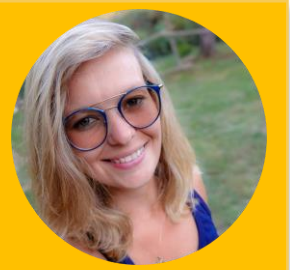
OPENCLASSROOMS



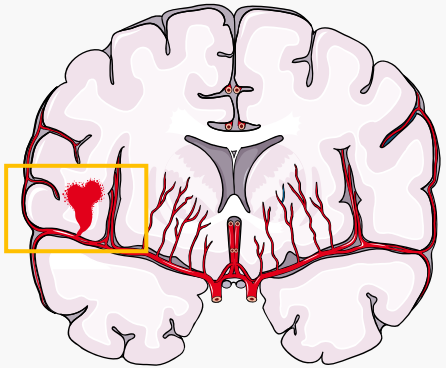
Lena Verboom

Détection des profils susceptibles d'être atteints par un AVC

Projet 8 : Réaliser un projet libre

Les AVC (Accident Vasculaire Cérébrale) :

- Rupture ou obstruction d'un **vaisseau sanguin**
- En France : **140 000** personnes touché par an
- 1 AVC toutes **les 4 minutes**
- 1ère cause **d'handicap**
- 2ème cause de **démence**
- 2ème cause de **mortalité**



Les conséquences d'un AVC :

- 20% des personnes atteintes décèdent 
- 50% Survivent mais ont des troubles de l'équilibre et de la mémoire 
- 30% des victimes souffrent de dépression 

- **Ces attaques cérébrales font des dommages irréversible sur le cerveau et sont donc très dangereuse pour les personnes** 

Comment lutter contre les AVC :

- La meilleure stratégie : La **prévention** (dépistage des facteurs de risque)

- Différents facteurs à **risques** :
  - Hypertension
  - Obésité
  - Alimentation
  - Manque d'activité physique
  - Consommation d'alcool
  - Diabète
  - Cholestérol

Objectif

- **Réduire au maximum le risque d'une personne d'être atteint par un AVC**

- **Stratégie** :

  - Identifier les caractéristiques qui favorisent l'apparition d'AVC dans un jeu de données
  - Créer un **modèle** de régression logistique pour la prédiction d'AVC
  - **Détection** des profils susceptibles d'être atteints par un **AVC**
  - Profils à risque pris en **consultation médicale**

Données à ma disposition :

kaggle

4861 : 0

↓

Sans AVC

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
...	...	...	...	...	...	...	...	...	...	...	...	...
5105	18234	Female	80.0	1	0	Yes	Private	Urban	83.75	NaN	never smoked	0
5106	44873	Female	81.0	0	0	Yes	Self-employed	Urban	125.20	40.0	never smoked	0
5107	19723	Female	35.0	0	0	Yes	Self-employed	Rural	82.99	30.6	never smoked	0
5108	37544	Male	51.0	0	0	Yes	Private	Rural	166.29	25.6	formerly smoked	0
5109	44679	Female	44.0	0	0	Yes	Govt_job	Urban	85.28	26.2	Unknown	0

249 : 1

↓

AVC

5110 rows x 12 columns

Remplacement des valeurs NaN :

```
id                   0
gender               0
age                  0
hypertension         0
heart_disease        0
ever_married         0
work_type            0
Residence_type       0
avg_glucose_level    0
bmi                201
smoking_status       0
stroke               0
dtype: int64
```

Remplacement de la valeur par la moyenne de bmi

Création de la colonne diabète :

Sélection des personne avec un taux de glucose > 126 mg/ld → **Diabète**

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke	diabete
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.600000	formerly smoked	1	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	28.893237	never smoked	1	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.500000	never smoked	1	0
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.400000	smokes	1	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.000000	never smoked	1	1

Préparation des données pour l'équilibrage :

Données déséquilibrées

Transformation log de bmi et glucose level

Transformation des données textes en binaire

Get_dummies

Création de dictionnaires

```
dict_gender={'Male':0,'Female':1}
dict_married={'No':0,'Yes':1}
dict_residence_type={'Urban':0,'Rural':1}
```

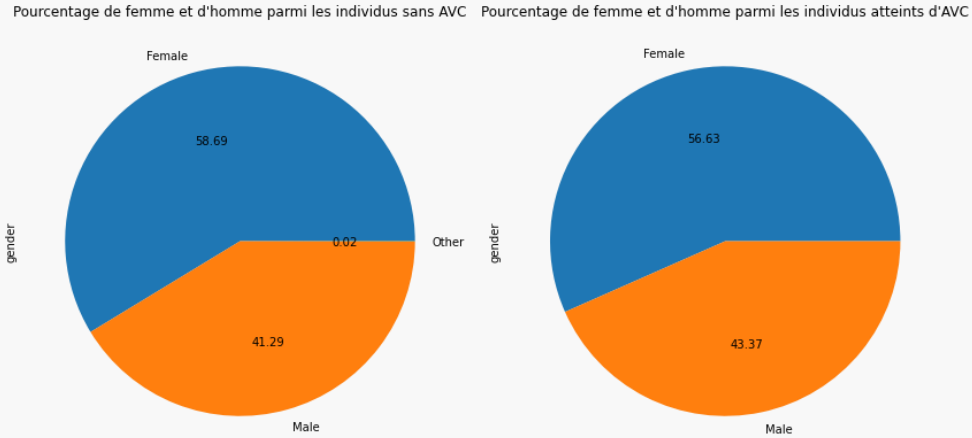
	Govt_job	Never_worked	Private	Self-employed	children
0	0	0	1	0	0
1	0	0	0	1	0
2	0	0	1	0	0
3	0	0	1	0	0
4	0	0	0	1	0

Visualisation des données finales :

	age	hypertension	heart_disease	stroke	diabete	gender_map	married_map	residence_map	smoke_map	Govt_job	Never_worked	Private	Self-employed
0	67.0	0	1	1	1	0.0	1	0	1	0	0	1	0
1	61.0	0	0	1	1	1.0	1	1	0	0	0	0	1
2	80.0	0	1	1	0	0.0	1	1	0	0	0	1	0
3	49.0	0	0	1	1	1.0	1	0	3	0	0	1	0
4	79.0	1	0	1	1	1.0	1	1	0	0	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...
5105	80.0	1	0	0	0	1.0	1	0	0	0	0	1	0
5106	81.0	0	0	0	0	1.0	1	0	0	0	0	0	1
5107	35.0	0	0	0	0	1.0	1	1	0	0	0	0	1

