Mathematical Methods for Artificial Intelligence Lab 3

Vytautas Kraujalis

2021-10-30

Contents

1 Reading Data **1**

2 Required packages **2**

3 Parallel processing **2**

4 EDA, first look at the dataset **2**

5 Pre-process **9**

5.1 Removal of correlated variables 9

6 Fitting models **10**

6.1 LDA 10

6.2 QDA 11

6.3 RDA 11

6.4 CART 15

6.5 Conditional Inference Tree 17

7 Comparison of models **19**

8 References **21**

1 Reading Data

```
set.seed(123)

data_original <- read.csv("Arrhythmia_Dataset.csv")
```

2 Required packages

```
library(SmartEDA)
library(dplyr)
library(ggplot2)
library(caret)
library(rattle)
library(partykit)
library(groupdata2)
library(cvms)

# Function for 3 class accuracy from confusion matrix
classAcc <- function(confusionMatrix) {
  class0 <- round(confusionMatrix$table[1, 1] / sum(confusionMatrix$table[, 1]) * 100, 1)
  class1 <- round(confusionMatrix$table[2, 2] / sum(confusionMatrix$table[, 2]) * 100, 1)
  class2 <- round(confusionMatrix$table[3, 3] / sum(confusionMatrix$table[, 3]) * 100, 1)
  acc <- c(class0, class1, class2 )
  names(acc) <- colnames(confusionMatrix$table)
  return(acc)
}
```

3 Parallel processing

```
library(parallel)
no_cores <- detectCores() - 1
library(doParallel)
cl <- makePSOCKcluster(no_cores)
registerDoParallel(cl)
```

4 EDA, first look at the dataset

```
ExpData(data_original,type=1)
```

```
##                                         Descriptions     Value
## 1                                 Sample size (nrow)    175729
## 2                            No. of variables (ncol)        34
## 3                  No. of numeric/interger variables        32
## 4                            No. of factor variables         0
## 5                              No. of text variables         2
## 6                           No. of logical variables         0
## 7                        No. of identifier variables         0
## 8                              No. of date variables         0
## 9               No. of zero variance variables (uniform)     0
## 10             %. of variables having complete cases 100% (34)
## 11   %. of variables having >0% and <50% missing cases    0% (0)
## 12 %. of variables having >=50% and <90% missing cases    0% (0)
## 13          %. of variables having >=90% missing cases    0% (0)
```

We have a dataset of 175729 observations with 34 variables, of which are 2 text variables and 32 - numerical. All variables have no missing values.

Let's look at the text variables:

```
data_original %>%
   group_by(record) %>%
   summarise(n = n()) %>%
   summary()
```

```
##     record                n
##  Length:75          Min.   :1452
##  Class :character   1st Qu.:1957
##  Mode  :character   Median :2323
##                     Mean   :2343
##                     3rd Qu.:2659
##                     Max.   :3909
```

Our first text variable record has 75 unique values (unique patients). Each patient has on average 2343 observations.

```
data_original %>%
   group_by(type) %>%
   summarise(n = n()) %>%
   mutate(n_prop = round(n / sum(n) * 100, 2))
```

```
## # A tibble: 5 x 3
##   type       n n_prop
##   <chr>  <int>  <dbl>
## 1 F        219   0.12
## 2 N     153546  87.4
## 3 Q          6   0
## 4 SVEB    1958   1.11
## 5 VEB    20000  11.4
```

The next text variable type is our target variable. It has 5 classes, but there's a huge class imbalance. 2 out of 5 classes takes up to 98.76% of all observations, 1 class has just 6 observations. We are planning to combine F, Q, SVEB classes into one class, as those 3 classes combined only takes up to 1.23% of observations.

N (Normal) - Normal beat.

SVEB (Supraventricular ectopic beat) - Supraventricular Ectopic Beats indicates atrial irritability. Isolated Supraventricular Ectopic Beats are generally not significant in nature but a high frequency can represent more risk. An increasing trend in Supraventricular Ectopic Beats may be an indicator or sign for atrial fibrillation. Atrial Fibrillation is considered to be significant as it can lead to heart attack or stroke.

VEB (Ventricular ectopic beat) - Ventricular ectopics are a type of arrhythmia or abnormal heart rhythm. It is caused by the electric signals in the heart starting in a different place and travelling a different way through the heart. If it happens occasionally, it should not cause any problems but if it happens a lot, you will need to have treatment.

F (Fusion beat) - A fusion beat occurs when a supraventricular and a ventricular impulse coincide to produce a hybrid complex. It indicates that there are two foci of pacemaker cells firing simultaneously: a supraventricular pacemaker (e.g. the sinus node) and a competing ventricular pacemaker (source of ventricular ectopics).

Q (Unknown beat) - Unknown beat.

```r
data <- data_original %>%
   mutate(type = case_when(
   type %in% c("F", "Q", "SVEB") ~ "F_Q_SVEB",
   TRUE ~ type
   ))

data %>%
   group_by(type) %>%
   summarise(n = n()) %>%
   mutate(n_prop = round(n / sum(n) * 100, 2))
```

```
## # A tibble: 3 x 3
##   type          n n_prop
##   <chr>     <int>  <dbl>
## 1 F_Q_SVEB   2183   1.24
## 2 N        153546  87.4
## 3 VEB       20000  11.4
```

We combined the 3 classes into 1 class, which has only 2183 observations (1.24%) which is still low, but still better for our basic models to classify.

We are not interested in the patient record variable, so we'll remove it:

```r
data <- data %>%
   select(-record)
```

