MRR : Projet

Lina EL HADDAJ & Ania POLIDORI

2023-12-10

```r
data <- read.csv(file = "data.csv", sep = ';', header = TRUE)

# remise au bon format des variables à facteurs
data <- data %>% mutate_at(vars(Marital.status,
                                Application.mode,
                                Application.order,
                                Course,
                                Daytime.evening.attendance.,
                                Previous.qualification,
                                Nacionality,
                                Mother.s.qualification,
                                Father.s.qualification,
                                Mother.s.occupation,
                                Father.s.occupation,
                                Displaced,
                                Educational.special.needs,
                                Debtor,
                                Tuition.fees.up.to.date,
                                Gender,
                                Scholarship.holder,
                                International,
                                Target
                                ), as.factor)

# changement des noms des facteurs du genre (pour que ce soit plus intuitif dans les analyses)
data$Gender <- factor(data$Gender, labels=c("F", "M"))

# création de la nouvelle variable cible
data$Ybin <- factor(ifelse(data$Target == "Graduate", 1, 0))
```

Introduction

Nous avons un jeu de données composé d'**étudiants qui suivent différents cursus universitaires**.

Dans celui-ci, de nombreuses informations sur l'étudiant au moment de son arrivée dans son cursus ainsi que des sur ses performances académiques aux premier et second semestres sont connues.

Grâce à toutes ces informations récoltées, nous allons tenter de répondre à la problématique suivante :

Comment utiliser les caractéristiques démographiques, académiques et socio-économiques des étudiants pour prédire avec précision les abandons d'étudiants et les réussites académiques dans un contexte éducatif donné, à un stade tôt de leur chemin académique ?

La variable cible choisie pour notre analyse est dérivée de la variable `Target`, qui représente le résultat académique des étudiants à la fin de la durée normale du cursus. Cette variable est formulée comme une tâche de classification à trois catégories, distinguant entre les étudiants qui ont abandonné (`dropout`), ceux qui sont inscrits (`enrolled`), et ceux qui ont obtenu leur diplôme (`graduate`).

Afin de simplifier la tâche de prédiction, nous avons créé une variable binaire, `Ybin`, où les étudiants diplômés sont représentés par 1, et les étudiants inscrits ou ayant abandonné par 0. Cette approche nous permet de concentrer notre modèle sur la prédiction de la réussite académique à la fin de la durée normale du cursus, en distinguant de manière binaire entre les étudiants qui obtiennent leur diplôme et ceux qui n'atteignent pas cet objectif, simplifiant ainsi la complexité de la tâche de classification initiale à trois catégories.

Dans une première partie (*cf. rendu 1*), nous avons effectué une étude préliminaire de nos données. Grâce à celle-ci, nous avons pu observer de la corrélation entre certaines de nos variables explicatives, ainsi qu'un lien probable entre certaines de nos variables explicatives et notre variable cible. On va donc modéliser nos données pour tenter d'expliquer et prédire au mieux notre variable cible.

