# HISTOLOGY PREDICTION

Adam YIMBI MARC

3 Decembre 2019 Pi

L'objectif de cette data challenge est d'arriver à prédire l'histologie du cancer. Nous disposons des informations ci-apres:

age, sexe et l'histologie os_months, dead, dead_at_24_months, t, n, m, tnm_stage, tnm_grade génétiques : 1000 gènes Nous allons debuter par létude des variables.

## Statistiques descriptives

Trouver un modèle en mesure de prédire le sexe d'une série de patients. Data egale 546 observations et 1011 variables, dont 10 variables discrètes et 1001 continues. ous avons 766 NA, où la variable tnm_grade est complètement manquante. Les femmes, la catégorie TCGA-LUAD est majoritaire, TCGA-LUSC chez les hommes. La variable AAK1 est significativement différente dans les deux catégories d'historique. Il existe une forte corrélation entre certaines variables continues. Nous devons donc choisir certaines variables qui sont essentielles.

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.5.3
```

```
## -- Attaching packages ---------------------------------------------------
---------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
## Warning: package 'tibble' was built under R version 3.5.3
```

```
## Warning: package 'tidyr' was built under R version 3.5.3
```

```
## Warning: package 'readr' was built under R version 3.5.3
```

```
## Warning: package 'purrr' was built under R version 3.5.3
```

```
## Warning: package 'dplyr' was built under R version 3.5.3
```

```
## Warning: package 'stringr' was built under R version 3.5.3
```

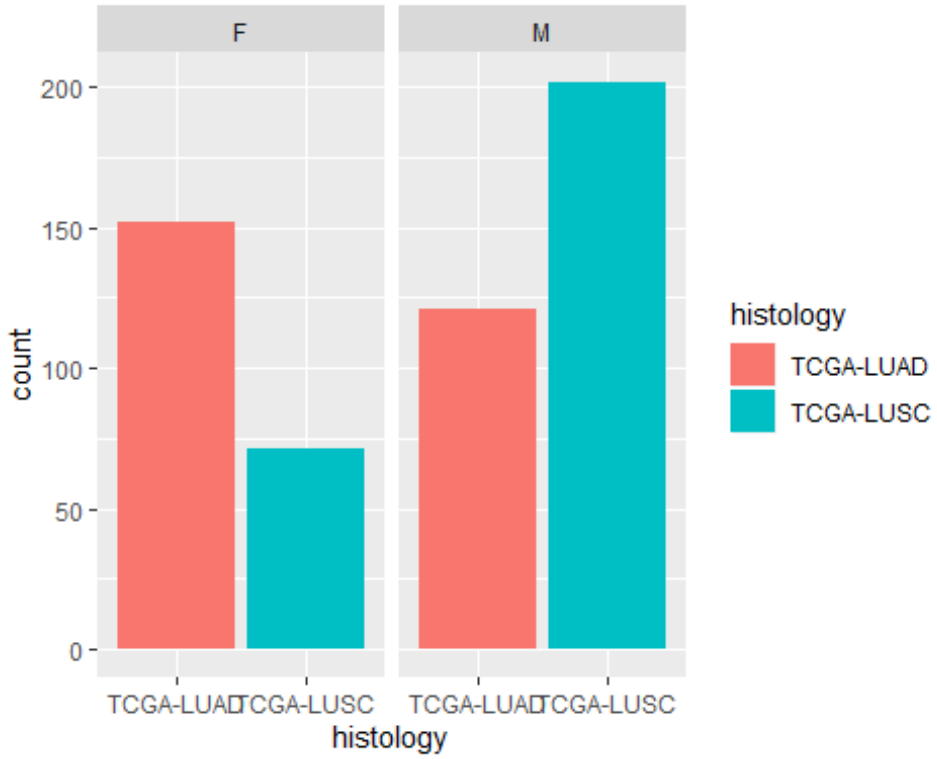
```
## Warning: package 'forcats' was built under R version 3.5.3
```

```
## -- Conflicts ------------------------------------------------------------
----------------- tidyverse_conflicts() --
```

```
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

data<-readRDS("C:/Users/ACER/Desktop/starting_kit_histology/starting_kit_histology/data_learn.rds")
dim(data)

## [1]  546 1012

table(data$histology)

##
## TCGA-LUAD TCGA-LUSC
##       273       273

str(data[,1:4])

