**Comparez les perfomances des méthodes CART et régression logistique**

Muyao GUO

## 1. Objectif

L'objectif de ce rapport est d'analyser les facteurs influençant le poids des nouveau-nés (inférieur ou supérieur à 2,5 kg) à l'aide de deux méthodes statistiques: l'arbre de décision (CART) et la régression logistique.
Ces deux approches seront comparées en termes de performance prédictive, notamment à travers le taux d'erreur et la courbe ROC, afin d'évaluer la précision et la capacité discriminante de chaque modèle.

## 2. Source des données

Le jeu de données *birthwt* contient 189 observations et 10 variables. Après un traitement préalable et une sélection des variables pertinentes, l'analyse a conservé les variables suivantes :

- low : indicateur du poids de naissance inférieur à 2,5 kg.(1 = low<2.5kg, 0 = low > 2.5kg)
- age : âge de la mère (en années).
- lwt : poids de la mère (en livres) lors de sa dernière période menstruelle.
- race : origine ethnique de la mère (1 = blanche, 2 = noire, 3 = autre).
- smoke : statut tabagique pendant la grossesse (1 = fumeuse, 0 = non fumeuse).
- ptd : nombre d'accouchements prématurés antérieurs.
- ht : antécédent d'hypertension (1 = oui, 0 = non).
- ui : présence d'irritabilité utérine (1 = oui, 0 = non).
- ftv : nombre de consultations médicales au premier trimestre.

## 3. Régression logistique

Dans cette étude, la régression logistique a été utilisée pour modéliser la variable *low*. Ce choix méthodologique se justifie par plusieurs raisons. Premièrement, la variable *low* est binaire (0 = poids normal, 1 = poids < 2,5 kg). La régression logistique, adaptée aux réponses binomiales, permet de relier la probabilité d'un événement à un ensemble de variables explicatives par le biais de la fonction logit. Deuxièmement, les coefficients du modèle peuvent être transformés en odds ratios, offrant une interprétation directe de l'effet de chaque facteur maternel sur la probabilité de faible poids de naissance.
Dans le cadre de cette étude, toutes les variables explicatives disponibles dans le jeu de données *birthwt* ont été intégrées dans le modèle de régression logistique.
L'objectif est d'évaluer l'effet simultané des caractéristiques maternelles sur la probabilité d'un faible poids de naissance. Le modèle estimé peut s'écrire sous la forme suivante :

$$\text{logit}\big(P(low = 1|x)\big) = \beta_0 + \beta_1 \times age + \beta_2 \times lwt + \beta_3 \times race + \beta_4 \times smoke + \beta_5 \times ptd + \beta_6 \times ht + \beta_7 \times ui + \beta_8 \times ftv$$

où $P(low = 1|x)$ représente la probabilité qu'un nouveau-né présente un poids inférieur à 2,5 kg, et $\beta_i$ désigne le coefficient associé à chaque variable explicative.

### 3.1 Erreur test

Afin d'évaluer la performance prédictive des modèles, le jeu de données *birthwt* a été divisé en deux sous-ensembles : Un jeu d'apprentissage (70%), composé de 129 observations, utilisé pour l'entraînement du modèle ; Un jeu de test (30%), composé de 60 observations, réservé à l'évaluation de la précision des prédictions.

Le modèle de régression logistique a d'abord été ajusté sur le jeu d'apprentissage. Ensuite, les performances du modèle ont été mesurées sur le jeu de test à l'aide du taux d'erreur (error test), qui permet d'évaluer la proportion de classifications incorrectes.
Il est important de souligner que les prédictions n'ont pas été réalisées sur le jeu d'apprentissage, car cela aurait conduit à une surestimation artificielle de la performance du modèle. En effet, un modèle testé sur les mêmes données que celles ayant servi à son ajustement tend à mémoriser les observations plutôt qu'à apprendre les relations générales (phénomène de *surapprentissage*).

| Reference | Low>2,5kg | Low<2,5kg |
|---|---|---|
| **Prédiction** | | |
| **Low>2.5kg** | 35 | 15 |
| **Low<2,5kg** | 6 | 4 |

Table 3.1 : Matrice de confusion du modèle de régression logistique sur le jeu de test.

