Détectez des faux billets

SOMMAIRE

- Introduction
- Analyses univariées et bivariées
- PCA
- Apprentissage non-supervisé
- Apprentissage supervisé
- Bilan

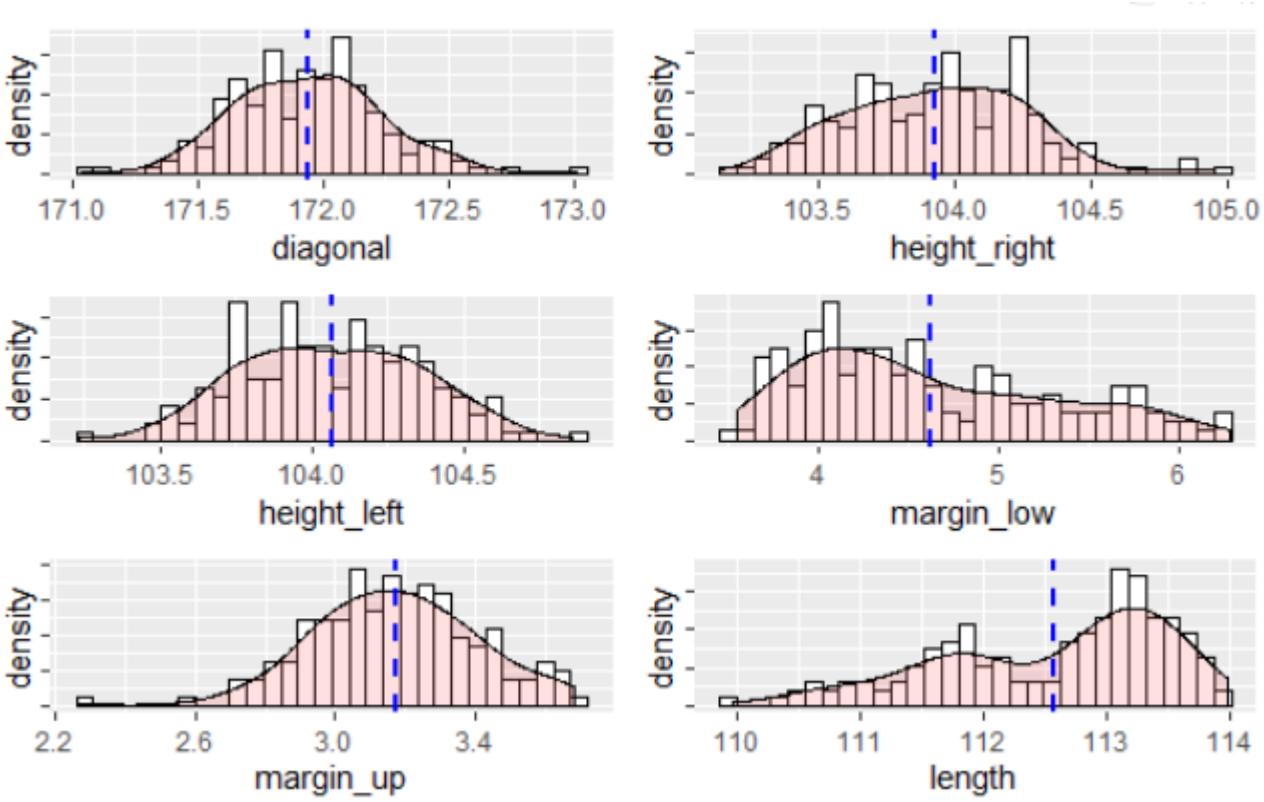
Introduction

Objectif : Créer un algorithme de vérification des billets performant

Taux de réussite > 99%

Analyse univariée

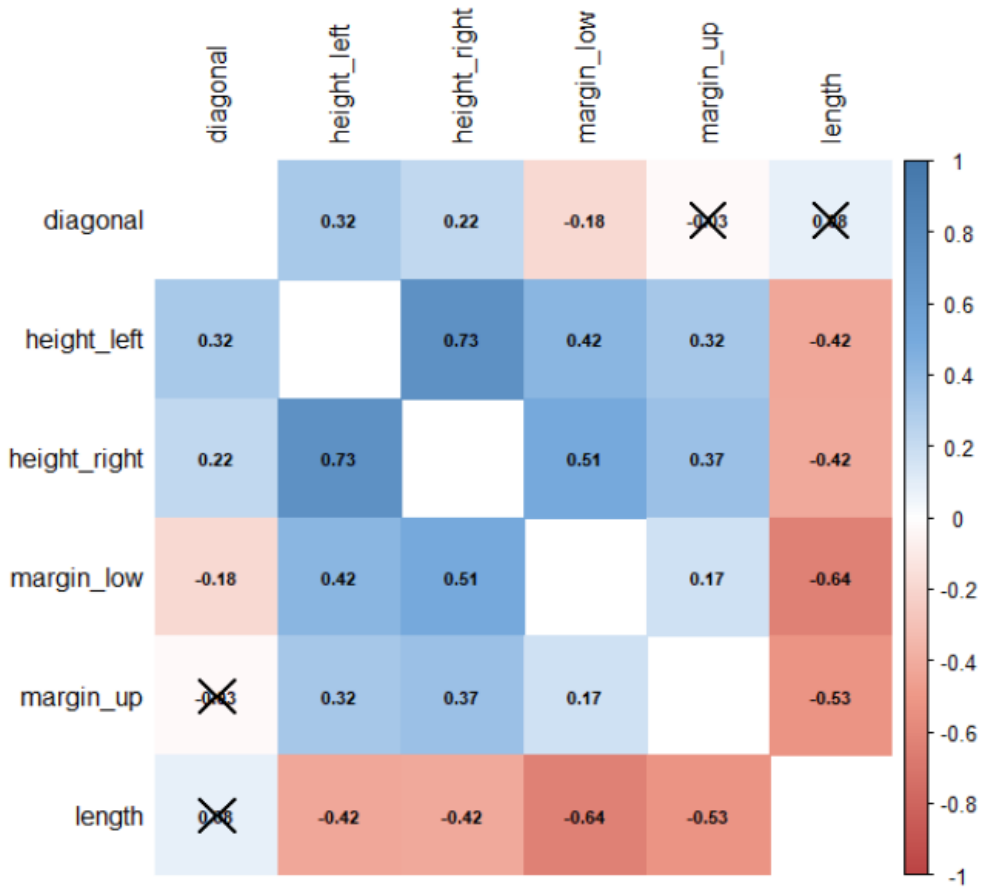
Courbes de densité



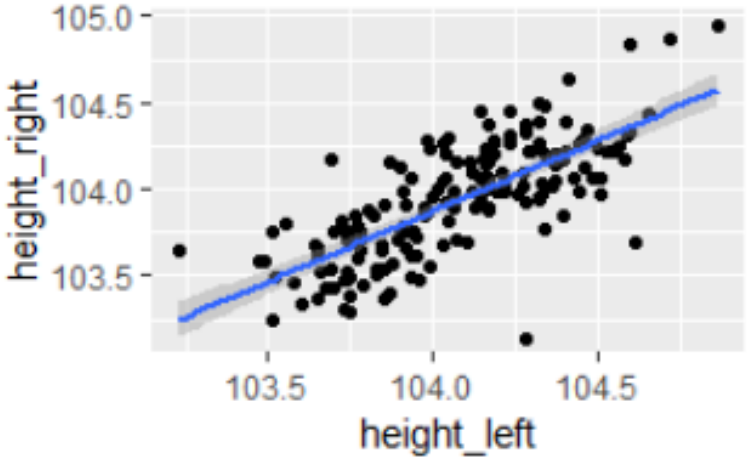
Shapiro test

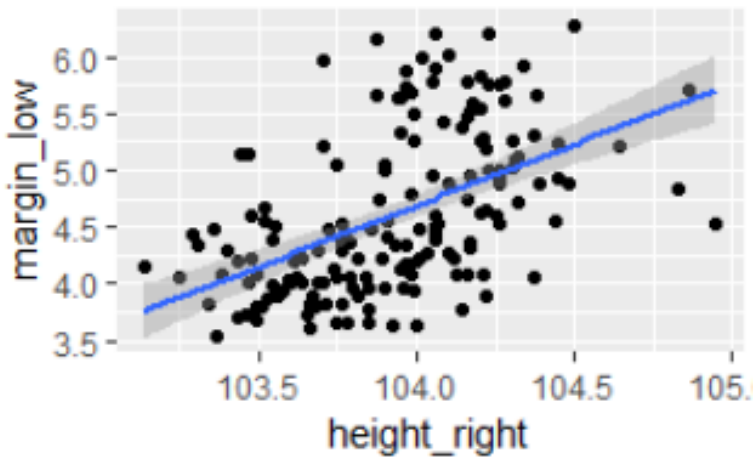
```
        Shapiro-Wilk normality test

data:  df$diagonal
W = 0.99318, p-value = 0.6107


        Shapiro-Wilk normality test

data:  df$height_left
W = 0.99272, p-value = 0.5533


        Shapiro-Wilk normality test

data:  df$height_right
W = 0.98812, p-value = 0.1625


        Shapiro-Wilk normality test

data:  df$margin_low
W = 0.9354, p-value = 6.226e-07


        Shapiro-Wilk normality test

data:  df$margin_up
W = 0.98892, p-value = 0.2044


        Shapiro-Wilk normality test

data:  df$length
W = 0.93246, p-value = 3.715e-07
```

