Decision Tree

> library(caret)

>

> eddat <- read.csv('http://math.mercyhurst.edu/~sousley/STAT\_139/data/EDDat3.csv', as.is = T);

>

> attach(eddat);

>

> sapply(eddat, function(x) sum(is.na(x))/length(x))

Rec RecT Grp i1 i2 i3 i4 i5

0.000000000000 0.000000000000 0.000000000000 0.116564417178 0.008179959100 0.008179959100 0.044989775051 0.096114519427

i6 i7 i8 i9 i10 i11 i12 i13

0.010224948875 0.038854805726 0.141104294479 0.020449897751 0.008179959100 0.290388548057 0.505112474438 0.858895705521

i14 i15 i16 i17 i18 i19 i20 i21

0.169734151329 0.171779141104 0.049079754601 0.075664621677 0.065439672802 0.063394683027 0.057259713701 0.030674846626

i22 i23 i24 i25 i26 i27 i28 i29

0.010224948875 0.110429447853 0.022494887526 0.194274028630 0.067484662577 0.253578732106 1.000000000000 0.024539877301

i30 i31 i32 i33 i34 i35 i36 i37

0.151329243354 1.000000000000 0.040899795501 0.177914110429 1.000000000000 0.028629856851 0.014314928425 0.002044989775

i38 i39 i40 i41 i42 i43 i44 i45

1.000000000000 0.024539877301 0.040899795501 0.051124744376 0.130879345603 0.615541922290 0.539877300613 0.038854805726

i46 i47 i48 i49 i50 i51 i52 i53

0.239263803681 0.149284253579 0.065439672802 0.128834355828 0.226993865031 0.177914110429 0.656441717791 0.034764826176

i54 i55 i56 i57 i58

0.233128834356 0.040899795501 0.239263803681 0.061349693252 0.165644171779

>

> sapply(eddat, function(x) sum(!is.na(x))/length(x))

Rec RecT Grp i1 i2 i3 i4 i5 i6 i7

1.0000000000 1.0000000000 1.0000000000 0.8834355828 0.9918200409 0.9918200409 0.9550102249 0.9038854806 0.9897750511 0.9611451943

i8 i9 i10 i11 i12 i13 i14 i15 i16 i17

0.8588957055 0.9795501022 0.9918200409 0.7096114519 0.4948875256 0.1411042945 0.8302658487 0.8282208589 0.9509202454 0.9243353783

i18 i19 i20 i21 i22 i23 i24 i25 i26 i27

0.9345603272 0.9366053170 0.9427402863 0.9693251534 0.9897750511 0.8895705521 0.9775051125 0.8057259714 0.9325153374 0.7464212679

i28 i29 i30 i31 i32 i33 i34 i35 i36 i37

0.0000000000 0.9754601227 0.8486707566 0.0000000000 0.9591002045 0.8220858896 0.0000000000 0.9713701431 0.9856850716 0.9979550102

i38 i39 i40 i41 i42 i43 i44 i45 i46 i47

0.0000000000 0.9754601227 0.9591002045 0.9488752556 0.8691206544 0.3844580777 0.4601226994 0.9611451943 0.7607361963 0.8507157464

i48 i49 i50 i51 i52 i53 i54 i55 i56 i57

0.9345603272 0.8711656442 0.7730061350 0.8220858896 0.3435582822 0.9652351738 0.7668711656 0.9591002045 0.7607361963 0.9386503067

i58

0.8343558282

>

> colSums(is.na(eddat))

Rec RecT Grp i1 i2 i3 i4 i5 i6 i7 i8 i9 i10 i11 i12 i13 i14 i15 i16 i17 i18 i19 i20 i21 i22 i23

0 0 0 57 4 4 22 47 5 19 69 10 4 142 247 420 83 84 24 37 32 31 28 15 5 54

i24 i25 i26 i27 i28 i29 i30 i31 i32 i33 i34 i35 i36 i37 i38 i39 i40 i41 i42 i43 i44 i45 i46 i47 i48 i49

11 95 33 124 489 12 74 489 20 87 489 14 7 1 489 12 20 25 64 301 264 19 117 73 32 63

i50 i51 i52 i53 i54 i55 i56 i57 i58

111 87 321 17 114 20 117 30 81

>

> table(na.omit(eddat[,c(3,5,6,9,13,25,39,40,12,27,32)])$Grp)

1 2 3

79 272 101

>

> eddat <- na.omit(eddat[,c(3,5,6,9,13,25,39,40,12,27,32)])

>

> Accuracies <- c(0.00)

>

> for (i in seq(1000))

+ {

+ inTest<-createDataPartition(eddat$Grp, p = .25, list = FALSE)

+ require(rpart)

+ edata1<-rpart(Grp~ ., data=eddat,method = "class",subset = inTest,

+ parms= list(split = "gini",prior = c(1/3,1/3,1/3)),

+ control = rpart.control(usesurrogate= 0, maxsurrogate= 0))

+

+ Accuracies[i] <- confusionMatrix(eddat[inTest,"Grp"],predict(edata1,newdata= eddat[inTest,],type = "class"))$overall["Accuracy"]

+

+ }

> summary(Accuracies)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.6315789 0.7456140 0.7719298 0.7680965 0.7894737 0.8771930