Analyse des variables :

Sexe des individus



Test d'indépendance entre les variables

Test Chi2
p-value =0.78

H0 : Il n'y a pas de différence significative entre les variables

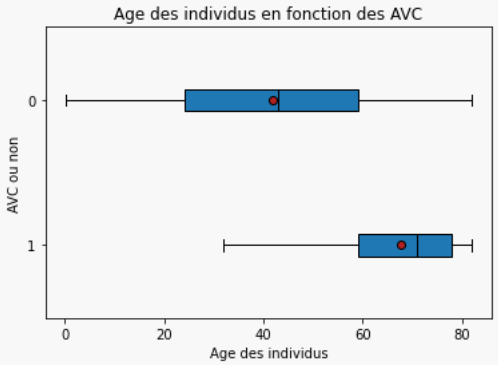
Ha : Il y a une différence significative entre les variables

p value > 0.05 : il n'y a de pas différence significative

Indépendance entre le sexe et l'apparition d'AVC

Analyse des variables :

Age des individus



Vérification des conditions ANOVA

Homoscédasticité : équivalence des variances

Test levene

p-value = 1

H0 : Les variances de population sont égales

Ha : Au moins une des variances est différente

p value > 0.05 : les variances sont égales

Test de normalité

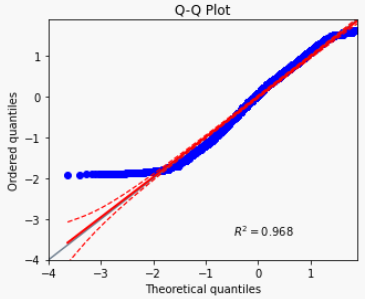
Test Jarque - Bera

p-value =0.0

H0 : Les échantillons suivent une distribution normale

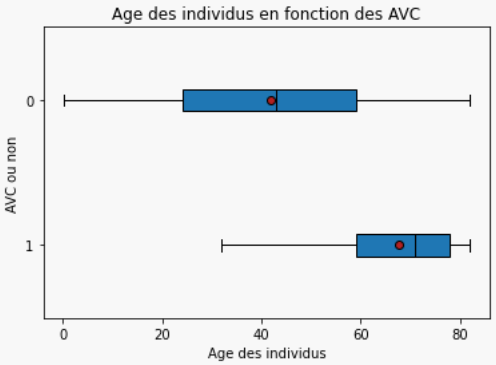
Ha : Les échantillons ne suivent pas une distribution normale

p value < 0.05 : Les échantillons ne suivent pas une loi normale



Analyse des variables :

Age des individus



Test d'indépendance des variables

Test paramétrique

Test ANOVA welch

p-value = 2.19e-95

H0 : Les variables sont indépendantes

Ha : Les variables sont dépendantes

p value < 0.05 : Les moyennes sont significativement différentes entres échantillons

Test non paramétrique

Test Kruskal - Wallis

p-value =3.72e-71

H0 : Les médianes des populations sont égales

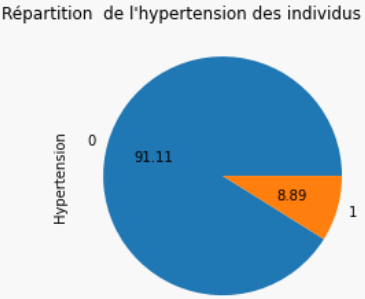
Ha : Les médianes des populations sont différentes

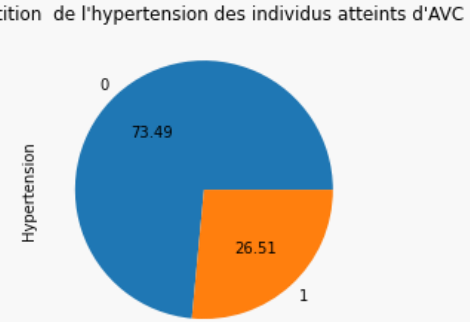
p value < 0.05 : Au moins un échantillon est différent

Dépendance entre l'âge et l'apparition d'AVC

Analyse des variables :

Hypertension des individus





Test d'indépendance entre les variables

Test Chi2
p-value =1.66e-19

H0 : Il n'y a pas de différence significative entre les variables

Ha : Il y a une différence significative entre les variables

p value < 0.05 : il y a une différence significative

Dépendance entre l'hypertension et l'apparition d'AVC

Analyse des variables :

Maladie cardiaque des individus

Test d'indépendance entre les variables

Test Chi2
p-value =2.08e-21

H0 : Il n'y a pas de différence significative entre les variables

Ha : Il y a une différence significative entre les variables

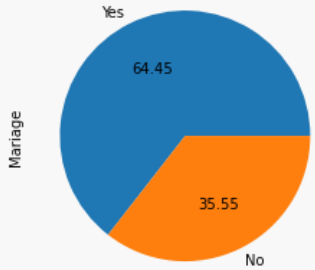
p value < 0.05 : il y a une différence significative

Dépendance entre les maladies cardiaques et l'apparition d'AVC

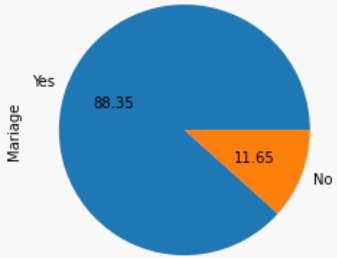
Analyse des variables :

Statut mariée des individus

Répartition des personnes sans AVC en fonction du mariage



Répartition des personnes atteintes d'AVC en fonction du mariage



Test d'indépendance entre les variables

Test Chi2
p-value =1.63e-14

H0 : Il n'y a pas de différence significative entre les variables

Ha : Il y a une différence significative entre les variables

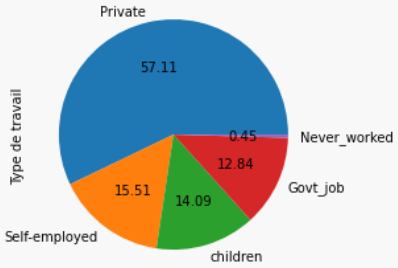
p value < 0.05 : il y a une différence significative