## 'data.frame':    546 obs. of  4 variables:
##  $ age          : num  62.3 82 60.6 46.9 60 ...
##  $ sex          : chr  "F" "M" "M" "F" ...
##  $ tissue_status: chr  "tumoral" "tumoral" "tumoral" "tumoral" ...
##  $ histology    : chr  "TCGA-LUAD" "TCGA-LUAD" "TCGA-LUAD" "TCGA-LUAD" ...

table(is.na(data[,1:4]))

##
## FALSE
##  2184

anyNA(data)

## [1] TRUE

sum(is.na(data))

## [1] 766

sum(is.na(data[,13:1011]))

## [1] 0
```

Les données biologiques ne contiennent aucune donnée manquante. De plus, tous les tissus que nous observons sont tumoraux, Mais les données cliniques présentaient des valeurs manquantes pour différentes variables et nous avons procede a l'imputation de ce variables.Et nous pouvons nous interresser a la variable que nous allons predire.

```
ggplot(data)+
  geom_bar(aes(x=histology,fill=histology))+
  facet_wrap(.~sex)
```




```
table(is.na(data[,13:1012]))
```

```
##
##   FALSE
## 546000
```

## Méthodes

Nous allons prédire l'histologie du patient à partir d'autres informations en utilisant le modele de régression logistique.L'histologie etant binaire. Nous partirons par verifier la correlation entre les variables, pour enfin realiser no modeles, dont l'un sera base a partir du test de vraissemblance. Creer un modèle à une variable et ensuite récupéré celui dont le coefficient est relativement le plus significatif

Et nous allons finir par une méthode de sélection de variable avec un stepwise.

Avant tout, remplissons les valeurs manquantes. Ici nous utilisons le package 'mice'

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 3.5.3
```

```
## corrplot 0.84 loaded
```

```
library(mice)
```

```
## Warning: package 'mice' was built under R version 3.5.3
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'mice'

## The following object is masked from 'package:tidyr':
##
##     complete

## The following objects are masked from 'package:base':
##
##     cbind, rbind

for ( v in c("sex", "tissue_status", "histology", "os_months", "dead", "dead_at_
24_months", "t", "n", "m", "tnm_stage", "tnm_grade") ) {
    data[[ v ]] <- as.factor(x = data[[ v ]])
}

str(data)

## 'data.frame':    546 obs. of  1012 variables:
##  $ age              : num  62.3 82 60.6 46.9 60 ...
##  $ sex              : Factor w/ 2 levels "F","M": 1 2 2 1 2 1 1 2 2 1 ...
##  $ tissue_status    : Factor w/ 1 level "tumoral": 1 1 1 1 1 1 1 1 1 1 ...
##  $ histology        : Factor w/ 2 levels "TCGA-LUAD","TCGA-LUSC": 1 1 1 1 1 1
1 1 1 1 ...
##  $ os_months        : Factor w/ 467 levels "0","0.03","0.07",..: 203 211 244
236 231 195 64 261 56 28 ...
##  $ dead             : Factor w/ 2 levels "false","true": 1 1 1 1 2 1 2 1 1 1
...
##  $ dead_at_24_months: Factor w/ 2 levels "false","true": NA NA NA NA 2 NA 2 1
NA NA ...
##  $ t                : Factor w/ 9 levels "T1","T1a","T1b",..: 1 5 3 7 4 4 4 4
4 5 ...
