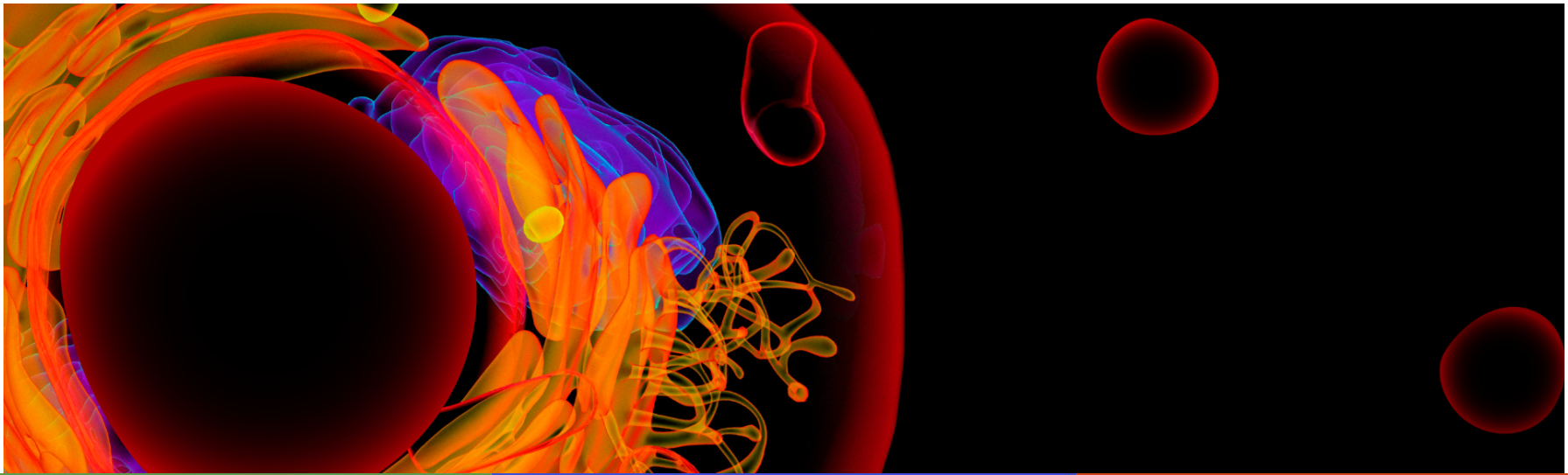
Classification du risque de cancer du col de l'utérus

Magistère d'économiste statisticien

Charlotte Bredy-Maux & Kayané Robach

20 juillet 2022



Introduction

- Le cancer du col de l'utérus est la principale tumeur gynécologique maligne dans le monde [Juneja et al., 2003]
- Il en existe plusieurs types, on se focalise sur le carcinome de l'épithélium pavimenteux:
  - représente 3/4 des cancers du col de l'utérus dans le monde
  - 2ème le plus commun
  - 3ème le plus mortel chez les humains dotés d'un utérus
- Première cause du cancer du col de l'utérus: infection au papilloma virus
- Plus ce cancer est détecté tôt, plus il y a de chance de survie
- L'objectif de notre travail est de développer un modèle de prédiction pour le résultat du test de biopsie

- C'est un challenge tiré d'un travail posté sur kaggle il y a 3 ans

Les données

Les 838 patientes sont identifiées et décrites via les 34 variables:

- comportement sexuel, [Liu, 2015]
- statut tabagique,
- contraceptif,
- IST,
- examens médicaux

La patiente moyenne a **26 ans**, a eu **2.5 partenaires sexuels** dans sa vie, a eu son **premier rapport à 17 ans**. Elle a entre **2 et 3 enfants**, a **fumé pendant 1.21 ans**, **fume 0.45 packets par an**. Elle suit une **contraception hormonale sur 2.35 ans**, a eu **très peu voir pas d'IST** et a reçu **peu d'examens médicaux**.
On a 98% de patientes saines contre 2% de patientes atteintes.