Let's look at descriptive statistics of each variable:

```r
ExpNumStat(data,by ="A",round= 2, gp = "type") %>%
  select(Vname, min, max, mean, median, SD)
```

```
##              Vname   min    max   mean median    SD
## 2       X0_post.RR 71.00 506.00 197.20 188.00 61.72
## 3         X0_pPeak -4.21  10.57   0.04   0.00  0.19
## 9   X0_pq_interval  1.00 113.00   8.71   6.00  7.89
## 1        X0_pre.RR 49.00 506.00 197.18 188.00 61.70
## 7         X0_qPeak -7.11   3.13  -0.17  -0.11  0.37
## 8  X0_qrs_interval  0.00 131.00  17.62  17.00 10.62
## 12   X0_qrs_morph0 -7.11   3.13  -0.17  -0.11  0.37
## 13   X0_qrs_morph1 -7.11   3.47  -0.08  -0.03  0.40
## 14   X0_qrs_morph2 -7.51   4.09   0.28   0.24  0.64
## 15   X0_qrs_morph3 -7.80   4.50   0.61   0.64  0.85
## 16   X0_qrs_morph4 -7.94   4.30   0.17   0.14  0.68
## 10  X0_qt_interval  3.00 247.00  37.46  30.00 23.64
## 5         X0_rPeak -7.11   4.60   0.86   0.91  0.87
## 6         X0_sPeak -7.97   3.89  -0.59  -0.63  0.80
## 11  X0_st_interval  1.00  89.00  11.13   6.00 14.30
## 4         X0_tPeak -7.96   4.63   0.12  -0.04  0.64
## 18      X1_post.RR 71.00 506.00 197.20 188.00 61.72
## 19        X1_pPeak -1.10   7.63   0.06   0.04  0.11
## 25  X1_pq_interval  1.00 121.00   7.11   5.00  7.65
## 17       X1_pre.RR 49.00 506.00 197.18 188.00 61.70
```

```
## 23        X1_qPeak -6.46   1.38  -0.22  -0.12  0.28
## 24 X1_qrs_interval  0.00 128.00  10.59   5.00  9.92
## 28   X1_qrs_morph0 -6.46   1.38  -0.22  -0.12  0.28
## 29   X1_qrs_morph1 -4.91   2.78  -0.25  -0.19  0.31
## 30   X1_qrs_morph2 -3.99   4.04  -0.32  -0.34  0.41
## 31   X1_qrs_morph3 -2.96   2.77  -0.39  -0.48  0.49
## 32   X1_qrs_morph4 -3.13   2.76  -0.54  -0.59  0.49
## 26  X1_qt_interval  4.00 290.00  28.05  24.00 14.83
## 21        X1_rPeak -2.39   4.06  -0.11  -0.09  0.40
## 22        X1_sPeak -3.27   2.82  -0.75  -0.75  0.42
## 27  X1_st_interval  1.00  86.00  10.35   9.00  4.83
## 20        X1_tPeak -2.02   3.65   0.37   0.32  0.28
```

```
variables_of_further_interest <- c("X0_post.RR", "X0_pq_interval", "X0_pre.RR", "X0_qrs_interval", "X0_q
```

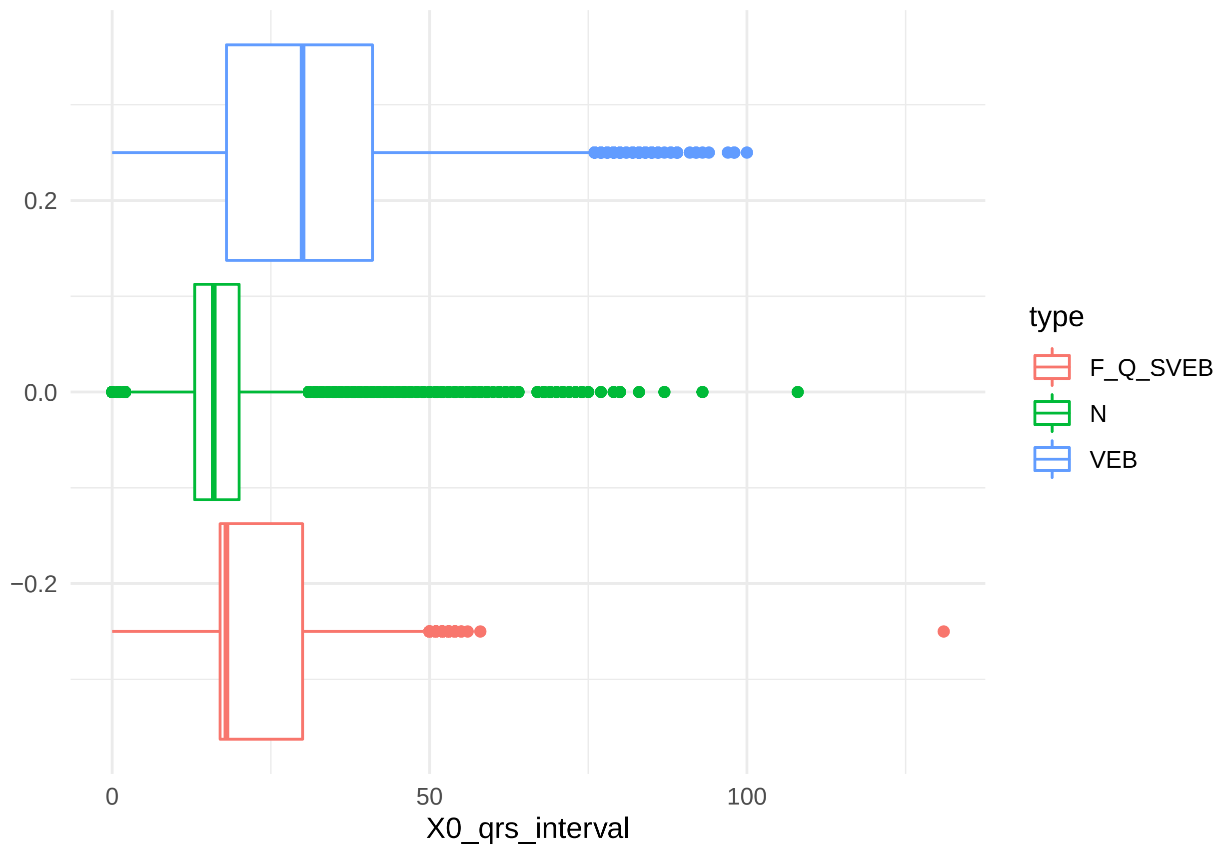
We noticed some variables which should require further analysis, those variables are:

```
variables_of_further_interest
```

```
## [1] "X0_post.RR"      "X0_pq_interval"  "X0_pre.RR"       "X0_qrs_interval"
## [5] "X0_qt_interval"  "X1_post.RR"      "X1_pre.RR"       "X1_qrs_interval"
## [9] "X1_qt_interval"
```

We will look through these variables more closely and report any irregularities.

```
ggplot(data, aes(x = X0_qrs_interval, color = type)) +
   geom_boxplot() +
   theme_minimal()
```