Dépendance entre le mariage et l'apparition d'AVC

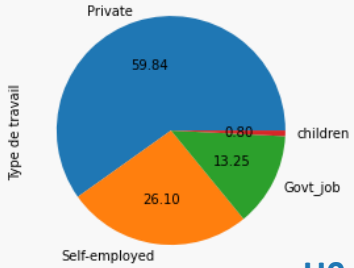
Analyse des variables :

Type de travail des individus

Répartition des personnes sans d'AVC en fonction du type de travail



Répartition des personnes atteintes d'AVC en fonction du type de travail



Test d'indépendance entre les variables

Test Chi2
p-value =5.39e-10

H0 : Il n'y a pas de différence significative entre les variables

Ha : Il y a une différence significative entre les variables

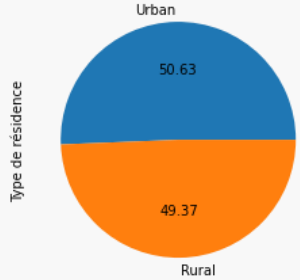
p value < 0.05 : il y a une différence significative

Dépendance entre le type de travail et l'apparition d'AVC

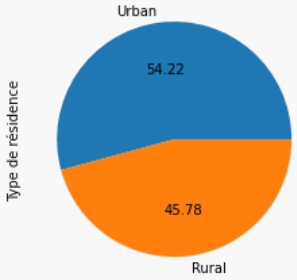
Analyse des variables :

Type de résidence des individus

Répartition des personnes sans d'AVC en fonction de la résidence



Répartition des personnes atteintes d'AVC en fonction de la résidence



Test d'indépendance entre les variables

Test Chi2
p-value =0.3

H0 : Il n'y a pas de différence significative entre les variables

Ha : Il y a une différence significative entre les variables

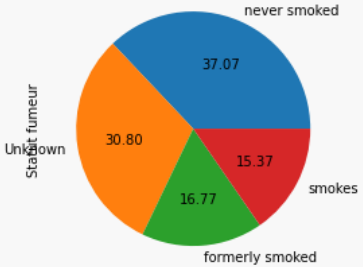
p value > 0.05 : il n'y a pas de différence significative

Indépendance entre le type de résidence et l'apparition d'AVC

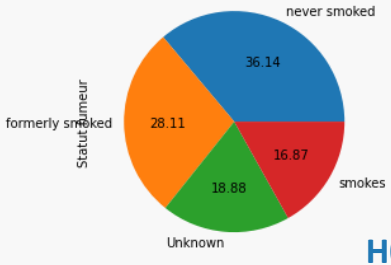
Analyse des variables :

Statut fumeur des individus

Répartition des personnes sans d'AVC en fonction du status fumeur



Répartition des personnes atteintes d'AVC en fonction du status fumeur



Test d'indépendance entre les variables

Test Chi2
p-value =0.00

H0 : Il n'y a pas de différence significative entre les variables

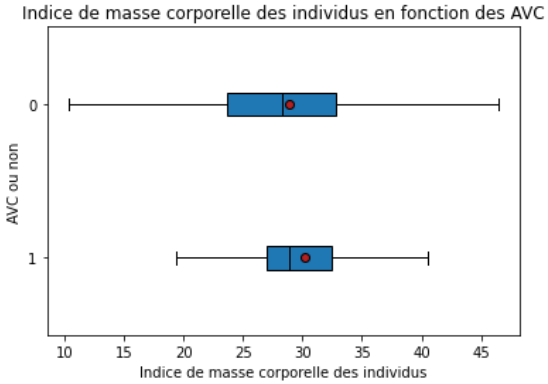
Ha : Il y a une différence significative entre les variables

p value < 0.05 : il y a une différence significative

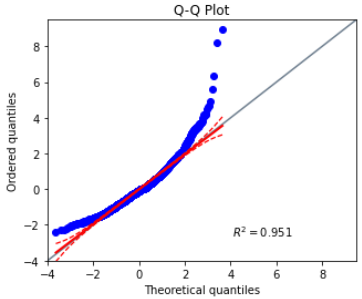
Les variables 'unknown' biaise les résultats de l'analyse

Analyse des variables :

Indice de masse corporelle des individus



Vérification des conditions ANOVA



Test de normalité

Test Jarque - Bera
p-value =0.0

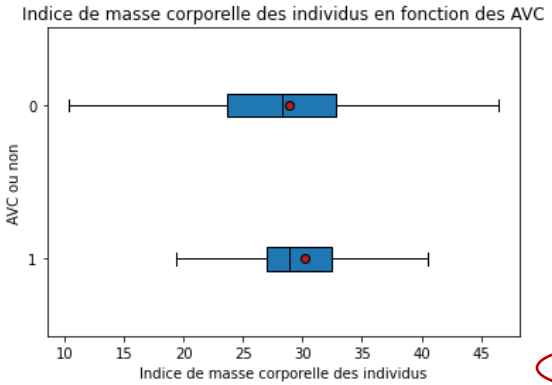
H0 : Les échantillons suivent une distribution normale

Ha : Les échantillons ne suivent pas une distribution normale

p value < 0.05 : Les échantillons ne suivent pas une loi normale

Analyse des variables :

Indice de masse corporelle des individus



Test d'indépendance des variables

Test paramétrique	Test non paramétrique
Test ANOVA welch	Test Kruskal - Wallis
p-value = 0.00	p-value = 0.00

H0 : Les variables sont indépendantes

Ha : Les variables sont dépendantes

p value < 0.05 : Les moyennes sont significativement différentes entres échantillons