La classification sur des données désequilibrées

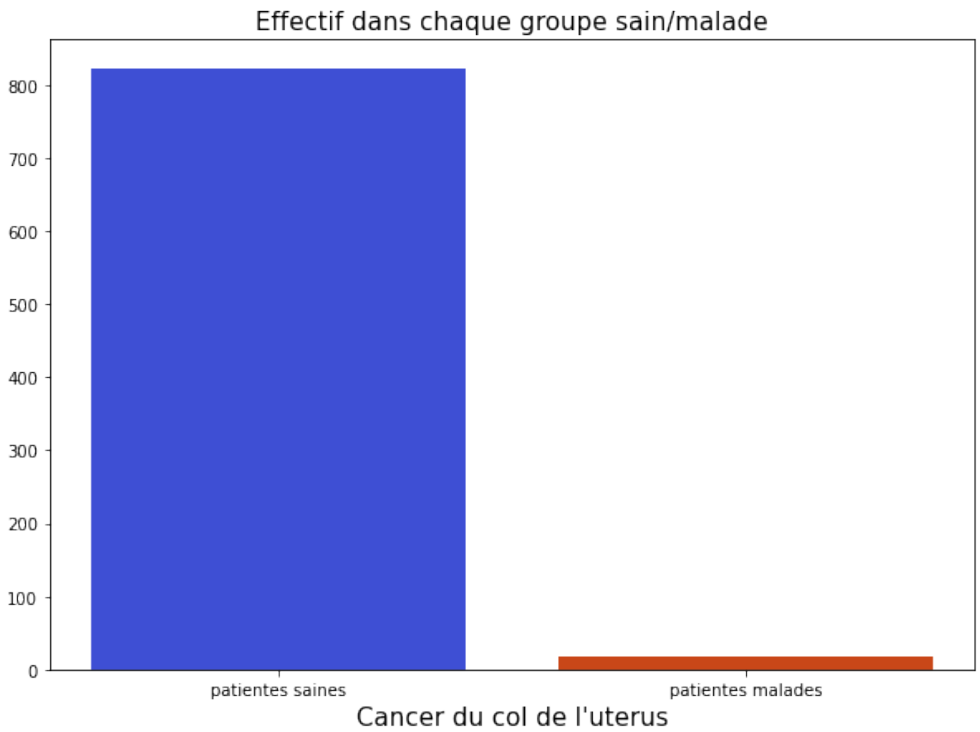


Figure: Désequilibre entre les effectifs de population atteinte ou saine

Pour pallier à ce problème nous allons générer des données synthétiques à partir de l'algorihtm de Gibbs. Cette méthode utilise les distributions conditionnelles entre variables explicatives corrélées.

Premiere selection sur les variables explicatives

Nous avons 3 variables concernant le tabagisme:

- oui/non
- nombre de paquets par an
- nombre d'années de tabagisme

Nous commencons l'étude avec le nombre d'années de tabagisme mais il sera intéressant de regarder le nombre de paquets par an car il existe un effet dose (fumer depuis 2 ans une cartouche dans l'année ou une cartouche par semaine c'est différent).
[Brinton et al., 1993], [Juneja et al., 2003]

Concernant la contraception: on regarde le nombre d'années passées sous contraceptif hormonal mais on peut aussi explorer les termes d'interaction contraception/tabagisme qui sont cités dans la littérature.

Les variables concernant le DIU étant à part nous supposons qu'il s'agit d'un dispositif en cuivre.

Nous rassemblons toutes les données sur les IST dans une seule variable car peu de femmes de la base de données ont été infectées donc différencier les maladies ne semble pas intéressant.

Nous ne prenons pas en compte le papillomavirus car 95% des personnes atteintes du cancer sont infectées par ce virus.

Il existe plusieurs patientes atteintes n'ayant pas fait d'examens médicaux donc on ne se focalise pas sur les variables cliniques dans l'étude.

De plus, certains examens ne sont fait uniquement sous suspision d'un cancer.

Les variables explicatives pour Gibbs

On regarde les variables les plus corrélées entre elles pour utiliser l'algorithme de sampling de Gibbs:

- Age
- Nombre de partenaires sexuels
- Nombre de grossesses
- Nombre d'années de tabagisme
- Nombre d'années de contraception (hormonale)
- Nombre d'années avec un DIU

