Ćwiczenia 9

2024-01-08

Zadanie 1

```
df <- MASS::painters

model_lda <- train(
  School ~ .,
  data = df,
  method="lda",
  trControl = trainControl(method = "LOOCV", search = "grid")
)
predict_class <- predict(model_lda)
```

Błąd resubstytucji

```
mean(substitution_error <- predict_class == df$School)
```

```
## [1] 0.5555556
```

Błąd cv:

```
model_lda$results$Accuracy
```

```
## [1] 0.3333333
```

```
confusionMatrix(
  data = predict_class,
  reference = df$School
)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction A B C D E F G H
##          A 5 4 0 0 0 1 1 0
##          B 0 1 2 0 0 0 0 0
##          C 1 1 2 0 0 0 0 1
##          D 2 0 0 9 1 0 1 0
##          E 0 0 2 0 4 0 1 0
##          F 0 0 0 0 0 2 0 0
##          G 0 0 0 1 1 1 4 0
##          H 2 0 0 0 1 0 0 3
##
## Overall Statistics
##
##                Accuracy : 0.5556
##                  95% CI : (0.414, 0.6908)
##     No Information Rate : 0.1852
##     P-Value [Acc > NIR] : 1.328e-09
```

```
##
##                   Kappa : 0.4812
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E Class: F
## Sensitivity           0.50000  0.16667  0.33333   0.9000  0.57143  0.50000
## Specificity           0.86364  0.95833  0.93750   0.9091  0.93617  1.00000
## Pos Pred Value        0.45455  0.33333  0.40000   0.6923  0.57143  1.00000
## Neg Pred Value        0.88372  0.90196  0.91837   0.9756  0.93617  0.96154
## Prevalence            0.18519  0.11111  0.11111   0.1852  0.12963  0.07407
## Detection Rate        0.09259  0.01852  0.03704   0.1667  0.07407  0.03704
## Detection Prevalence  0.20370  0.05556  0.09259   0.2407  0.12963  0.03704
## Balanced Accuracy     0.68182  0.56250  0.63542   0.9045  0.75380  0.75000
##                      Class: G Class: H
## Sensitivity           0.57143  0.75000
## Specificity           0.93617  0.94000
## Pos Pred Value        0.57143  0.50000
## Neg Pred Value        0.93617  0.97917
## Prevalence            0.12963  0.07407
## Detection Rate        0.07407  0.05556
## Detection Prevalence  0.12963  0.11111
## Balanced Accuracy     0.75380  0.84500
```