##  $ n                : Factor w/ 5 levels "N0","N1","N2",..: 1 1 1 1 2 1 3 1 2
1 ...
##  $ m                : Factor w/ 5 levels "M0","M1","M1a",..: 1 1 1 5 1 1 1 1
1 1 ...
##  $ tnm_stage        : Factor w/ 10 levels "I","IA","IB",..: 2 3 2 6 6 3 8 3 6
3 ...
##  $ tnm_grade        : Factor w/ 0 levels: NA NA NA NA NA NA NA NA NA NA ...
##  $ MIR107           : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ CT45A10          : num  0 0 0 0 5.1 ...
##  $ ATP10B           : num  9.14 7.77 5.38 4.37 6.61 ...
##  $ TMEM134          : num  9.83 10.64 9.94 11.48 9.47 ...
##  $ EMC10            : num  12.7 12.6 12.7 12.3 13.2 ...
##  $ FAHD2B           : num  7.67 7.86 8.5 8.07 9.14 ...
##  $ QKI              : num  12.6 12.4 12 12.4 11.7 ...
##  $ ZDHHC24          : num  10.23 10.5 10.29 10.27 9.58 ...
##  $ GFI1B            : num  1.87 3.34 4.7 0 1.59 ...
##  $ BMP1             : num  11.4 12.2 10.3 11 10.9 ...
##  $ MIR6826          : num  1.87 1.41 1.03 0 0 ...
##  $ ATP1A1-AS1       : num  6.16 5.02 7.73 6.55 6.5 ...
##  $ COX14            : num  10.7 10.6 10.4 10.9 10.8 ...
##  $ TADA2A           : num  9.57 10.34 10.19 9.06 8.97 ...
```

```
##  $ KLRC3            : num  2.32 1.41 0 2.64 2.33 ...
##  $ SGF29            : num  8.51 9.22 9.13 8.33 9.1 ...
##  $ TBC1D31          : num  8.1 8.78 7.63 8.96 9.46 ...
##  $ KIF2C            : num  9.99 9.29 5.89 11.54 10.99 ...
##  $ SNORD115-41      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ TRG-AS1          : num  7.97 7.97 6.1 5.97 4.65 ...
##  $ TAF2             : num  10.5 10.1 10.2 10.9 10.6 ...
##  $ MIR6719          : num  0 1.41 1.03 1.21 0 ...
##  $ ZNF561-AS1       : num  6.6 7.4 7.99 8.37 6.98 ...
##  $ CTD-2350J17.1    : num  0 0 0 0 0 ...
##  $ UBE2G1           : num  10.9 10.8 11 11.3 11.6 ...
##  $ NT5DC2           : num  10.8 11.4 11.6 11.4 10.2 ...
##  $ TNFRSF14-AS1     : num  7.18 9.44 7.53 7.79 5.14 ...
##  $ SLC4A1AP         : num  10.8 10.4 10.4 10.7 11.3 ...
##  $ MIR1272          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ BCO1             : num  7.34 6.43 4.52 7.18 7.95 ...
##  $ ZNF792           : num  8.14 7.76 8.36 7.54 8.49 ...
##  $ LINC01777        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ SDHAF4           : num  7.88 9.15 8.67 9.47 8.15 ...
##  $ MIR483           : num  0 1.41 0 0 0 ...
##  $ RPL28            : num  14.7 14.1 14.4 15 14.9 ...
##  $ CEBPE            : num  2.94 3.83 3.22 1.21 3.18 ...
##  $ MINDY4B          : num  1.22 1.8 1.03 0 1.59 ...
##  $ STK17B           : num  12.5 11.8 10.9 11.2 11.2 ...
##  $ CHEK2            : num  8.75 8.78 8.2 10.4 10.69 ...
##  $ CHRNB1           : num  8.48 8.64 8.83 9.24 9.7 ...
##  $ MIR4425          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ POTEG            : num  0 0 0 1.21 2.01 ...
##  $ KCNQ4            : num  5.26 6.65 5.61 8.08 3.47 ...
