1. **Casse-tête n°1 :**

- On ne garde qu’un seul client par famille dans la base (on retire les autres clients de sa famille)

Avec cette réponse, on supprime les liens d’influences entre familles, mais le dataset perdra trop d’information selon moi. Ce n’est pas une méthode que j’utiliserai. De plus les liens entre familles peuvent être des informations importantes.

- On ne garde qu’un seul client de chaque famille dans train, et un dans test

Avec cette réponse on fait l’hypothèse que tous les membres d’une même famille se comportent de la même manière. Ce n’est pas une méthode que j’utiliserai.

- On split train-test sur les clients (un split train-test normal)

Avec cette réponse, on suppose que les membres de chaque famille agissent de manière indépendante. Je testerai cette méthode mais ce n’est pas le premier choix que je ferai.

- On split train-test sur les familles (chaque famille est soit intégralement dans train, soit intégralement dans test)

Avec cette réponse, on suppose que toutes les familles agissent de la même manière. Ce n’est pas une méthode que j’utiliserai.

- On doit régulariser fortement le modèle

C’est cette méthode que je choisirai un premier. En effet, on ne connaissant pas l’impact et l’influence que génère le fait d’être dans une même famille. Pour éviter un overfiting, je modifierai les données en faisant une régularisation.

- On n’est pas obligé de régulariser fortement le modèle

Si en faisant des stats, on observe que les membres d’une même famille ne s’influencent pas entre eux. C’est une bonne option. On peut alors appliquer cette réponse ainsi que la 3è.

1. **Casse-tête n°2 :**

Selon moi plusieurs facteurs expliquent cette différence de résultats :

* Les données d’input : si les données utilisées sont différentes, il est évident que les résultats le seront aussi sachant que le niveau de santé est sujet à beaucoup de facteurs (âge, antécédents médicaux, sexe, lieu de résidence, lieu d’origine, salaire, statuts social…)
* Les méthodes de soin + la temporalité des études : imaginons que la population soit identique, le cancer a différentes phases. Les méthodes par médicament utilisées par la clinique A sont peut-être plus efficaces au début de la maladie. Inversement pour la clinique B.