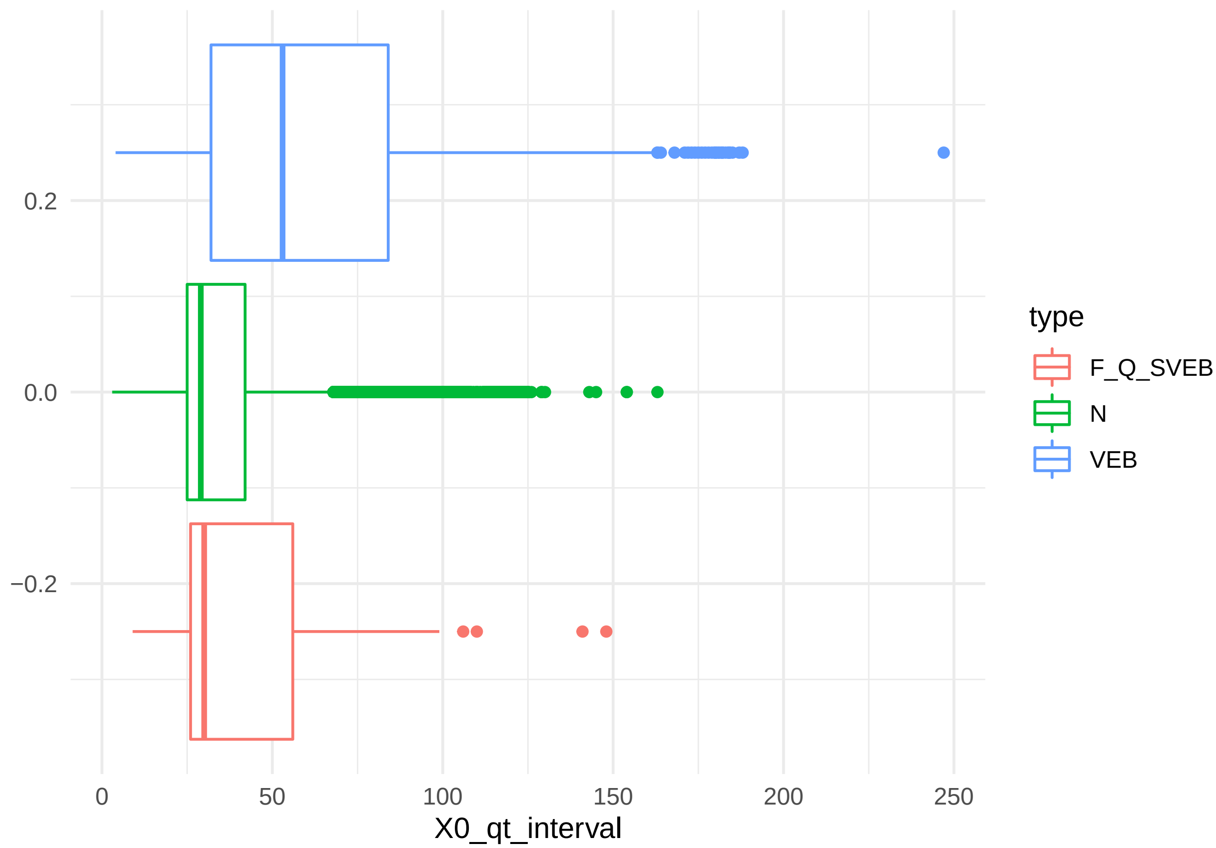

```
data_original %>%
  filter(X0_qrs_interval > 125)
```

```
##   record type X0_pre.RR X0_post.RR   X0_pPeak X0_tPeak X0_rPeak X0_sPeak
## 1    I43    F       335        115 -0.1446686 1.394598 1.178978 1.178978
##     X0_qPeak X0_qrs_interval X0_pq_interval X0_qt_interval X0_st_interval
## 1 -0.1450307             131              2            135              2
##   X0_qrs_morph0 X0_qrs_morph1 X0_qrs_morph2 X0_qrs_morph3 X0_qrs_morph4
## 1    -0.1450307   -0.09808338   -0.04288349  -0.004493709    0.01901518
##   X1_pre.RR X1_post.RR  X1_pPeak X1_tPeak  X1_rPeak  X1_sPeak  X1_qPeak
## 1       335        115 0.1182595 1.325323 -2.114941 -2.323128 -2.114941
##   X1_qrs_interval X1_pq_interval X1_qt_interval X1_st_interval X1_qrs_morph0
## 1               2              8             26             16     -2.114941
##   X1_qrs_morph1 X1_qrs_morph2 X1_qrs_morph3 X1_qrs_morph4
## 1     -2.114941     -2.114941     -2.306017     -2.306017
```

There's quite an unusual observation in X0_qrs_interval variable, where our new class F_Q_SVEB has a value of > 125. We also see from the original dataset that this observation was classified as F. We are going to classify this observation as an outlier and remove it.

```
data <- data %>%
  filter(X0_qrs_interval < 125)
```

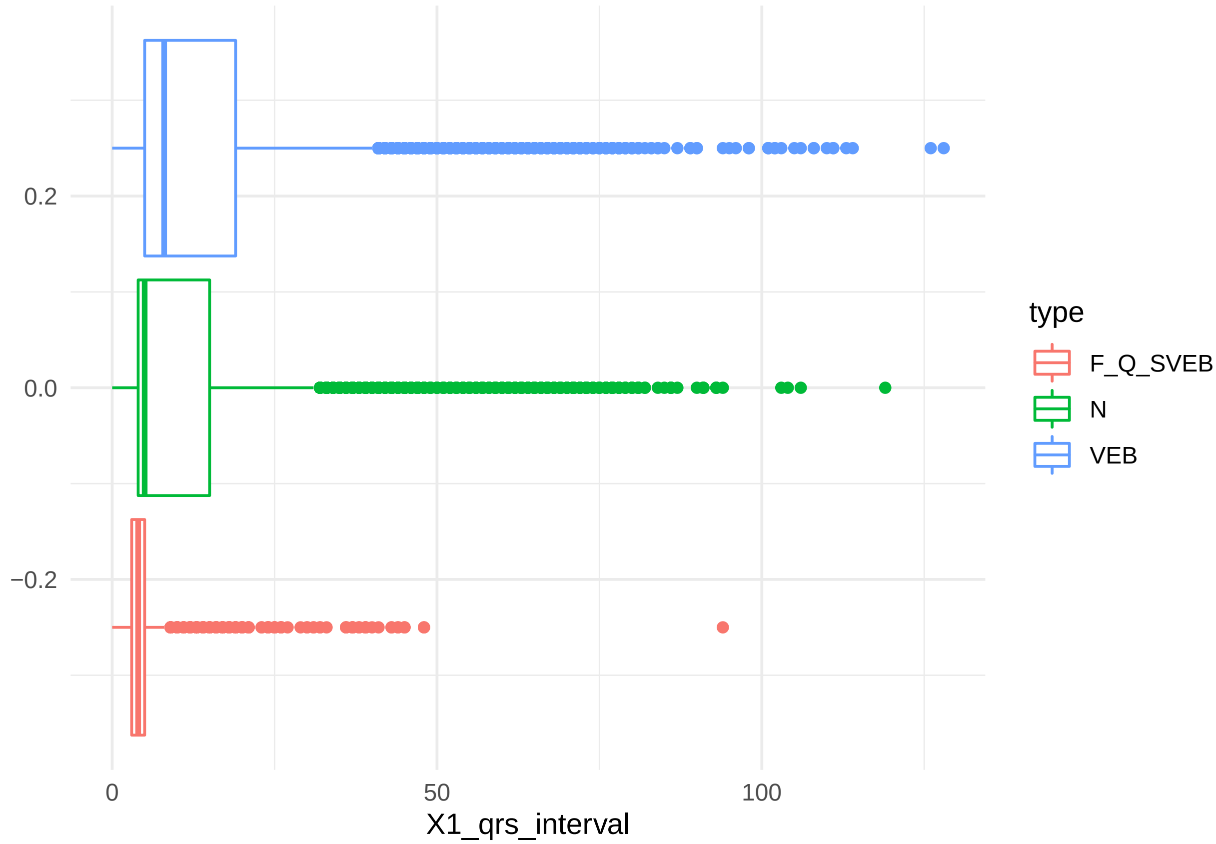
```r
ggplot(data, aes(x = X0_qt_interval, color = type)) +
  geom_boxplot() +
  theme_minimal()
```