##  $ CARD6            : num  10.8 9.12 9.5 8.87 9.58 ...
##  $ LOC100506178     : num  5.8 4.74 3.22 6.84 3.71 ...
##  $ LLGL2            : num  12.8 11.6 12.8 12.1 12.8 ...
##  $ PDZK1IP1         : num  14.37 12.07 9.98 10.53 9.7 ...
##  $ TENM3            : num  9.69 8.11 9.21 7.96 9.1 ...
##  $ LOC100996447     : num  0 0 1.63 0 0 ...
##  $ CLDN17           : num  1.22 0 0 0 0 ...
##  $ SLC28A3          : num  8.05 8.33 8.95 7.76 7.42 ...
##  $ MIR30D           : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ MOCS1            : num  8.39 9.94 9.74 7.17 8.06 ...
##  $ SMIM11B          : num  0 0.871 1.031 0 1.005 ...
##  $ LINC01179        : num  0 0 0 1.85 3.47 ...
##  $ LOC101929353     : num  0 0 0 0 2.01 ...
##  $ LOC105376671     : num  3.17 2.11 3.51 1.85 4.65 ...
##  $ IFITM10          : num  10.37 10 7.69 7.84 8.39 ...
##  $ MIR3136          : num  1.22 1.8 0 0 0 ...
##  $ ARRDC5           : num  5.49 4.95 4.7 3.67 3.18 ...
##  $ HORMAD2          : num  2.32 3.92 0 5.9 2.01 ...
##  $ SULT1C2P1        : num  2.32 3.92 4.64 0 1.01 ...
##  $ BLM              : num  8.77 8.25 6.29 8.94 10.16 ...
##  $ FRG2C            : num  0 0 1.03 0 0 ...
##  $ CHMP6            : num  9.97 10.36 10.13 10.07 9.63 ...
##  $ ZCCHC14          : num  11.1 11 11.3 11.3 10.9 ...
```

```
##  $ DNAJB13          : num  7.18 7.86 9.07 4.17 7.91 ...
##  $ CCL28            : num  8.44 8.43 8.92 9.76 9.87 ...
##  $ PTCH1            : num  8.72 9.35 12.08 8.9 8.57 ...
##  $ ART4             : num  5.96 6.18 6.01 3.52 2.59 ...
##  $ FER              : num  9.48 9.75 10.31 9.13 9.92 ...
##  $ MIR8071-1        : num  4.19 6.04 2.64 0 1.59 ...
##  $ ANKRD28          : num  11.2 10.9 11 11.7 11 ...
##  $ TIAF1            : num  5.9 8.02 6.95 6.38 5.82 ...
##  $ PRAMEF36P        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ RBMX2            : num  10.06 9.67 9.5 11.01 10.02 ...
##  $ SPATA2           : num  10.22 9.79 9.8 10.24 9.82 ...
##  $ DCAF13P3         : num  1.22 0.871 1.031 1.207 2.594 ...
##  $ EFCAB7           : num  8.42 8.09 9.11 7.98 8.25 ...
##  $ PHIP             : num  11.4 11 12 11.1 11.6 ...
##  $ PXMP2            : num  8.79 8.12 8.38 8.69 8.06 ...
##  $ TMEFF2           : num  2.66 1.8 4.45 0 1.59 ...
##  $ LGALS3           : num  13.5 12.8 13.2 14.6 14 ...
##  $ MARCH11          : num  0 0 4.58 0 0 ...
##  $ YBX3P1           : num  2.32 2.36 1.03 2.91 4.1 ...
##  $ NCAM2            : num  7.81 8.19 9.13 4.96 7.3 ...
##  $ GSC2             : num  0 0 0 0 0 0 0 0 0 0 ...
##   [list output truncated]
```

```
colnames(data)<-str_replace_all(colnames(data),c('[^A-Za-z0-9]'=''))

colnames(data[ ,1:11])
```

```
##  [1] "age"            "sex"            "tissuestatus"   "histology"
##  [5] "osmonths"       "dead"           "deadat24months" "t"
##  [9] "n"              "m"              "tnmstage"
```

