ROCHETTE LEVEQUE

# Rapport MERR

Nous avons pour objectif de prédire si un patient va mourir ou non (à savoir la variable `hospdead`) à partir du dataset.

## Préparation des données

### Variables numériques

Nous avons utilisé ce site afin d'obtenir des valeurs moyennes pour les variables `pafi`, `alb`, `bun`, `urine`. Nous avons rempli les cellules manquantes pour ces colonnes avec ces valeurs.

Pour les autres variables numériques manquantes, nous avons rempli les cellules par la médianne de ces valeurs.

### Variables non numériques

Pour les variables non numériques qui comportent des cellules manquantes, à savoir `income`, `race`, `dnr` et `sfdm2` nous avons :

- `income` : supprimé la colonne, car le format est difficilement traitable ('under 11k') et possédant 2982 valeurs manquantes, soit presque un tiers du nombre d'observations.
- `race` et `dnr` : supprimé les lignes comportant des valeurs manquantes pour ces variables, car à elles deux il n'y a que 72 lignes à supprimer.
- `sfdm2` : supprimé temporairement la colonne, car il y a 1400 valeurs manquantes, et que cette colonne nécessite un traitement spécifique (variable catégorielle).

## Régression logistique

Nous avons effectué une régression à partir du dataset ainsi préparé. Cette première régression donne un AIC de 63555. Cet AIC étant assez élevé, nous avons affiché la matrice de confusion associée :

| **Actual \ Predicted** | **0** | **1** |
|---|---|---|
| 0 | 5263 | 791 |
| 1 | 89 | 1503 |

Nous avons une précision de 88%. Néanmoins, cette bonne précision est relative, car le nombre de faux positifs est très important (plus que le nombre de faux négatifs), or c'est le nombre de faux positifs qui est le plus préjudiciable dans ce contexte.

Nous avons donc essayé d'améliorer le modèle en effectuant une sélection stepwise, qui nous a permis de refaire une régression logistique en ne prenant que les variables données par la sélection stepwise.

Nous avons obtenu un AIC très bas, à savoir 76. De plus, le modèle donne une précision de 100%, ce qui résulte d'un problème d'overfitting, car le dataset n'est pas séparé.

## Séparation

Nous avons donc séparé le dataset en `train` et `test`, avec respectivement 70% et 30% des observations. Après avoir rééffectué une régression logistique avec les variables données par la sélection stepwise, sur le dataset `train`, nous avons obtenu un AIC de 43112. C'est une amélioration notable par rapport à la précédente régression (sans stepwise).

Après avoir testé le modèle sur le dataset `test`, nous avons obtenu cette matrice de confusion :

| **Actual \ Predicted** | **0** | **1** |
|---|---|---|
| 0 | 1544 | 43 |
| 1 | 225 | 482 |

On peut observer qu'il y a de manière générale plus de vivants que de morts.

De plus, on cherche à minimiser le nombre de faux positifs, car un faux positif revient à considérer un patient vivant comme mort, ce qui est plus préjudiciable qu'un faux négatif.

## Pondération

Nous allons donc pondérer notre dataset, afin d'accorder un poids plus important aux observations de morts, car il y a de manière générale plus de vivants que de mort.

Pour calculer ce poids, on calcule les proportions relatives des deux états (vivant/mort), puis on calcule l'inverse de la proportion pour les morts. Enfin on multiplie cela par la proportion de vivants. C'est le poids que nous allons attribuer aux observations de morts.

Pour garder les mêmes proportions que dans le dataset complet, il convient de faire une partition respectant ces proportion à l'aide de la bibliothèque `caret` et de la fonction `createDataPartition`.

En effectuant la régression logistique sur le dataset pondéré, nous obtenons un AIC de 47149. Ce résultat est plus élevé que le précédent AIC, ce qui est attendu, car on sacrifie de la précision théorique, pour une précision plus réaliste.

## Prochaines optimisations

Nous projetons d'améliorer le modèle, en rajoutant `sfdm2` et `income`. Nous allons essayer de traiter les valeurs pour les rendre utilisables par la régression. Le but est de déterminer si elles sont plus bénéfiques au modèle de prendre en compte les deux variables (en retirant les lignes manquantes) ou de les supprimer.

Nous prévoyons aussi d'effectuer différentes modifications sur le dataset afin d'ajouter plus de variables, et de voir si cela améliore le modèle.