Projet final-Rapport

Cours : PSY4016
Nom : Jacqueline Nguyen Phuong Trieu

Description
Les données ont été collectées à l'hôpital universitaire de Caracas au Venezuela. L'ensemble de données comprend les informations démographiques, les habitudes et les dossiers médicaux historiques de 858 patientes. Plusieurs patients ont décidé de ne pas répondre à certaines questions pour des raisons de confidentialité (valeurs manquantes).

Objectifs

1. Investiguer la relation entre les facteurs de risque du cancer cervical avec le nombre d'infection transmissible sexuellement par le sang (ITSS).
2. Investiguer l'impact des facteurs de risque du cancer cervical sur la décision de faire une biopsie.
3. Investiguer les facteurs de risque du nombre d'ITSS

Hypothèses
1. L'ensemble des facteurs de risque peuvent prédire le nombre d'ITSS.
2. Il existerait des différences significatives entre la décision de faire une biopsie et les facteurs de risque du cancer cervical.

Variables d'intérêts

Les variables dépendantes incluent le nombre d'ITSS (Objectif 1) et la présence d'une biopsie (Objectif 2).
Les variables indépendantes de l'objectif 1 incluent des facteurs de risque du cancer cervical tels que l'âge, le nombre de partenaires sexuels, le nombre de grossesse, le nombre de cigarette fumées par année et la durée d'utilisation de contraceptif.
Les variables indépendantes de l'objectif 2 incluent des variables en lien avec la prise de décision de faire une biopsie, soit l'examen Hinselmann (Objectif 2).

Dans la base de données fournies, les variables présentant un astérisque représentent les variables d'intérêts.

Nom de la variable	Description	Type
Age*	Âge du participant (année)	entier
Number of sexual partners*	Nombre de partenaire sexuel	entier
First sexual intercourse	Premier rapport sexuel (année)	entier
Num of pregnancies*	Nombre de grossesse	entier
Smoke	Fumeur (oui-1/non-0)	catégorielle
Smokes (years)	Durée à fumer (année)	flottante
Smokes (packs/year)*	Nombre de paquets par année	flottante
Hormonal Contraceptive	Utilisation contraceptif oraux (oui-1/non-0)	catégorielle
Hormonal contra.(years)	Durée utilisation contraceptif (année)	flottante

IUD	Utilisation stérilet (oui/non)	catégorielle
IUD (years)	Durée utilisation stérilet (année)	entier
STDs	Infection sexuellement transmises (ITSS)	catégorielle
STDs (numbers)*	Nombre d'ITSS	entier
STDs:condylomatosis	ITSS de type condylomatosis	catégorielle
STDs:cervical condylomatosis	ITSS de type condylomatosis cervical	catégorielle
STDs:vaginal condylomatosis	ITSS de type condylomatosis vaginal	catégorielle
STDs:vulvo-perineal condylomatosis	ITSS de type condylomatosis vulvo-périnéal	catégorielle
STDs:syphilis	ITSS de type syphilis	catégorielle
STDs:pelvic inflammatory disease	ITSS de type inflammation pelvis	catégorielle
STDs:genital herpes	ITSS de type herpès génital	catégorielle
STDs:molluscum contagiosum	ITSS de type molluscum contagiosum	catégorielle
STDs:AIDS	ITSS de type AIDS	catégorielle
STDs:HIV	ITSS de type HIV	catégorielle
STDs:Hepatitis B	ITSS de type Hépatite B	catégorielle
STDs:HPV	ITSS de type Virus Papillome Humain	catégorielle
STDs: Number of diagnosis	Nombre de diagnostic d'ITSS	entier
STDs: Time since first diagnosis	Durée depuis le 1er diagnostic	entier
STDs: Time since last diagnosis	Durée depuis le dernier diagnostic	entier
Dx:Cancer	Diagnostic cancer cervical antérieur	catégorielle
Dx:CIN	Dx cervical intrepithelial neoplasia (CIN)	catégorielle
Dx:HPV	human papilloma virus (HPV), high risk cervical cancer	catégorielle
Dx	Nombre de diagnostic	catégorielle
Hinselmann*	Examen Hinselmann	catégorielle
Schiller	Test Schiller	catégorielle
Citology	Examen cytologie	catégorielle
Biopsie*	Examen biopsie	catégorielle