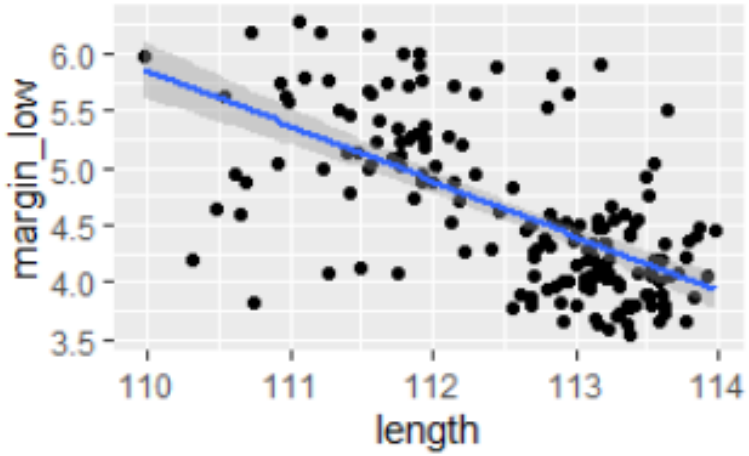
Corrélations

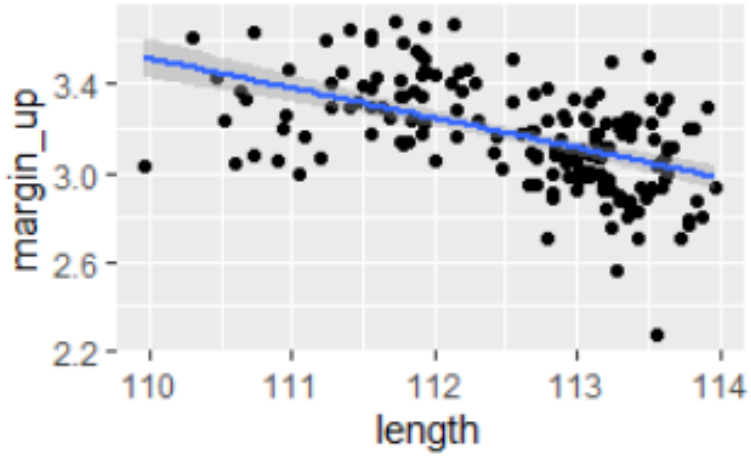


Analyse bivariée



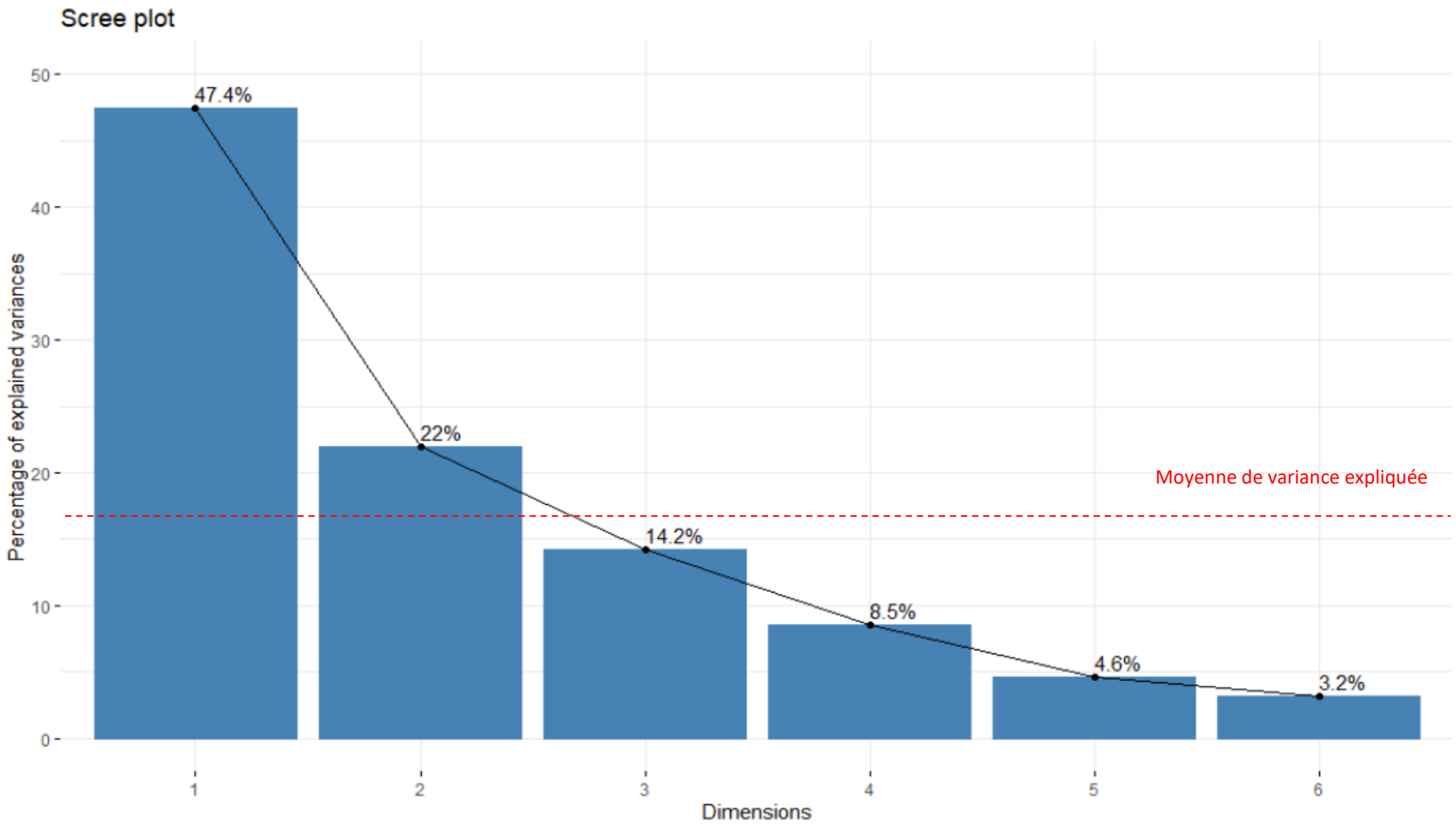