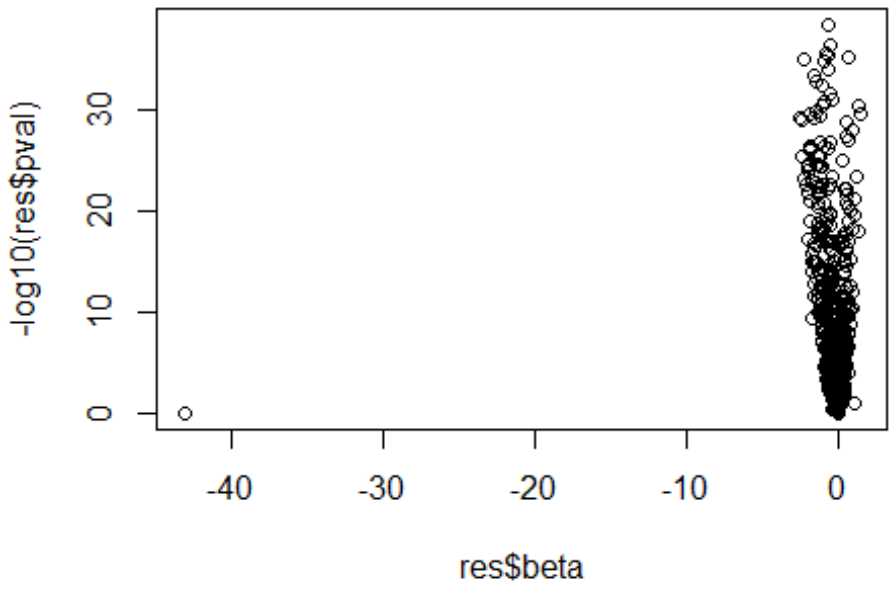
## Evaluation de l'évolution du score

On va creer des modèles avec une variable pour un debut, et conserver la pvaleur du test de la variable concernée. Pour enfin comparer la significativité des coefficients.

```
# Récupération des noms de gènes
gs <- names(data)[13:1012]

# Variable
histology <- as.numeric(data$histology == "TCGA-LUAD")

# Calcul des pvaleurs des coefficients pour les gènes
res <- sapply(gs, function(g) {
  m = glm(c(histology,0,0,1,1)~c(data[[g]],0,max(data[,gs]),0,max(data[,gs])),
    family = binomial(logit))
  b = m$coefficients[[2]]
  pv = summary(m)$coefficients[2,4]
  c(pval = pv,beta = b)
})
```

```
## Warning: glm.fit: algorithm did not converge
```

```r
# Calcul des pvaleur des coefficients pour les auters variables
res_bis <- sapply(c(c(1,2),c(5:11)), function(g) {
  m = glm(histology~data[[g]],
    family = binomial(logit))
  b = m$coefficients[[2]]
  pv = summary(m)$coefficients[2,4]
  c(pval = pv,beta = b)
})

# Mise en forme de toutes les pvaleurs
colnames(res_bis) <- colnames(data[,c(c(1,2),c(5:11))])
res <- cbind(res,res_bis)
res <- t(res)
res <- as.data.frame(res)
plot(res$beta,-log10(res$pval))
```




```r
head(res)
```

```
##                  pval         beta
## MIR107  9.644035e-01 -0.003028164
## CT45A10 3.353450e-12 -0.262419324
## ATP10B  2.530902e-01  0.042087232
## TMEM134 1.751987e-06 -0.624654516
## EMC10   7.866760e-01 -0.020734017
## FAHD2B  6.254652e-06 -0.453096992
```

Etant donne le nombre de genes est considerable, il est impossible de savoir sur le graphique quel gène sélectionner. Beaucoup de gènes semblent avoir des P-valeur très faibles. Il semble

raisonnable en se basant sur ce volcan plot de choisir les critères pour sélectionner les gènes les plus pertinents.

```
# rownames(res)[order(res$pval)][1]

# Méthode avec la meilleure variable 0.8
mod_1 <- glm(
    formula = histology ~ LINC02428,
    data = data,
    family = binomial(link = 'logit')
)
imp<-mice(data[ ,1:11],method = "rf")

##
##  iter imp variable
##   1   1  deadat24months  n  m  tnmstage
##   1   2  deadat24months  n  m  tnmstage
##   1   3  deadat24months  n  m  tnmstage
##   1   4  deadat24months  n  m  tnmstage
##   1   5  deadat24months  n  m  tnmstage
##   2   1  deadat24months  n  m  tnmstage
##   2   2  deadat24months  n  m  tnmstage
##   2   3  deadat24months  n  m  tnmstage
##   2   4  deadat24months  n  m  tnmstage
##   2   5  deadat24months  n  m  tnmstage
##   3   1  deadat24months  n  m  tnmstage
##   3   2  deadat24months  n  m  tnmstage
##   3   3  deadat24months  n  m  tnmstage
##   3   4  deadat24months  n  m  tnmstage
##   3   5  deadat24months  n  m  tnmstage
##   4   1  deadat24months  n  m  tnmstage
##   4   2  deadat24months  n  m  tnmstage
##   4   3  deadat24months  n  m  tnmstage
##   4   4  deadat24months  n  m  tnmstage
##   4   5  deadat24months  n  m  tnmstage
##   5   1  deadat24months  n  m  tnmstage
##   5   2  deadat24months  n  m  tnmstage
##   5   3  deadat24months  n  m  tnmstage
##   5   4  deadat24months  n  m  tnmstage
##   5   5  deadat24months  n  m  tnmstage

## Warning: Number of logged events: 106
```

Nous avons sélectionne uniquement la variable admettant une faible p-valeur.

```
output<-mice::complete(imp)

data[ ,1:11]<-output
anyNA(data)

## [1] TRUE
```

```
data=data[ ,-12]
anyNA(data)

## [1] FALSE
```