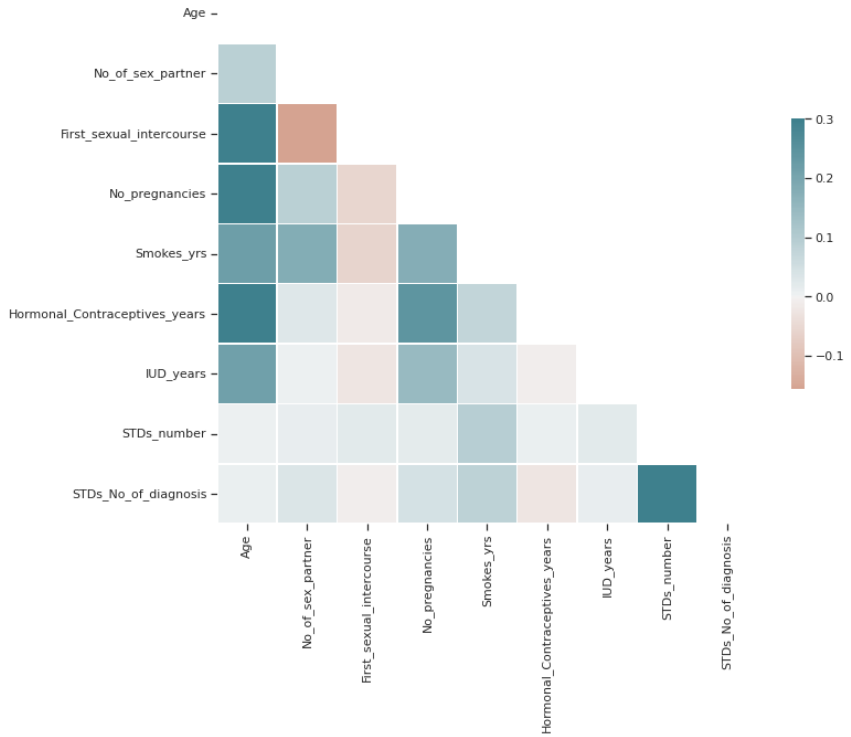


Figure: Matrice de corrélation entre les variables pré-sélectionnées

La méthode de Gibbs

- Parmis les variables explicatives à disposition (en excluant celles qui ne nous intéressent pas, citées précédemment)
- On regarde les paires de variables corrélées fortement (seuil ≥ 0.2)
- On suppose alors que la distribution jointe entre ces paires fait sens et on 'se propose' un distribution (souvent la gaussienne)
- On prend un point d'initialisation des paramètres (choisi par nos précautions selon les données)
- On construit itérativement la distribution jointe en calculant les nouveaux paramètres à chaque itération (l'espérance et la variance pour une loi normale) pour générer une nouvelle donnée que l'on standardise grâce aux moments calculés afin d'obtenir une nouvelle valeur synthétique

La méthode de Gibbs

- Parmis les variables explicatives à disposition (en excluant celles qui ne nous intéressent pas, citées précédemment)
- On regarde les paires de variables corrélées fortement (seuil ≥ 0.2)
- On suppose alors que la distribution jointe entre ces paires fait sens et on 'se propose' un distribution (souvent la gaussienne)
- On prend un point d'initialisation des paramètres (choisi par nos précautions selon les données)
- On construit itérativement la distribution jointe en calculant les nouveaux paramètres à chaque itération (l'espérance et la variance pour une loi normale) pour générer une nouvelle donnée que l'on standardise grâce aux moments calculés afin d'obtenir une nouvelle valeur synthétique

La méthode de Gibbs

- Parmis les variables explicatives à disposition (en excluant celles qui ne nous intéressent pas, citées précédemment)
- On regarde les paires de variables corrélées fortement (seuil ≥ 0.2)
- On suppose alors que la distribution jointe entre ces paires fait sens et on 'se propose' un distribution (souvent la gaussienne)
- On prend un point d'initialisation des paramètres (choisi par nos précautions selon les données)
- On construit itérativement la distribution jointe en calculant les nouveaux paramètres à chaque itération (l'espérance et la variance pour une loi normale) pour générer une nouvelle donnée que l'on standardise grâce aux moments calculés afin d'obtenir une nouvelle valeur synthétique

La méthode de Gibbs

- Parmis les variables explicatives à disposition (en excluant celles qui ne nous intéressent pas, citées précédemment)
- On regarde les paires de variables corrélées fortement (seuil ≥ 0.2)
- On suppose alors que la distribution jointe entre ces paires fait sens et on 'se propose' un distribution (souvent la gaussienne)
- On prend un point d'initialisation des paramètres (choisi par nos précautions selon les données)
- On construit itérativement la distribution jointe en calculant les nouveaux paramètres à chaque itération (l'espérance et la variance pour une loi normale) pour générer une nouvelle donnée que l'on standardise grâce aux moments calculés afin d'obtenir une nouvelle valeur synthétique