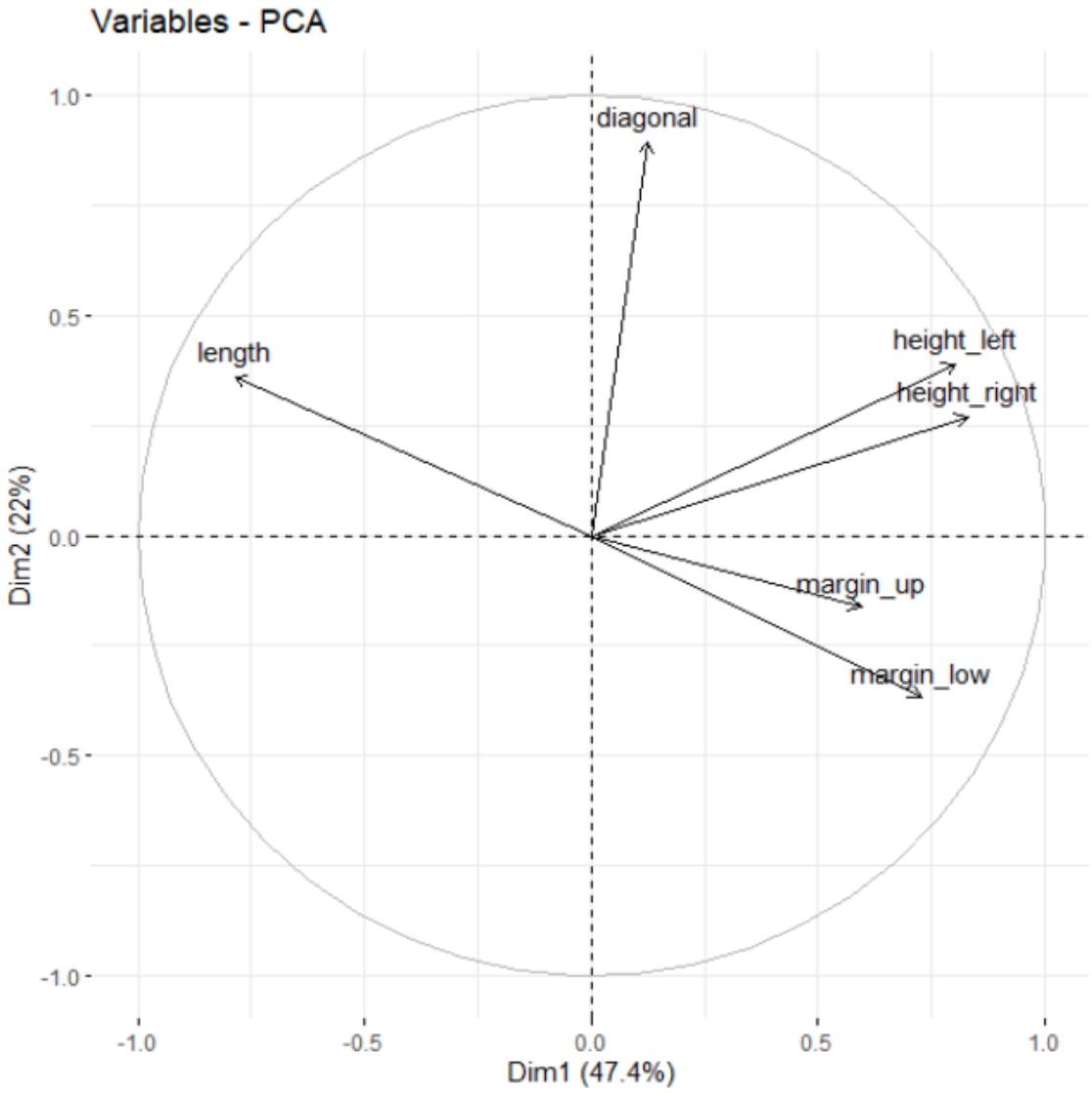


PCA

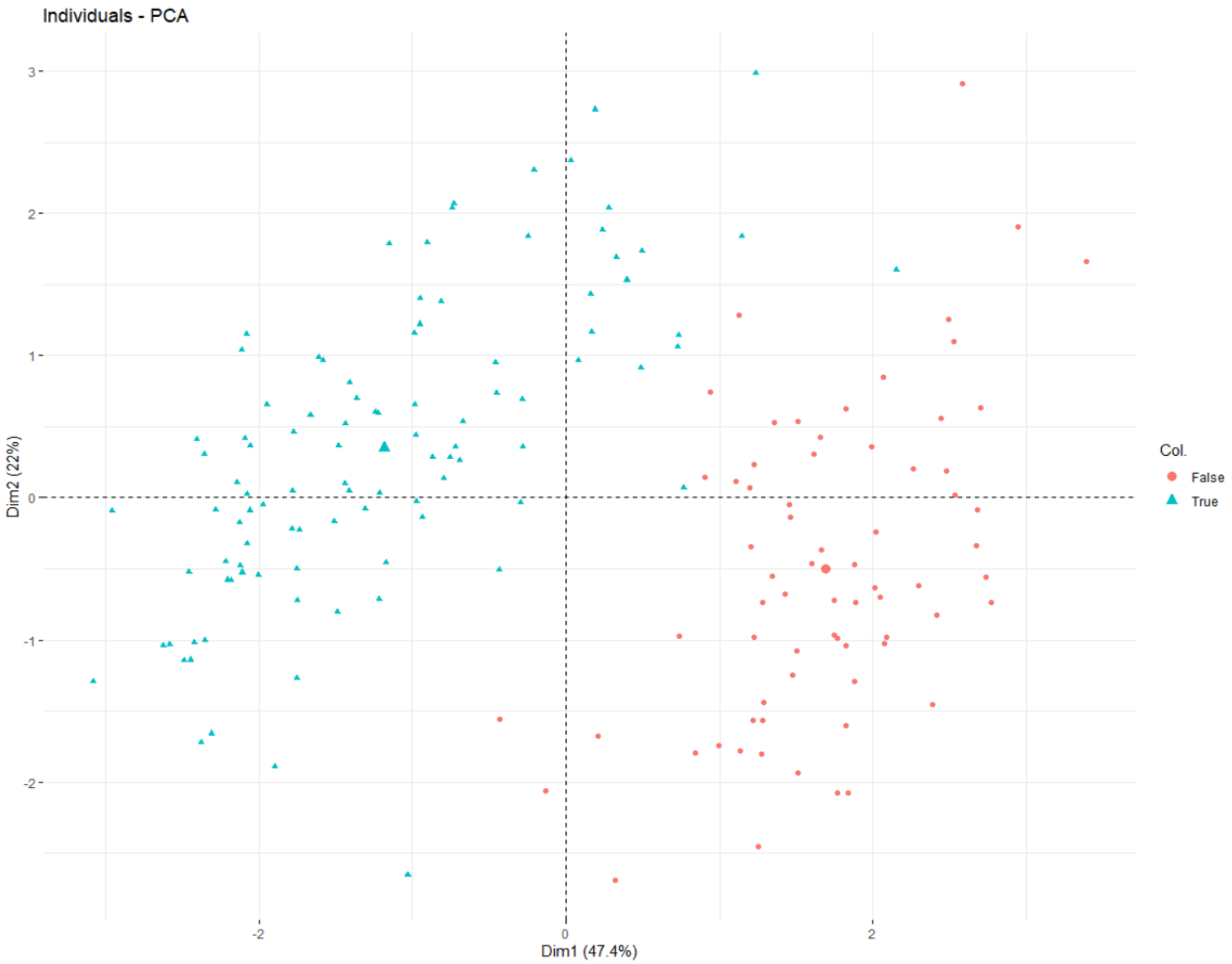
Éblouis des valeurs propres



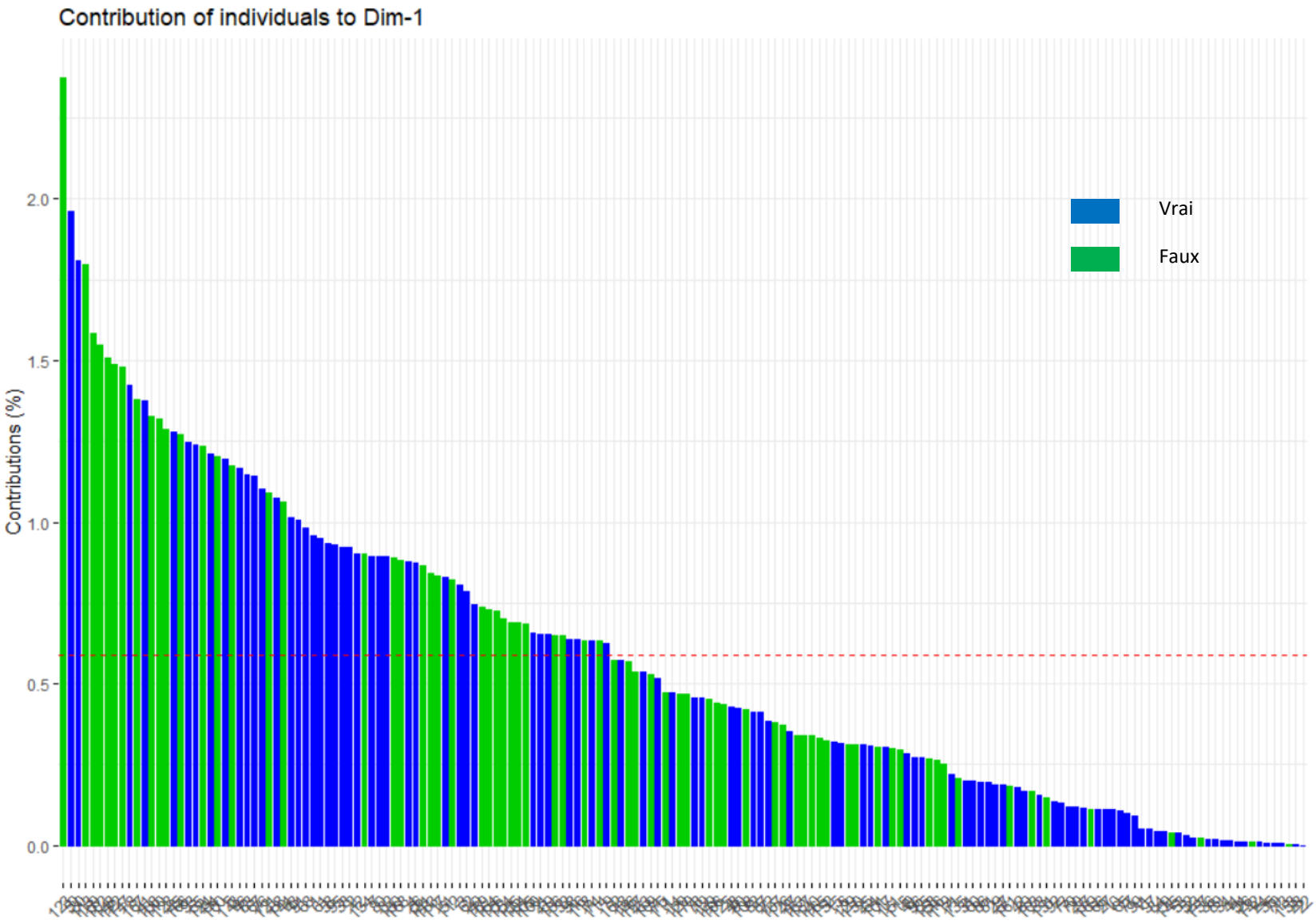
Cercle de corrélation