We noticed another outlier, for the X0_qt_interval variable, when the type is VEB, the value is close to 250 which is quite unusual in this dataset. We'll remove this observation.

```r
data <- data %>%
  filter(X0_qt_interval < 225)
```

```r
ggplot(data, aes(x = X1_qrs_interval, color = type)) +
  geom_boxplot() +
  theme_minimal()
```




```
data_original %>%
  filter(X1_qrs_interval > 75, type %in% c("F", "Q", "SVEB"))
```

```
##   record type X0_pre.RR X0_post.RR   X0_pPeak  X0_tPeak   X0_rPeak   X0_sPeak
## 1    I05    F       177        367 0.02885405 0.4625619 -0.1868355 -0.2210357
##     X0_qPeak X0_qrs_interval X0_pq_interval X0_qt_interval X0_st_interval
## 1 -0.1868355               2             12             27             13
##   X0_qrs_morph0 X0_qrs_morph1 X0_qrs_morph2 X0_qrs_morph3 X0_qrs_morph4
## 1    -0.1868355    -0.1868355    -0.1868355    -0.2103347    -0.2103347
##   X1_pre.RR X1_post.RR   X1_pPeak  X1_tPeak  X1_rPeak   X1_sPeak   X1_qPeak
## 1       177        367 0.08331537 0.3157339 0.1065414 -0.1558735 -0.1257657
##   X1_qrs_interval X1_pq_interval X1_qt_interval X1_st_interval X1_qrs_morph0
## 1              94             22            123              7    -0.1257657
##   X1_qrs_morph1 X1_qrs_morph2 X1_qrs_morph3 X1_qrs_morph4
## 1   -0.09340949   -0.04265345   -0.01324232   0.006886856
```

Seems like another outlier was detected in X1_qrs_interval variable, where the type is F_Q_SVEB and the value is close to 100. From the original dataset, this observation had F type. We'll remove it.

```
data %>%
  mutate(row = row_number()) %>%
  filter(X1_qrs_interval > 75, type == "F_Q_SVEB") %>%
  select(row)
```

```
##     row
## 1 10934
```

```r
data <- data %>%
   filter(row_number() != 10934)
```

We should look at the correlation between variables

```r
# Correlation

corr_simple <- function(df,sig=0.5){
  corr <- cor(df)
  #prepare to drop duplicates and correlations of 1
  corr[lower.tri(corr,diag=TRUE)] <- NA
  #drop perfect correlations
  corr[corr == 1] <- NA
  #turn into a 3-column table
  corr <- as.data.frame(as.table(corr))
  #remove the NA values from above
  corr <- na.omit(corr)
  #select significant values
  corr <- subset(corr, abs(Freq) > sig)
  #sort by highest correlation
  corr <- corr[order(-abs(corr$Freq)),]
  return(corr)
}

correlation_matrix = cor(data %>% select(-type))

length(findCorrelation(correlation_matrix, cutoff = 0.99))
```

```
## [1] 4
```

```r
length(findCorrelation(correlation_matrix, cutoff = 0.95))
```

```
## [1] 6
```

```r
length(findCorrelation(correlation_matrix, cutoff = 0.9))
```

```
## [1] 8
```

We have 4 variables with higher than .99 correlation. We have 6 variables with higher than .95 correlation. We have 8 variables with higher than .9 correlation.

5 Pre-process

5.1 Removal of correlated variables

We've found some highly correlated variables. We don't have a lot of variables (32) and because of that, we'll only remove the variables with >.99 correlation.

```
data <- data %>%
  select(-findCorrelation(correlation_matrix, cutoff = 0.99))
```

6 Fitting models

6.1 LDA

```
fitControl = trainControl(## 5-fold CV
                           method = "cv",
                           number = 5,
                           ## Estimate class probabilities
                           classProbs = TRUE,
                           ## Evaluate performance using
                           ## the following function
                           summaryFunction = multiClassSummary)
lda.fit = train(type ~ ., data = data,
                 method = "lda",
                 preProcess=c("center","scale"),
                 trControl = fitControl)

print(lda.fit$results %>% t())
```

```
##                            1
## parameter                  "none"
## logLoss                    "0.1569349"
## AUC                        "0.9660774"
## prAUC                      "0.8503791"
## Accuracy                   "0.9682119"
## Kappa                      "0.8479544"
## Mean_F1                    "0.8095788"
## Mean_Sensitivity           "0.7998741"
## Mean_Specificity           "0.9366059"
## Mean_Pos_Pred_Value        "0.8271019"
## Mean_Neg_Pred_Value        "0.9781875"
## Mean_Precision             "0.8271019"
## Mean_Recall                "0.7998741"
## Mean_Detection_Rate        "0.3227373"
## Mean_Balanced_Accuracy     "0.86824"
## logLossSD                  "0.003783519"
## AUCSD                      "0.002409343"
## prAUCSD                    "0.006933635"
## AccuracySD                 "0.0004901414"
## KappaSD                    "0.002839515"
## Mean_F1SD                  "0.003479908"
## Mean_SensitivitySD         "0.002722242"
## Mean_SpecificitySD         "0.002274744"
## Mean_Pos_Pred_ValueSD      "0.005635765"
## Mean_Neg_Pred_ValueSD      "0.0006510096"
## Mean_PrecisionSD           "0.005635765"
## Mean_RecallSD              "0.002722242"
```

```
## Mean_Detection_RateSD    "0.0001633805"
## Mean_Balanced_AccuracySD "0.00183603"
```