Modèle complet

Dans un premier temps, notre variable cible étant maintenant une variable binaire, nous allons effectuer un modèle logistique complet, c'est-à-dire prenant en compte toutes les variables du jeu de données. Cette modélisation nous permettra déjà de savoir si au moins une des variables est significative dans notre modèle, et donc s'il est pertinent de tenter de modéliser la variable cible avec au moins une de ces variables prédictives.

```
# on retire la variable Target pour ne pas avoir de conflit avec notre nouvelle variable cible
data <- data %>% select(-Target)

# modèle complet
mod_log <- glm(Ybin ~ . , data, family = "binomial")
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(mod_log)
```

```
##
## Call:
## glm(formula = Ybin ~ ., family = "binomial", data = data)
##
## Coefficients: (4 not defined because of singularities)
##                                                 Estimate Std. Error z value
## (Intercept)                                    8.197e+00  3.956e+03   0.002
## Marital.status2                                1.469e-01  2.644e-01   0.555
## Marital.status3                               -2.159e-01  1.528e+00  -0.141
## Marital.status4                                4.505e-01  4.319e-01   1.043
## Marital.status5                                9.153e-01  6.976e-01   1.312
## Marital.status6                               -1.995e+00  1.472e+00  -1.356
## Application.mode2                             -5.693e-01  1.920e+00  -0.297
## Application.mode5                             -1.109e+00  7.593e-01  -1.461
## Application.mode7                              1.824e-01  6.180e-01   0.295
## Application.mode10                            -1.285e+00  1.303e+00  -0.986
## Application.mode15                             1.392e+00  1.051e+00   1.324
## Application.mode16                             6.247e-02  4.902e-01   0.127
## Application.mode17                            -1.802e-01  1.365e-01  -1.320
```

```
## Application.mode18                          2.934e-01  3.577e-01   0.820
## Application.mode26                         -1.700e+01  3.956e+03  -0.004
## Application.mode27                         -1.988e+01  3.956e+03  -0.005
## Application.mode39                         -1.998e-01  2.445e-01  -0.817
## Application.mode42                          2.267e-01  4.422e-01   0.513
## Application.mode43                         -1.456e-01  2.460e-01  -0.592
## Application.mode44                          5.177e-01  6.705e-01   0.772
## Application.mode51                         -4.530e-01  4.675e-01  -0.969
## Application.mode53                         -1.111e+00  2.547e+00  -0.436
## Application.mode57                          1.747e+01  3.956e+03   0.004
## Application.order1                         -1.377e+01  3.956e+03  -0.003
## Application.order2                         -1.357e+01  3.956e+03  -0.003
## Application.order3                         -1.411e+01  3.956e+03  -0.004
## Application.order4                         -1.346e+01  3.956e+03  -0.003
## Application.order5                         -1.354e+01  3.956e+03  -0.003
## Application.order6                         -1.342e+01  3.956e+03  -0.003
## Application.order9                         -3.194e+01  5.595e+03  -0.006
## Course171                                   3.455e+00  1.854e+00   1.863
## Course8014                                  3.489e+00  1.785e+00   1.955
## Course9003                                  3.020e+00  1.788e+00   1.689
## Course9070                                  2.488e+00  1.782e+00   1.396
## Course9085                                  3.292e+00  1.784e+00   1.845
## Course9119                                  5.957e-01  1.809e+00   0.329
## Course9130                                  2.500e+00  1.798e+00   1.391
## Course9147                                  2.515e+00  1.777e+00   1.415
## Course9238                                  3.877e+00  1.781e+00   2.176
## Course9254                                  2.989e+00  1.778e+00   1.681
## Course9500                                  2.586e+00  1.778e+00   1.454
## Course9556                                  1.742e+00  1.824e+00   0.955
## Course9670                                  2.617e+00  1.777e+00   1.473
## Course9773                                  2.571e+00  1.777e+00   1.447
## Course9853                                  1.265e+00  1.782e+00   0.710
## Course9991                                  1.848e+00  1.782e+00   1.037
## Daytime.evening.attendance.1                       NA         NA      NA
## Previous.qualification2                    -6.771e-01  9.063e-01  -0.747
## Previous.qualification3                     5.292e-02  5.495e-01   0.096
## Previous.qualification4                    -8.947e-01  1.006e+00  -0.889
## Previous.qualification5                    -6.140e+00  3.956e+03  -0.002
## Previous.qualification6                    -4.175e-01  8.192e-01  -0.510
## Previous.qualification9                    -1.029e+01  1.031e+03  -0.010
## Previous.qualification10                    1.809e+00  9.654e+00   0.187
## Previous.qualification12                    1.836e-02  6.394e-01   0.029
## Previous.qualification14                   -9.422e+00  3.956e+03  -0.002
## Previous.qualification15                    4.262e+00  4.128e+00   1.032
## Previous.qualification19                    4.154e-01  3.736e-01   1.112
## Previous.qualification38                   -2.503e-01  1.322e+00  -0.189
## Previous.qualification39                   -4.573e-01  6.478e-01  -0.706
## Previous.qualification40                   -3.630e-01  6.181e-01  -0.587
## Previous.qualification42                    1.766e+00  2.504e+00   0.705
## Previous.qualification43                    5.050e+00  3.984e+00   1.267
## Previous.qualification..grade.             2.203e-03  5.051e-03   0.436
## Nacionality2                               1.445e+01  2.788e+03   0.005
## Nacionality6                               -1.352e-01  8.866e-01  -0.153
## Nacionality11                              1.598e+01  2.137e+03   0.007
```

```
## Nacionality13                          1.312e+01  3.956e+03   0.003
## Nacionality14                          1.508e+01  3.956e+03   0.004
## Nacionality17                         -1.316e+01  3.956e+03  -0.003
## Nacionality21                         -1.357e+01  3.956e+03  -0.003
## Nacionality22                          2.237e+00  1.217e+00   1.837
## Nacionality24                          5.408e+00  1.873e+00   2.888
## Nacionality25                          2.542e+00  1.232e+01   0.206
## Nacionality26                          8.249e-01  1.372e+00   0.601
## Nacionality32                          7.997e+00  4.485e+03   0.002
## Nacionality41                         -2.231e-01  5.602e-01  -0.398
## Nacionality62                          1.423e+00  8.600e+00   0.166
## Nacionality100                        -1.544e+01  1.884e+03  -0.008
## Nacionality101                        -5.827e-01  1.527e+00  -0.382
## Nacionality103                         3.664e+00  7.615e+00   0.481
## Nacionality105                        -1.809e+01  2.038e+03  -0.009
## Nacionality108                        -8.075e+00  7.912e+03  -0.001
## Nacionality109                        -2.042e+01  3.956e+03  -0.005
## Mother.s.qualification2                1.723e-01  3.689e-01   0.467
## Mother.s.qualification3                4.502e-02  2.332e-01   0.193
## Mother.s.qualification4                4.470e-01  5.304e-01   0.843
## Mother.s.qualification5                1.991e-01  7.626e-01   0.261
## Mother.s.qualification6                1.361e+01  2.417e+03   0.006
## Mother.s.qualification9                8.058e-01  1.837e+00   0.439
## Mother.s.qualification10               2.876e+00  8.635e+00   0.333
## Mother.s.qualification11              -3.909e+00  4.044e+00  -0.966
## Mother.s.qualification12              -3.376e-01  5.482e-01  -0.616
## Mother.s.qualification14               3.400e-01  6.707e+00   0.051
## Mother.s.qualification18              -1.512e+01  3.956e+03  -0.004
## Mother.s.qualification19              -8.705e-05  1.582e-01  -0.001
## Mother.s.qualification22               3.022e+01  4.510e+03   0.007
## Mother.s.qualification26               3.416e+01  4.510e+03   0.008
## Mother.s.qualification27              -1.497e+01  3.956e+03  -0.004
## Mother.s.qualification29               1.592e+01  1.687e+03   0.009
## Mother.s.qualification30               6.441e-01  2.532e+00   0.254
## Mother.s.qualification34              -3.984e-01  8.787e-01  -0.453
## Mother.s.qualification35               1.615e+01  3.956e+03   0.004
## Mother.s.qualification36               9.586e-01  2.046e+00   0.468
## Mother.s.qualification37              -6.264e-02  1.961e-01  -0.319
## Mother.s.qualification38               1.528e-01  1.962e-01   0.779
## Mother.s.qualification39              -2.326e-01  1.578e+00  -0.147
## Mother.s.qualification40              -1.935e+00  1.276e+00  -1.517
## Mother.s.qualification41               2.046e+00  1.272e+00   1.608
## Mother.s.qualification42               4.354e+00  1.397e+01   0.312
## Mother.s.qualification43              -2.158e+00  2.318e+00  -0.931
## Mother.s.qualification44              -1.207e+01  3.956e+03  -0.003
## Father.s.qualification2               -9.620e-01  4.040e-01  -2.381
## Father.s.qualification3                1.906e-02  2.654e-01   0.072
## Father.s.qualification4                5.712e-02  5.392e-01   0.106
## Father.s.qualification5               -5.397e-01  9.057e-01  -0.596
## Father.s.qualification6               -2.248e+01  3.561e+03  -0.006
## Father.s.qualification9                3.910e+00  3.087e+00   1.267
## Father.s.qualification10              -2.694e+01  2.683e+03  -0.010
## Father.s.qualification11               2.958e+00  3.348e+00   0.883
## Father.s.qualification12               4.353e-01  5.633e-01   0.773
```

```
## Father.s.qualification13                -7.703e+00  3.956e+03  -0.002
## Father.s.qualification14                 1.248e+00  2.007e+00   0.622
## Father.s.qualification18                        NA         NA      NA
## Father.s.qualification19                 1.917e-01  1.601e-01   1.197
## Father.s.qualification20                -1.082e+01  3.956e+03  -0.003
## Father.s.qualification22                -1.179e+01  1.813e+03  -0.007
## Father.s.qualification25                -1.675e+01  3.956e+03  -0.004
## Father.s.qualification26                 3.369e+00  1.397e+01   0.241
## Father.s.qualification27                -2.487e+01  4.636e+03  -0.005
## Father.s.qualification29                -1.298e+01  1.799e+03  -0.007
## Father.s.qualification30                -1.649e+01  2.165e+03  -0.008
## Father.s.qualification31                -1.699e+01  3.956e+03  -0.004
## Father.s.qualification33                -6.837e+00  3.956e+03  -0.002
## Father.s.qualification34                -7.228e-01  9.491e-01  -0.762
## Father.s.qualification35                -3.324e+01  4.547e+03  -0.007
## Father.s.qualification36                 2.132e-01  1.461e+00   0.146
## Father.s.qualification37                 1.913e-01  1.866e-01   1.025
## Father.s.qualification38                 3.768e-01  1.862e-01   2.024
## Father.s.qualification39                -2.149e+00  8.065e-01  -2.665
## Father.s.qualification40                -2.889e-01  1.528e+00  -0.189
## Father.s.qualification41                -8.421e+00  2.019e+03  -0.004
## Father.s.qualification42                 1.769e+01  3.956e+03   0.004
## Father.s.qualification43                 1.904e+01  2.490e+03   0.008
## Father.s.qualification44                -1.008e+01  3.956e+03  -0.003
## Mother.s.occupation1                    -3.461e-02  7.477e-01  -0.046
## Mother.s.occupation2                    -3.597e-01  7.149e-01  -0.503
## Mother.s.occupation3                    -3.054e-01  6.917e-01  -0.441
## Mother.s.occupation4                    -1.447e-01  6.727e-01  -0.215
## Mother.s.occupation5                    -1.384e-02  6.776e-01  -0.020
## Mother.s.occupation6                     4.918e-01  7.749e-01   0.635
## Mother.s.occupation7                    -2.940e-01  6.942e-01  -0.424
## Mother.s.occupation8                    -1.314e+00  8.414e-01  -1.562
## Mother.s.occupation9                    -6.074e-02  6.701e-01  -0.091
## Mother.s.occupation10                    4.047e-02  1.461e+00   0.028
## Mother.s.occupation90                    7.786e-01  1.074e+00   0.725
## Mother.s.occupation99                   -1.875e+00  5.298e+00  -0.354
## Mother.s.occupation122                  -8.535e+00  6.852e+03  -0.001
## Mother.s.occupation123                   7.045e+00  5.595e+03   0.001
## Mother.s.occupation125                   1.271e+01  3.956e+03   0.003
## Mother.s.occupation131                  -2.629e+01  7.240e+03  -0.004
## Mother.s.occupation132                   5.368e+00  5.595e+03   0.001
## Mother.s.occupation134                   3.935e+00  5.595e+03   0.001
## Mother.s.occupation141                   3.305e+00  5.595e+03   0.001
## Mother.s.occupation143                   4.293e+00  5.595e+03   0.001
## Mother.s.occupation144                   5.934e+00  5.595e+03   0.001
## Mother.s.occupation151                   1.868e+01  6.143e+03   0.003
## Mother.s.occupation152                   5.100e+00  5.595e+03   0.001
## Mother.s.occupation153                   2.324e+01  6.852e+03   0.003
## Mother.s.occupation171                   2.199e+01  6.852e+03   0.003
## Mother.s.occupation173                  -1.062e+01  6.852e+03  -0.002
## Mother.s.occupation175                   6.220e+00  5.595e+03   0.001
## Mother.s.occupation191                   4.497e+00  5.595e+03   0.001
## Mother.s.occupation192                   1.512e+00  5.595e+03   0.000
## Mother.s.occupation193                   5.118e+00  5.595e+03   0.001
```

```
## Mother.s.occupation194                3.405e+00  5.595e+03   0.001
## Father.s.occupation1                 -1.034e+00  7.269e-01  -1.422
## Father.s.occupation2                 -1.236e+00  7.327e-01  -1.687
## Father.s.occupation3                 -8.517e-01  6.856e-01  -1.242
## Father.s.occupation4                 -1.449e+00  6.870e-01  -2.109
## Father.s.occupation5                 -1.180e+00  6.796e-01  -1.736
## Father.s.occupation6                 -1.030e+00  7.130e-01  -1.445
## Father.s.occupation7                 -1.362e+00  6.782e-01  -2.008
## Father.s.occupation8                 -1.253e+00  6.924e-01  -1.809
## Father.s.occupation9                 -1.359e+00  6.762e-01  -2.010
## Father.s.occupation10                -1.318e+00  6.905e-01  -1.908
## Father.s.occupation90                -1.671e+00  9.897e-01  -1.688
## Father.s.occupation99                 1.558e+00  5.147e+00   0.303
## Father.s.occupation101                1.233e+01  6.852e+03   0.002
## Father.s.occupation102               -5.122e+00  5.595e+03  -0.001
## Father.s.occupation103               -7.396e+00  5.595e+03  -0.001
## Father.s.occupation112               -2.395e+01  5.990e+03  -0.004
## Father.s.occupation114                1.048e+01  6.852e+03   0.002
## Father.s.occupation121               -2.535e+01  6.852e+03  -0.004
## Father.s.occupation122               -1.052e+01  5.595e+03  -0.002
## Father.s.occupation123               -5.193e+00  5.595e+03  -0.001
## Father.s.occupation124                       NA         NA      NA
## Father.s.occupation131                1.293e+01  6.852e+03   0.002
## Father.s.occupation132               -2.381e+01  6.852e+03  -0.003
## Father.s.occupation134               -1.378e+01  6.852e+03  -0.002
## Father.s.occupation135               -8.001e+00  5.595e+03  -0.001
## Father.s.occupation141                9.504e+00  6.852e+03   0.001
## Father.s.occupation143                9.033e+00  6.852e+03   0.001
## Father.s.occupation144               -4.580e+00  5.595e+03  -0.001
## Father.s.occupation151                8.974e+00  6.143e+03   0.001
## Father.s.occupation152               -2.538e+01  5.981e+03  -0.004
## Father.s.occupation153               -9.471e+00  7.912e+03  -0.001
## Father.s.occupation154                1.147e+01  6.852e+03   0.002
## Father.s.occupation161               -3.396e+01  7.306e+03  -0.005
## Father.s.occupation163               -5.318e+00  5.595e+03  -0.001
## Father.s.occupation171               -5.975e+00  5.595e+03  -0.001
## Father.s.occupation172               -4.757e+00  5.595e+03  -0.001
## Father.s.occupation174                1.050e+01  6.852e+03   0.002
## Father.s.occupation175               -6.476e+00  5.595e+03  -0.001
## Father.s.occupation181                8.519e+00  6.064e+03   0.001
## Father.s.occupation182               -5.258e+00  5.595e+03  -0.001
## Father.s.occupation183               -7.324e+00  5.595e+03  -0.001
## Father.s.occupation192               -3.539e+00  5.595e+03  -0.001
## Father.s.occupation193               -5.848e+00  5.595e+03  -0.001
## Father.s.occupation194               -2.098e+00  5.595e+03   0.000
## Father.s.occupation195                1.055e+01  6.852e+03   0.002
## Admission.grade                       8.573e-03  4.971e-03   1.725
## Displaced1                           -4.728e-02  1.221e-01  -0.387
## Educational.special.needs1           -1.207e-01  4.335e-01  -0.279
## Debtor1                              -1.126e+00  2.125e-01  -5.301
## Tuition.fees.up.to.date1              2.317e+00  2.916e-01   7.945
## GenderM                              -1.545e-01  1.191e-01  -1.298
## Scholarship.holder1                   7.337e-01  1.255e-01   5.844
## Age.at.enrollment                    -1.282e-02  1.398e-02  -0.917
```