La méthode de Gibbs

- Parmis les variables explicatives à disposition (en excluant celles qui ne nous intéressent pas, citées précédemment)
- On regarde les paires de variables corrélées fortement (seuil ≥ 0.2)
- On suppose alors que la distribution jointe entre ces paires fait sens et on 'se propose' un distribution (souvent la gaussienne)
- On prend un point d'initialisation des paramètres (choisi par nos précautions selon les données)
- On construit itérativement la distribution jointe en calculant les nouveaux paramètres à chaque itération (l'espérance et la variance pour une loi normale) pour générer une nouvelle donnée que l'on standardise grâce aux moments calculés afin d'obtenir une nouvelle valeur synthétique

Tests statistiques

Etant donné le désequilibre entre effectifs de patientes saines et atteintes, on utilise le test statistique suivant:

- Mann Whitney U test

H0: Les distributions sont similaires dans les différentes classes.

C'est la version non paramétrique du T-test. Elle est plus adaptée à nos données déséquilibrées dans lesquelles on trouve peu de patientes atteintes. Cela nous permet d'éviter les erreurs de type I.

Boîtes à moustache des variables pour lesquelles H0 est rejetée

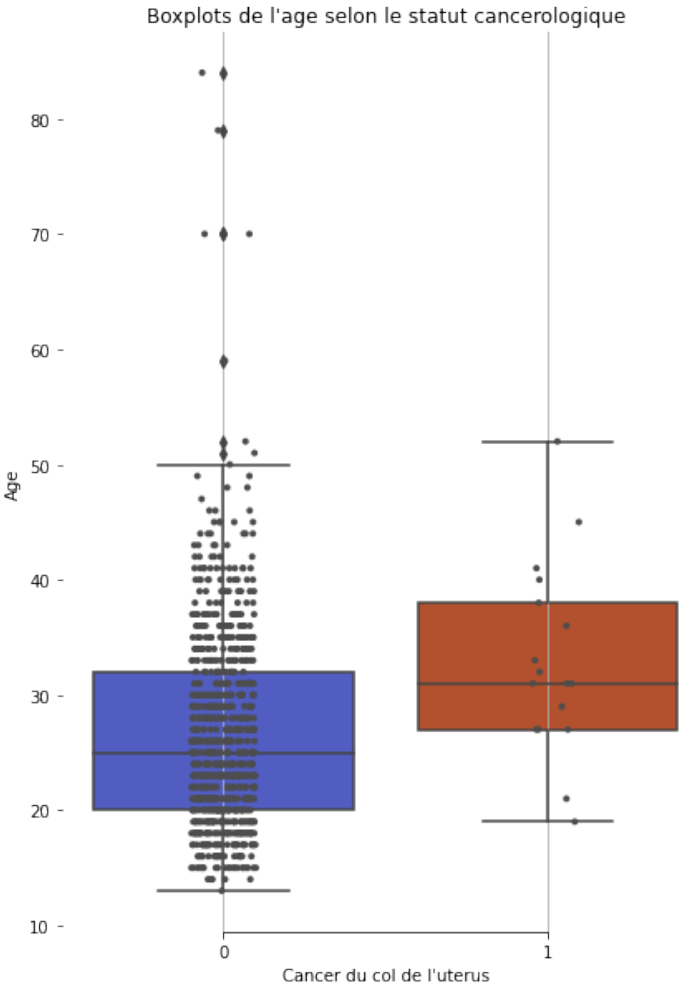


Figure: Distribution de l'age

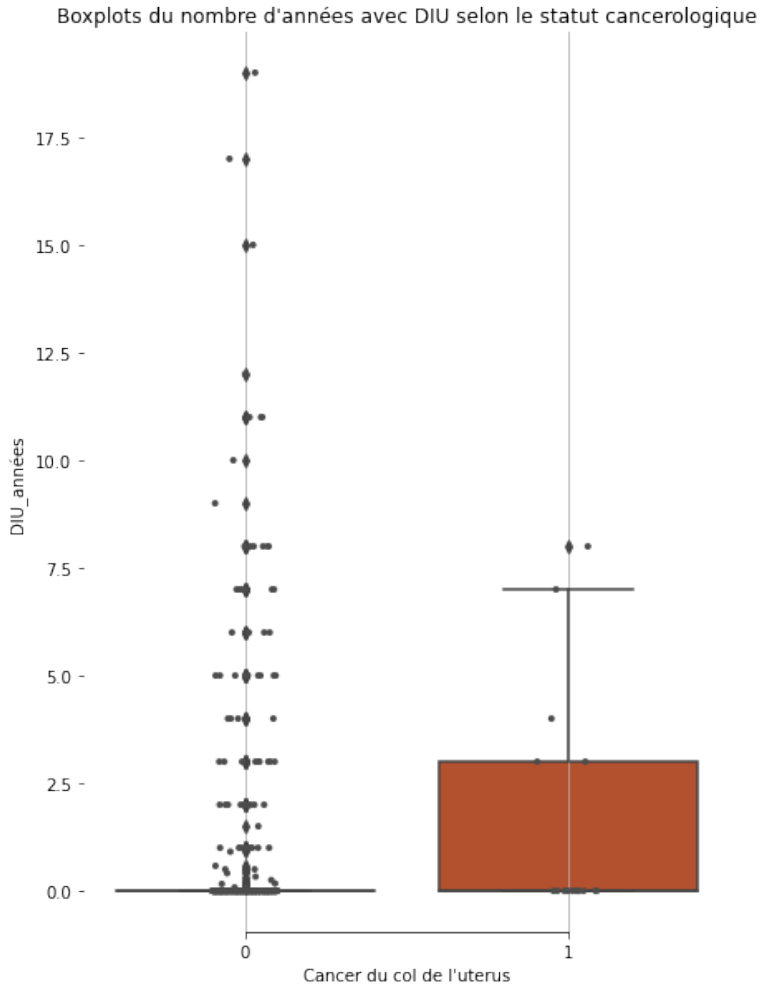


Figure: Distribution de la duree du DIU (années)

Générer des données synthétiques à partir des données

Comment contrer le déséquilibre de notre base de données ?

$\longrightarrow$ Plutôt que de recycler des observations pour augmenter l'echantillon des patientes atteintes...

$\longrightarrow$ On créé des données à partir de la distribution de chacune des variables:

- On utilise ensuite l'algorithme de Gibbs pour générer des données à partir des distributions conditionnelles pour le groupe de patientes atteintes [Gupta, 2020]
- La matrice de corrélation nous aide à vérifier que les données synthétiques gardent des corrélations en accord avec les données réelles