Le modèle de régression logistique a correctement classé 39 observations sur 60 (*Table 3.1*), soit un taux d'erreur de 35 %. La spécificité (0.70) indique que le modèle identifie correctement la majorité des naissances de poids normal. En revanche, la sensibilité (0.40) reste plus faible, ce qui signifie que le modèle a plus de difficulté à détecter les cas de faible poids de naissance.
En résumé, le modèle tend à sous-estimer le risque de faible poids de naissance : il prédit bien les cas normaux, mais il manque une partie des cas positifs (Low < 2.5 kg).

## 3.2 Validation croisée du modèle de régression logistique

Il est important de noter que l'évaluation de la performance d'un modèle uniquement à partir d'un jeu de test unique peut conduire à une estimation instable du taux d'erreur, en raison de la variabilité liée à la partition aléatoire des données.
La validation croisée permet de réduire cette variance en répétant le processus d'évaluation sur plusieurs sous-échantillons indépendants. Plus précisément, une validation croisée à 10 plis (10-fold cross-validation) a été mise en œuvre sur le jeu de données *birthwt*. Le taux d'erreur moyen obtenu au terme de cette procédure est de 30,6 %. Ce résultat indique que, bien que la régression logistique parvienne à distinguer partiellement les cas de faible poids de naissance, sa performance prédictive reste modérée et laisse une marge d'amélioration.

# 4. CART

Pour étudier le même problème avec une approche non paramétrique, nous appliquons la méthode CART. Dans chaque nœud terminal, on estime la probabilité $P(low = 1|x)$ et l'observation est affectée à la classe dont la probabilité est la plus élevée. Le jeu de données *birthwt* a été séparé selon la même procédure que précédemment.

## 4.1 Construction de l'arbre maximal et sélection du paramètre de complexité

Dans un premier temps, un arbre de classification maximal a été construit à partir du jeu d'apprentissage (129 observations), en utilisant la fonction rpart() avec un paramètre cp = 0 et minsplit = 3, de manière à ne pas limiter la croissance de l'arbre.
L'arbre maximal (*Figure S1*) obtenu à l'étape initiale permet une séparation quasi parfaite des observations du jeu d'apprentissage. Cependant, ce modèle présente un risque important de surapprentissage (overfitting). Pour limiter ce phénomène, il est nécessaire d'introduire un paramètre de complexité (cp), qui pénalise la croissance excessive de l'arbre.
La *Figure S2* présente la courbe d'évolution de l'erreur relative de validation croisée (xerror) en fonction du paramètre de complexité (cp). La ligne pointillée horizontale indique la règle du "1-SE", minimum xerreur + écart type.
Dans notre cas, en raison du faible effectif du jeu d'apprentissage, la courbe cp–xerror n'affiche pas une forme parfaitement régulière, ce qui reflète une certaine variabilité statistique. Néanmoins,

en appliquant la règle du 1-SE, le premier cp inférieur à la ligne de référence a été retenu, soit cp = 0.029, pour réaliser le pruning (élagage) de l'arbre maximal.

### 4.2 Pruning

Après la sélection du paramètre de complexité optimal (cp = 0.029), l'arbre maximal a été élagué afin de réduire sa complexité et d'améliorer sa capacité de généralisation.
Le modèle obtenu conserve les divisions les plus pertinentes, tout en supprimant les branches redondantes ou peu informatives (*Figure S3*).