H0 : La médianes des populations sont égales

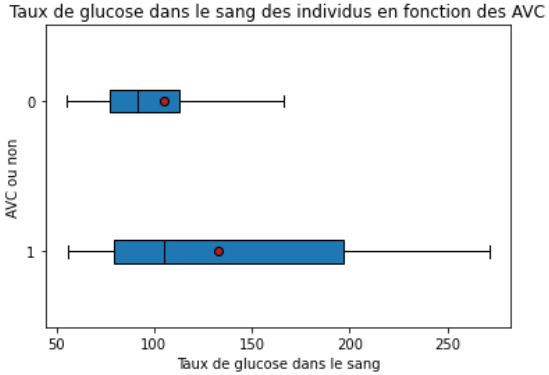
Ha : Les médianes des populations sont différentes

p value < 0.05 : Au moins un échantillon est différent

Dépendance entre l'indice de masse corporelle et l'apparition d'AVC

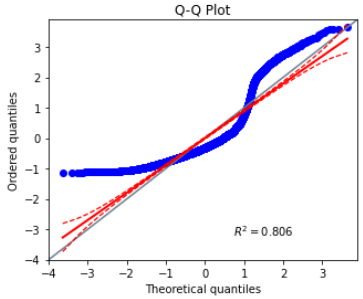
Analyse des variables :

Taux de glucose dans le sang des individus



Vérification des conditions ANOVA

Test de normalité



Test Jarque - Bera
p-value =0.0

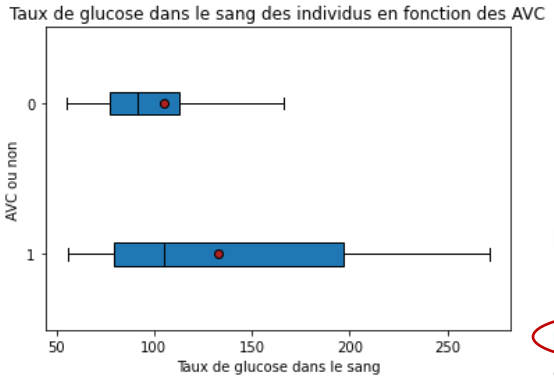
H0 : Les échantillons suivent une distribution normale

Ha : Les échantillons ne suivent pas une distribution normale

p value < 0.05 : Les échantillons ne suivent pas une loi normale

Analyse des variables :

Taux de glucose dans le sang des individus



Test d'indépendance des variables

Test paramétrique

Test ANOVA welch

p-value = 2.40e-11

H0 : Les variables sont indépendantes

Ha : Les variables sont dépendantes

p value < 0.05 : Les moyennes sont significativement différentes entres échantillons

Test non paramétrique

Test Kruskal - Wallis

p-value = 3.64e-9

H0 : La médianes des populations sont égales

Ha : Les médianes des populations sont différentes

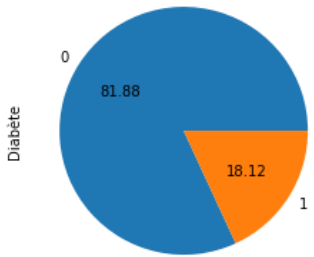
p value < 0.05 : Au moins un échantillon est différent

Dépendance entre le taux de glucose dans le sang et l'apparition d'AVC

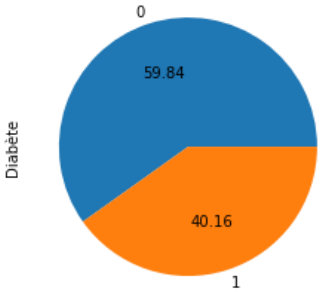
Analyse des variables :

Diabète des individus

Répartition des personnes sans AVC en fonction du diabète



Répartition des personnes atteints d'AVC en fonction du diabète



Test d'indépendance entre les variables

Test Chi2
p-value =1.47e-17

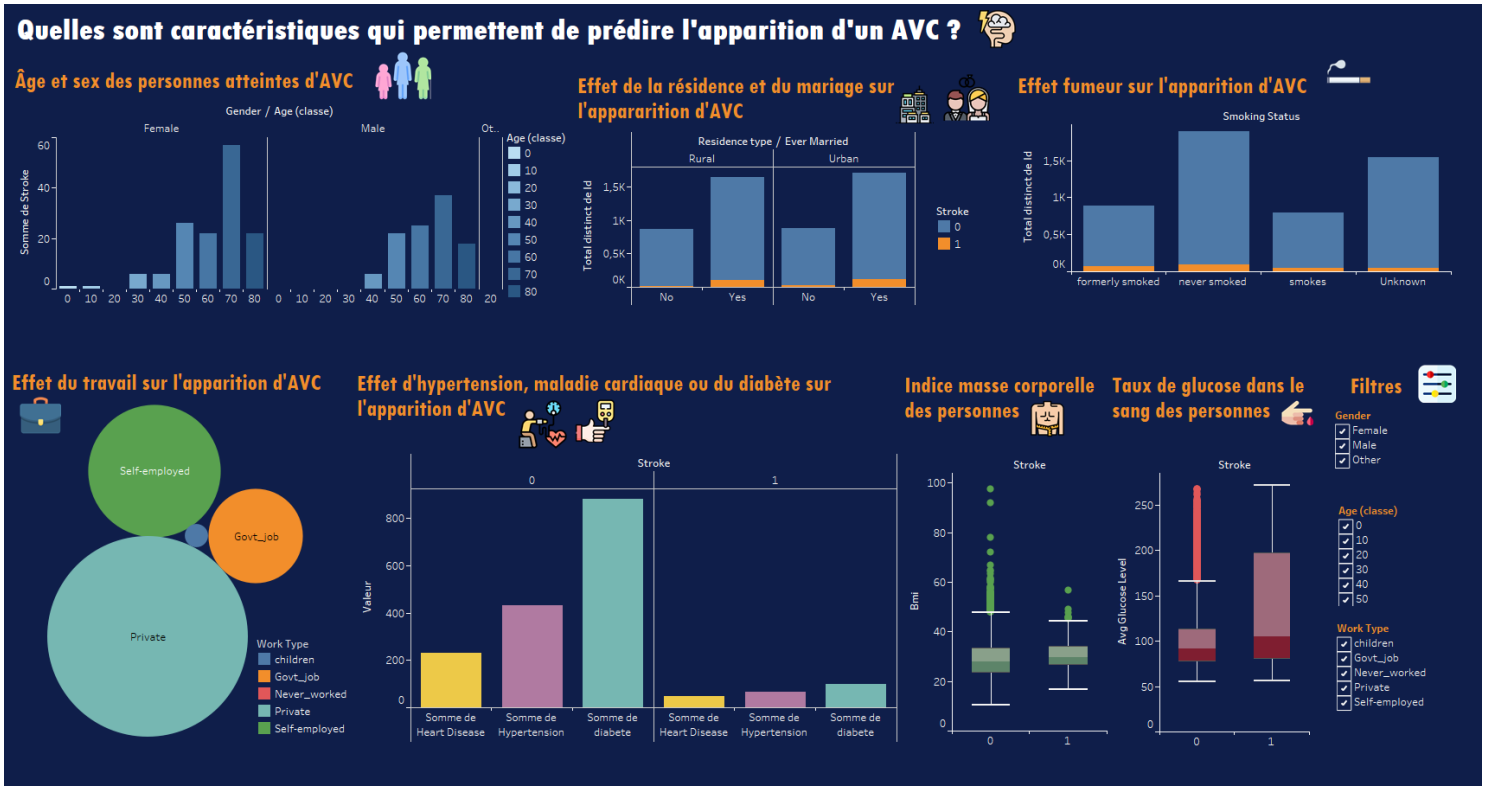
H0 : Il n'y a pas de différence significative entre les variables

