

Rapport Projet 1 - Données cliniques cancer du sein

Candice AUBERTIN - Ewen EXPUESTO (Binôme n°2)

24 Novembre 2025

1 Introduction

La survie des patientes atteintes d'un cancer du sein dépend d'interactions entre leurs données cliniques et caractéristiques génomiques. Dans ce rapport, nous allons aborder l'analyse de ces données et ainsi comprendre quels modèles utiliser pour prédire la variable ('vital_status').

2 Présentation du jeu de données

2.1 Données

Le jeu de données consiste en deux tableaux de données pour 1 231 patientes :

- **Expression génique (GeneX)** : matrice 1231×5000 . Les lignes correspondent aux patients avec le nom de chaque ligne l'identifiant du patient comme 'EW-A2FS-01A-11R-A17B'. Les colonnes correspondent aux gènes avec chaque nom de colonne un gène comme 'CLEC3A', qui sont les 5000 gènes les plus variables d'une personne à l'autre. Chaque case correspond à la valeur d'expression (un flottant) d'un certain gène (colonne) pour un certain patient (ligne) -> variables continues
- **Profil clinique (clinical_data)** : matrice 1231×24 . Les lignes correspondent aux noms des patients comme pour 'GeneX'. Les colonnes sont les noms des données cliniques (genre, âge au moment du diagnostic, morphologie et 'vital_status'...). Chaque case correspond à la valeur de la donnée clinique pour chaque patient, donc ce peut être un entier ou une chaîne de caractères, un booléen, une liste ou un numeric.

2.2 Variable cible : vital_status

La variable binaire `vital_status` encode l'issue 'Alive' ou 'Dead'.

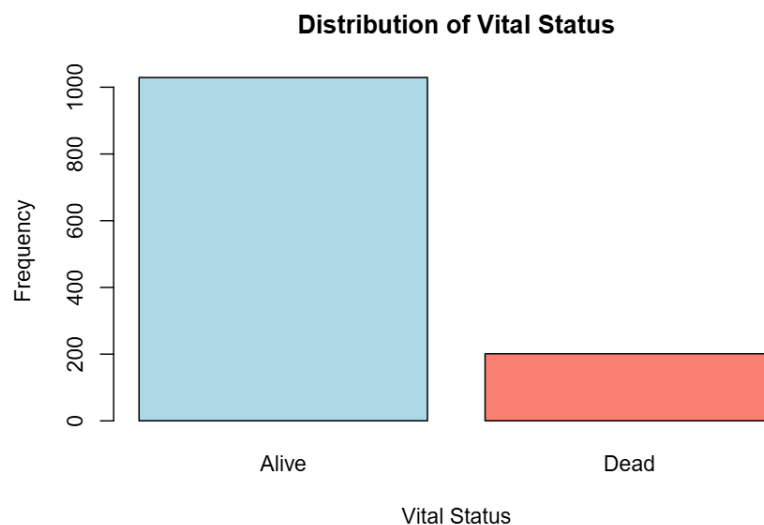


Figure 1: Distribution de la variable `vital_status`

3 Problématique

Nous cherchons à déterminer quelles caractéristiques cliniques et de profils d'expression génique améliore la prédiction du statut vital et quantifier cela. Ainsi, nos principales interrogations sont :

- Quelles variables sont utiles pour prédire 'vital_status' ?
- Quels modèles utiliser (kNN, k-means, régression logistique, lasso, group lasso, ridge, elastic net) ?
- Quelle fiabilité est associée à chaque modèle ?
- Quelle interprétabilité est associée à notre modèle prédictif ?

La variable cible est une variable binaire. Il y a deux plus deux tableaux de données qui sont de natures différentes. Le tableau de gènes est très large donc il y a potentiellement beaucoup de gènes superflus. C'est pour cela que nous utiliserons en premier lieu une régression LASSO logistique ou elastic net logistique pour éliminer les composantes qui n'aident pas à prédire. Dans le tableau clinique, il y a des variables qualitatives et quantitatives. Nous pouvons transformer les variables qualitatives en variables quantitatives (one-hot encoding) pour pouvoir les utiliser dans un unique modèle de classification comme la régression logistique RIDGE voire groupe RIDGE, et aussi kNN et k-means.

4 Premiers pas de notre étude

Nous avons construit un premier modèle de régression logistique en utilisant uniquement les données cliniques les plus pertinentes : l'âge au diagnostic, le poids initial du patient, le stade de la tumeur et la morphologie du patient.

La matrice de confusion présentée ci-dessous permet d'évaluer la capacité de ce modèle à prédire le `vital_status`. Le modèle parvient globalement à identifier correctement la majorité des patientes.

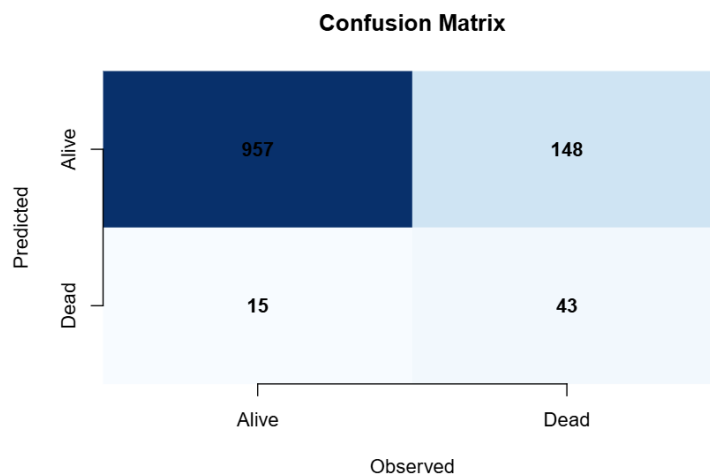


Figure 2: Matrice de confusion du modèle simple

Cette première étape met en évidence que les variables cliniques apportent une information utile, mais on va chercher à obtenir une meilleure performance prédictive.