Zadanie 2

```
df <- DAAG::leafshape
model_lda <- train(
  location ~ . - arch - latitude,
  data = df,
  method = "lda",
  trControl = trainControl(method = "LOOCV", search = "grid")
)
model_qda <- train(
  location ~ . - arch - latitude,
  data = df,
  method = "qda",
  trControl = trainControl(method = "LOOCV", search = "grid")
)
predict_class_qda <- predict(model_qda)
predict_class_lda <- predict(model_lda)
```

```
model_qda$results
```

```
##   parameter  Accuracy     Kappa
## 1      none 0.3461538 0.1543278
```

```
model_lda$results
```

```
##   parameter  Accuracy     Kappa
## 1      none 0.3601399 0.1644769
```

```
confusionMatrix(
```

```
  data = predict_class_qda,
  reference = df$location
)
```

```
## Confusion Matrix and Statistics
##
##               Reference
## Prediction     Sabah Panama Costa Rica N Queensland S Queensland Tasmania
##   Sabah           51     23         23           12            1        0
##   Panama           0      3          1            1            1        0
##   Costa Rica       4      4          8            3            1        0
##   N Queensland    24     25         18           44           21        0
##   S Queensland     1      0          0            0            5        0
##   Tasmania         0      0          0            1            2        9
##
## Overall Statistics
##
##                Accuracy : 0.4196
##                  95% CI : (0.3617, 0.4791)
##     No Information Rate : 0.2797
##     P-Value [Acc > NIR] : 2.724e-07
##
##                   Kappa : 0.2502
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: Sabah Class: Panama Class: Costa Rica
## Sensitivity                0.6375       0.05455           0.16000
## Specificity                0.7136       0.98701           0.94915
## Pos Pred Value             0.4636       0.50000           0.40000
## Neg Pred Value             0.8352       0.81429           0.84211
## Prevalence                 0.2797       0.19231           0.17483
## Detection Rate             0.1783       0.01049           0.02797
## Detection Prevalence       0.3846       0.02098           0.06993
## Balanced Accuracy          0.6755       0.52078           0.55458
##                      Class: N Queensland Class: S Queensland Class: Tasmania
## Sensitivity                       0.7213             0.16129         1.00000
## Specificity                       0.6089             0.99608         0.98917
## Pos Pred Value                    0.3333             0.83333         0.75000
## Neg Pred Value                    0.8896             0.90714         1.00000
## Prevalence                        0.2133             0.10839         0.03147
## Detection Rate                    0.1538             0.01748         0.03147
## Detection Prevalence              0.4615             0.02098         0.04196
## Balanced Accuracy                 0.6651             0.57868         0.99458
```

```
confusionMatrix(
  data = predict_class_lda,
  reference = df$location
)
```

```
## Confusion Matrix and Statistics
##
##               Reference
```

```
## Prediction     Sabah Panama Costa Rica N Queensland S Queensland Tasmania
##   Sabah           64     33         30           23            2        0
##   Panama           0      2          0            2            0        0
##   Costa Rica       5      4         10            1            3        0
##   N Queensland    10     15          9           31           18        2
##   S Queensland     0      1          1            4            5        2
##   Tasmania         1      0          0            0            3        5
##
## Overall Statistics
##
##                Accuracy : 0.4091
##                  95% CI : (0.3516, 0.4685)
##     No Information Rate : 0.2797
##     P-Value [Acc > NIR] : 1.734e-06
##
##                   Kappa : 0.2279
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: Sabah Class: Panama Class: Costa Rica
## Sensitivity                0.8000      0.036364           0.20000
## Specificity                0.5728      0.991342           0.94492
## Pos Pred Value             0.4211      0.500000           0.43478
## Neg Pred Value             0.8806      0.812057           0.84791
## Prevalence                 0.2797      0.192308           0.17483
## Detection Rate             0.2238      0.006993           0.03497
## Detection Prevalence       0.5315      0.013986           0.08042
## Balanced Accuracy          0.6864      0.513853           0.57246
##                      Class: N Queensland Class: S Queensland Class: Tasmania
## Sensitivity                       0.5082             0.16129         0.55556
## Specificity                       0.7600             0.96863         0.98556
## Pos Pred Value                    0.3647             0.38462         0.55556
## Neg Pred Value                    0.8507             0.90476         0.98556
## Prevalence                        0.2133             0.10839         0.03147
## Detection Rate                    0.1084             0.01748         0.01748
## Detection Prevalence              0.2972             0.04545         0.03147
## Balanced Accuracy                 0.6341             0.56496         0.77056
```

Zadanie 3

```
df <- tibble(
  group = rep(1:3, each = 5) |> factor(),
  depression = c(
      6,  4, 0, 4,  0,
     11, 11, 5, 8,  4,
     12,  8, 9, 8, 11
  ),
  anxiety = c(
     8, 3, 2,  1, 8,
     9, 6, 7,  6, 9,
     11, 8, 6, 10, 4
  ),
```

```
  social_unrest = c(
    9, 3, 8, 6, 4,
    8, 6, 4, 5, 4,
    6, 5, 7, 8, 3
  )
)
```

```
model_lda <- train(
  group ~ .,
  data = df,
  method = "lda",
  trControl = trainControl(method = "LOOCV", search = "grid")
)
model_qda <- train(
  group ~ .,
  data = df,
  method = "qda",
  trControl = trainControl(method = "LOOCV", search = "grid")
)
model_naive_bayes <- train(
  group ~ .,
  data = df,
  method = "naive_bayes",
  trControl = trainControl(method = "LOOCV", search = "grid")
)
```

```
cbind(rbind(
  model_lda$results[, -1],
  model_qda$results[, -1],
  model_naive_bayes$results[1:2 ,4:5]
), classificators = c("lda", "qda", "naive_bayes", "kernel_bayes"))
```

```
##    Accuracy Kappa classificators
## 1 0.3333333   0.0            lda
## 2 0.5333333   0.3            qda
## 3 0.4000000   0.1    naive_bayes
## 4 0.4666667   0.2   kernel_bayes
```