Ha : Il y a une différence significative entre les variables

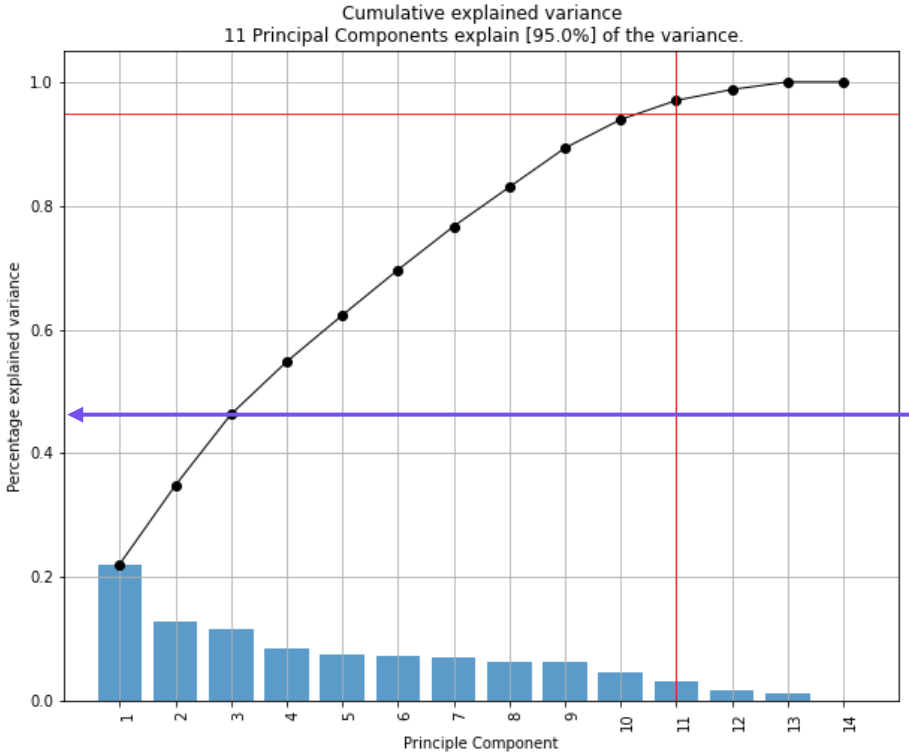
p value < 0.05 : il y a une différence significative

Dépendance entre le diabète et l'apparition d'AVC

Présentation du Dashboard avec tableau :



Détermination du nombre de composantes :