### 4.3 Évaluation du modèle CART sur le jeu de test et validation croisée

L'exactitude globale (Accuracy) du modèle est de 66.7 %**,** avec un intervalle de confiance de 95 % compris entre 0.53 et 0.78. La sensibilité (taux de détection correcte des cas *Low < 2.5 kg*) est de 0.41, tandis que la spécificité (capacité à identifier les cas normaux) atteint 0.77**.** Ces résultats indiquent que le modèle CART présente une bonne capacité de reconnaissance des cas normaux**,** mais reste moins performant pour détecter les cas de faible poids de naissance**,** probablement en raison du déséquilibre de classes et du nombre limité d'observations positives**.** Le taux d'erreur global s'élève ainsi à environ 33 %, ce qui reste cohérent avec la taille de l'échantillon et la nature bruitée des variables socio-cliniques.
Comme pour la régression logistique, une validation croisée à 10 plis a été effectuée afin d'évaluer la performance du modèle CART sur l'ensemble du jeu de données birthwt. Les données ont été réparties aléatoirement en dix sous-ensembles : à chaque itération, neuf sous-ensembles ont été utilisés pour l'apprentissage du modèle, et le dernier pour le test. Le taux d'erreur a été calculé à chaque pli, puis moyenné sur l'ensemble des dix itérations. Le taux d'erreur moyen obtenu est de 35.53 %, ce qui correspond à une précision moyenne d'environ 64.5 %.

| Reference | Low>2,5kg | Low<2,5kg |
|---|---|---|
| **Prédiction** | | |
| **Low>2.5kg** | 33 | 10 |
| **Low<2,5kg** | 10 | 7 |

Table 4.3 : Matrice de confusion du modèle de CART sur le jeu de test.

## 5. La courbe ROC : Régression Logistique vs CART

La *figure S4* présente la comparaison des courbes ROC des deux modèles (régression logistique en bleu et CART en rouge) appliqués aux données *birthwt*. L'axe des abscisses correspond au taux de faux positifs (FPR = 1 – spécificité) et l'axe des ordonnées au taux de vrais positifs (TPR = sensibilité). La diagonale pointillée représente le modèle aléatoire (absence de pouvoir discriminant).
On observe que la courbe bleu (logistique) se situe globalement au-dessus de la courbe rough (CART), ce qui indique une meilleure capacité de discrimination pour le modèle logistique sur ce jeu de test. Autrement dit, pour un même taux de faux positifs, le modèle logistique obtient en moyenne une sensibilité plus élevée. Cependant, les deux courbes restent relativement proches de la diagonale, ce qui suggère que la performance globale de classification reste modérée.

## 6. Conclusion

Dans cette étude, deux approches de classification ont été appliquées au jeu de données *birthwt*: la régression logistique et la méthode CART. Les résultats montrent que la régression logistique

présente une légère supériorité en performance globale, avec un taux d'erreur moyen d'environ 31% obtenu par validation croisée, contre 33 % pour le modèle CART testé sur le jeu de test.
Cependant, les deux modèles présentent une sensibilité limitée dans la détection des cas de faible poids de naissance, principalement à cause du déséquilibre de classes et du nombre restreint d'observations positives**.**
L'analyse des courbes ROC confirme ces résultats : la courbe de la régression logistique (en rouge) se situe globalement au-dessus de celle du CART (en bleu), indiquant une meilleure capacité discriminante du modèle logistique.
En conclusion, la régression logistique apparaît plus performante en termes de précision moyenne, tandis que CART constitue un outil complémentaire particulièrement utile pour visualiser la structure décisionnelle et identifier les facteurs de risque dominants.

## 7. Améliorations des modèles

Pour la régression logistique, nous pourrions améliorer la modélisation en ajustant des paramètres plus adaptés et en renforçant la validation par une validation croisée avec un nombre de plis plus élevé (par ex. $k = 20$, voire validation croisée répétée). Une telle procédure réduit la variance de l'estimation des performances et permet une sélection de modèle plus robuste.
Pour CART, compte tenu de la taille d'échantillon limitée, il est utilisable de réaliser une forêt aléatoire (Random Forest), qui agrège de nombreuses arbres bootstrapés et réduit la variance par bagging. Cette approche offre en général une meilleure capacité de généralisation, fournit une estimation out-of-bag (OOB) de l'erreur et des importances de variables plus stables, ce qui est particulièrement pertinent dans notre contexte de petit échantillon.

