2297 Median :-0.2688   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.3554 3rd Qu.: 0.2884 3rd Qu.: 0.5216   
## Max. : 7.0657 Max. : 9.8429 Max. : 4.0906   
## texure\_worst perimeter\_worst area\_worst smoothness\_worst   
## Min. :-2.22204 Min. :-1.6919 Min. :-1.2213 Min. :-2.6803   
## 1st Qu.:-0.74797 1st Qu.:-0.6890 1st Qu.:-0.6416 1st Qu.:-0.6906   
## Median :-0.04348 Median :-0.2857 Median :-0.3409 Median :-0.0468   
## Mean : 0.00000 Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.65776 3rd Qu.: 0.5398 3rd Qu.: 0.3573 3rd Qu.: 0.5970   
## Max. : 3.88249 Max. : 4.2836 Max. : 5.9250 Max. : 3.9519   
## compactness\_worst concavity\_worst concave\_points\_worst  
## Min. :-1.4426 Min. :-1.3047 Min. :-1.7435   
## 1st Qu.:-0.6805 1st Qu.:-0.7558 1st Qu.:-0.7557   
## Median :-0.2693 Median :-0.2180 Median :-0.2233   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.5392 3rd Qu.: 0.5307 3rd Qu.: 0.7119   
## Max. : 5.1084 Max. : 4.6965 Max. : 2.6835   
## symmetry\_worst fractal\_dimension\_worst  
## Min. :-2.1591 Min. :-1.6004   
## 1st Qu.:-0.6413 1st Qu.:-0.6913   
## Median :-0.1273 Median :-0.2163   
## Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.4497 3rd Qu.: 0.4504   
## Max. : 6.0407 Max. : 6.8408

#normalization data  
normalize <- function(x) {  
 return((x-min(x))/(max(x)-min(x)))  
}  
cancer.normalized <- as.data.frame(lapply(cancer\_data[,c(3:32)],normalize))  
head(cancer.normalized)

## radius\_mean texure\_mean perimeter\_mean area\_mean smoothness\_mean  
## 1 0.32746462 0.2336828 0.31221063 0.19338282 0.1412837  
## 2 0.32273179 0.2056138 0.32229977 0.18689290 0.4339623  
## 3 0.20393772 0.1126141 0.19653099 0.10371156 0.4126569  
## 4 0.09555587 0.1586067 0.08686338 0.04360551 0.1572628  
## 5 0.25931185 0.3442678 0.25319605 0.13904560 0.2878036  
## 6 0.21056368 0.2570172 0.20641283 0.10795334 0.5106978  
## compactness\_mean concavity\_mean concave\_points\_mean symmetry\_mean  
## 1 0.10370529 0.052108716 0.06655070 0.3803030  
## 2 0.33316974 0.182497657 0.25193837 0.3040404  
## 3 0.17391571 0.076499531 0.13692843 