```
## International1                                        NA         NA      NA
## Curricular.units.1st.sem..credited.           -1.165e-01  9.671e-02  -1.205
## Curricular.units.1st.sem..enrolled.           -3.316e-01  1.318e-01  -2.515
## Curricular.units.1st.sem..evaluations.        -3.965e-02  3.144e-02  -1.261
## Curricular.units.1st.sem..approved.            6.660e-01  6.934e-02   9.605
## Curricular.units.1st.sem..grade.              -5.701e-02  4.748e-02  -1.201
## Curricular.units.1st.sem..without.evaluations. -6.713e-02  1.416e-01  -0.474
## Curricular.units.2nd.sem..credited.            1.441e-02  9.944e-02   0.145
## Curricular.units.2nd.sem..enrolled.           -7.077e-01  1.404e-01  -5.040
## Curricular.units.2nd.sem..evaluations.        -1.102e-01  2.942e-02  -3.746
## Curricular.units.2nd.sem..approved.            9.007e-01  6.578e-02  13.694
## Curricular.units.2nd.sem..grade.               1.453e-01  4.983e-02   2.915
## Curricular.units.2nd.sem..without.evaluations.  2.991e-01  1.170e-01   2.557
## Unemployment.rate                             -3.348e-02  2.333e-02  -1.435
## Inflation.rate                                 2.425e-02  3.848e-02   0.630
## GDP                                           -1.223e-02  2.690e-02  -0.455
##                                               Pr(>|z|)
## (Intercept)                                    0.99835
## Marital.status2                                0.57861
## Marital.status3                                0.88761
## Marital.status4                                0.29682
## Marital.status5                                0.18949
## Marital.status6                                0.17522
## Application.mode2                              0.76682
## Application.mode5                              0.14413
## Application.mode7                              0.76784
## Application.mode10                             0.32404
## Application.mode15                             0.18541
## Application.mode16                             0.89859
## Application.mode17                             0.18671
## Application.mode18                             0.41207
## Application.mode26                             0.99657
## Application.mode27                             0.99599
## Application.mode39                             0.41388
## Application.mode42                             0.60817
## Application.mode43                             0.55382
## Application.mode44                             0.44006
## Application.mode51                             0.33257
## Application.mode53                             0.66265
## Application.mode57                             0.99648
## Application.order1                             0.99722
## Application.order2                             0.99726
## Application.order3                             0.99715
## Application.order4                             0.99729
## Application.order5                             0.99727
## Application.order6                             0.99729
## Application.order9                             0.99545
## Course171                                      0.06241 .
## Course8014                                     0.05059 .
## Course9003                                     0.09119 .
## Course9070                                     0.16274
## Course9085                                     0.06497 .
## Course9119                                     0.74194
## Course9130                                     0.16432
```

```
## Course9147                                0.15699
## Course9238                                0.02952 *
## Course9254                                0.09269 .
## Course9500                                0.14581
## Course9556                                0.33968
## Course9670                                0.14077
## Course9773                                0.14798
## Course9853                                0.47801
## Course9991                                0.29960
## Daytime.evening.attendance.1                   NA
## Previous.qualification2                   0.45498
## Previous.qualification3                   0.92327
## Previous.qualification4                   0.37375
## Previous.qualification5                   0.99876
## Previous.qualification6                   0.61030
## Previous.qualification9                   0.99204
## Previous.qualification10                  0.85135
## Previous.qualification12                  0.97710
## Previous.qualification14                  0.99810
## Previous.qualification15                  0.30190
## Previous.qualification19                  0.26611
## Previous.qualification38                  0.84982
## Previous.qualification39                  0.48027
## Previous.qualification40                  0.55701
## Previous.qualification42                  0.48080
## Previous.qualification43                  0.20499
## Previous.qualification..grade.            0.66269
## Nacionality2                              0.99586
## Nacionality6                              0.87878
## Nacionality11                             0.99403
## Nacionality13                             0.99735
## Nacionality14                             0.99696
## Nacionality17                             0.99735
## Nacionality21                             0.99726
## Nacionality22                             0.06614 .
## Nacionality24                             0.00388 **
## Nacionality25                             0.83655
## Nacionality26                             0.54777
## Nacionality32                             0.99858
## Nacionality41                             0.69049
## Nacionality62                             0.86854
## Nacionality100                            0.99346
## Nacionality101                            0.70268
## Nacionality103                            0.63041
## Nacionality105                            0.99292
## Nacionality108                            0.99919
## Nacionality109                            0.99588
## Mother.s.qualification2                   0.64037
## Mother.s.qualification3                   0.84696
## Mother.s.qualification4                   0.39932
## Mother.s.qualification5                   0.79407
## Mother.s.qualification6                   0.99551
## Mother.s.qualification9                   0.66098
## Mother.s.qualification10                  0.73908
```

```
## Mother.s.qualification11                  0.33383
## Mother.s.qualification12                  0.53807
## Mother.s.qualification14                  0.95957
## Mother.s.qualification18                  0.99695
## Mother.s.qualification19                  0.99956
## Mother.s.qualification22                  0.99465
## Mother.s.qualification26                  0.99396
## Mother.s.qualification27                  0.99698
## Mother.s.qualification29                  0.99247
## Mother.s.qualification30                  0.79923
## Mother.s.qualification34                  0.65024
## Mother.s.qualification35                  0.99674
## Mother.s.qualification36                  0.63946
## Mother.s.qualification37                  0.74936
## Mother.s.qualification38                  0.43605
## Mother.s.qualification39                  0.88285
## Mother.s.qualification40                  0.12924
## Mother.s.qualification41                  0.10782
## Mother.s.qualification42                  0.75524
## Mother.s.qualification43                  0.35191
## Mother.s.qualification44                  0.99757
## Father.s.qualification2                   0.01726 *
## Father.s.qualification3                   0.94274
## Father.s.qualification4                   0.91564
## Father.s.qualification5                   0.55124
## Father.s.qualification6                   0.99496
## Father.s.qualification9                   0.20532
## Father.s.qualification10                  0.99199
## Father.s.qualification11                  0.37699
## Father.s.qualification12                  0.43968
## Father.s.qualification13                  0.99845
## Father.s.qualification14                  0.53397
## Father.s.qualification18                       NA
## Father.s.qualification19                  0.23128
## Father.s.qualification20                  0.99782
## Father.s.qualification22                  0.99481
## Father.s.qualification25                  0.99662
## Father.s.qualification26                  0.80940
## Father.s.qualification27                  0.99572
## Father.s.qualification29                  0.99424
## Father.s.qualification30                  0.99392
## Father.s.qualification31                  0.99657
## Father.s.qualification33                  0.99862
## Father.s.qualification34                  0.44632
## Father.s.qualification35                  0.99417
## Father.s.qualification36                  0.88395
## Father.s.qualification37                  0.30528
## Father.s.qualification38                  0.04302 *
## Father.s.qualification39                  0.00771 **
## Father.s.qualification40                  0.85003
## Father.s.qualification41                  0.99667
## Father.s.qualification42                  0.99643
## Father.s.qualification43                  0.99390
## Father.s.qualification44                  0.99797
```

```
## Mother.s.occupation1                    0.96308
## Mother.s.occupation2                    0.61486
## Mother.s.occupation3                    0.65885
## Mother.s.occupation4                    0.82970
## Mother.s.occupation5                    0.98370
## Mother.s.occupation6                    0.52565
## Mother.s.occupation7                    0.67192
## Mother.s.occupation8                    0.11824
## Mother.s.occupation9                    0.92777
## Mother.s.occupation10                   0.97791
## Mother.s.occupation90                   0.46832
## Mother.s.occupation99                   0.72348
## Mother.s.occupation122                  0.99901
## Mother.s.occupation123                  0.99900
## Mother.s.occupation125                  0.99744
## Mother.s.occupation131                  0.99710
## Mother.s.occupation132                  0.99923
## Mother.s.occupation134                  0.99944
## Mother.s.occupation141                  0.99953
## Mother.s.occupation143                  0.99939
## Mother.s.occupation144                  0.99915
## Mother.s.occupation151                  0.99757
## Mother.s.occupation152                  0.99927
## Mother.s.occupation153                  0.99729
## Mother.s.occupation171                  0.99744
## Mother.s.occupation173                  0.99876
## Mother.s.occupation175                  0.99911
## Mother.s.occupation191                  0.99936
## Mother.s.occupation192                  0.99978
## Mother.s.occupation193                  0.99927
## Mother.s.occupation194                  0.99951
## Father.s.occupation1                    0.15498
## Father.s.occupation2                    0.09158 .
## Father.s.occupation3                    0.21414
## Father.s.occupation4                    0.03495 *
## Father.s.occupation5                    0.08263 .
## Father.s.occupation6                    0.14841
## Father.s.occupation7                    0.04469 *
## Father.s.occupation8                    0.07040 .
## Father.s.occupation9                    0.04445 *
## Father.s.occupation10                   0.05634 .
## Father.s.occupation90                   0.09139 .
## Father.s.occupation99                   0.76206
## Father.s.occupation101                  0.99856
## Father.s.occupation102                  0.99927
## Father.s.occupation103                  0.99895
## Father.s.occupation112                  0.99681
## Father.s.occupation114                  0.99878
## Father.s.occupation121                  0.99705
## Father.s.occupation122                  0.99850
## Father.s.occupation123                  0.99926
## Father.s.occupation124                       NA
## Father.s.occupation131                  0.99849
## Father.s.occupation132                  0.99723
```

```
## Father.s.occupation134                           0.99840
## Father.s.occupation135                           0.99886
## Father.s.occupation141                           0.99889
## Father.s.occupation143                           0.99895
## Father.s.occupation144                           0.99935
## Father.s.occupation151                           0.99883
## Father.s.occupation152                           0.99661
## Father.s.occupation153                           0.99904
## Father.s.occupation154                           0.99866
## Father.s.occupation161                           0.99629
## Father.s.occupation163                           0.99924
## Father.s.occupation171                           0.99915
## Father.s.occupation172                           0.99932
## Father.s.occupation174                           0.99878
## Father.s.occupation175                           0.99908
## Father.s.occupation181                           0.99888
## Father.s.occupation182                           0.99925
## Father.s.occupation183                           0.99896
## Father.s.occupation192                           0.99950
## Father.s.occupation193                           0.99917
## Father.s.occupation194                           0.99970
## Father.s.occupation195                           0.99877
## Admission.grade                                  0.08460 .
## Displaced1                                       0.69848
## Educational.special.needs1                       0.78061
## Debtor1                                         1.15e-07 ***
## Tuition.fees.up.to.date1                        1.94e-15 ***
## GenderM                                          0.19446
## Scholarship.holder1                             5.08e-09 ***
## Age.at.enrollment                                0.35907
## International1                                       NA
## Curricular.units.1st.sem..credited.              0.22814
## Curricular.units.1st.sem..enrolled.              0.01189 *
## Curricular.units.1st.sem..evaluations.           0.20727
## Curricular.units.1st.sem..approved.              < 2e-16 ***
## Curricular.units.1st.sem..grade.                 0.22981
## Curricular.units.1st.sem..without.evaluations.   0.63552
## Curricular.units.2nd.sem..credited.              0.88479
## Curricular.units.2nd.sem..enrolled.             4.66e-07 ***
## Curricular.units.2nd.sem..evaluations.           0.00018 ***
## Curricular.units.2nd.sem..approved.              < 2e-16 ***
## Curricular.units.2nd.sem..grade.                 0.00355 **
## Curricular.units.2nd.sem..without.evaluations.   0.01056 *
## Unemployment.rate                                0.15134
## Inflation.rate                                   0.52858
## GDP                                              0.64944
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6133.0  on 4423  degrees of freedom
## Residual deviance: 2632.2  on 4183  degrees of freedom
## AIC: 3114.2
```

```
##
## Number of Fisher Scoring iterations: 16
```