Apres remplissage de valeurs manquantes, nous avons faits la selection de variables importantes avec le package Boruta.Au final, nous avons 138 variables significatives avec lesquelles nous allons creer notre modele. Voici quelques unes “TMEM134”, “TBC1D31” ,“KIF2C”, “TAF2”, “TMEM134”,“TBC1D31”,“TAF2”,“CHEK2”,“BCO1”,“KIF2C”,“UBE2G1”,“CLPSL2”, “TAB2”, “MEPCE”, “SERPINB2”, “EXOSC5”, “ASCC2”,“KPNA2”,...

```
library(Boruta)

## Warning: package 'Boruta' was built under R version 3.5.3

## Loading required package: ranger

## Warning: package 'ranger' was built under R version 3.5.3

set.seed(5)
boruta<-Boruta(histology~.,data=data,doTrace=2)

##  1. run of importance source...

##  2. run of importance source...

##  3. run of importance source...

##  4. run of importance source...

##  5. run of importance source...

##  6. run of importance source...

##  7. run of importance source...

##  8. run of importance source...

##  9. run of importance source...

##  10. run of importance source...

##  11. run of importance source...

##  12. run of importance source...

##  13. run of importance source...

##  14. run of importance source...

##  15. run of importance source...

##  16. run of importance source...

##  17. run of importance source...
```

```
## After 17 iterations, +6 mins:
##  confirmed 80 attributes: ARHGAP23, BBOX1AS1, BLM, CACNA1D, CAPN5 and 75 more
;
##  rejected 647 attributes: ABCB5, ABTB1, ABTB2, ACKR4, ACOX3 and 642 more;
##  still have 283 attributes left.
##  18. run of importance source...
##  19. run of importance source...
##  20. run of importance source...
##  21. run of importance source...
##  22. run of importance source...
## After 22 iterations, +6.7 mins:
##  confirmed 2 attributes: CCAT1, SEPT10;
##  rejected 62 attributes: AADACL3, ALKBH5, ARL2, ASB9, ATL2 and 57 more;
##  still have 219 attributes left.
##  23. run of importance source...
##  24. run of importance source...
##  25. run of importance source...
## After 25 iterations, +7 mins:
##  confirmed 9 attributes: AAK1, ATP6V1A, DEPTOR, GSC2, PINLYP and 4 more;
##  rejected 36 attributes: ADAM10, ADAM29, ALCAM, ANKRD28, ANKRD35 and 31 more;
##  still have 174 attributes left.
##  26. run of importance source...
##  27. run of importance source...
##  28. run of importance source...
##  29. run of importance source...
## After 29 iterations, +7.3 mins:
##  confirmed 6 attributes: EIPR1, EXOSC5, KPNA2, PIM1, RAB35 and 1 more;
##  rejected 15 attributes: AOAHIT1, CRY1, FRY, GALNT4, GGT1 and 10 more;
##  still have 153 attributes left.
##  30. run of importance source...
```

```
##  31. run of importance source...
##  32. run of importance source...
## After 32 iterations, +7.5 mins:
##  confirmed 2 attributes: CDAN1, ERCC3;
##  rejected 11 attributes: ALPG, BEND3, C12orf75, CHRNB1, ESF1 and 6 more;
##  still have 140 attributes left.
##  33. run of importance source...
##  34. run of importance source...
##  35. run of importance source...
##  36. run of importance source...
## After 36 iterations, +7.8 mins:
##  rejected 7 attributes: CCND2AS1, EPO, FZD4, NT3, SEMA3B and 2 more;
##  still have 133 attributes left.
##  37. run of importance source...
##  38. run of importance source...
##  39. run of importance source...
## After 39 iterations, +8 mins:
##  confirmed 2 attributes: CIAPIN1, HORMAD2;
##  rejected 8 attributes: ADAM20, COL22A1, CYSTM1, PIWIL3, RAD21AS1 and 3 more;
##  still have 123 attributes left.
##  40. run of importance source...
##  41. run of importance source...
##  42. run of importance source...
## After 42 iterations, +8.2 mins:
##  rejected 3 attributes: ABCG1, LINC00574, RBMX2;
##  still have 120 attributes left.
##  43. run of importance source...
##  44. run of importance source...
##  45. run of importance source...
## After 45 iterations, +8.4 mins:
```

```
