PROJET – MODÈLES LINÉAIRES ET SES GÉNÉRALISATIONS :
PRÉDICTION DU DIABÈTE PAR UN MODÈLE DE RÉGRESSION LOGISTIQUE

Ismaël Bendib – Paul Caillere – Adrien Passuello – Axel Sauvaget

Janvier 2023

▷ Introduction

Le but de cette étude est de déterminer si des variables cliniques peuvent être utilisées pour prédire le diagnostic de diabète (variable à expliquer : `class`). Les données utilisées dans cette analyse proviennent de deux échantillons suivant : `Diabetes_train.csv` et `Diabetes_test.csv`

▷ Partie I - Analyse empirique descriptive des données

Avant de construire un modèle de prédiction, une analyse empirique descriptive des données a été effectuée. Il n'a pas été trouvé de lien significatif entre l'âge des patients et le diagnostic de diabète. Les variables `class` et `Gender` ont été converties en 0 et 1 pour être utilisées dans un modèle linéaire généralisé. Les autres variables ont également été converties en 0 et 1 pour une analyse empirique.

◇ Analyse des corrélations

Une analyse de la corrélation entre les variables a été effectuée. Les facteurs `Polyuria` et `Polydipsia` sont fortement corrélés entre eux et avec la variable `class`, et le genre (variable `Gender`) est négativement corrélé avec la variable `class`.

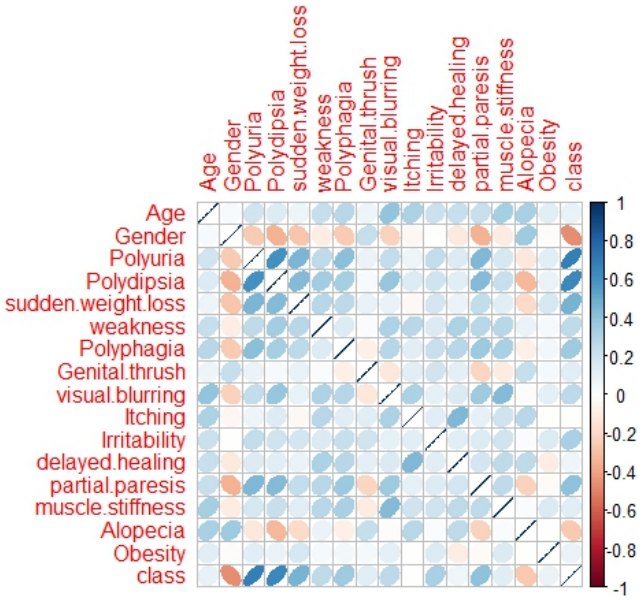

FIGURE 1 – Représentation graphique de la matrice de corrélation entre les variables

▷ Partie II - Sélection et validation du modèle

On pratique une régression logistique pour expliquer la variable `class` (patient diabétique ou non) en utilisant toutes les autres variables disponibles. Seules les variables `Age`, `Gender`, `Polyuria`, `Polydipsia`, `Genital.thrush`, `Itching` et `Irritability` sont significatives pour prédire la variable réponse `class`. Des méthodes ont été utilisées pour enlever des variables pour affiner le modèle de régression logistique.

Dans cette étude, nous avons utilisé la fonction `step()` de R pour sélectionner les variables les plus pertinentes pour notre modèle de régression logistique. Nous avons utilisé les méthodes `forward`, `backward` et `both` pour sélectionner les variables, en comparant les modèles obtenus selon le critère `AIC`.

Les résultats obtenus montrent que les variables les plus pertinentes pour expliquer la variable réponse `class` sont : `Polyuria`, `Gender`, `Polydipsia`, `Irritability`, `Itching`, `Genital.thrush`, `Age`, `visual.blurring`, weakness et partial.paresis. Ce résultat est identique pour les trois méthodes de sélection utilisées.

Nous avons également ajouté des interactions entre certaines variables qui semblaient pertinentes, comme entre `Polyuria` (quantités importantes d'urine) et `Polydipsia` (soif excessive), entre l'âge (`Age`) et la `visual.blurring` (vision floue) et entre `weakness` (faiblesses) et toutes les variables liées à des maladies en général.

Après avoir réaliser une `anova` de type I et une `Anova` de type III. On en déduit le modèle final suivant qui semble le mieux expliquer la variable réponse `class` :

```
modf <- glm(class ~ Gender + Irritability + Polyuria + Polydipsia  + Itching +
        (Genital.thrush + visual.blurring) * weakness + Age:visual.blurring, data=train,
         family='binomial')
```

Nous allons maintenant faire des prédictions de la variable d'intérêt, évaluer la performance de ce modèle par rapport au modèle complet `mod_complet` et interpréter les résultats obtenus.

```
mod_complet <- glm(class ~., data=train, family='binomial')
```

▷ Partie III - Prédiction de la variable d'intérêt et évaluation du modèle sur les données

Pour évaluer la performance de notre modèle par rapport au modèle complet `mod_complet`, nous allons utiliser (sur les deux échantillons `train` et `test`) plusieurs métriques de comparaison telles que l'erreur quadratique moyenne (MSE) et des matrices de confusion. Nous allons également utiliser des courbes ROC (Receiver Operating Characteristic) pour visualiser les performances de notre modèle en fonction de différents paliers de classification puis calculer les aires sous ces courbes (AUC).