Plusieurs des variables présentes dans ce modèle ont une p-value $< \alpha = 5\%$, tels que **Curricular.units.2nd.sem..approved.** ou **Tuition.fees.up.to.date**, les analyses ont donc pu être continuées.

Mais ce modèle affiche un message permettant de comprendre qu'il y a des problèmes de convergence dans le modèle ce qui va entraîner des problèmes dans le calcul des coefficients. 4 des coefficients n'ont pas pu être définis à cause de singularités, et d'autres affichent des coefficients NA.

Mais il va quand même être possible d'effectuer des modèles logistiques à l'aide d'autres méthodes.

Echantillonnage

Avant de modéliser les données, le jeu de données a du être découpé aléatoirement en 2 échantillons : un échantillon apprentissage (train) avec 70% du jeu de données initial et un échantillon test avec 30% du jeu de données initial. Cela permettra d'effectuer la modélisation sur l'échantillon d'apprentissage et ensuite de tester ce modèle et effectuer des prédictions sur l'échantillon test qui ne sera pas encore touché.

```
# indices aléatoires pour split le dataset
indices <- sample(1:nrow(data), size = floor(0.70 * nrow(data)))

# échantillon train
train_set <- data[indices,]

#échantillon test
test_set <- data[-indices,]
```

Il a fallu vérifier que le jeu de données a été découpé de façon assez uniforme sur la variable cible afin de ne pas se retrouver avec un manque de données pour créer les modèles.

```
# summary pour voir les répartitions de toutes les variables
# summary(train_set)
# summary(test_set)
```

```
# proportions de la variable cible dans les deux nouveaux jeux de données
# pour l'échantillon train
round(prop.table(table(train_set$Ybin)) * 100, 2)
```

```
##
##     0     1
## 49.55 50.45
```

```
# pour l'échantillon test
round(prop.table(table(test_set$Ybin)) * 100, 2)
```

```
##
##     0     1
## 51.28 48.72
```

Modélisation

Afin de trouver un modèle permettant de prédire au mieux notre variable cible, nous allons effectuer deux types de modélisations différentes : la régression logistique pénalisée RIDGE (ℓ_2) et la régression logistique pénalisée LASSO (ℓ_1).

Avec la régression logistique, on peut rencontrer des problèmes de sur-entraînement (*overfitting*) et de colinéarité entre les variables explicatives.

C'est pour cela que l'on va faire les modèles RIDGE et LASSO : ce sont des régressions qui vont pénaliser les coefficients afin d'avoir un compromis acceptable entre la performance du modèle et la pénalisation des coefficients, en minimisant l'erreur de prédiction tout en homogéninisant les valeurs des paramètres.

Dans les méthodes de régression pénalisées, λ représente l'hyperparamètre clé utilisé pour régulariser les modèles.

En somme, λ reste un outil clé dans la boîte à outils de la régularisation, mais son application dépend du modèle et de ses caractéristiques spécifiques, justifiant son utilisation dans des contextes comme la régression RIDGE et LASSO, mais pas nécessairement dans tous les types de modélisation, notamment la modélisation logistique classique.

Pour LASSO, λ favorise la parcimonie, forçant certains coefficients à zéro pour une sélection automatique de variables.

RIDGE ne va pas vraiment servir à faire de la sélection de variables, mais plutôt à enlever la multi colinéarité des variables.

Contrairement à RIDGE, qui prévient le sur-ajustement en restreignant les coefficients, LASSO offre une solution plus épurée

Dans la modélisation logistique, l'utilisation de λ n'est pas courante. Les régularisations se font souvent via d'autres mécanismes, comme la pénalité Elastic Net, rendant λ spécifique aux approches RIDGE et LASSO.

Les pénalités étant définies en terme de valeurs numériques des coefficients, on va devoir transformer toutes nos variables factorielles (sauf la variable cible) en variables binaires.

```
# jeu train et test avec toutes les variables factorielles transformées en binaire
# sans la variable cible
x_train <- model.matrix(Ybin ~ . -1 , data = train_set)
x_test <- model.matrix(Ybin ~ . -1 , data = test_set)
```

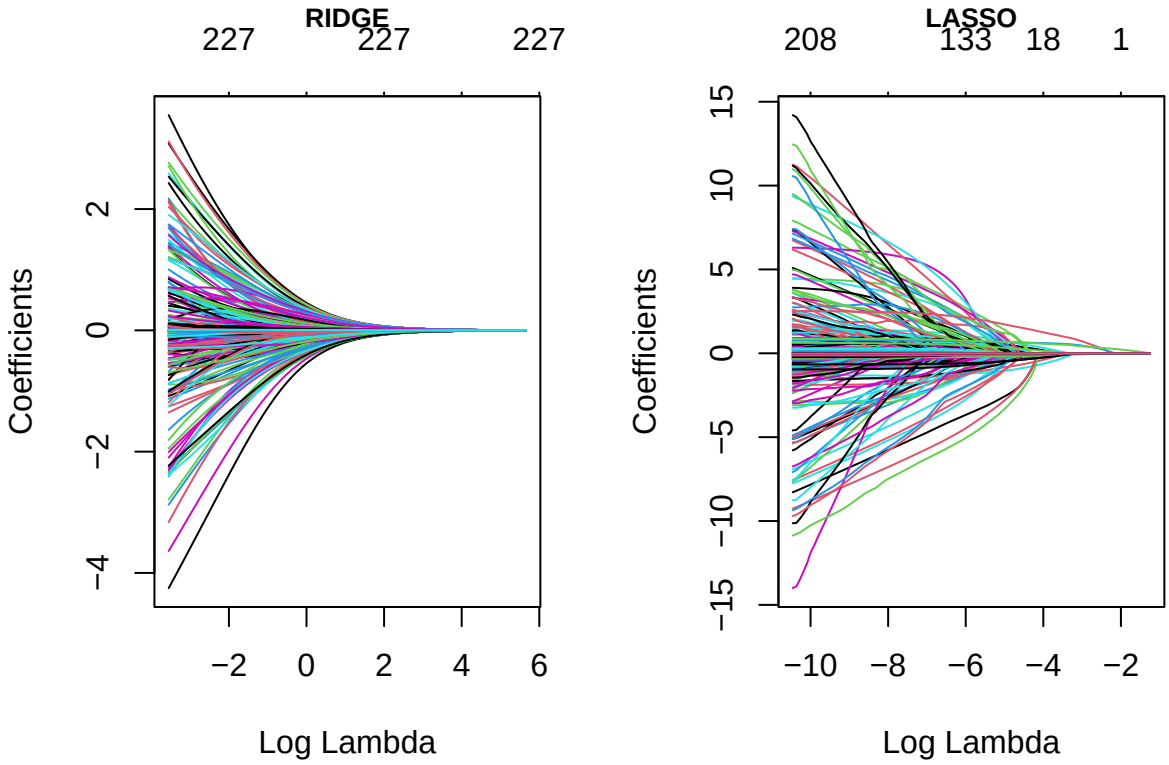
```
# ajustement du modèle ridge
mod_ridge <- glmnet(x = x_train , y = train_set$Ybin, family = "binomial", alpha = 0)

# ajustement du modèle lasso
mod_lasso <- glmnet(x = x_train, y = train_set$Ybin, family = "binomial", alpha = 1)

# pour avoir les coefficients de chaque variable par régression
# pour chacun des 100 lambdas pris
## coef.glmnet(mod_ridge)
## coef.glmnet(mod_lasso)
```