AUC is 0.96, which is pretty good, the Mean F1 is 0.8 which also indicated quite a good model. The variance for AUC is 0.002, for Mean F1 - 0.003, this indicated a stable model.

```
confusion_matrix <- confusionMatrix(lda.fit)

confusion_matrix
```

```
## Cross-Validated (5 fold) Confusion Matrix
##
## (entries are percentual average cell counts across resamples)
##
##             Reference
## Prediction F_Q_SVEB    N  VEB
##   F_Q_SVEB      0.8  0.2  0.5
##   N             0.4 87.0  1.8
##   VEB           0.0  0.2  9.1
##
##  Accuracy (average) : 0.9682
```

```
classAcc(confusion_matrix)
```

```
## F_Q_SVEB        N      VEB
##     60.8     99.6     79.6
```

The overall accuracy is pretty good (96.8%), the accuracies for each class are: F_Q_SVEB = 60.9%, N = 99.6%, VEB = 79.6%.

6.2 QDA

```
fitControl <- trainControl(# 5-fold CV
                           method = "cv",
                           number = 5,
                           # Estimate class probabilities
                           classProbs = TRUE,
                           # Evaluate performance using
                           # the following function
                           summaryFunction = multiClassSummary,
                           verboseIter = TRUE)
# qda.fit <- train(type ~ ., data = data,
#                  method = "qda",
#                  preProcess=c("center","scale"),
#                  trControl = fitControl)
```

Can't fit a QDA model because of our small class F_Q_SVEB.

6.3 RDA

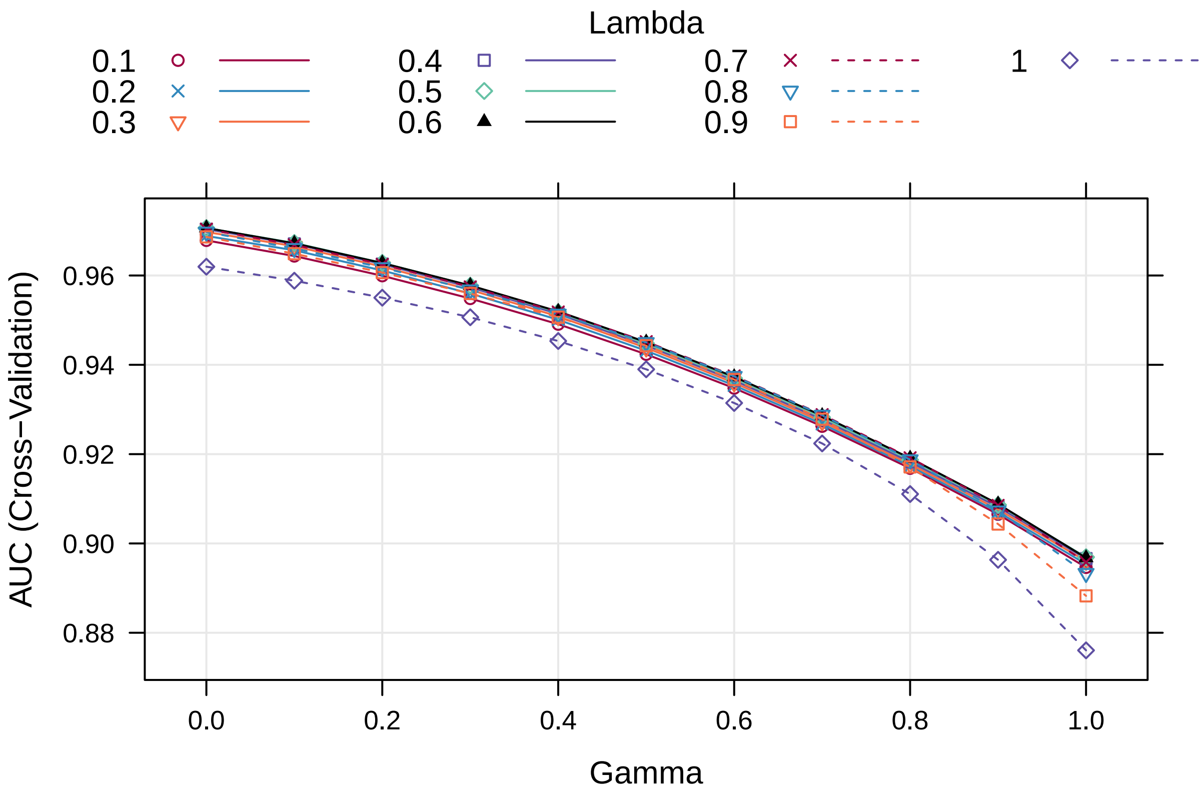
```r
library(tictoc)
tic()
fitControl <- trainControl(## 5-fold CV
                           method = "cv",
                           number = 5,
                           ## Estimate class probabilities
                           classProbs = TRUE,
                           ## Evaluate performance using
                           ## the following function
                           summaryFunction = multiClassSummary,
                           savePredictions="all",
                           verboseIter = TRUE)
rdaGrid =  expand.grid(lambda = seq(0.1,1,0.1),
                        gamma = seq(0,1,0.1))
rda.fit <- train(type ~ ., data = data,
                 method = "rda",
                 trControl = fitControl,
                 preProcess=c("center", "scale","pca"),
                 ## Now specify the exact models
                 ## to evaluate:
                 tuneGrid = rdaGrid)
```

```
## Aggregating results
## Selecting tuning parameters
## Fitting gamma = 0, lambda = 0.9 on full training set
```