Résultats

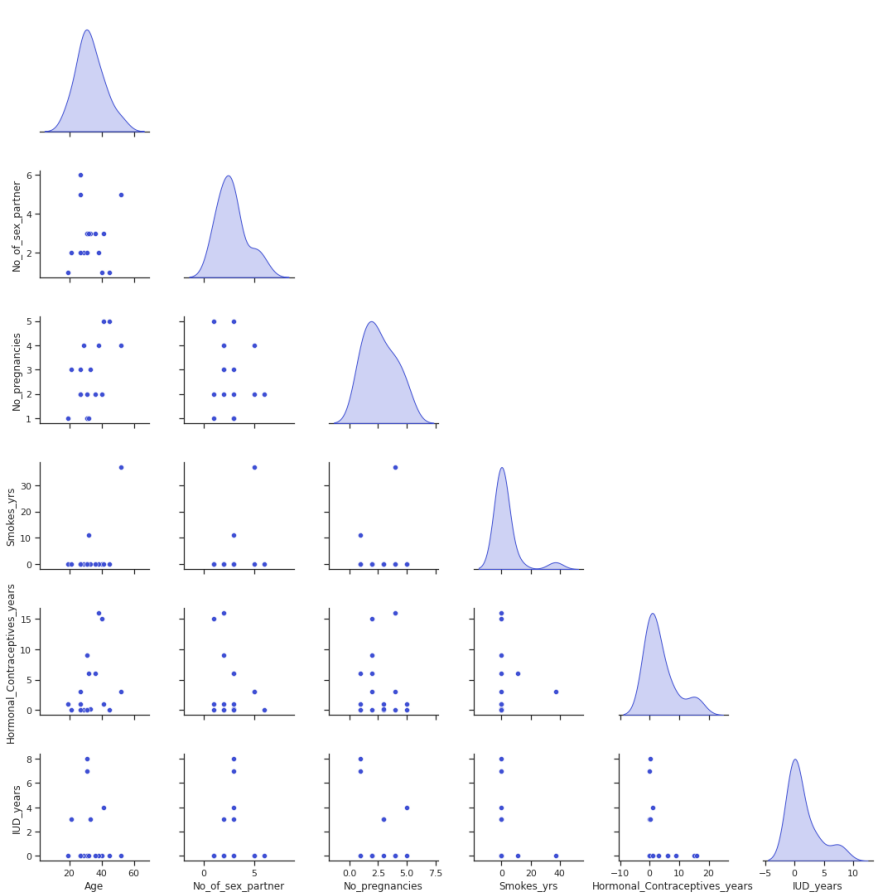


Figure: Données réelles de patientes atteintes

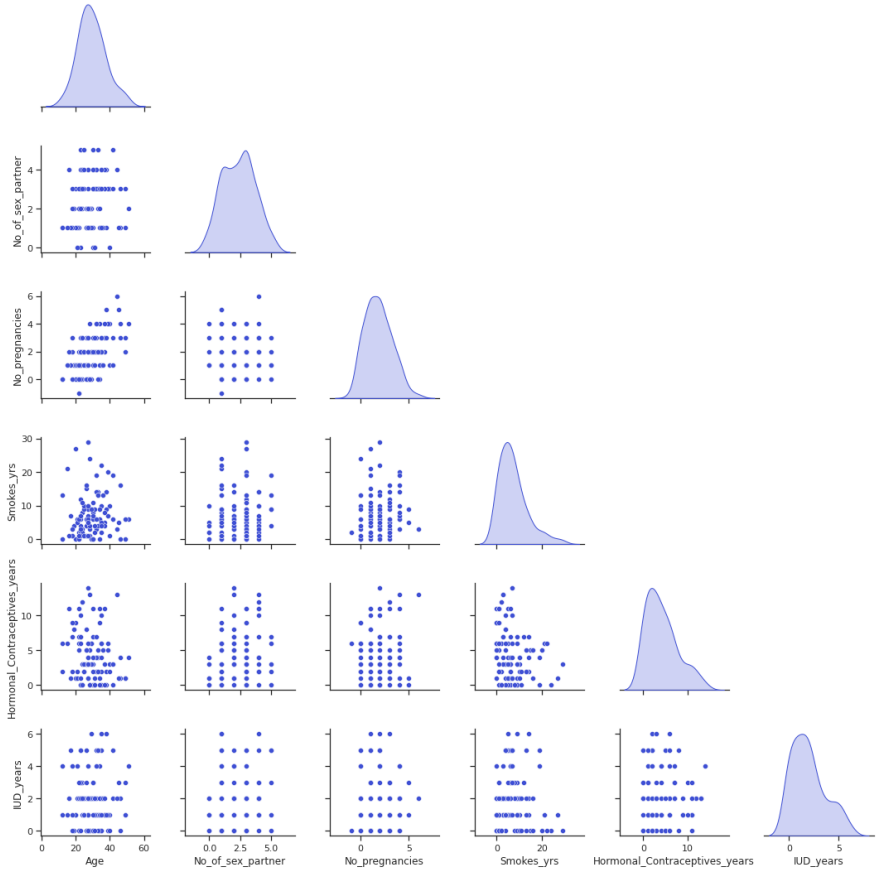


Figure: Données générées pour des patientes atteintes

Affaire à suivre...

- Les médecins ne sont pas statisticiens et se basent souvent uniquement sur des ratios de chance
- Dans la littérature les variables utilisées sont qualitatives tandis que nous avons accès à des variables quantitatives nous permettant d'extraire plus de résultats
- L'algo de Gibbs mets à l'épreuve les hypothèses de correlation
- Les distributions générées ne collent pas complètement avec les distributions réelles, nous pourrons ajouter des priors pour améliorer la génération


Brinton, L. A., Herrero, R., Reeves, W. C., de Britton, R. C., Gaitan, E., and Tenorio, F. (1993).
Risk factors for cervical cancer by histology.
Gynecologic Oncology, 51(3):301–306.


Gupta, R. (2020).
Implementing gibbs sampling in python.


Juneja, A., Sehgal, A., Mitra, B., and Pandey, A. (2003).
A survey on risk factors associated with cervical cancer.
Indian journal of cancer, 40:15–22.


Liu, Z.-C., L. W.-D. L. Y.-H. Y. X.-H. . C. S.-D. (2015).
Multiple sexual partners as a potential independent risk factor for cervical cancer: a meta-analysis of epidemiological studies.
Asian Pacific Journal of Cancer Prevention.

Merci !