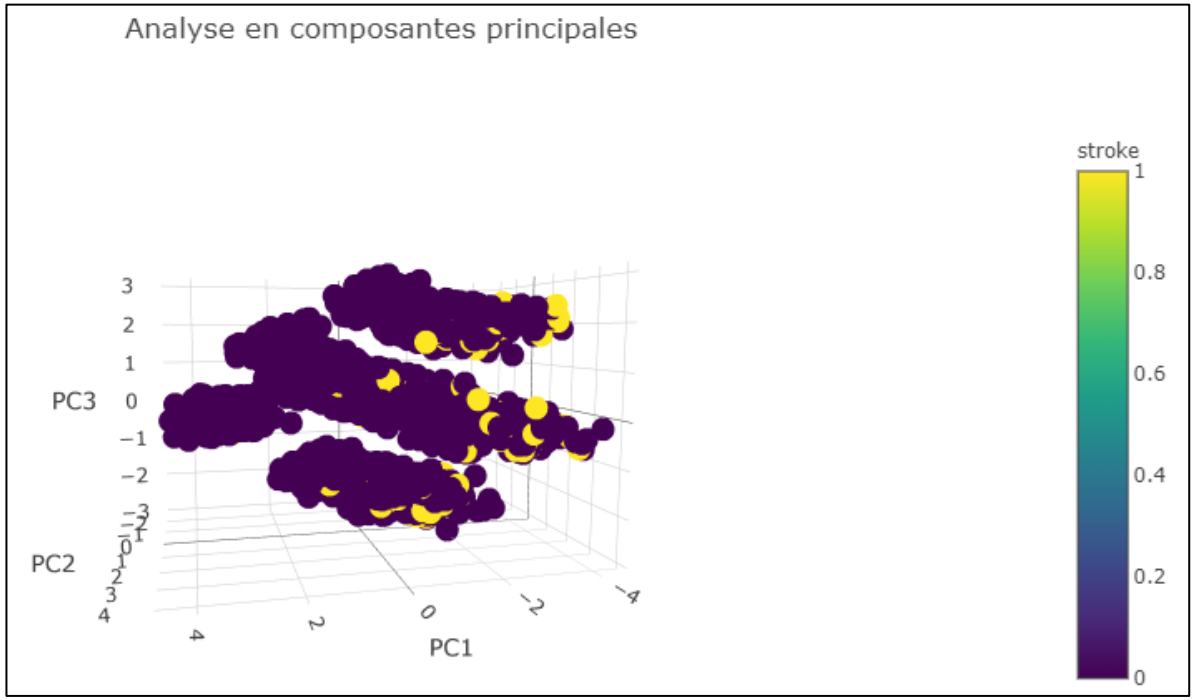


Courbe de variance accumulée

45 % avec 3 composantes → **3 axes**

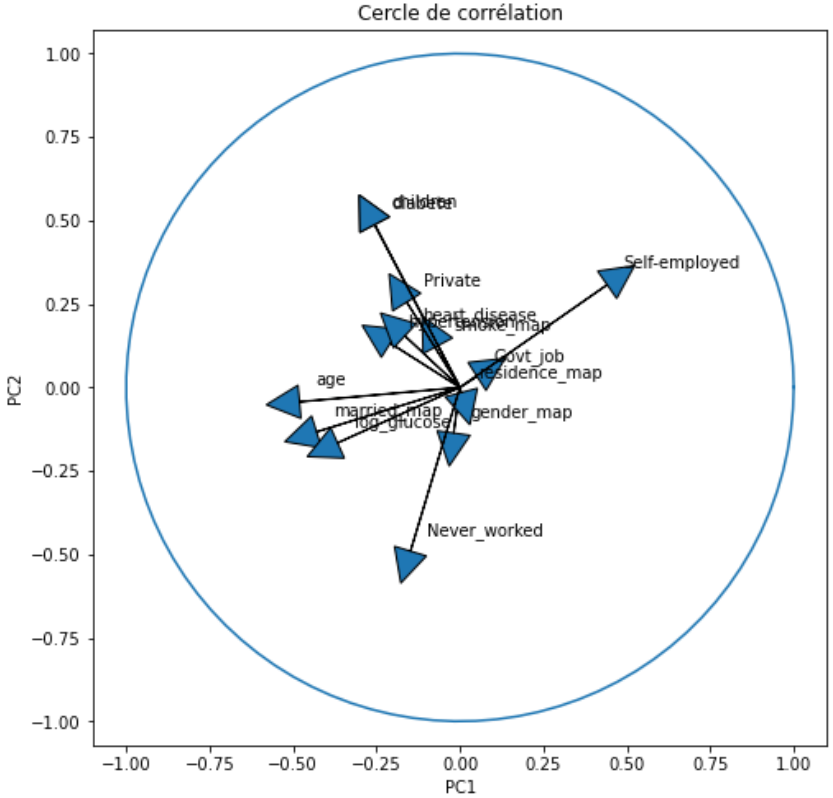
Variance pour chaque axe principal

Analyse en composantes principales :



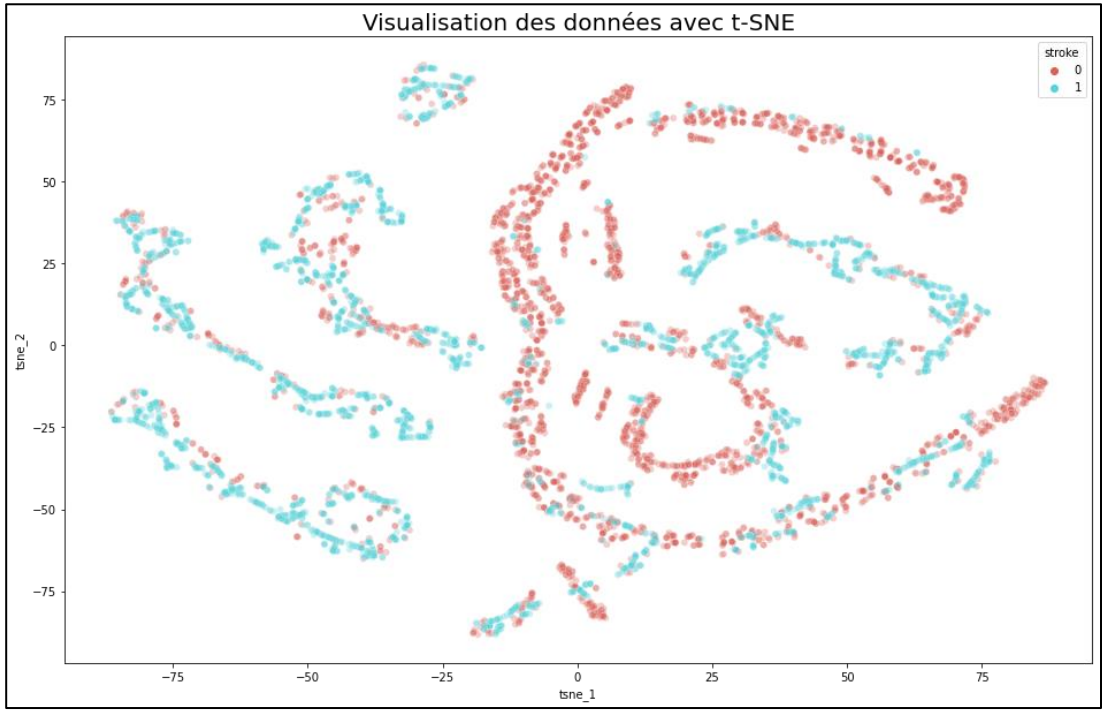
Pas de groupes distincts

Cercle de corrélation :



Trop de variables pour conclure sur des corrélations

Visualisation à l'aide de la technique t-SNE :



Deux groupes assez distincts avec des chevauchements

Régression à partir des données non équilibrées :

$$\frac{Vrai_{positif}}{Vrai_{positif}\ et\ faux_{positif}}$$

$$\frac{Vrai_{positif}}{Vrai\ positif\ +\ faux_{négatif}}$$

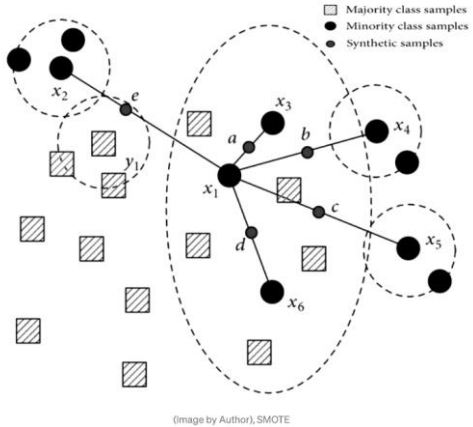
	precision	recall	f1-score	support
0	0.94	1.00	0.97	1590
1	0.00	0.00	0.00	96
accuracy			0.94	1686
macro avg	0.47	0.50	0.49	1686
weighted avg	0.89	0.94	0.92	1686

$$\frac{nb\ predictions\ correctes}{nb\ prédictions\ total}$$

Le modèle ne permet pas de prédire les AVC

Régression à partir des données équilibrées avec SMOTE-Tomek :