En effectuant ces deux régressions avec la commande `glmnet`, un objet en ressort avec plusieurs informations. La valeur de λ n'ayant pas été spécifiée, cet objet va contenir 100 valeurs de λ différentes. Pour chacun de ces λ, il y aura les coefficients associés après la pénalisation.

```r
# graphiques des coefficients en fonction des log des lambdas
par(mfrow = c(1,2))
plot(mod_ridge, xvar = "lambda", main = "RIDGE", cex.main = 0.8)
plot(mod_lasso, xvar = "lambda", main = "LASSO", cex.main = 0.8)
```



Les valeurs des $log(\lambda)$ prises dans les modélisations ne sont pas les mêmes : celles utilisées dans la modélisation RIDGE sont en générales plus grandes que celles prises dans la modélisation LASSO. Mais ce qui en ressort pour les 2 modélisations est que plus le λ pris est grand, plus les coefficients ont tendance à avoir une valeur nulle. C'est une des **particularités** du paramètre λ : plus il est grand, moins il y aura de coefficients. Mais ce qui change de RIDGE à LASSO et qui est visible à travers ces deux graphiques est que les coefficients RIDGE vont converger vers 0 tous à peu près au même λ quand il devient très grand, tandis que les coefficients LASSO vont converger vers 0 pour n'importe quel λ même s'ils vont presque tous être nuls quand le λ est élevé.

Mais afin d'avoir un modèle pénalisant permettant quand même d'avoir un bon pouvoir prédictif, il faut trouver le **meilleur lambda**. Il est donc nécessaire de faire de la validation croisée.

Validation croisée

La validation croisée va permettre de partitionner les données en plusieurs sous-ensembles grâce à la méthode des **k-folds**. Cette méthode divise les données en plis (*folds*) de taille égale. On utilisera les données d'apprentissage pour cela.

Pour chaque pli, la validation croisée va ajuster le modèle sur les *k-1* plis restants (*ensemble d'entraînement*) et évalué sur le pli retenu (*ensemble de validation*). Ce processus est répété k fois, chaque pli servant une fois comme ensemble de validation.

La validation croisée va permettre de trouver le λ optimal, cest-à-dire le λ qui maximise les performances moyennes sur les ensembles de validation, c'est à dire sur tous les plis.

Avec la commande `cv.glmnet`, le λ choisi est celui qui minimise la déviance binomiale. C'est une mesure du **mauvais ajustement** du modèle par rapport aux données et une déviance basse pour un modèle logistique indique un bon ajustement du modèle aux données.

$$D = -2 \times \ell(\hat{\beta})$$

avec $\ell(\hat{\beta})$ la log-vraisemblance sur notre modèle avec les coefficients $\hat{\beta}$.

```
# on prend 10 k folds, le nombre de folds par défaut

# cross validation ridge
cv_ridge <- cv.glmnet(x = x_train, y = train_set$Ybin, family = "binomial", alpha = 0)
# lambda optimal ridge
best_lambda_ridge <- cv_ridge$lambda.min
cat("Lambda optimal ridge :", best_lambda_ridge, "\n")
```

```
## Lambda optimal ridge : 0.02868925
```

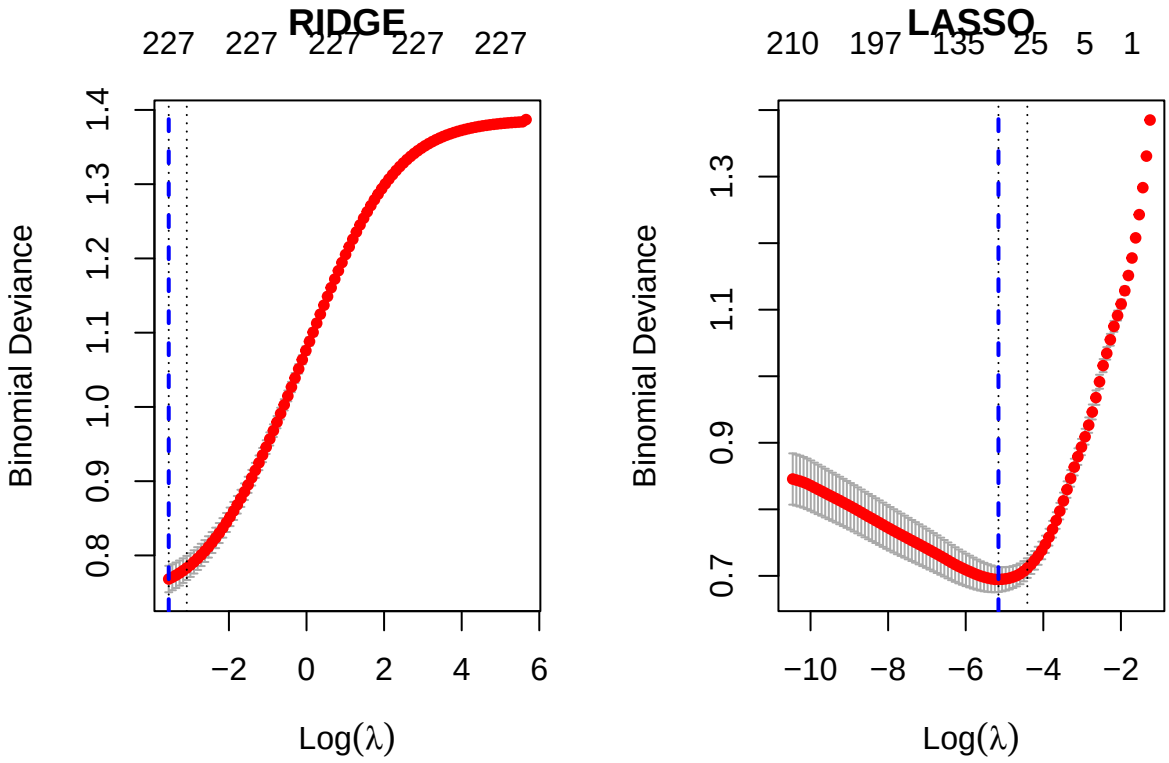
```
# cross validation lasso
cv_lasso <- cv.glmnet(x = x_train, y = train_set$Ybin, family = "binomial", alpha = 1)

# lambda optimal lasso
best_lambda_lasso <- cv_lasso$lambda.min
cat("Lambda optimal lasso :", best_lambda_lasso, "\n")
```

```
## Lambda optimal lasso : 0.005764339
```

```
# graphiques déviance binomiale en fonction des log lambda
# droite pointillé bleue correspond au lambda optimal

par(mfrow = c(1,2))
# pour la cross validation ridge
plot(cv_ridge, main = "RIDGE")
abline(v = log(best_lambda_ridge), col = "blue", lty = 2, lwd = 2)
# pour la cross validation lasso
plot(cv_lasso, main = "LASSO")
abline(v = log(best_lambda_lasso), col = "blue", lty = 2, lwd = 2)
```



Il s'agit de graphiques représentant tous les λ qui ont été testés avec la cross validation, en affichant leur log, par rapport à la déviance binomiale.

Les λ optimals en fonction de la méthode ne vont pas du tout être les mêmes : pour RIDGE, le λ optimal est d'environ 0.0287 , qui est également un des plus petits λ pris en compte dans la modélisation RIDGE initiale (avec 100 valeurs de λ différentes) tandis que pour LASSO, le lambda optimal est d'environ 0.0052 et n'est pas un des plus petits pris en compte dans la modélisation, pourtant c'est bien celui qui minimise la déviance comme on peut le voir dans ce graphique.

Le **choix** de ces λ va permettre de passer à la dernière étape pour le choix du modèle final : les prédictions sur l'échantillon test.

Prédictions sur l'échantillon test

Les prédictions vont permettre de mesurer la performance des deux modèles sur un échantillon non entraîné.