## 8. Annexes

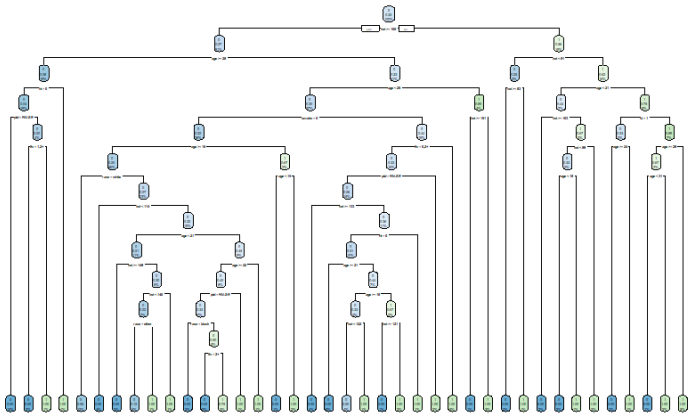

*Figure S1: Arbre maximum*

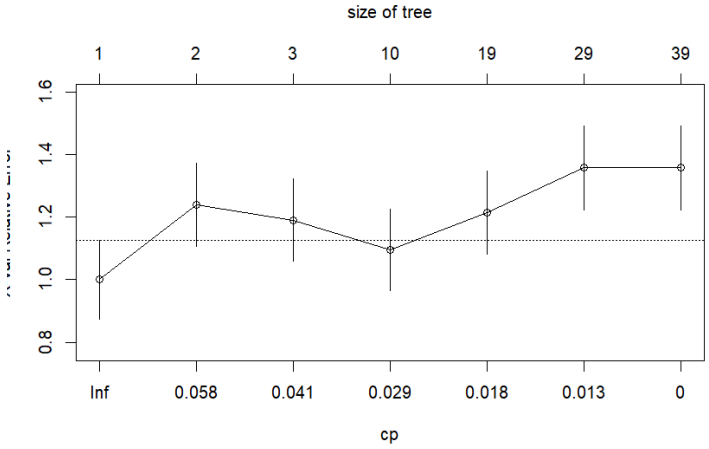


*Figure S2: Graphique des cp*

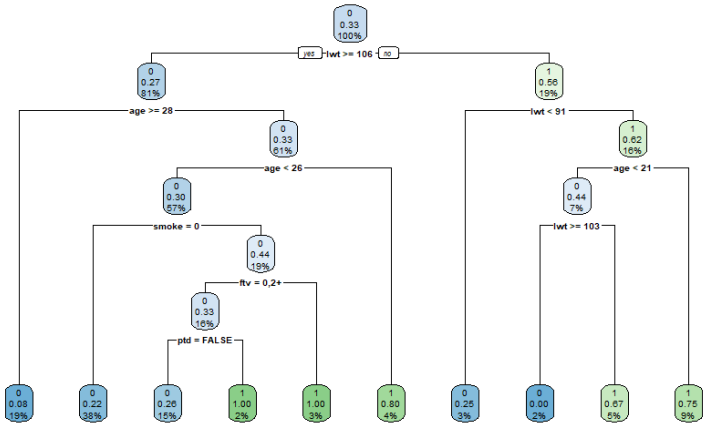


*Figure S3: Arbre après pruning*

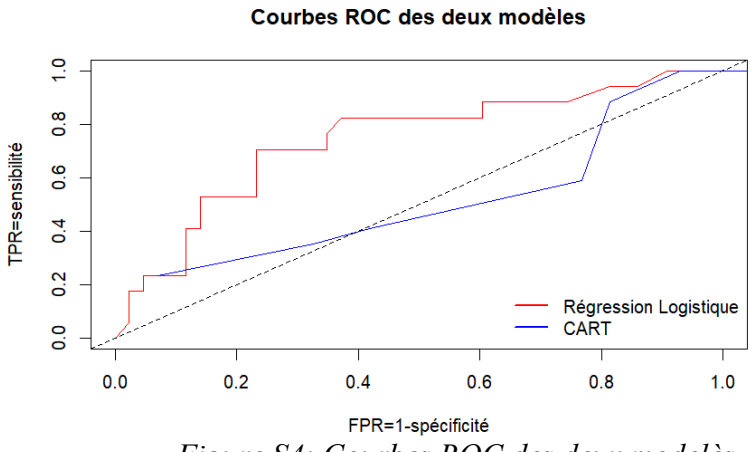


*Figure S4: Courbes ROC des deux modelès*