Défis du projet et gestion

Tout d'abord, il existait plusieurs défis dans la base de données. Il a fallu changer les éléments de chaînes en nombre. De plus, l'entrée de données comporte également des erreurs à corriger, tels que « 26y » dans la variable « Âge ». Les données manquantes sont identifiées par des points d'interrogations. J'ai dû remplacer ces valeurs en NaN pour pouvoir faire un code pour modifier les valeurs manquantes. Ces tâches constituent l'étape de la correction des erreurs dans la base de données. Il était également difficile d'implémenter les codes d'apprentissage automatique. Beaucoup de temps a été dédié à cela. Je ne suis pas sûre ce qu'il a fallu faire pour la structure pipeline, mais j'aurais transformé mes variables en z-score.

1. Régression multiple -Objectif 1

OLS Regression Results

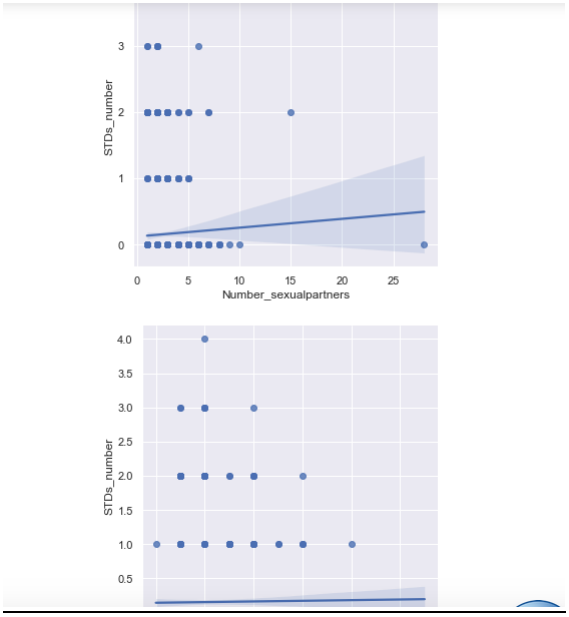
Dep. Variable:	STDs_number	R-squared:	0.003
Model:	OLS	Adj. R-squared:	-0.003
Method:	Least Squares	F-statistic:	0.4288
Date:	Tue, 19 Apr 2022	Prob (F-statistic):	0.829
Time:	13:45:26	Log-Likelihood:	-670.52
No. Observations:	858	AIC:	1353.
Df Residuals:	852	BIC:	1382.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.1364	0.065	2.094	0.037	0.009	0.264
Age	-0.0010	0.003	-0.397	0.692	-0.006	0.004
Number_sexualpartners	0.0120	0.011	1.069	0.285	-0.010	0.034
Num_pregnancies	0.0058	0.015	0.390	0.697	-0.023	0.035
Smokes_year	0.0058	0.008	0.696	0.487	-0.011	0.022
Hormonal_years	0.0003	0.005	0.050	0.960	-0.010	0.011

Omnibus:	671.037	Durbin-Watson:	2.018
Prob(Omnibus):	0.000	Jarque-Bera (JB):	8642.352
Skew:	3.673	Prob(JB):	0.00
Kurtosis:	16.703	Cond. No.	103.

Aucun résultat significatif. L'âge, le nombre de partenaire sexuel, le nombre de grossesse, le nombre de cigarette par année et le nombre d'année avec contraceptif ne prédisent pas le nombre d'ITSS.