0.3580808  
## 4 0.03613275 0.008624649 0.01725646 0.3676768  
## 5 0.25157966 0.160028116 0.15402584 0.3641414  
## 6 0.23151954 0.047586692 0.09249503 0.2954545  
## fractal\_dimension\_mean radius\_SE texure\_SE perimeter\_SE area\_SE  
## 1 0.1137321 0.01593337 0.04773692 0.02992037 0.013500237  
## 2 0.3306655 0.06054680 0.05752740 0.05654243 0.030907848  
## 3 0.2683235 0.02857143 0.03737182 0.01917731 0.011893956  
## 4 0.3868997 0.02284990 0.20880481 0.01823493 0.005579027  
## 5 0.2639006 0.09150824 0.15023868 0.08585026 0.040190662  
## 6 0.3298231 0.06257469 0.34207037 0.05560006 0.027228342  
## smoothness\_SE compactness\_SE concavity\_SE concave\_points\_SE symmetry\_SE  
## 1 0.05422035 0.08184877 0.025010101 0.09153249 0.07426690  
## 2 0.10211782 0.13817707 0.044419192 0.17438909 0.06075871  
## 3 0.16177720 0.06735362 0.032373737 0.16762644 0.12717397  
## 4 0.19294965 0.01975997 0.009295455 0.06577003 0.26915067  
## 5 0.16374885 0.23611320 0.118989899 0.26577003 0.27463837  
## 6 0.16830404 0.11339262 0.029242424 0.18033719 0.21019305  
## fractal\_dimension\_SE radius\_worst texure\_worst perimeter\_worst  
## 1 0.04146456 0.25649235 0.2606610 0.25295084  
## 2 0.07722178 0.30736393 0.2356077 0.29827183  
## 3 0.06640825 0.15617218 0.1108742 0.13875193  
## 4 0.04346835 0.06293134 0.2145522 0.05224364  
## 5 0.12976936 0.19672714 0.2945096 0.18785796  
## 6 0.09090281 0.17253646 0.3928571 0.16061557  
## area\_worst smoothness\_worst compactness\_worst concavity\_worst  
## 1 0.13116890 0.1497061 0.16814623 0.11054313  
## 2 0.15414864 0.4617975 0.31716972 0.22196486  
## 3 0.06692391 0.4267979 0.11973300 0.10183706  
## 4 0.02465100 0.1812058 0.02428423 0.01175719  
## 5 0.08997739 0.2234036 0.18289335 0.15207668  
## 6 0.07810657 0.4631183 0.14738384 0.06156550  
## concave\_points\_worst symmetry\_worst fractal\_dimension\_worst  
## 1 0.21381443 0.2195939 0.1439066  
## 2 0.47525773 0.2018529 0.3145743  
## 3 0.30068729 0.2302385 0.1651581  
## 4 0.04773196 0.2810960 0.1508592  
## 5 0.26202749 0.2207767 0.1482356  
## 6 0.22000000 0.2008673 0.1700118