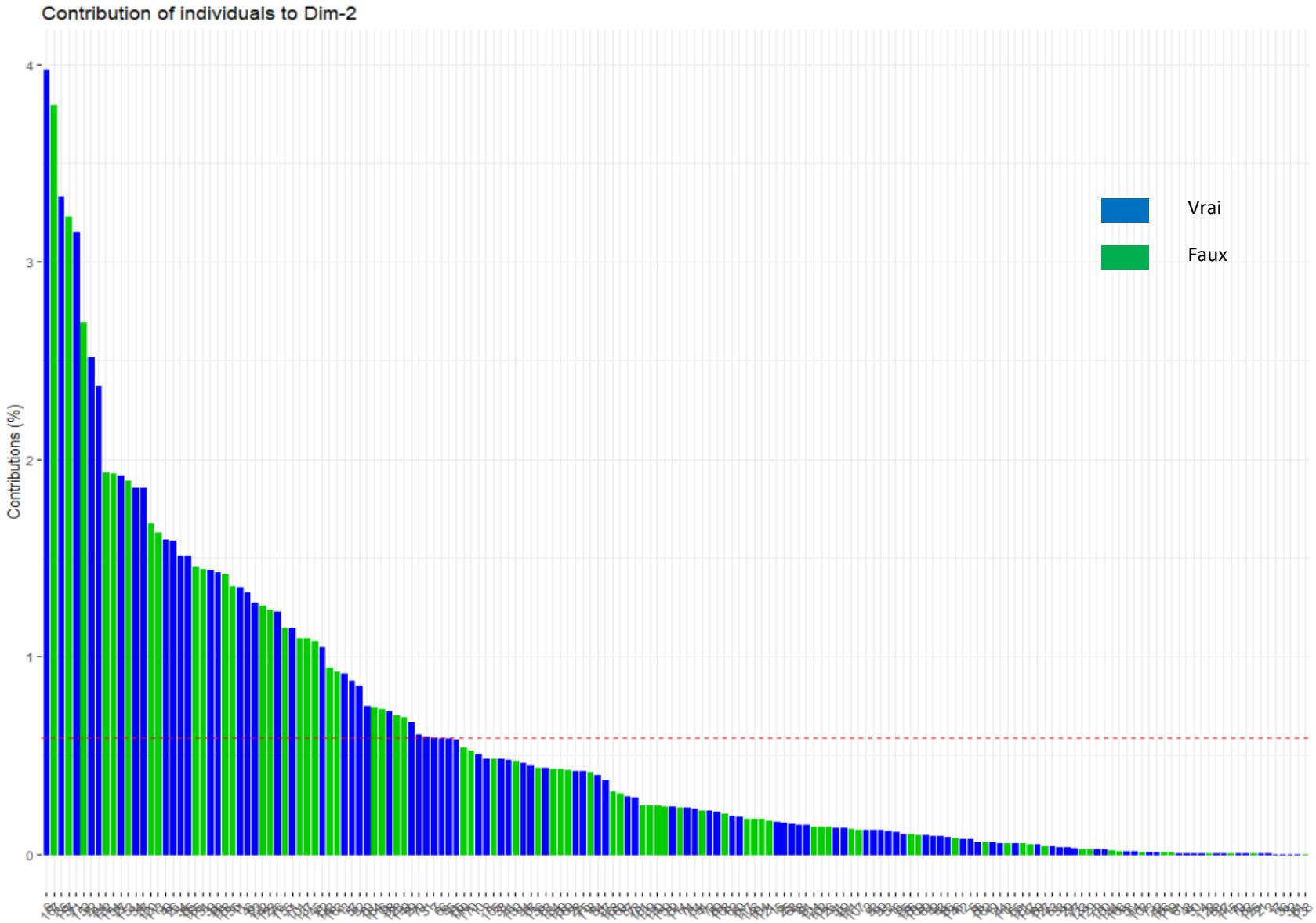
Analyse du premier plan factoriel



Contribution des individus

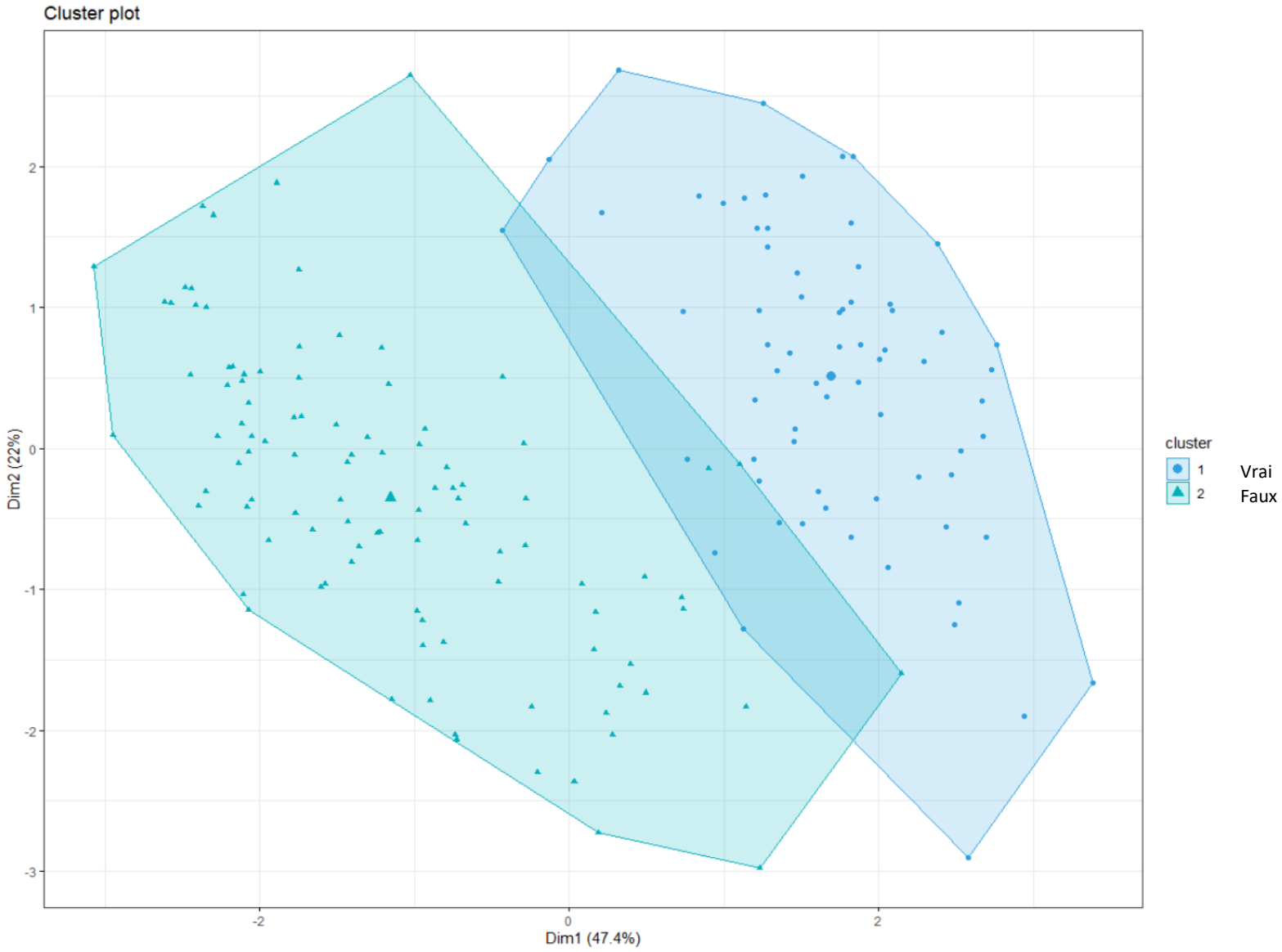


Contribution des individus



Apprentissage non-supervisé

K-means


Cluster plot
Dim2 (22%)
Dim1 (47.4%)
2
1
0
-1
-2
-3
-2
0
2
cluster
1
2
Vrai
Faux

Analyse de la classification

```
###Je calcul la réussite de cette classification
mean(res.km$cluster == df$is_genuine%>%as.numeric())
```

```
[1] 0.9823529
```

Taux de réussite : 98,2 %

167 billets sur 170 ont été classé comme on le souhaite.

