Diabetes Analysis

Arthur Gurupatham

28/07/2020

Question 3

Basic information about the data set

THe diabetes data set is a data set that contains 3 measurements (glucose, insulin and steady state plasma glucose) to determine which type of diabetes the adult has (Normal, Overt or Chemical). The data contains the information about 145 non-obese adults. The source of this data is: Reaven, G. M. and Miller, R. G. (1979). An attempt to define the nature of chemical diabetes using a multidimensional analysis. Diabetologia 16:17-24.

Splitting Data set and applying mixture discriminant analysis

```
library(mclust)
```

```
## Warning: package 'mclust' was built under R version 3.6.3
```

```
## Package 'mclust' version 5.4.6
## Type 'citation("mclust")' for citing this R package in publications.
```

```
data("diabetes")
x<-diabetes[,-1]
x<-scale(x)

#I chose 36, for the unlabelled split, since 145*0.25 is approximately 36.
diabetes_delete<-rep(0,36)
k<-1
for(i in 1:dim(diabetes)[1]){
  if(i%%4==0){diabetes_delete[k]<-i; k<-k+1}
}

diabetesMclustDA <- MclustDA(x[-diabetes_delete,], diabetes[-diabetes_delete,1])
summary(diabetesMclustDA, parameters = TRUE)
```

```
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
## MclustDA model summary:
##
##  log-likelihood   n df       BIC
##       -106.4722 109 44 -419.3636
##
## Classes     n     % Model G
##   Chemical 29 26.61   EVE 2
##   Normal   55 50.46   XXI 1
##   Overt    25 22.94   EEV 3
##
## Class prior probabilities:
##  Chemical    Normal     Overt
## 0.2660550 0.5045872 0.2293578
##
## Class = Chemical
##
## Mixing probabilities: 0.7243276 0.2756724
##
## Means:
##               [,1]        [,2]
## glucose -0.3625259 -0.36738537
## insulin -0.2023485 -0.01937982
## sspg     0.3063090  2.27083730
##
## Variances:
## [,,1]
##            glucose     insulin        sspg
## glucose 0.01883928  0.02103161  0.01594272
## insulin 0.02103161  0.03408880 -0.01341426
## sspg    0.01594272 -0.01341426  0.37529684
## [,,2]
##             glucose      insulin        sspg
## glucose 0.010363982  0.003161283  0.09757478
## insulin 0.003161283  0.014102438 -0.08400769
## sspg    0.097574780 -0.084007694  2.30016618
##
## Class = Normal
##
## Mixing probabilities: 1
##
## Means:
##                [,1]
## glucose -0.48866812
## insulin -0.58962212
## sspg    -0.09207763
##
## Variances:
## [,,1]
##            glucose    insulin     sspg
## glucose 0.01655959 0.00000000 0.000000
## insulin 0.00000000 0.01224078 0.000000
## sspg    0.00000000 0.00000000 0.383819
## 
## Class = Overt
## 
## Mixing probabilities: 0.2821021 0.2009324 0.5169655
## 
## Means:
##              [,1]       [,2]       [,3]
## glucose  3.100281 0.09578814  1.0974366
## insulin  2.860062 0.32644882  1.3346995
## sspg    -1.329241 0.25067881 -0.7181817
## 
## Variances:
## [,,1]
##             glucose     insulin        sspg
## glucose  0.47538715  0.10290897 -0.06433021
## insulin  0.10290897  0.09064625 -0.02458185
## sspg    -0.06433021 -0.02458185  0.01936142
## [,,2]
##              glucose     insulin         sspg
## glucose  0.012959848  0.01558889 -0.008539637
## insulin  0.015588887  0.06549157 -0.039810525
## sspg    -0.008539637 -0.03981052  0.506943399
## [,,3]
##             glucose     insulin        sspg
## glucose  0.35228490  0.22838049 -0.03500172
## insulin  0.22838049  0.16379612 -0.03705799
## sspg    -0.03500172 -0.03705799  0.06931379
## 
## Training confusion matrix:
##           Predicted
## Class      Chemical Normal Overt
##   Chemical       28      0     1
##   Normal          1     54     0
##   Overt           0      0    25
## Classification error = 0.0183 
## Brier score          = 0.02
```

```
summary(diabetesMclustDA, newdata = x[diabetes_delete,], newclass = diabetes[diabetes_delete,1])
```

```
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
## MclustDA model summary:
##
##  log-likelihood   n df       BIC
##       -106.4722 109 44 -419.3636
##
## Classes     n     % Model G
##   Chemical 29 26.61   EVE 2
##   Normal   55 50.46   XXI 1
##   Overt    25 22.94   EEV 3
##
## Training confusion matrix:
##           Predicted
## Class      Chemical Normal Overt
##   Chemical       28      0     1
##   Normal          1     54     0
##   Overt           0      0    25
## Classification error = 0.0183
## Brier score          = 0.02
##
## Test confusion matrix:
##           Predicted
## Class      Chemical Normal Overt
##   Chemical        5      0     2
##   Normal          0     21     0
##   Overt           0      0     8
## Classification error = 0.0556
## Brier score          = 0.0381
```