summary(cancer.normalized)

## radius\_mean texure\_mean perimeter\_mean area\_mean   
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.2233 1st Qu.:0.2185 1st Qu.:0.2168 1st Qu.:0.1174   
## Median :0.3024 Median :0.3088 Median :0.2933 Median :0.1729   
## Mean :0.3382 Mean :0.3240 Mean :0.3329 Mean :0.2169   
## 3rd Qu.:0.4164 3rd Qu.:0.4089 3rd Qu.:0.4168 3rd Qu.:0.2711   
## Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :1.0000   
## smoothness\_mean compactness\_mean concavity\_mean concave\_points\_mean  
## Min. :0.0000 Min. :0.0000 Min. :0.00000 Min. :0.0000   
## 1st Qu.:0.3046 1st Qu.:0.1397 1st Qu.:0.06926 1st Qu.:0.1009   
## Median :0.3904 Median :0.2247 Median :0.14419 Median :0.1665   
## Mean :0.3948 Mean :0.2606 Mean :0.20806 Mean :0.2431   
## 3rd Qu.:0.4755 3rd Qu.:0.3405 3rd Qu.:0.30623 3rd Qu.:0.3678   
## Max. :1.0000 Max. :1.0000 Max. :1.00000 Max. :1.0000   
## symmetry\_mean fractal\_dimension\_mean radius\_SE   
## Min. :0.0000 Min. :0.0000 Min. :0.00000   
## 1st Qu.:0.2823 1st Qu.:0.1630 1st Qu.:0.04378   
## Median :0.3697 Median :0.2439 Median :0.07702   
## Mean :0.3796 Mean :0.2704 Mean :0.10635   
## 3rd Qu.:0.4530 3rd Qu.:0.3404 3rd Qu.:0.13304   
## Max. :1.0000 Max. :1.0000 Max. :1.00000   
## texure\_SE perimeter\_SE area\_SE smoothness\_SE   
## Min. :0.0000 Min. :0.00000 Min. :0.00000 Min. :0.0000   
## 1st Qu.:0.1047 1st Qu.:0.04000 1st Qu.:0.02064 1st Qu.:0.1175   
## Median :0.1653 Median :0.07209 Median :0.03311 Median :0.1586   
## Mean :0.1893 Mean :0.09938 Mean :0.06264 Mean :0.1811   
## 3rd Qu.:0.2462 3rd Qu.:0.12251 3rd Qu.:0.07170 3rd Qu.:0.2187   
## Max. :1.0000 Max. :1.00000 Max. :1.00000 Max. :1.0000   
## compactness\_SE concavity\_SE concave\_points\_SE symmetry\_SE   
## Min. :0.00000 Min. :0.00000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.08132 1st Qu.:0.03811 1st Qu.:0.1447 1st Qu.:0.1024   
## Median :0.13667 Median :0.06538 Median :0.2070 Median :0.1526   
## Mean :0.17444 Mean :0.08054 Mean :0.2235 Mean :0.1781   
## 3rd Qu.:0.22680 3rd Qu.:0.10619 3rd Qu.:0.2787 3rd Qu.:0.2195   
## Max. :1.00000 Max. :1.00000 Max. :1.0000 Max. :1.0000   
## fractal\_dimension\_SE radius\_worst texure\_worst perimeter\_worst   
## Min. :0.00000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.04675 1st Qu.:0.1807 1st Qu.:0.2415 1st Qu.:0.1678   
## Median :0.07919 Median :0.2504 Median :0.3569 Median :0.2353   
## Mean :0.10019 Mean :0.2967 Mean :0.3640 Mean :0.2831   
## 3rd Qu.:0.12656 3rd Qu.:0.3863 3rd Qu.:0.4717 3rd Qu.:0.3735   
## Max. :1.00000 Max. :1.0000 Max. :1.0000 Max. :1.0000   
## area\_worst smoothness\_worst compactness\_worst concavity\_worst   
## Min. :0.00000 Min. :0.0000 Min. :0.0000 Min. :0.00000   
## 1st Qu.:0.08113 1st Qu.:0.3000 1st Qu.:0.1163 1st Qu.:0.09145   
## Median :0.12321 Median :0.3971 Median :0.1791 Median :0.18107   
## Mean :0.17091 Mean :0.4041 Mean :0.2202 Mean :0.21740   
## 3rd Qu.:0.22090 3rd Qu.:0.4942 3rd Qu.:0.3025 3rd Qu.:0.30583   
## Max. :1.00000 Max. :1.0000 Max. :1.0000 Max. :1.00000   
## concave\_points\_worst symmetry\_worst fractal\_dimension\_worst  
## Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.2231 1st Qu.:0.1851 1st Qu.:0.1077   
## Median :0.3434 Median :0.2478 Median :0.1640   
## Mean :0.3938 Mean :0.2633 Mean :0.1896   
## 3rd Qu.:0.5546 3rd Qu.:0.3182 3rd Qu.:0.2429   
## Max. :1.0000 Max. :1.0000 Max. :1.0000