Voici un tableau récapitulatif des résultats obtenus sur `R` :

ERREUR QUADRATIQUE MOYENNE (MSE)

Modèle \ Échantillon	train	test
modf	1,029489	0,6927533
mod_complet	1,085621	0,6234274

MATRICES DE CONFUSION (PALIER DE CLASSIFICATION : 0,1)

modf

Classif. \ Préd.	0	1
0	108	52
1	2	254

Classif. \ Préd.	0	1
0	27	13
1	2	62

mod_complet

Classif. \ Préd.	0	1
0	121	39
1	4	252

Classif. \ Préd.	0	1
0	29	11
1	2	62

AIRE SOUS LA COURBE ROC (AUC)

Modèle \ Échantillon	train	test
modf	0,9871826	0,9552734
mod_complet	0,9842041	0,9667969

FIGURE 2 – Tableau récapitulatif des résultats R

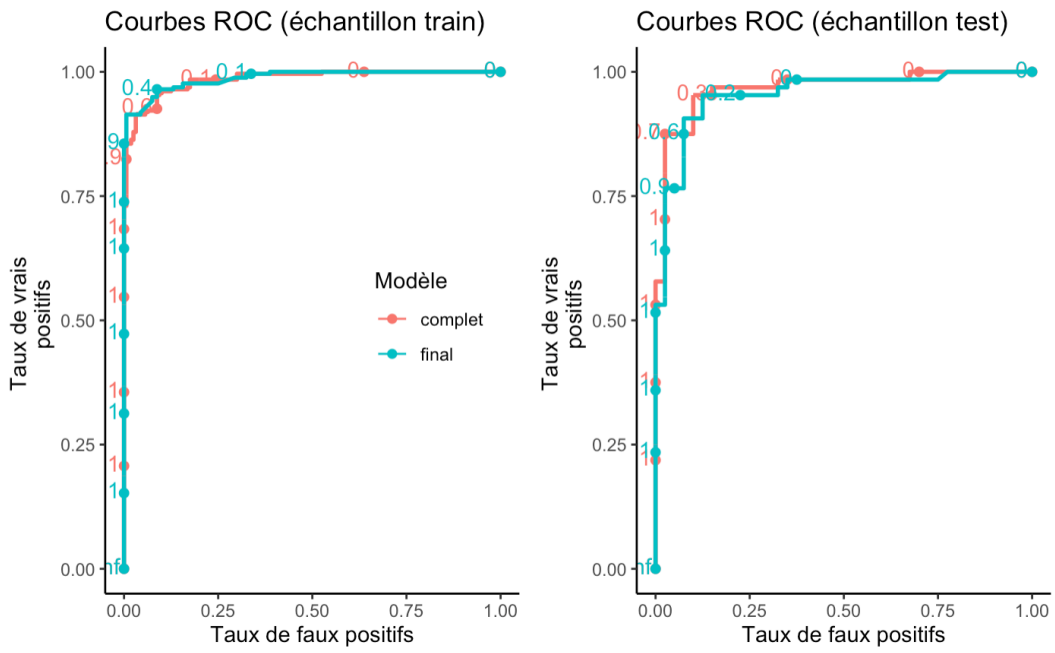

FIGURE 3 – Représentation graphique simultanée des courbes ROC selon l'échantillon et le modèle

▷ Conclusion

Pour le modèle final `modf` déterminé dans la partie II et pour le modèle complet, les résultats de prédiction sont à peu près les mêmes. Puisque le modèle final `modf` a une Residual Deviance plus petite que celle du modèle complet, le modèle final `modf` semble être un modèle pratique pour réaliser les prédictions sans avoir trop de faux négatifs mais possède un taux de faux positifs un peu trop conséquent.

Critiques :

- ◇ Certaines variables des tableaux de données ne sont pas significatives alors qu'elles devraient à priori jouer un rôle dans le diagnostic d'après des études cliniques (exemples : obésité, polyphagie...).
- ◇ Il ne semble pas y avoir de différences significatives entre le modèle final et le modèle complet en terme de prédiction malgré le fait que le modèle complet présente de nombreuses variables non significatives.
- ◇ Pour améliorer le modèle, on pourrait ajouter des variables quantitatives aux données comme par exemple la glycémie des patients.
- ◇ De plus, on ne sait pas de quel type de diabète il s'agit (1 ou 2). Le diabète de type 1 est une maladie immunitaire avec comme syndrome cardinal la polyurodipsie, l'amaigrissement et la polyphagie tandis que le diabète de type 2 est une maladie métabolique dans laquelle l'obésité est un facteur de risque très important. D'après les résultats, il semblerait donc qu'il s'agit ici d'une étude sur le diabète de type 1.