Apprentissage supervisé

Régression logistique

1) La longueur du billet

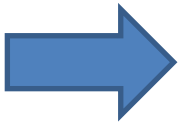
2) La hauteur du billet mesurée sur le côté gauche

3) La hauteur du billet mesurée sur le côté droit

4) La marge entre le bord supérieur du billet et l'image de celui-ci

5) La marge entre le bord inférieur du billet et l'image de celui-ci

6) La diagonale du billet



97.5 %

Régression logistique

Critère d'information d'Akaike

```
Start:  AIC=14
is_genuine ~ diagonal + height_left + height_right + margin_low +
    margin_up + length

               Df Deviance    AIC
- diagonal      1    0.000 12.000
- height_right  1    0.000 12.000
- height_left   1    0.000 12.000
<none>                0.000 14.000
- margin_up     1    8.265 20.265
- length        1   11.198 23.198
- margin_low    1   42.342 54.342

Step:  AIC=12
is_genuine ~ height_left + height_right + margin_low + margin_up +
    length

               Df Deviance    AIC
- height_right  1    0.000 10.000
- height_left   1    0.000 10.000
<none>                0.000 12.000
- margin_up     1    8.568 18.568
- length        1   12.462 22.462
- margin_low    1   47.782 57.782

Step:  AIC=10
is_genuine ~ height_left + margin_low + margin_up + length

              Df Deviance    AIC
- height_left  1    0.000  8.000
<none>               0.000 10.000
- margin_up    1    8.585 16.585
- length       1   12.716 20.716
- margin_low   1   53.624 61.624

Step:  AIC=8
is_genuine ~ margin_low + margin_up + length

             Df Deviance    AIC
<none>            0.000  8.000
- margin_up   1    8.586 14.586
- length      1   12.721 18.721
- margin_low  1   57.812 63.812
```

Régression logistique sur les variables choisis

1) La longueur du billet

2) La marge entre le bord supérieur du billet et l'image de celui-ci

3) La marge entre le bord inférieur du billet et l'image de celui-ci

98.4 %

Régression logistique

Analyse discriminante linéaire

1) La longueur du billet

2) La hauteur du billet mesurée sur le côté gauche

3) La hauteur du billet mesurée sur le côté droit

4) La marge entre le bord supérieur du billet et l'image de celui-ci

5) La marge entre le bord inférieur du billet et l'image de celui-ci

6) La diagonale du billet

99.3 %

```
Coefficients of linear discriminants:
                    LD1
diagonal      -0.07883636
height_left    0.06650273
height_right  -0.13381362
margin_low    -1.65636378
margin_up     -1.07126350
length         0.98662629
```

Analyse discriminante linéaire

Conclusion

Au vue de mon échantillon de billet,

la meilleur méthode parmi celles que

j'ai testé est l'Analyse discriminante linéaire.

Elle atteint bien l'objectif que je m'étais ciblé,

avoir un taux de réussite supérieur à 99%

Bilan

Ce projet m’a permis de me familiariser avec les différents types d’apprentissages supervisés et non supervisés et il m’a aussi donné l’occasion de créer un programme.

Le projet demandait d’utiliser seulement la régression logistique et je trouve ça dommage..

Slides additionnelles

Matrice de confusion kmeans

```
Confusion Matrix and Statistics

          Reference
Prediction FALSE TRUE
     FALSE    68    1
     TRUE      2   99

               Accuracy : 0.9824
                 95% CI : (0.9493, 0.9963)
    No Information Rate : 0.5882
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9635

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9900
            Specificity : 0.9714
         Pos Pred Value : 0.9802
         Neg Pred Value : 0.9855
             Prevalence : 0.5882
         Detection Rate : 0.5824
   Detection Prevalence : 0.5941
      Balanced Accuracy : 0.9807

       'Positive' Class : TRUE
```

Matrice de confusion (cross validation)

1000 cross validation de proportion :
0.7 (train) / 0.3 (test)

Pour chaque cross-validation :
- 30 vrais billets
- 21 faux billets

Ce qui donne en tout :
- 30000 vrais billets
- 21000 faux billets

Matrice de confusion GLM (6 variables)

```
Confusion Matrix and Statistics

          Reference
Prediction FALSE  TRUE
     FALSE 20386   614
     TRUE    614 29386

               Accuracy : 0.9759
                 95% CI : (0.9746, 0.9772)
    No Information Rate : 0.5882
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9503

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9795
            Specificity : 0.9708
         Pos Pred Value : 0.9795
         Neg Pred Value : 0.9708
             Prevalence : 0.5882
         Detection Rate : 0.5762
   Detection Prevalence : 0.5882
      Balanced Accuracy : 0.9751

       'Positive' Class : TRUE
```

Matrice de confusion GLM (3 variables)

```
Confusion Matrix and Statistics

          Reference
Prediction FALSE  TRUE
     FALSE 20418   288
     TRUE    582 29712

               Accuracy : 0.9829
                 95% CI : (0.9818, 0.984)
    No Information Rate : 0.5882
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9647

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9904
            Specificity : 0.9723
         Pos Pred Value : 0.9808
         Neg Pred Value : 0.9861
             Prevalence : 0.5882
         Detection Rate : 0.5826
   Detection Prevalence : 0.5940
      Balanced Accuracy : 0.9813

       'Positive' Class : TRUE
```

Matrice de confusion LDA

```
Confusion Matrix and Statistics

          Reference
Prediction FALSE  TRUE
     FALSE 20842   181
     TRUE    158 29819

               Accuracy : 0.9934
                 95% CI : (0.9926, 0.994)
    No Information Rate : 0.5882
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9863

 Mcnemar's Test P-Value : 0.2321

            Sensitivity : 0.9940
            Specificity : 0.9925
         Pos Pred Value : 0.9947
         Neg Pred Value : 0.9914
             Prevalence : 0.5882
         Detection Rate : 0.5847
   Detection Prevalence : 0.5878
      Balanced Accuracy : 0.9932

       'Positive' Class : TRUE
```