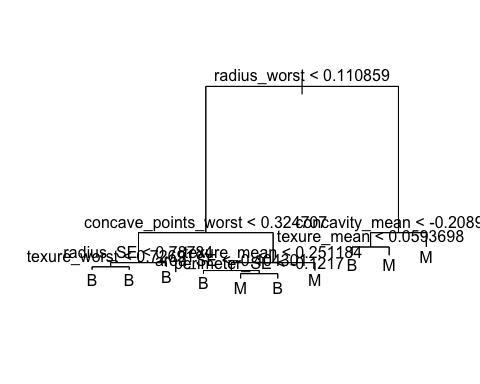
#unscaling data  
cancer.unscaled <- as.data.frame(cancer\_data[,c(3:32)])  
head(cancer.unscaled)

## radius\_mean texure\_mean perimeter\_mean area\_mean smoothness\_mean  
## 478 13.90 16.62 88.97 599.4 0.06828  
## 74 13.80 15.79 90.43 584.1 0.10070  
## 272 11.29 13.04 72.23 388.0 0.09834  
## 308 9.00 14.40 56.36 246.3 0.07005  
## 302 12.46 19.89 80.43 471.3 0.08451  
## 143 11.43 17.31 73.66 398.0 0.10920  
## compactness\_mean concavity\_mean concave\_points\_mean symmetry\_mean  
## 478 0.05319 0.022240 0.013390 0.1813  
## 74 0.12800 0.077890 0.050690 0.1662  
## 272 0.07608 0.032650 0.027550 0.1769  
## 308 0.03116 0.003681 0.003472 0.1788  
## 302 0.10140 0.068300 0.030990 0.1781  
## 143 0.09486 0.020310 0.018610 0.1645  
## fractal\_dimension\_mean radius\_SE texure\_SE perimeter\_SE area\_SE  
## 478 0.05536 0.1555 0.5762 1.392 14.030  
## 74 0.06566 0.2787 0.6205 1.957 23.350  
## 272 0.06270 0.1904 0.5293 1.164 13.170  
## 308 0.06833 0.1746 1.3050 1.144 9.789  
## 302 0.06249 0.3642 1.0400 2.579 28.320  
## 143 0.06562 0.2843 1.9080 1.937 21.380  
## smoothness\_SE compactness\_SE concavity\_SE concave\_points\_SE  
## 478 0.003308 0.013150 0.009904 0.004832  
## 74 0.004717 0.020650 0.017590 0.009206  
## 272 0.006472 0.011220 0.012820 0.008849  
## 308 0.007389 0.004883 0.003681 0.003472  
## 302 0.006530 0.033690 0.047120 0.014030  
## 143 0.006664 0.017350 0.011580 0.009520  
## symmetry\_SE fractal\_dimension\_SE radius\_worst texure\_worst  
## 478 0.01316 0.002095 15.140 21.80  
## 74 0.01220 0.003130 16.570 20.86  
## 272 0.01692 0.002817 12.320 16.18  
## 308 0.02701 0.002153 9.699 20.07  
## 302 0.02740 0.004651 13.460 23.07  
## 143 0.02282 0.003526 12.780 26.76  
## perimeter\_worst area\_worst smoothness\_worst compactness\_worst  
## 478 101.20 718.9 0.09384 0.20060  
## 74 110.30 812.4 0.14110 0.35420  
## 272 78.27 457.5 0.13580 0.15070  
## 308 60.90 285.5 0.09861 0.05232  
## 302 88.13 551.3 0.10500 0.21580  
## 143 82.66 503.0 0.14130 0.17920  
## concavity\_worst concave\_points\_worst symmetry\_worst  
## 478 0.13840 0.06222 0.2679  
## 74 0.27790 0.13830 0.2589  
## 272 0.12750 0.08750 0.2733  
## 308 0.01472 0.01389 0.2991  
## 302 0.19040 0.07625 0.2685  
## 143 0.07708 0.06402 0.2584  
## fractal\_dimension\_worst  
## 478 0.07698  
## 74 0.10300  
## 272 0.08022  
## 308 0.07804  
## 302 0.07764  
## 143 0.08096

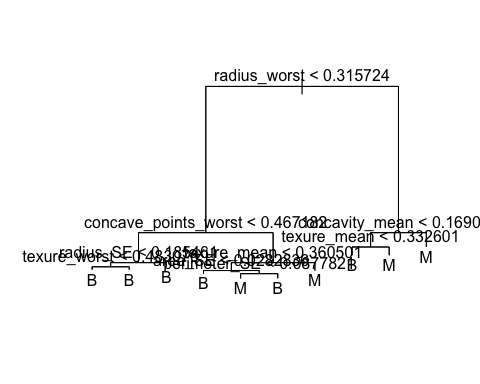
summary(cancer.unscaled)