##  rejected 5 attributes: ARHGAP28, EFCAB7, FLGAS1, GSTM1, MAPKAPK2;
##  still have 115 attributes left.
##  46. run of importance source...
##  47. run of importance source...
##  48. run of importance source...
## After 48 iterations, +8.6 mins:
##  confirmed 3 attributes: NCL, RHBDD1, TSPAN13;
##  rejected 3 attributes: RRP1B, SULT4A1, ZNF74;
##  still have 109 attributes left.
##  49. run of importance source...
##  50. run of importance source...
##  51. run of importance source...
## After 51 iterations, +8.8 mins:
##  confirmed 3 attributes: AGBL5, SPRYD4, TMEM134;
##  rejected 3 attributes: GBP5, KMT2EAS1, OTUD1;
##  still have 103 attributes left.
##  52. run of importance source...
##  53. run of importance source...
##  54. run of importance source...
## After 54 iterations, +9 mins:
##  confirmed 1 attribute: LCE2B;
##  rejected 1 attribute: BIRC2;
##  still have 101 attributes left.
##  55. run of importance source...
##  56. run of importance source...
##  57. run of importance source...
## After 57 iterations, +9.2 mins:
##  rejected 1 attribute: QKI;
##  still have 100 attributes left.
##  58. run of importance source...
```

```
##  59. run of importance source...
## After 59 iterations, +9.4 mins:
##  confirmed 1 attribute: CPNE1;
##  rejected 2 attributes: CASP6, LINC01770;
##  still have 97 attributes left.
##  60. run of importance source...
##  61. run of importance source...
##  62. run of importance source...
## After 62 iterations, +9.6 mins:
##  rejected 3 attributes: FHOD3, NR1D1, RFK;
##  still have 94 attributes left.
##  63. run of importance source...
##  64. run of importance source...
##  65. run of importance source...
##  66. run of importance source...
##  67. run of importance source...
##  68. run of importance source...
## After 68 iterations, +10 mins:
##  confirmed 1 attribute: SLC44A1;
##  rejected 1 attribute: KRT33B;
##  still have 92 attributes left.
##  69. run of importance source...
##  70. run of importance source...
## After 70 iterations, +10 mins:
##  confirmed 1 attribute: WIPI1;
##  still have 91 attributes left.
##  71. run of importance source...
##  72. run of importance source...
##  73. run of importance source...
## After 73 iterations, +10 mins:
```

```
##  confirmed 1 attribute: TAB2;
##  rejected 2 attributes: GSTM4, PRKAB2;
##  still have 88 attributes left.
##  74. run of importance source...
##  75. run of importance source...
##  76. run of importance source...
## After 76 iterations, +10 mins:
##  confirmed 1 attribute: RPUSD4;
##  rejected 3 attributes: ACP4, DAPK1, NUDCD1;
##  still have 84 attributes left.
##  77. run of importance source...
##  78. run of importance source...
## After 78 iterations, +11 mins:
##  rejected 2 attributes: AP5B1, LOC101927136;
##  still have 82 attributes left.
##  79. run of importance source...
##  80. run of importance source...
##  81. run of importance source...
##  82. run of importance source...
##  83. run of importance source...
##  84. run of importance source...
## After 84 iterations, +11 mins:
##  rejected 1 attribute: DCAF13P3;
##  still have 81 attributes left.
##  85. run of importance source...
##  86. run of importance source...
## After 86 iterations, +11 mins:
##  rejected 1 attribute: PDZRN3;
##  still have 80 attributes left.
##  87. run of importance source...
```

```
##  88. run of importance source...

##  89. run of importance source...

## After 89 iterations, +11 mins:

##  confirmed 1 attribute: DAW1;

##  still have 79 attributes left.

##  90. run of importance source...

##  91. run of importance source...

##  92. run of importance source...

##  93. run of importance source...

##  94. run of importance source...

## After 94 iterations, +11 mins:

##  rejected 1 attribute: WSB1;

##  still have 78 attributes left.

##  95. run of importance source...

##  96. run of importance source...

##  97. run of importance source...

## After 97 iterations, +12 mins:

##  confirmed 1 attribute: HNF4A;

##  still have 77 attributes left.

##  98. run of importance source...

##  99. run of importance source...

## After 99 iterations, +12 mins:

##  confirmed 1 attribute: EDAR;

##  still have 76 attributes left.

print(Boruta)

## function (x, ...)
## UseMethod("Boruta")
## <bytecode: 0x0000000014abf980>
## <environment: namespace:Boruta>

final.boruta<-TentativeRoughFix(boruta)
print(final.boruta)
```

```
## Boruta performed 99 iterations in 11.76633 mins.
## Tentatives roughfixed over the last 99 iterations.
##  140 attributes confirmed important: AAK1, ABHD17C, AGBL5, ALG1L,
## ARHGAP23 and 135 more;
##  870 attributes confirmed unimportant: AADACL3, ABCB5, ABCG1, ABTB1,
## ABTB2 and 865 more;
```

```
selected<-getSelectedAttributes(final.boruta,withTentative = F)

dftrain<-data[c("histology",selected)]
```