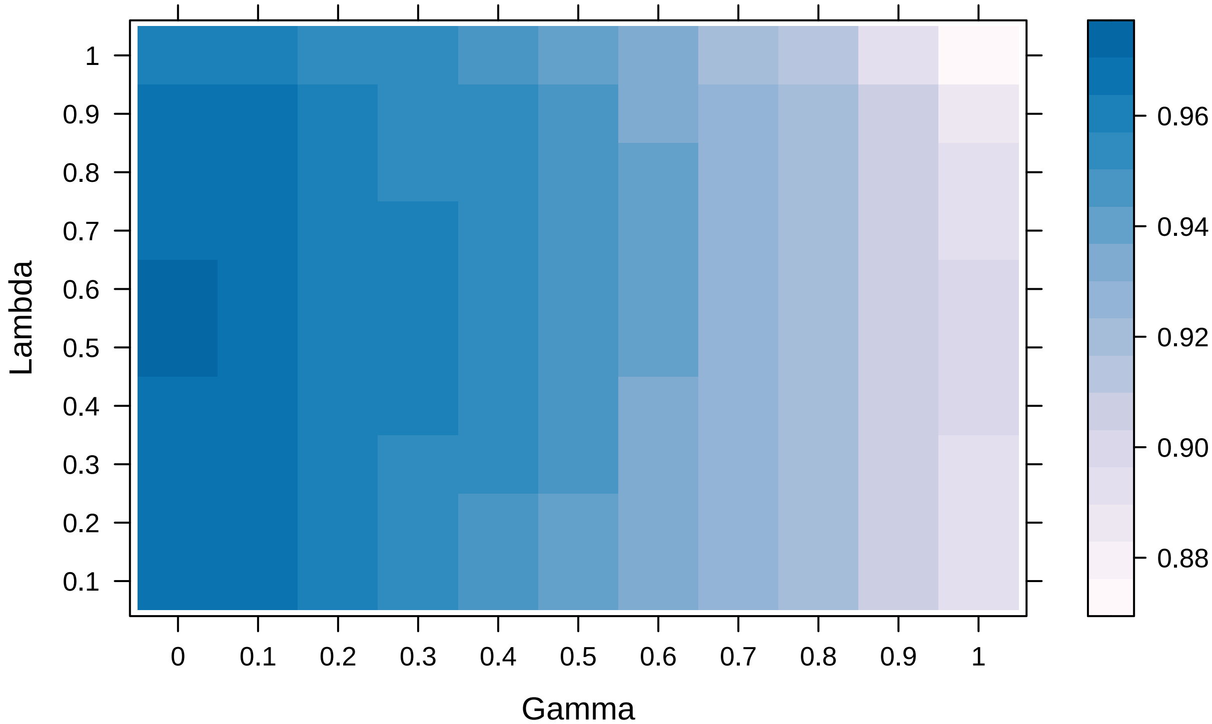
```r
toc()
```

```
## 1989.33 sec elapsed
```

```r
trellis.par.set(caretTheme())
plot(rda.fit, metric = "AUC")
```




```
plot(rda.fit, metric = "AUC", plotType = "level")
```



AUC (Cross−Validation)

From those 2 graphs, we can see that the best cross-validation AUC is achieved using gamma - 0 and Lambda 0.5 or 0.6.

```
resamp = rda.fit$pred %>%
    filter(gamma == 0 & lambda == 0.6)
confusion_matrix <- confusionMatrix(resamp$pred, resamp$obs)
confusion_matrix
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction F_Q_SVEB      N    VEB
##   F_Q_SVEB     1309    334    943
##   N             725 150212   2902
##   VEB           147   3000  16154
##
## Overall Statistics
##
##                Accuracy : 0.9542
##                  95% CI : (0.9532, 0.9552)
##     No Information Rate : 0.8738
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.794
##
##  Mcnemar's Test P-Value : < 2.2e-16
```

```
##
## Statistics by Class:
##
##                      Class: F_Q_SVEB Class: N Class: VEB
## Sensitivity                 0.600183   0.9783    0.80774
## Specificity                 0.992642   0.8365    0.97979
## Pos Pred Value              0.506187   0.9764    0.83695
## Neg Pred Value              0.994964   0.8477    0.97542
## Prevalence                  0.012411   0.8738    0.11381
## Detection Rate              0.007449   0.8548    0.09193
## Detection Prevalence        0.014716   0.8754    0.10984
## Balanced Accuracy           0.796413   0.9074    0.89377
```

```
classAcc(confusion_matrix)
```

```
## F_Q_SVEB        N      VEB
##     60.0     97.8     80.8
```

Overall accuracy using gamma - 0 and lambda - 0.6 is 95.4%, while accuracies for each class are: F_Q_SVEB = 60.1%, N = 97.8% and VEB = 80.8%.

6.4 CART

```
fitControl <- trainControl(## 5-fold CV
                           method = "cv",
                           number = 5,
                           ## Estimate class probabilities
                           classProbs = TRUE,
                           ## Evaluate performance using
                           ## the following function
                           summaryFunction = multiClassSummary,
                           savePredictions="all",
                           verboseIter = TRUE)
cart.fit <- train(type ~ ., data = data,
                 method = "rpart",
                 trControl = fitControl,
                 preProcess=c("center", "scale"))
```

```
## Aggregating results
## Selecting tuning parameters
## Fitting cp = 0.0357 on full training set
```

```
cart.fit
```

```
## CART
##
## 175726 samples
##     28 predictor
##      3 classes: 'F_Q_SVEB', 'N', 'VEB'
##
```

```
## Pre-processing: centered (28), scaled (28)
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 140581, 140581, 140581, 140581, 140580
## Resampling results across tuning parameters:
##
##   cp          logLoss    AUC        prAUC       Accuracy   Kappa      Mean_F1
##   0.03566276  0.1825564  0.8196115  0.30644074  0.9594482  0.7983500  NaN
##   0.04332732  0.1957116  0.8106880  0.21700151  0.9552940  0.7732096  NaN
##   0.31169973  0.3345831  0.6220519  0.03361721  0.9051024  0.3017712  NaN
##   Mean_Sensitivity  Mean_Specificity  Mean_Pos_Pred_Value  Mean_Neg_Pred_Value
##   0.5953798         0.9126904         NaN                  0.9723515
##   0.5820672         0.9004816         NaN                  0.9714621
##   0.4291232         0.7569596         NaN                  0.9696970
##   Mean_Precision  Mean_Recall  Mean_Detection_Rate  Mean_Balanced_Accuracy
##   NaN             0.5953798    0.3198161            0.7540351
##   NaN             0.5820672    0.3184313            0.7412744
##   NaN             0.4291232    0.3017008            0.5930414
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was cp = 0.03566276.
```

```
confusion_matrix <- confusionMatrix(cart.fit)

confusion_matrix
```