## radius\_mean texure\_mean perimeter\_mean area\_mean   
## Min. : 6.981 Min. : 9.71 Min. : 43.79 Min. : 143.5   
## 1st Qu.:11.700 1st Qu.:16.17 1st Qu.: 75.17 1st Qu.: 420.3   
## Median :13.370 Median :18.84 Median : 86.24 Median : 551.1   
## Mean :14.127 Mean :19.29 Mean : 91.97 Mean : 654.9   
## 3rd Qu.:15.780 3rd Qu.:21.80 3rd Qu.:104.10 3rd Qu.: 782.7   
## Max. :28.110 Max. :39.28 Max. :188.50 Max. :2501.0   
## smoothness\_mean compactness\_mean concavity\_mean concave\_points\_mean  
## Min. :0.05263 Min. :0.01938 Min. :0.00000 Min. :0.00000   
## 1st Qu.:0.08637 1st Qu.:0.06492 1st Qu.:0.02956 1st Qu.:0.02031   
## Median :0.09587 Median :0.09263 Median :0.06154 Median :0.03350   
## Mean :0.09636 Mean :0.10434 Mean :0.08880 Mean :0.04892   
## 3rd Qu.:0.10530 3rd Qu.:0.13040 3rd Qu.:0.13070 3rd Qu.:0.07400   
## Max. :0.16340 Max. :0.34540 Max. :0.42680 Max. :0.20120   
## symmetry\_mean fractal\_dimension\_mean radius\_SE texure\_SE   
## Min. :0.1060 Min. :0.04996 Min. :0.1115 Min. :0.3602   
## 1st Qu.:0.1619 1st Qu.:0.05770 1st Qu.:0.2324 1st Qu.:0.8339   
## Median :0.1792 Median :0.06154 Median :0.3242 Median :1.1080   
## Mean :0.1812 Mean :0.06280 Mean :0.4052 Mean :1.2169   
## 3rd Qu.:0.1957 3rd Qu.:0.06612 3rd Qu.:0.4789 3rd Qu.:1.4740   
## Max. :0.3040 Max. :0.09744 Max. :2.8730 Max. :4.8850   
## perimeter\_SE area\_SE smoothness\_SE compactness\_SE   
## Min. : 0.757 Min. : 6.802 Min. :0.001713 Min. :0.002252   
## 1st Qu.: 1.606 1st Qu.: 17.850 1st Qu.:0.005169 1st Qu.:0.013080   
## Median : 2.287 Median : 24.530 Median :0.006380 Median :0.020450   
## Mean : 2.866 Mean : 40.337 Mean :0.007041 Mean :0.025478   
## 3rd Qu.: 3.357 3rd Qu.: 45.190 3rd Qu.:0.008146 3rd Qu.:0.032450   
## Max. :21.980 Max. :542.200 Max. :0.031130 Max. :0.135400   
## concavity\_SE concave\_points\_SE symmetry\_SE   
## Min. :0.00000 Min. :0.000000 Min. :0.007882   
## 1st Qu.:0.01509 1st Qu.:0.007638 1st Qu.:0.015160   
## Median :0.02589 Median :0.010930 Median :0.018730   
## Mean :0.03189 Mean :0.011796 Mean :0.020542   
## 3rd Qu.:0.04205 3rd Qu.:0.014710 3rd Qu.:0.023480   
## Max. :0.39600 Max. :0.052790 Max. :0.078950   
## fractal\_dimension\_SE radius\_worst texure\_worst perimeter\_worst   
## Min. :0.0008948 Min. : 7.93 Min. :12.02 Min. : 50.41   
## 1st Qu.:0.0022480 1st Qu.:13.01 1st Qu.:21.08 1st Qu.: 84.11   
## Median :0.0031870 Median :14.97 Median :25.41 Median : 97.66   
## Mean :0.0037949 Mean :16.27 Mean :25.68 Mean :107.26   
## 3rd Qu.:0.0045580 3rd Qu.:18.79 3rd Qu.:29.72 3rd Qu.:125.40   
## Max. :0.0298400 Max. :36.04 Max. :49.54 Max. :251.20   
## area\_worst smoothness\_worst compactness\_worst concavity\_worst   
## Min. : 185.2 Min. :0.07117 Min. :0.02729 Min. :0.0000   
## 1st Qu.: 515.3 1st Qu.:0.11660 1st Qu.:0.14720 1st Qu.:0.1145   
## Median : 686.5 Median :0.13130 Median :0.21190 Median :0.2267   
## Mean : 880.6 Mean :0.13237 Mean :0.25427 Mean :0.2722   
## 3rd Qu.:1084.0 3rd Qu.:0.14600 3rd Qu.:0.33910 3rd Qu.:0.3829   
## Max. :4254.0 Max. :0.22260 Max. :1.05800 Max. :1.2520   
## concave\_points\_worst symmetry\_worst fractal\_dimension\_worst  
## Min. :0.00000 Min. :0.1565 Min. :0.05504   
## 1st Qu.:0.06493 1st Qu.:0.2504 1st Qu.:0.07146   
## Median :0.09993 Median :0.2822 Median :0.08004   
## Mean :0.11461 Mean :0.2901 Mean :0.08395   
## 3rd Qu.:0.16140 3rd Qu.:0.3179 3rd Qu.:0.09208   
## Max. :0.29100 Max. :0.6638 Max. :0.20750