Dans un premier temps, il faut prédire les probabilités de la variable cible, qui vont ensuite être transformées en prédictions binaires de la variable cible. Pour cela, on va utiliser un seuil $S = 0.5$ (choisi par le critère MAP) mais cela ne veut pas dire que c'est le seuil le plus approprié pour mesurer les performances des modèles. Il sera vu juste après comment trouver un seuil acceptable. A partir de ce seuil, si l'estimation de la probabilité $>$ S, alors la prédiction de la variable cible $\hat{Y} = 1$. Sinon, $\hat{Y} = 0$.

```
# estimations des probabilités de la variable cible
pred_probs_ridge <- predict(mod_ridge, newx = x_test, s = best_lambda_ridge, type = "response")
pred_probs_lasso <- predict(mod_lasso, newx = x_test, s = best_lambda_lasso, type = "response")
```

```r
# transformations des probabilités en prédictions binaires (0 ou 1)
pred_classes_ridge <- ifelse(pred_probs_ridge > 0.5, 1, 0)
pred_classes_lasso <- ifelse(pred_probs_lasso > 0.5, 1, 0)

# stockage des prédictions et probas dans un dataframe
result_ridge <- cbind.data.frame(pred_probs_ridge, pred_classes_ridge)
colnames(result_ridge) <- c("Probabilités", "Prédictions")

result_lasso <- cbind.data.frame(pred_probs_lasso, pred_classes_lasso)
colnames(result_lasso) <- c("Probabilités", "Prédictions")
```

C'est à partir de ces prédictions que les performances du modèle vont être mesurées.

Performances

Matrice de confusion

```r
# créations des matrices de confusions
conf_matrix_ridge <- addmargins(table(Prédictions = pred_classes_ridge, Réalité = test_set$Ybin))
conf_matrix_lasso <- addmargins(table(Prédictions = pred_classes_lasso, Réalité = test_set$Ybin))

# affichage matrice de confusion ridge
cat("\n", "RIDGE", "\n")
```

```
##
##  RIDGE
```

```r
print(conf_matrix_ridge)
```

```
##            Réalité
## Prédictions    0    1  Sum
##         0    548   69  617
##         1    133  578  711
##         Sum  681  647 1328
```

```r
# affichage matrice de confusion lasso
cat("\n", "LASSO", "\n")
```

```
##
##  LASSO
```

```r
print(conf_matrix_lasso)
```

```
##            Réalité
## Prédictions    0    1  Sum
##         0    557   63  620
##         1    124  584  708
##         Sum  681  647 1328
```

Ces matrices de confusions contiennent plusieurs informations :

- **Vrais Positifs (VP) :** Cela représente le nombre d'étudiants pour lesquels le modèle a correctement prédit la réussite. ($\hat{Y} = 1$ quand $Y = 1$)

  Le modèle RIDGE contient 578 VP contre 581 pour le modèle LASSO

- **Vrais Négatifs (VN) :** Cela représente le nombre d'étudiants pour lesquels le modèle a correctement prédit l'échec ($\hat{Y} = 0$ quand $Y = 0$)

  548 pour RIDGE contre 559 pour LASSO

- **Faux Positifs (FP) :** Cela signifie que le modèle a prédit à tort la réussite pour ces étudiants. ($\hat{Y} = 1$ quand $Y = 0$)

  133 pour RIDGE contre 122 pour LASSO

- **Faux Négatifs (FN) :** Cela signifie que le modèle a omis de prédire la réussite pour ces étudiants. ($\hat{Y} = 0$ quand $Y = 1$)

  69 pour RIDGE contre 66 pour LASSO

Le modèle LASSO a une meilleure capacité à prédire à la fois les succès (VP élevés) et les échecs (VN élevés) que le modèle RIDGE. Les faux positifs (prédire à tort le succès) sont relativement faibles pour les deux modèles mais plus pour LASSO, ce qui suggère que le modèle n'est pas trop optimiste dans ses prédictions. Les faux négatifs (manquer la prédiction de réussite) sont deux fois plus faibles, indiquant que le modèle pourrait être assez sensible à ne pas manquer les réussites.

Mais ces matrices de confusions peuvent également permettre de trouver des indicateurs sur le modèle permettant de connaître le bon ajustement ou non des modèles.

```r
# Création d'une fonction pour afficher plusieurs métriques d'évaluation
evaluate_model <- function(actual, predicted) {
  confusion_matrix <- table(Prédiction = predicted, Réalité = actual)
  accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
  precision <- confusion_matrix[2, 2] / sum(confusion_matrix[, 2])
  recall <- confusion_matrix[2, 2] / sum(confusion_matrix[2, ])
  f1_score <- 2 * (precision * recall) / (precision + recall)
  specificity = confusion_matrix[1,1] / sum(confusion_matrix[,1])

  cat("Précision :", precision, "\n")
  cat("Rappel :", recall, "\n")
  cat("F-mesure :", f1_score, "\n")
  cat("Exactitude (Accuracy) :", accuracy, "\n")
  cat("Spécificité :", specificity, "\n")

  return(confusion_matrix)
}

# Utilisation de la fonction
cat("\n", "RIDGE", "\n")
```

```
##
##  RIDGE
```

```r
eval_ridge <- evaluate_model(test_set$Ybin, pred_classes_ridge)
```

```
## Précision : 0.8933539
## Rappel : 0.8129395
## F-mesure : 0.8512518
## Exactitude (Accuracy) : 0.8478916
## Spécificité : 0.804699
```

```r
cat("\n", "LASSO", "\n")
```

```
##
##  LASSO
```

```r
eval_lasso <- evaluate_model(test_set$Ybin, pred_classes_lasso)
```

```
## Précision : 0.9026275
## Rappel : 0.8248588
## F-mesure : 0.8619926
## Exactitude (Accuracy) : 0.8591867
## Spécificité : 0.8179148
```

- **Précision (Precision) :**

  Elle mesure la proportion de vrais positifs parmi les instances prédites comme positives. Elle indique ici les étudiants prédits comme réussissant (diplômés) qui le sont réellement. Elle est plus élevée pour le modèle LASSO à 89.8% que pour le modèle RIDGE a 89.33%, mais on ne remarque pas réellement de grosses distinctions entre les deux.

- **Rappel (Recall) / Sensibilité :**

```
Également appelé sensibilité, il mesure la proportion de vrais positifs parmi toutes les instances réel
```

- **F-mesure :**

  C'est une moyenne harmonique de la précision et du rappel. Elle donne une mesure équilibrée entre les deux.

- **Exactitude (Accuracy) :**

  Elle mesure la proportion totale de prédictions correctes (vrais positifs + vrais négatifs).

- **Spécificité :**

  Elle mesure la capacité du modèle à reconnaître les étudiants non diplômés parmi les étudiants non diplomés.

En comparant toutes ces mesures, on voit que LASSO a de meilleures performances que RIDGE.

Courbes ROC et AUC

Pour finir, il a fallu calculer l'AUC. C'est une mesure de performance permettant de mesurer le pouvoir prédictif du modèle. Elle peut se situer entre 1 et 0.5 : plus l'AUC est proche de 1, plus la qualité de prédiction du modèle est bien. Pour 0.5, c'est une prédiction aléatoire, ça sera une performance médiocre.

```r
pred_ridge <- prediction(pred_probs_ridge, test_set$Ybin)
pred_lasso <- prediction(pred_probs_lasso, test_set$Ybin)

# AUC
auc_ridge <- performance(pred_ridge, "auc")@y.values[[1]]
cat("RIDGE - AUC sur l'ensemble de test :", auc_ridge, "\n")
```

```
## RIDGE - AUC sur l'ensemble de test : 0.9204257
```

```r
auc_lasso <- performance(pred_lasso, "auc")@y.values[[1]]
cat("LASSO - AUC sur l'ensemble de test :", auc_lasso, "\n")
```

```
## LASSO - AUC sur l'ensemble de test : 0.9344836
```

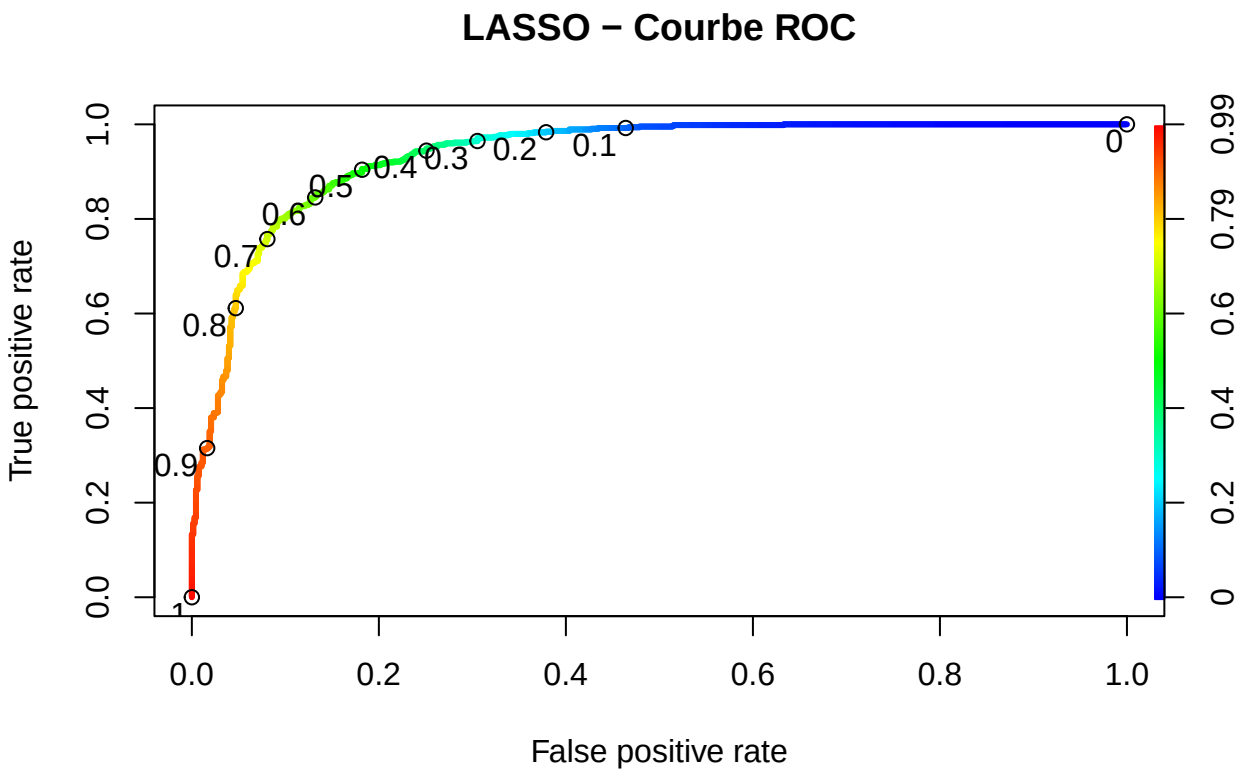
Pour cette dernière mesure utilisée, l'AUC est meilleure pour LASSO que pour RIDGE. Sa valeur est une indication positive de la capacité discriminante de notre modèle et suggère que le modèle a une excellente capacité à distinguer entre les étudiants qui réussissent et ceux qui échouent. En d'autres termes, il est capable de classer correctement la plupart des étudiants en fonction des facteurs inclus dans le modèle. Il paraît donc évident après ces analyses et les indicateurs de performances vus que la régression LASSO convient mieux à nos données pour prédire notre variable cible.

