Réunion Scrum - Sprint 2

- **/T (fausse le dtype de certaines variables) = supprimer le motif ? Redemander à la réunion Scrum pour être sûr**
  - ok
- **NaN = qu'en faire ? Les supprimer, les remplacer (par la moyenne ou la médiane par ex) ou autre ? Demander au client**
  - Toujours la Médiane mais être sûr que ce soit nécessaire

Dans le dataset CKD, les valeurs NaN représentent 60.5% du jeu de données, est-ce vraiment intelligent de les supprimer ? Quelle serait la meilleure stratégie à adopter ?

Médiane

- **Supprimer, remplacer, transformer (transformation logarithmique) ou segmenter les données en plusieurs ensembles homogène pour atténuer l'impact des valeurs aberrantes ? Demander au client**
- **Pertinence à travailler avec l'IQR (écart interquartiles) ? Demander au client**
  - Oui
- **Que faire des 0 quand il ne s'agit pas de valeurs binaires ou ordinales ? Demander au client**
  -
- **Comment savoir à quoi correspondent chacune des valeurs binaires et ordinales (cf 'su' (sucre) et 'al' (albumine) dans le dataset CKD) ? Demander au client**
  - C'était à poser avant

Dans le dataset cancer breast, on a 13 lignes avec des valeurs histopatho égales à 0 (sur 569), à supprimer ?

Oui

Dans le dataset CKD, 3 variables sont concernées : id = 1 valeur à 0, ce qui n'est pas gênant en soit / 'su' (sucre) et 'al' (albumine) qui semblent être des

catégories allant de 0 à 5. À quoi correspondent ces catégories ? Que signifient-elles ?

- **Si doublons, est-ce qu'on les supprime ? Demander au client**
  - Oui
- **Pertinent de faire les analyses en distinguant la cohorte (malade/seins) ? Demander au client**
  - Oui

CKD

- HHH

Le Diabète

- Dans le traitement médiacale on prend toujours la Médiane et non pas la Moyenne. Comprendre les 0. Est ce une erreur ?
- Glucose / Sucre / Insuline — à voir
- Que sont les 0 et quelles sont leur signification ?
- Attention bien regarder où sont sur les 0 ? ou NULL ? Y-en-a-t-il ? Combien de lignes, et les enlever, ce sont des patients non concluant.

Maladies Cardiaques

La Maladie Rénale Chronique

NaN erreur de typing, on ne va amputer aucune des variables !!!

Il y a deux valeurs qui ne devraient pas exister

Faire une df avec leur signification et la liste des valeurs uniques

Maladies du foie

Le Cancer du Sein

Quand on nous donne une variable diagnostique, c’est en double aveugle.

CHEST PAIN → échelle 0(min) 3(max)

ELECTRO CARDIO →

FLOROSCOPIE → PAS OUTLIER !! (Juste des gens spéciaux)

BIAIS FEMME / COMMENT ON PEUT REGLER CA POUR LE DATASET

Modèle qui vient d’Inde

Diabète (2 types) → fortement corrélée, est

Il faudra un seul NoteBook

Les OUTLIERS sont ok sauf les erreurs