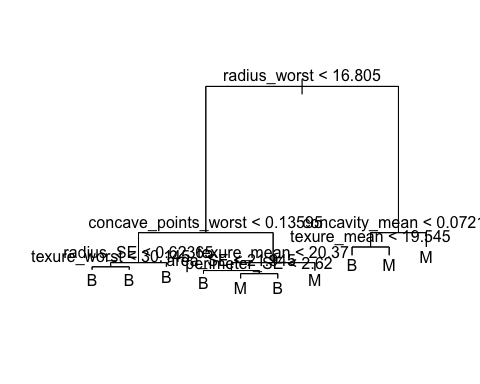
#make tree using scaled data  
cancer.scaled.train <- cancer.scaled[1:500,]  
cancer.scaled.test <- cancer.scaled[501:569,]  
cancer.scaled.train.target<- cancer\_data[1:500,2]  
cancer.scaled.test.target <- cancer\_data[501:569,2]  
cancer.tree.scaled <- tree(cancer\_data[1:500,2]~., data=cancer.scaled.train)  
plot(cancer.tree.scaled)  
text(cancer.tree.scaled)



#make tree using normalized data  
cancer.normalized.train <- cancer.normalized[1:500,]  
cancer.normalized.test <- cancer.normalized[501:569,]  
cancer.normalized.train.target<- cancer\_data[1:500,2]  
cancer.normalized.test.target <- cancer\_data[501:569,2]  
cancer.tree.normalized <- tree(cancer\_data[1:500,2]~., data=cancer.normalized.train)  
plot(cancer.tree.normalized)  
text(cancer.tree.normalized)



#make tree using unscaled data  
cancer.unscaled.train <- cancer.unscaled[1:500,]  
cancer.unscaled.test <- cancer.unscaled[501:569,]  
cancer.unscaled.train.target<- cancer\_data[1:500,2]  
cancer.unscaled.test.target <- cancer\_data[501:569,2]  
cancer.tree.unscaled <- tree(cancer\_data[1:500,2]~., data=cancer.unscaled.train)  
plot(cancer.tree.unscaled)  
text(cancer.tree.unscaled)



#evaluate the performence of these three models  
  
#model using scaled data   
scaled.pred <- predict(cancer.tree.scaled, cancer.scaled.test, type= "class")  
CrossTable(cancer.scaled.test.target, scaled.pred, prop.chisq = F, prop.c = F, prop.r = F, dnn = c('actual type', 'predicted type'))

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 69   
##   
##   
## | predicted type   
## actual type | B | M | Row Total |   
## -------------|-----------|-----------|-----------|  
## B | 38 | 1 | 39 |   
## | 0.551 | 0.014 | |   
## -------------|-----------|-----------|-----------|  
## M | 4 | 26 | 30 |   
## | 0.058 | 0.377 | |   
## -------------|-----------|-----------|-----------|  
## Column Total | 42 | 27 | 69 |   
## -------------|-----------|-----------|-----------|  
##   
##

#model using normalized data  
normalized.pred <- predict(cancer.tree.normalized, cancer.normalized.test, type="class")  
CrossTable(cancer.normalized.test.target, normalized.pred, prop.chisq = F, prop.c = F, prop.r = F, dnn = c('actual type', 'predicted type'))

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 69   
##   
##   
## | predicted type   
## actual type | B | M | Row Total |   
## -------------|-----------|-----------|-----------|  
## B | 38 | 1 | 39 |   
## | 0.551 | 0.014 | |   
## -------------|-----------|-----------|-----------|  
## M | 4 | 26 | 30 |   
## | 0.058 | 0.377 | |   
## -------------|-----------|-----------|-----------|  
## Column Total | 42 | 27 | 69 |   
## -------------|-----------|-----------|-----------|  
##   
##

#model using unscaled data  
unscaled.pred <- predict(cancer.tree.unscaled, cancer.unscaled.test, type="class")  
CrossTable(cancer.unscaled.test.target, unscaled.pred, prop.chisq = F, prop.c = F, prop.r = F, dnn = c('actual type', 'predicted type'))

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 69   
##   
##   
## | predicted type   
## actual type | B | M | Row Total |   
## -------------|-----------|-----------|-----------|  
## B | 38 | 1 | 39 |   
## | 0.551 | 0.014 | |   
## -------------|-----------|-----------|-----------|  
## M | 4 | 26 | 30 |   
## | 0.058 | 0.377 | |   
## -------------|-----------|-----------|-----------|  
## Column Total | 42 | 27 | 69 |   
## -------------|-----------|-----------|-----------|  
##   
##

The misclassification of three models are same. The transformation of data will not effect the outputs of the decision tree. The tree structure remain the same with or without the data transformation. So scaling, normalization or leaving the data unscaled makes no difference.