Pour finir, on va tracer la courbe ROC. Cette courbe relie les points avec comme abscisse le taux de faux positifs (FPR) = 1-Spécifité au seuil S et en ordonnée la sensibilité (TPR) au seuil S, pour une grille de plusieurs seuils. L'air sous la courbe correspond à l'AUC, et cette courbe permet de trouver le seuil le plus adapté afin d'avoir le meilleur compromis possible entre Spécificité et Sensibilité.

Plus la courbe est proche du coin supérieur gauche du carré, meilleur est le modèle. Cela représente le seuil permettant d'avoir les meilleurs performances vus grâce à la matrice de confusion. Il permet de capturer le plus possible de vrais évenements avec le moins possible de faux évenements.

```r
roc_perf_lasso <- performance(pred_lasso, measure="tpr", x.measure="fpr")

# courbe ROC
plot(roc_perf_lasso, colorize = TRUE, main = "LASSO - Courbe ROC", print.cutoffs.at = seq(0, 1,by = 0.1
```



Le seuil semblant le plus pertinent est de 0.7 ou bien de 0.6. Il peut être judicieux de calculer à nouveau la matrice de confusion et de s'intéresser à nouveau aux mesures vues telles que la performance, la sensibilité et la spécificité, pour voir si le modèle est devenu plus performant en changeant le seuil de décision.

Éxtension de l'étude

Malgré les diverses options d'extension possibles pour notre étude de jeu de données, nous avons délibérément choisi d'explorer une approche différente en adoptant un arbre de classification plutôt que d'opter pour un modèle de régression Elastic Net ou un classificateur k plus proches voisins (k-NN). Cette décision repose sur plusieurs considérations clés :

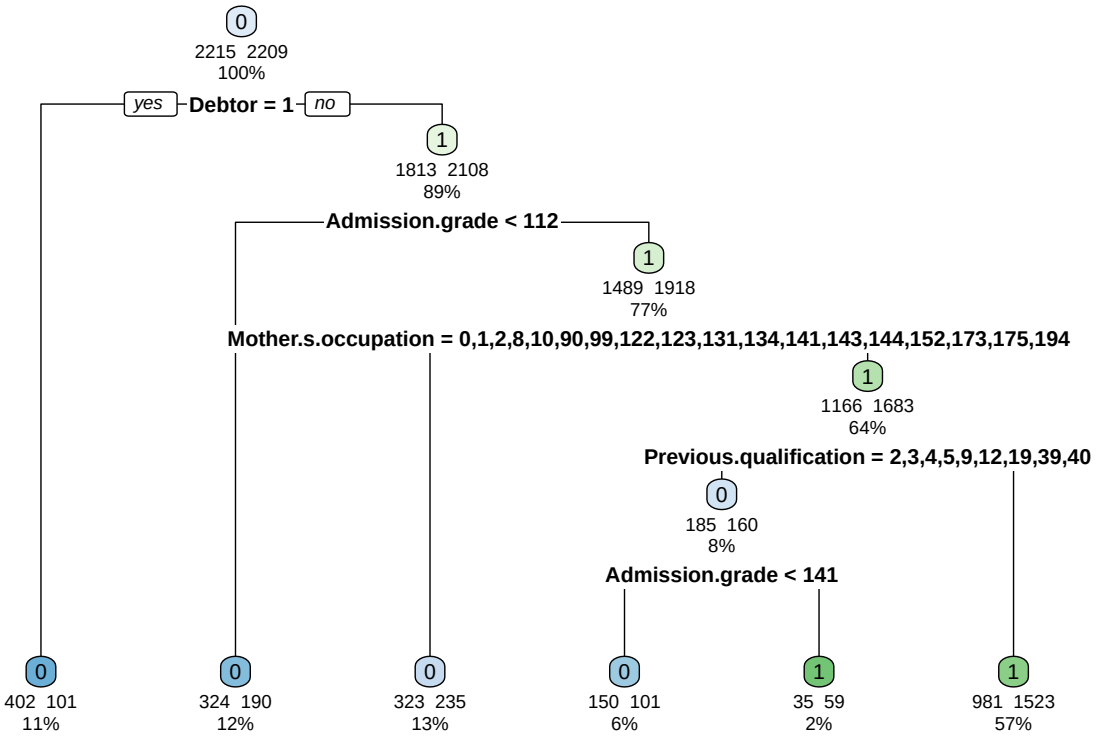
1. **Interprétabilité :** Les arbres de classification se distinguent par leur facilité d'interprétation et de visualisation. Leur logique sous-jacente est rendue transparente grâce à des règles de décision simples.
2. **Non-linéarité :** Les arbres sont capables de saisir des relations non linéaires entre les caractéristiques et la variable cible sans nécessiter une spécification explicite de la forme fonctionnelle.
3. **Adaptabilité :** Les arbres peuvent traiter efficacement des jeux de données mixtes, comprenant à la fois des caractéristiques catégorielles et numériques, sans nécessiter un prétraitement intensif.
4. **Robustesse aux valeurs aberrantes :** Comparativement à certains modèles linéaires, les arbres de classification démontrent une relative robustesse face aux valeurs aberrantes.
5. **Découverte de quelque chose de différent :** En plus des considérations techniques, notre choix reflète également un désir d'exploration et de découverte dans le cadre de notre projet, et l'opportunité de tirer parti des particularités de notre ensemble de données.

```r
# Choix des variables pertinentes
selected_vars <- c("Previous.qualification", "Admission.grade", "Debtor", "Mother.s.occupation")

# Création d'un sous-ensemble avec les variables sélectionnées
subset_data <- data[, c(selected_vars, "Ybin")]

# Construction de l'arbre de classification
tree_model <- rpart(Ybin ~ ., data = subset_data, method = "class")

# Visualisation de l'arbre avec des informations détaillées
rpart.plot(tree_model, extra = 101, under = TRUE, type = 2, fallen.leaves = TRUE)
```




```r
summary(tree_model)
```

```
## Call:
## rpart(formula = Ybin ~ ., data = subset_data, method = "class")
##   n= 4424
##
##           CP nsplit rel error    xerror       xstd
## 1 0.13354459      0 1.0000000 1.0475328 0.01503902
## 2 0.06066093      1 0.8664554 0.8664554 0.01491779
## 3 0.03983703      2 0.8057945 0.8325034 0.01483945
## 4 0.01131734      3 0.7659574 0.8044364 0.01476106
## 5 0.01086464      4 0.7546401 0.7781802 0.01467636
## 6 0.01000000      5 0.7437755 0.7686736 0.01464295
```

```
##
## Variable importance
##                 Debtor        Admission.grade    Mother.s.occupation
##                     55                     22                     15
## Previous.qualification
##                      7
##
## Node number 1: 4424 observations,    complexity param=0.1335446
##   predicted class=0  expected loss=0.4993219  P(node) =1
##     class counts:  2215  2209
##    probabilities: 0.501 0.499
##   left son=2 (503 obs) right son=3 (3921 obs)
##   Primary splits:
##       Debtor                 splits as  RL, improve=101.15390, (0 missing)
##       Admission.grade        < 112.25 to the left,  improve= 38.69350, (0 missing)
##       Previous.qualification splits as  RLLLLRLLLLRLRRRRR, improve= 32.51868, (0 missing)
##       Mother.s.occupation    splits as  LRRRRRRRLRLLLLRRLLLRLRRRRRLLRLRL, improve= 27.68553, (0 missi
##   Surrogate splits:
##       Mother.s.occupation splits as  RRRRRRRRRRRRRRRRRRRRRRRRLRRRRRR, agree=0.887, adj=0.002, (0 spl
##
## Node number 2: 503 observations
##   predicted class=0  expected loss=0.2007952  P(node) =0.113698
##     class counts:   402   101
##    probabilities: 0.799 0.201
##
## Node number 3: 3921 observations,    complexity param=0.06066093
##   predicted class=1  expected loss=0.462382  P(node) =0.886302
##     class counts:  1813  2108
##    probabilities: 0.462 0.538
##   left son=6 (514 obs) right son=7 (3407 obs)
##   Primary splits:
##       Admission.grade        < 112.25 to the left,  improve=33.37888, (0 missing)
##       Mother.s.occupation    splits as  LRRRRRRRLRLLLLLRLLLRRRRRR-LLRRRL, improve=32.60579, (0 missi
##       Previous.qualification splits as  RLLRLRLLLLRLRRRRR, improve=25.17284, (0 missing)
##   Surrogate splits:
##       Previous.qualification splits as  RRRRRRRRRLLRRRRRR, agree=0.870, adj=0.006, (0 split)
##       Mother.s.occupation    splits as  RRRRRRRRRRRRRRRRRLRRRRRRR-RRRRRR, agree=0.869, adj=0.004, (0
##
## Node number 6: 514 observations
##   predicted class=0  expected loss=0.3696498  P(node) =0.1161844
##     class counts:   324   190
##    probabilities: 0.630 0.370
##
## Node number 7: 3407 observations,    complexity param=0.03983703
##   predicted class=1  expected loss=0.4370414  P(node) =0.7701175
##     class counts:  1489  1918
##    probabilities: 0.437 0.563
##   left son=14 (558 obs) right son=15 (2849 obs)
##   Primary splits:
##       Mother.s.occupation    splits as  LLLRRRRRLRLLLLLRL-LLLLRLR-LLRRRL, improve=26.83909, (0 missi
##       Previous.qualification splits as  RLLLLRLLL--LLLRRR, improve=17.89895, (0 missing)
##       Admission.grade        < 130.25 to the left,  improve=10.55333, (0 missing)
##   Surrogate splits:
##       Previous.qualification splits as  RRRLRRLLR--RRRRRR, agree=0.839, adj=0.016, (0 split)
```

```
##       Admission.grade        < 181.75 to the right, agree=0.837, adj=0.007, (0 split)
##
## Node number 14: 558 observations
##   predicted class=0  expected loss=0.421147  P(node) =0.1261302
##     class counts:   323   235
##    probabilities: 0.579 0.421
##
## Node number 15: 2849 observations,    complexity param=0.01131734
##   predicted class=1  expected loss=0.4092664  P(node) =0.6439873
##     class counts:  1166  1683
##    probabilities: 0.409 0.591
##   left son=30 (345 obs) right son=31 (2504 obs)
##   Primary splits:
##       Previous.qualification splits as  RLLLLRL-L--LRLLRR, improve=12.655480, (0 missing)
##       Admission.grade        < 141.15 to the left,  improve= 9.300823, (0 missing)
##       Mother.s.occupation    splits as  ---LLRRL-R-----R------R-R---LRR-, improve= 1.548626, (0 miss
##   Surrogate splits:
##       Admission.grade < 179    to the right, agree=0.88, adj=0.006, (0 split)
##
## Node number 30: 345 observations,    complexity param=0.01086464
##   predicted class=0  expected loss=0.4637681  P(node) =0.07798373
##     class counts:   185   160
##    probabilities: 0.536 0.464
##   left son=60 (251 obs) right son=61 (94 obs)
##   Primary splits:
##       Admission.grade        < 141.35 to the left,  improve=6.940901, (0 missing)
##       Mother.s.occupation    splits as  ---LLLRL-L------------L-----RR--, improve=3.372679, (0 missi
##       Previous.qualification splits as  -LLRL-L-L--L-RR--, improve=1.687132, (0 missing)
##   Surrogate splits:
##       Mother.s.occupation splits as  ---LLLLL-L------------R-----RR--, agree=0.739, adj=0.043, (0 sp
##
## Node number 31: 2504 observations
##   predicted class=1  expected loss=0.3917732  P(node) =0.5660036
##     class counts:   981  1523
##    probabilities: 0.392 0.608
##
## Node number 60: 251 observations
##   predicted class=0  expected loss=0.4023904  P(node) =0.05673599
##     class counts:   150   101
##    probabilities: 0.598 0.402
##
## Node number 61: 94 observations
##   predicted class=1  expected loss=0.3723404  P(node) =0.02124774
##     class counts:    35    59
##    probabilities: 0.372 0.628
```