Applying the classification tree model to the data set

```
library(rpart)
library(mclust)
library(rattle)
```

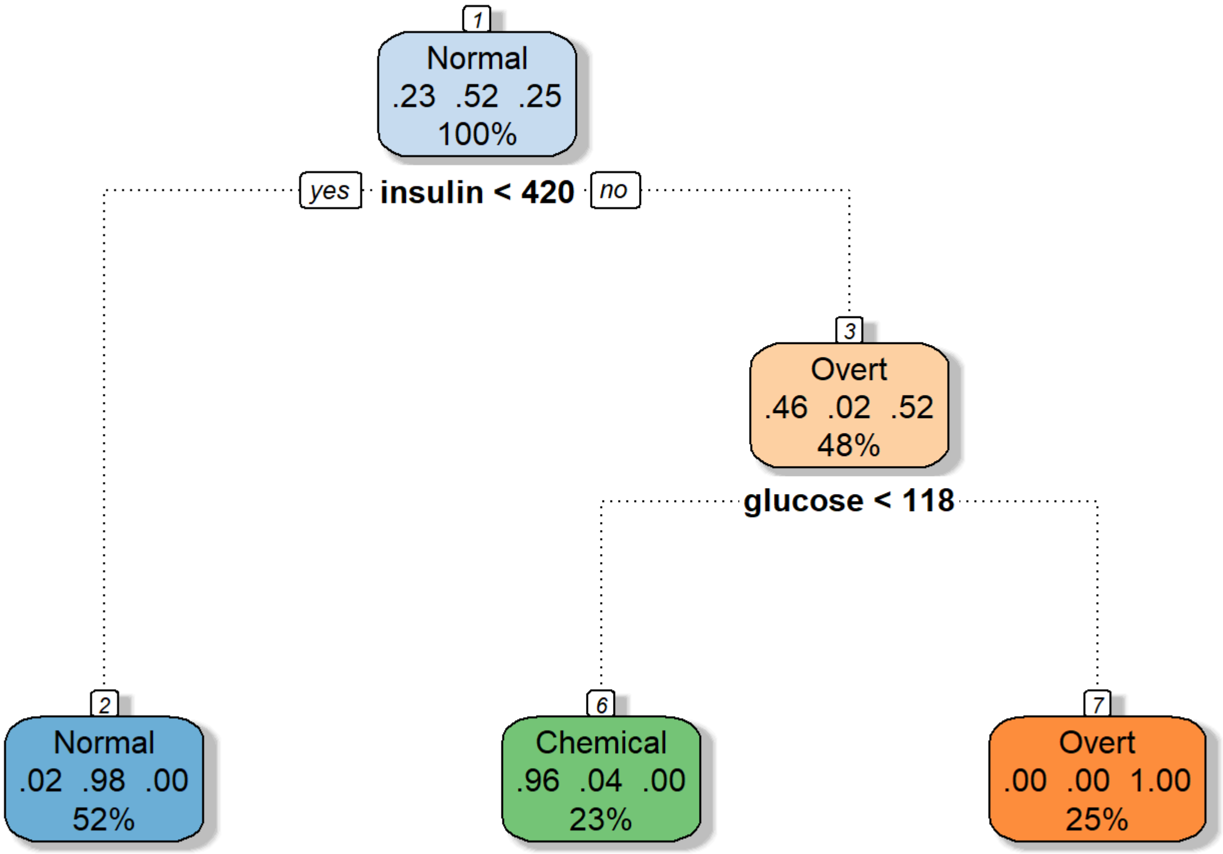
```
## Warning: package 'rattle' was built under R version 3.6.3
```

```
## Loading required package: tibble
```

```
## Loading required package: bitops
```

```
## Rattle: A free graphical interface for data science with R.
## Version 5.4.0 Copyright (c) 2006-2020 Togaware Pty Ltd.
## Type 'rattle()' to shake, rattle, and roll your data.
```

```
data("diabetes", package = "mclust")
# split data into a 75% training set, 25% test set
train <- sample(1:nrow(diabetes), size=nrow(diabetes)*0.75)
test=diabetes[-train]
diabetes_tree <- rpart(class ~ ., data = diabetes, subset = train)
fancyRpartPlot(diabetes_tree)
```



Rattle 2020-Jul-29 00:05:36 arthu

```
table(predict(diabetes_tree, diabetes[-train,], type = "class"),
      diabetes[-train, "class"])
```

```
##
##              Chemical Normal Overt
##    Chemical        10      0     0
##    Normal           1     20     0
##    Overt            0      0     6
```

Analysis of the methods

From the MDA, we can see that 55.5% of the data was Normal cases, 29.26% were Chemical and 25.2% of cases were Overt. Whereas, in the Classification tree model, the distribtuion of classes is different, as seen in the third level of the rattle plot. In MDA, the classification error of the test matrix was 5.56%. This was calculated by counting the misclassified data points (2) and dividing it by the total number of data points (36) and

2/36=0.0556=5.56%. Using a similar approach, we can see that for classification tress, the misclassification rate is 1/37=0.027=2.7%. Based on this fact, we can conclude that Classification tress is a better model approach for the diabetes data set than MDA.