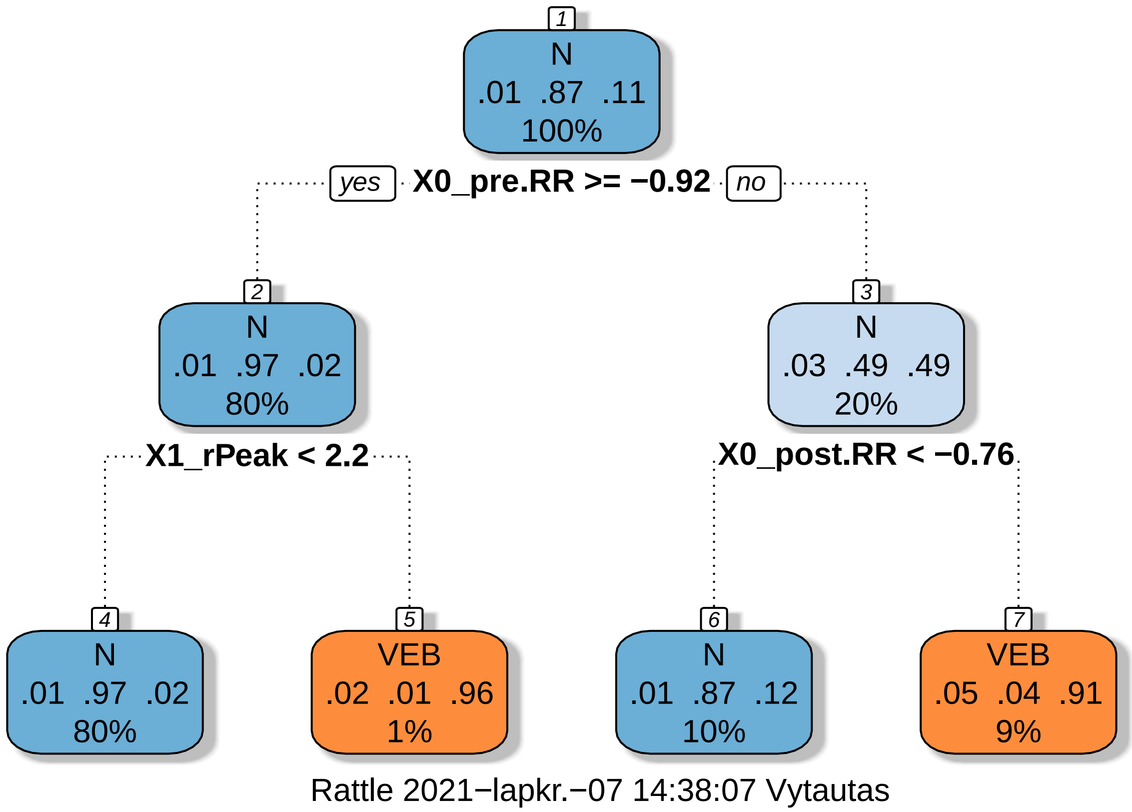
```
## Cross-Validated (5 fold) Confusion Matrix
##
## (entries are percentual average cell counts across resamples)
##
##           Reference
## Prediction F_Q_SVEB    N  VEB
##   F_Q_SVEB      0.0  0.0  0.0
##   N             0.8 86.9  2.4
##   VEB           0.4  0.4  9.0
##
##  Accuracy (average) : 0.9594
```

```
classAcc(confusion_matrix)
```

```
## F_Q_SVEB        N      VEB
##      0.0     99.5     79.1
```

Our overall accuracy is 95.9%, but if we would look at each class accuracies: F_Q_SVEB = 0%, N = 99.5% and VEB = 79.3% we could see that our newly created class has 0% accuracy.

```
fancyRpartPlot(cart.fit$finalModel)
```



We can see from the final model, that our simple decision tree is too simple for our data.

6.5 Conditional Inference Tree

```r
# Create 5 fold columns with 3 folds each
data_cv <- fold(
  data,
  k = 5,
  cat_col = "type",
  parallel = TRUE # set to TRUE to run in parallel
) %>%
   mutate(type = as.factor(type))

data_cv %>%
   count(.folds, type)
```

```
## # A tibble: 15 x 3
## # Groups:   .folds [5]
##     .folds type         n
##     <fct>  <fct>    <int>
##   1 1      F_Q_SVEB   436
##   2 1      N        30709
##   3 1      VEB       3999
##   4 2      F_Q_SVEB   436
```

```
##   5 2      N        30709
##   6 2      VEB       4000
##   7 3      F_Q_SVEB   436
##   8 3      N        30709
##   9 3      VEB       4000
## 10 4      F_Q_SVEB   436
## 11 4      N        30709
## 12 4      VEB       4000
## 13 5      F_Q_SVEB   437
## 14 5      N        30710
## 15 5      VEB       4000
```

We created fold for 5-fold cross validation with somewhat equal distributions of target variable in each fold.

```r
ctree_model_fn <- function(train_data, formula, hyperparameters){
   partykit::ctree(formula = as.formula(formula),
         data = train_data)
}

ctree_predict_fn <- function(test_data, model, formula, hyperparameters, train_data){
   stats::predict(object = model,
           newdata = test_data,
           allow.new.levels = TRUE,
           type = "prob")
}

formula <- paste("type ~ ", paste(colnames(data %>% select(-type)), collapse= "+"))

ctree.fit <- cross_validate_fn(
   data = data_cv,
   formulas = formula,
   type = "multinomial",
   model_fn = ctree_model_fn,
   predict_fn = ctree_predict_fn,
   fold_cols = ".folds",
   parallel = TRUE
)
```

```r
confusion_matrix_all <- ctree.fit$`Confusion Matrix`[[1]]

confusion_matrix <- confusion_matrix_all %>%
   select(-`Fold Column`) %>%
   mutate(N_perc = round(N / sum (N) * 100, 2)) %>%
   select(-N) %>%
   tidyr::pivot_wider(names_from = Target, values_from = N_perc) %>%
   remove_rownames %>%
   tibble::column_to_rownames(var = "Prediction")

class0 <- round(confusion_matrix[1, 1] / sum(confusion_matrix[, 1]) * 100, 1)
class1 <- round(confusion_matrix[2, 2] / sum(confusion_matrix[, 2]) * 100, 1)
class2 <- round(confusion_matrix[3, 3] / sum(confusion_matrix[, 3]) * 100, 1)
class_acc <- c(class0, class1, class2 )
names(class_acc) <- colnames(confusion_matrix)
```

```r
overall_acc <- sum(diag(as.matrix(confusion_matrix)))

overall_acc
```

```
## [1] 99.16
```

```r
class_acc
```

```
## F_Q_SVEB        N      VEB
##     79.7     99.7     97.5
```

Our conditional inference tree gave us overall accuracy of 99.15%, while the accuracies of each class are: F_Q_SVEB - 78.4%, N - 99.7%, VEB - 97.4%.