1. **Variables importantes :**

   - La variable la plus importante est “Debtor”, contribuant à 55 % de l’importance globale. Cela suggère que le fait d’avoir des dettes a un impact significatif sur la prédiction de la variable cible “Ybin” et donc sur la diplomation des étudiants.
   - Ensuite, “Admission.grade” contribue à 22 %, ce qui indique que les notes d’admission sont également un facteur crucial dans la prise de décision.

- "Mother.s.occupation" contribue à 15 %, ce qui suggère que le métier de la mère de l'étudiant joue un rôle important dans la prédiction.
- "Previous.qualification" contribue à 7 %, indiquant que le niveau d'éducation précédent a une influence moindre mais non négligeable.

2. **Nœuds de décision :**

   - Le nœud 3 (à gauche du nœud principal) utilise principalement la variable "Admission.grade" pour décider de la classification.
   - Le nœud 7 (à droite du nœud principal) utilise la variable "Mother.s.occupation" pour prendre des décisions.

3. **Feuilles :**

   - Les feuilles fournissent des probabilités de classe, par exemple, le nœud 2 (feuille) prédit la classe 0 (non diplomé) avec une probabilité de 0,799, indiquant une forte confiance dans la prédiction.
   - Le nœud 15 (feuille) prédit la classe 1 (diplômé) avec une probabilité de 0,591.

4. **Erreurs de classification :**

   - L'erreur de classification diminue à mesure que l'arbre progresse. Les feuilles ont des erreurs de classification spécifiques, par exemple, le nœud 15 a une erreur de 40,93 %.

5. **Variables de division :**

   - L'arbre utilise des variables spécifiques pour diviser les nœuds, telles que "Debtor", "Admission.grade", "Mother.s.occupation", et "Previous.qualification".

Maintenant, examinons les résultats de l'arbre de classification pour chaque variable une à une :

1. **Debtor (Endetté) :**

   - **Variable Importance :** 55
   - L'arbre commence par diviser les données en fonction de la variable "Debtor". La division la plus significative est basée sur le fait d'être endetté ou non.
   - **Feuille 2 (Endetté) :** Il y a 503 observations dans cette feuille, avec une probabilité de 20% de diplomation (Ybin = 0) et 80% de non-diplomation (Ybin = 1).
   - **Feuille 3 (Non-endetté) :** Il y a 3921 observations dans cette feuille, avec une probabilité de diplomation de 46% (Ybin = 1) et de non-diplomation de 54% (Ybin = 0).

2. **Admission.grade (Note d'admission) :**

   - **Variable Importance :** 22
   - L'arbre effectue une division supplémentaire en fonction de la note d'admission, en particulier, si la note d'admission est inférieure à 112.25.
   - **Feuille 6 (Note d'admission < 112.25) :** Il y a 514 observations dans cette feuille, avec une probabilité de diplomation de 37% (Ybin = 1) et de non-diplomation de 63% (Ybin = 0).
   - **Feuille 7 (Note d'admission >= 112.25) :** Il y a 3407 observations dans cette feuille, avec une probabilité de diplomation de 44% (Ybin = 1) et de non-diplomation de 56% (Ybin = 0).

3. **Mother's Occupation (Profession de la mère) :**

   - **Variable Importance :** 15

- L'arbre effectue une division basée sur la profession de la mère.
- **Feuille 14 (Profession de la mère catégorie 0, 1 ou 2) :** Il y a 558 observations dans cette feuille, avec une probabilité de diplomation de 42% (Ybin = 1) et de non-diplomation de 58% (Ybin = 0).
- **Feuille 15 (Profession de la mère catégorie 3 à 194) :** Il y a 2849 observations dans cette feuille, avec une probabilité de diplomation de 41% (Ybin = 1) et de non-diplomation de 59% (Ybin = 0).

4. **Previous qualification (Qualification précédente) :**

   - **Variable Importance :** 7
   - L'arbre effectue une division basée sur la qualification précédente.
   - **Feuille 30 (Qualification précédente catégorie 1 à 15) :** Il y a 345 observations dans cette feuille, avec une probabilité de diplomation de 46% (Ybin = 1) et de non-diplomation de 54% (Ybin = 0).
   - **Feuille 31 (Qualification précédente catégorie 19 à 43) :** Il y a 2504 observations dans cette feuille, avec une probabilité de diplomation de 39% (Ybin = 1) et de non-diplomation de 61% (Ybin = 0).

Ces résultats fournissent des informations sur la manière dont chaque variable influence la prédiction de diplomation. Par exemple, être endetté semble avoir un impact significatif, tout comme la note d'admission. La profession de la mère et la qualification précédente ont également un certain effet, bien que moins important.

En effet, l'accès à l'éducation supérieure représente souvent un défi financier important pour de nombreuses personnes, avec le coût annuel des études universitaires dans leur pays nécessitant parfois d'importants sacrifices financiers. Dans ce contexte, l'emprunt devient une nécessité incontournable pour certains individus cherchant à poursuivre des études supérieures.

Par ailleurs, un facteur tout aussi crucial, bien que moins prégnant que l'endettement financier, réside dans l'emploi de la mère. Appartenir à une catégorie sociale spécifique en raison de l'emploi des parents peut véritablement servir de levier pour les enfants. Que cet emploi fournisse à la famille des ressources financières supplémentaires ou qu'il requière un niveau d'études élevé, offrant ainsi à l'enfant un modèle de réussite scolaire et des conseils précieux, le statut social des parents exerce un impact significatif sur le parcours éducatif des enfants.

Il convient également de noter que des parents ayant suivi des études poussées peuvent créer un environnement familial plus structuré, avec une attention particulière portée à la scolarité de l'enfant. Cette dynamique familiale peut se traduire par un soutien accru, des attentes éducatives élevées et une sensibilisation accrue à l'importance de la réussite scolaire, contribuant ainsi de manière positive au développement académique de l'enfant.