Zadanie 4

```
model_1nn <- train(
  chd ~ .,
  data = df,
  method = "knn",
  trControl = trainControl(method = "boot", number = 100),
  tuneGrid = data.frame(k = 1)
)
model_rf <- train(
  chd ~ .,
  data = df,
  method = "ranger",
  trControl = trainControl(method = "boot", number = 100),
  tuneGrid = expand.grid(
    mtry = 2,
```

```
    splitrule = c("gini", "extratrees"),
    min.node.size = 1:4
  )
)
```

Bootstrap error

```
rbind(
  cbind(method = "1-nn", mtry = NA, splitrule = NA, min.node.size = NA,
        model_1nn$results[, -1], resub = mean(predict(model_1nn, df) != df$chd)),
  cbind(method = "Random forest", model_rf$results, resub = mean(predict(model_rf, df) != df$chd))
)
```

```
##          method mtry  splitrule min.node.size  Accuracy      Kappa AccuracySD
## 1          1-nn   NA       <NA>            NA 0.5839750 0.05771001 0.02823736
## 2 Random forest    2       gini             1 0.6864488 0.26274514 0.02989525
## 3 Random forest    2       gini             2 0.6868168 0.26361230 0.02844939
## 4 Random forest    2       gini             3 0.6872906 0.26613087 0.02679312
## 5 Random forest    2       gini             4 0.6871738 0.26559091 0.02816378
## 6 Random forest    2 extratrees             1 0.7003878 0.28450798 0.02820179
## 7 Random forest    2 extratrees             2 0.7016066 0.28765579 0.02909324
## 8 Random forest    2 extratrees             3 0.7005833 0.28538238 0.02726898
## 9 Random forest    2 extratrees             4 0.7033798 0.29179598 0.02814672
##      KappaSD       resub
## 1 0.06020750 0.000000000
## 2 0.06326014 0.008658009
## 3 0.06320679 0.008658009
## 4 0.06003348 0.008658009
## 5 0.06107560 0.008658009
## 6 0.06442685 0.008658009
## 7 0.06393261 0.008658009
## 8 0.05969748 0.008658009
## 9 0.06406198 0.008658009
```

```
confusionMatrix(model_rf)
```

```
## Bootstrapped (100 reps) Confusion Matrix
##
## (entries are percentual average cell counts across resamples)
##
##           Reference
## Prediction    0    1
##          0 56.2 20.8
##          1  8.9 14.1
##
##  Accuracy (average) : 0.7033
```

```
confusionMatrix(predict(model_rf, df), df$chd)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 302   4
##          1   0 156
##
```

```
##                Accuracy : 0.9913
##                  95% CI : (0.978, 0.9976)
##     No Information Rate : 0.6537
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9808
##
##  Mcnemar's Test P-Value : 0.1336
##
##             Sensitivity : 1.0000
##             Specificity : 0.9750
##          Pos Pred Value : 0.9869
##          Neg Pred Value : 1.0000
##              Prevalence : 0.6537
##          Detection Rate : 0.6537
##    Detection Prevalence : 0.6623
##       Balanced Accuracy : 0.9875
##
##        'Positive' Class : 0
##
```