7 Comparison of models

```r
resamps = resamples(list(LDA = lda.fit, RDA = rda.fit, CART = cart.fit))

accuracies <- data.frame() %>%
   bind_rows(
      classAcc(confusionMatrix(lda.fit)),
      classAcc(confusionMatrix(resamp$pred, resamp$obs)),
      classAcc(confusionMatrix(cart.fit)),
      class_acc
             ) %>%
   bind_cols(
      model = c("LDA", "RDA", "CART", "Conditional Inference Tree"),
      overall_accuracy = round(c(
         sum(diag(as.matrix(confusionMatrix(lda.fit)$table))),
         sum(diag(as.matrix(confusionMatrix(resamp$pred, resamp$obs)$table))) / nrow(data) * 100,
         sum(diag(as.matrix(confusionMatrix(cart.fit)$table))),
         overall_acc
      ), 2)
      ) %>%
   tibble::column_to_rownames(var = "model")


summary(resamps)
```

```
##
## Call:
## summary.resamples(object = resamps)
##
## Models: LDA, RDA, CART
## Number of resamples: 5
##
## Accuracy
##          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## LDA  0.9676777 0.9678484 0.9681311 0.9682119 0.9685304 0.9688718    0
```

```
## RDA  0.9616162 0.9619292 0.9621568 0.9626066 0.9636648 0.9636659    0
## CART 0.9572059 0.9573481 0.9580038 0.9594482 0.9618438 0.9628397    0
##
## AUC
##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## LDA  0.9635871 0.9636174 0.9668114 0.9660774 0.9673088 0.9690626    0
## RDA  0.9645315 0.9684363 0.9694333 0.9686474 0.9698999 0.9709360    0
## CART 0.8077285 0.8182562 0.8186461 0.8196115 0.8199856 0.8334413    0
##
## Kappa
##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## LDA  0.8451663 0.8456669 0.8472608 0.8479544 0.8498086 0.8518692    0
## RDA  0.8170104 0.8174471 0.8192346 0.8212942 0.8263744 0.8264045    0
## CART 0.7847392 0.7852633 0.7904257 0.7983500 0.8124632 0.8188587    0
##
## logLoss
##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## LDA  0.1512249 0.1562215 0.1564785 0.1569349 0.1600049 0.1607446    0
## RDA  0.1526979 0.1527323 0.1568849 0.1574844 0.1577424 0.1673643    0
## CART 0.1716706 0.1779383 0.1860728 0.1825564 0.1871610 0.1899393    0
##
## Mean_Balanced_Accuracy
##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## LDA  0.8656041 0.8670207 0.8692831 0.8682400 0.8695879 0.8697044    0
## RDA  0.8533618 0.8568336 0.8570949 0.8579671 0.8603835 0.8621618    0
## CART 0.7457662 0.7473393 0.7503022 0.7540351 0.7612619 0.7655060    0
##
## Mean_Detection_Rate
##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## LDA  0.3225592 0.3226161 0.3227104 0.3227373 0.3228435 0.3229573    0
## RDA  0.3205387 0.3206431 0.3207189 0.3208689 0.3212216 0.3212220    0
## CART 0.3190686 0.3191160 0.3193346 0.3198161 0.3206146 0.3209466    0
##
## Mean_F1
##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## LDA  0.8066834 0.8069488 0.8076621 0.8095788 0.8123586 0.8142412    0
## RDA  0.7891354 0.7916615 0.7964622 0.7957094 0.7973791 0.8039086    0
## CART        NA        NA        NA       NaN        NA        NA    5
##
## Mean_Neg_Pred_Value
##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## LDA  0.9778328 0.9778440 0.9778449 0.9781875 0.9780785 0.9793374    0
## RDA  0.9670877 0.9677082 0.9687852 0.9688488 0.9703084 0.9703545    0
## CART 0.9712582 0.9714106 0.9728544 0.9723515 0.9729029 0.9733313    0
##
## Mean_Pos_Pred_Value
##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## LDA  0.8211892 0.8218708 0.8268752 0.8271019 0.8326699 0.8329045    0
## RDA  0.8059583 0.8067503 0.8111167 0.8119861 0.8153332 0.8207721    0
## CART        NA        NA        NA       NaN        NA        NA    5
##
## Mean_Precision
##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## LDA  0.8211892 0.8218708 0.8268752 0.8271019 0.8326699 0.8329045    0
```

```
## RDA  0.8059583 0.8067503 0.8111167 0.8119861 0.8153332 0.8207721    0
## CART        NA        NA        NA       NaN        NA        NA    5
##
## Mean_Recall
##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## LDA  0.7957643 0.7997502 0.7998441 0.7998741 0.8006556 0.8033562    0
## RDA  0.7798595 0.7836094 0.7847503 0.7869854 0.7926574 0.7940505    0
## CART 0.5872881 0.5886834 0.5913900 0.5953798 0.6029916 0.6065460    0
##
## Mean_Sensitivity
##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## LDA  0.7957643 0.7997502 0.7998441 0.7998741 0.8006556 0.8033562    0
## RDA  0.7798595 0.7836094 0.7847503 0.7869854 0.7926574 0.7940505    0
## CART 0.5872881 0.5886834 0.5913900 0.5953798 0.6029916 0.6065460    0
##
## Mean_Specificity
##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## LDA  0.9342911 0.9352099 0.9354438 0.9366059 0.9387532 0.9393316    0
## RDA  0.9268641 0.9281097 0.9289170 0.9289489 0.9302731 0.9305804    0
## CART 0.9042443 0.9059951 0.9092144 0.9126904 0.9195322 0.9244660    0
##
## prAUC
##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## LDA  0.8429418 0.8447308 0.8498277 0.8503791 0.8547689 0.8596263    0
## RDA  0.8536096 0.8539713 0.8543330 0.8543330 0.8546947 0.8550564    3
## CART 0.3038983 0.3045366 0.3070263 0.3064407 0.3071344 0.3096082    0
```

```
accuracies
```

```
##                            F_Q_SVEB    N  VEB overall_accuracy
## LDA                            60.8 99.6 79.6            96.82
## RDA                            60.0 97.8 80.8            95.42
## CART                            0.0 99.5 79.1            95.94
## Conditional Inference Tree     79.7 99.7 97.5            99.16
```

Conditional Inference Tree gave us the best result, the overall accuracy reaches 99.1%, while the accuracies for each class are: F_Q_SVEB = 78.4%, N = 99.7% and VEB = 97.4%.

Both LDA and RDA models gave very similar results, resulting in ~60% accuracy for F_Q_SVEB, ~99% for N and ~80% for VEB classes.

Decision tree performed worse, the model could not predict F_Q_SVEB class at all. Seems like the decision tree was too simple to capture our new class.

8 References

https://cran.r-project.org/web/packages/cvms/vignettes/cross_validating_custom_functions.html https://cran.r-project.org/web/packages/cvms/readme/README.html#main-functions https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4980381/