Pour la construire de notre modèle de régression logistique, nous allons utiliser le package caret, ainsi que la cross validationpour estimer l'erreur generale. Et ce modele sera constitue de variables importantes selectionnees par Boruta.

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.5.3
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
for ( v in c("sex", "tissue_status", "histology", "os_months", "dead", "dead_at_
24_months", "t", "n", "m", "tnm_stage", "tnm_grade") ) {
  }
    data[[ v ]] <- (x = data[[ v ]])

    test <- readRDS("C:/Users/ACER/Desktop/starting_kit_histology/starting_kit_h
istology/data_learn.rds")

for ( v in c("sex", "tissue_status", "histology", "os_months", "dead", "dead_at_
24_months", "t", "n", "m", "tnm_stage", "tnm_grade") ){
}
    test[[ v ]] <- as.factor(x = test[[ v ]])

model <- glm(
    formula = histology ~ age + TMEM134+TBC1D31+TAF2+CHEK2+BCO1+KIF2C+UBE2G1+CLP
SL2+TAB2+MEPCE+SERPINB2+EXOSC5+ASCC2+KPNA2
  , data = data
  , family = binomial(link = 'logit')
)
summary(object = model)
```

```
##
## Call:
## glm(formula = histology ~ age + TMEM134 + TBC1D31 + TAF2 + CHEK2 +
##     BCO1 + KIF2C + UBE2G1 + CLPSL2 + TAB2 + MEPCE + SERPINB2 +
##     EXOSC5 + ASCC2 + KPNA2, family = binomial(link = "logit"),
##     data = data)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3959  -0.1481  -0.0027   0.0947   2.2512
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -137.10977   17.35661  -7.900 2.80e-15 ***
## age           -0.00195    0.02557  -0.076  0.93922
## TMEM134        1.98803    0.37996   5.232 1.67e-07 ***
## TBC1D31        1.13452    0.49793   2.278  0.02270 *
## TAF2           0.07578    0.55283   0.137  0.89097
## CHEK2          0.55369    0.36139   1.532  0.12549
## BCO1          -0.05192    0.10304  -0.504  0.61436
## KIF2C         -0.52147    0.35130  -1.484  0.13770
## UBE2G1         2.00920    0.50860   3.950 7.80e-05 ***
## CLPSL2        -0.20545    0.12009  -1.711  0.08714 .
## TAB2           3.18923    0.62519   5.101 3.37e-07 ***
## MEPCE          1.10954    0.35769   3.102  0.00192 **
## SERPINB2       0.48719    0.09034   5.393 6.94e-08 ***
## EXOSC5         0.77128    0.36852   2.093  0.03636 *
## ASCC2          1.21053    0.37933   3.191  0.00142 **
## KPNA2          0.51744    0.39108   1.323  0.18580
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 756.92  on 545  degrees of freedom
## Residual deviance: 159.87  on 530  degrees of freedom
## AIC: 191.87
##
## Number of Fisher Scoring iterations: 8

pred <- predict.glm(object = model, newdata = test, type = "response")
idx  <- pred <= 0.5
pred[  idx ] <- levels(x = data$histology)[ 1 ]
pred[ !idx ] <- levels(x = data$histology)[ 2 ]
table(pred, useNA = "ifany")

## pred
## TCGA-LUAD TCGA-LUSC
##       276       270
```

Nous retenons enfin ce modele, car il nous permet d'obtenir un score de 0.95 sur Codalab. Mais nous pouvons encore creer d'autres modeles avec nos autres variables, et peut etre nous pourrions encore ameliorer notre score.

## Conclusion

Nous remarquons que le gène LINC01503 est tres important, car il s'associe à la présence d'un carcinome épidermoïde. Le gène EXOSC5 est associé à des dystrophies. Ces maladies sont observable dans le cas de carcinome épidermoïde du poumon.

Et le gène ABTB1 est présent dans différents cancers ce qui semble normal quand on sait que la protéine pour laquelle il code fait partie de la grande famille des suppresseurs de tumeurs.