2. Graphique- -Objectif 1

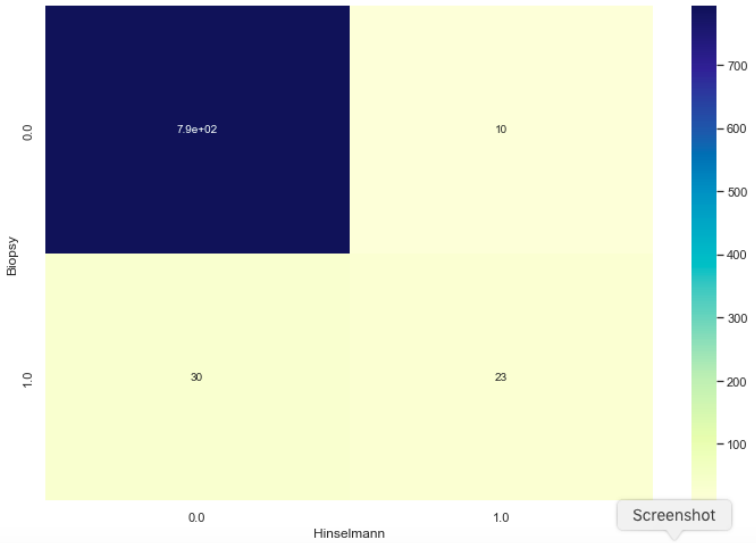


Il ne semble pas avoir de tendance dans les courbes avec les variables

3. Chi-Carré- Objectif2

```
In [28]: #graphique tableau chi carré
         plt.figure(figsize=(12,8))
         sns.heatmap(contigency, annot=True, cmap="YlGnBu")
```

```
Out[28]: <AxesSubplot:xlabel='Hinselmann', ylabel='Biopsy'>
```

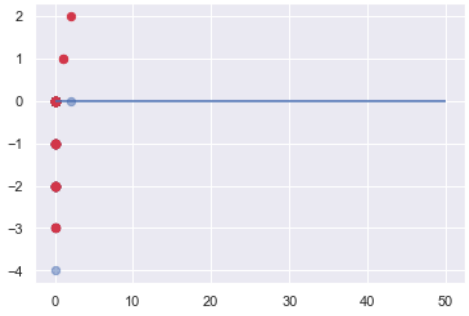


Voici le heatmap pour l'objectif 2, et il n'existe pas de corrélation forte. p=2.75

3.Apprentissage automatique supervisé -Objectif 3

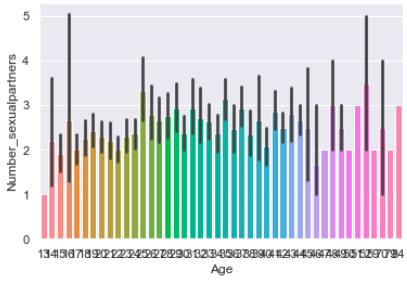
```
In [50]: plt.scatter(analyse.predict(X_train), analyse.predict(X_train)-y_train, c = 'b', s=40, alpha=0.5)
         plt.scatter(analyse.predict(X_test), analyse.predict(X_test)-y_test, c='r', s=40)
         plt.hlines(y=0, xmin=0, xmax=50)
```

```
Out[50]: <matplotlib.collections.LineCollection at 0x11ba4ee00>
```



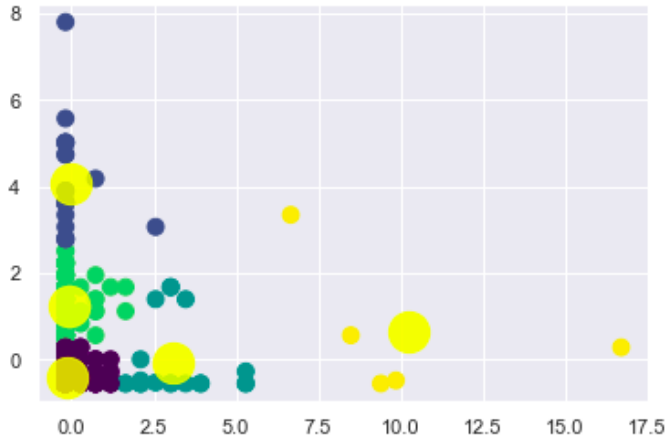
```
In [54]:
import seaborn as sns
sns.barplot(x = df_cervicalnew["Age"], y = df_cervicalnew["Number_sexualpartners"])
```

```
Out[54]: <AxesSubplot:xlabel='Age', ylabel='Number_sexualpartners'>
```



Le nombre de partenaire sexuel selon l'âge. Le niveau de précision est de 0,91.

4. Apprentissage automatique non-supervisé -Objectif 3



Il ne semble pas avoir de clusters. Les groupes sont mis ensemble.