Suréchantillonage et suppression des échantillons proches de la limites des deux classes



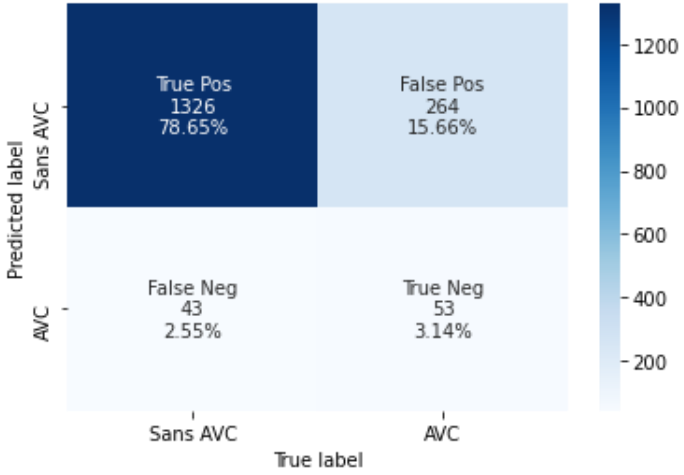
(Image by Author), SMOTE

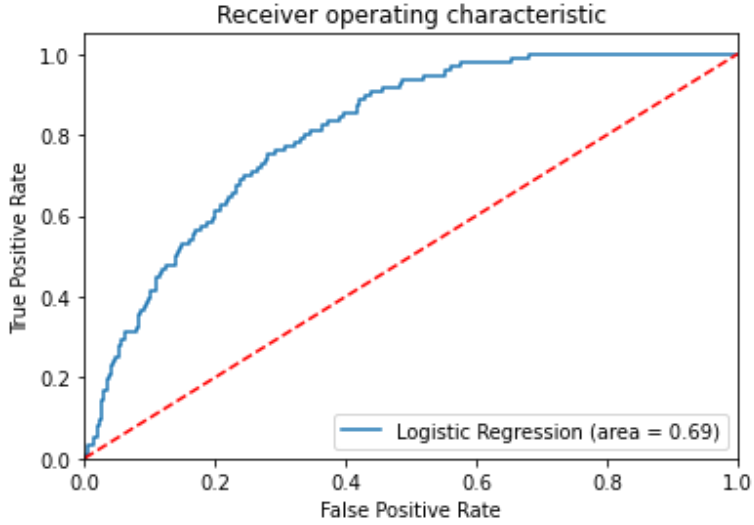
	precision	recall	f1-score	support
0	0.97	0.84	0.90	1590
1	0.17	0.54	0.25	96
accuracy			0.82	1686
macro avg	0.57	0.69	0.58	1686
weighted avg	0.92	0.82	0.86	1686

```
1    3265
0    3265
Name: stroke, dtype: int64
```

L'équilibrage des données a permis d'améliorer le modèle

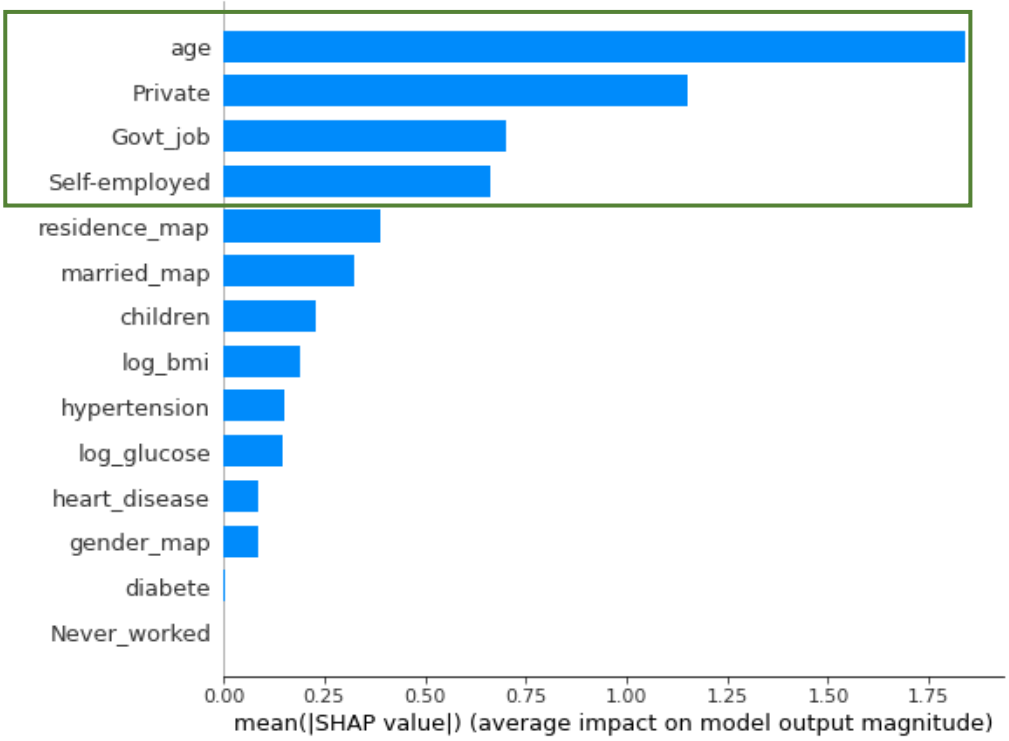
Evaluation des performances du modèle LogisticRegression:





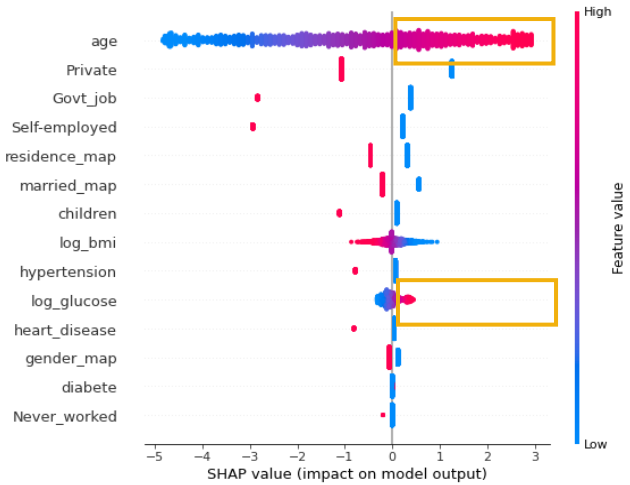
ROC AUC score 0.69

Interprétation des variable du modèle LogisticRegression avec SHAP :



4 variables qui ont le plus d'impact pour la prédiction

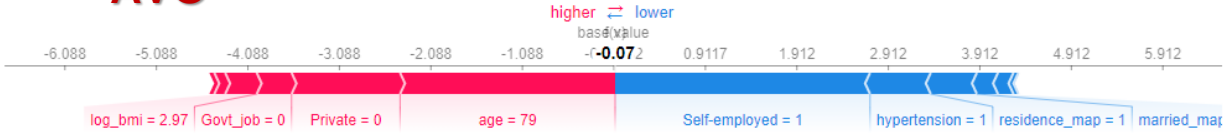
Interprétation des variable du modèle LogisticRegression avec SHAP :



Plus c'est élevé plus les personnes sont prédit en AVC

Non AVC

AVC



Comparaison de plusieurs algorithmes de classification :

Random Forest Classifier

	precision	recall	f1-score	support
0	0.96	0.87	0.91	1590
1	0.16	0.44	0.24	96
accuracy			0.84	1686
macro avg	0.56	0.65	0.58	1686
weighted avg	0.92	0.84	0.87	1686

K Nearest Neighbors

	precision	recall	f1-score	support
0	0.96	0.84	0.90	1590
1	0.15	0.47	0.23	96
accuracy			0.82	1686
macro avg	0.56	0.66	0.57	1686
weighted avg	0.92	0.82	0.86	1686

Performances entre les classifications équivalentes

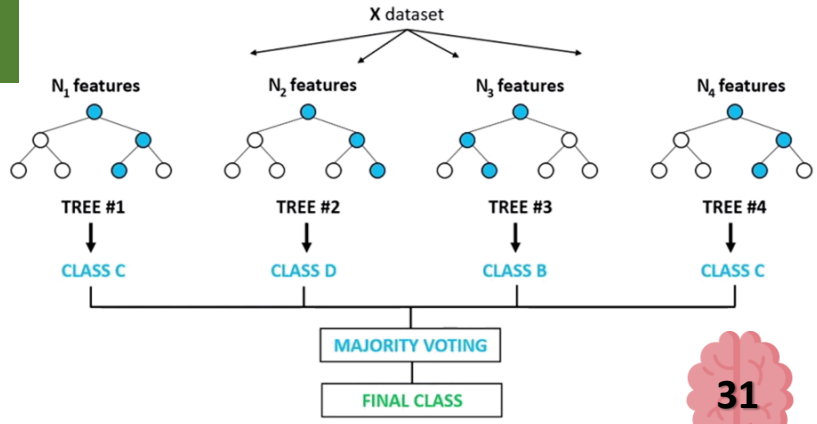
Sélection du modèle final:

	Accuracy	f1-score : 1	Recall : 1
LogisticRegression	0.82	0.25	0.54
Random Forest Classifier	**0.84**	**0.24**	**0.44**
K Nearest Neighbors	0.82	0.23	0.47
RFE 4 attributs	0.82	0.27	0.59

Random Forrest Classifier

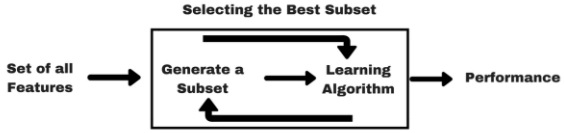
Paramètres

```
rfc1=RandomForestClassifier(criterion= 'gini',
 max_depth= 8,
 max_features= 'auto',
 n_estimators= 500)
```



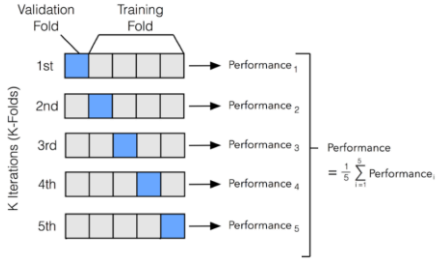
Amélioration du modèle de classification Random Forest Classifier :

RFE : Sélection des variables les plus pertinentes



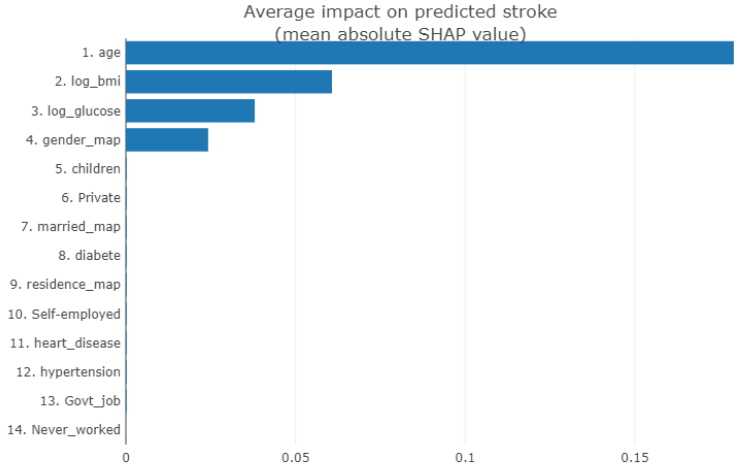

```
number of features : 2 cross_val_score : 0.8860617651560165 recall of positive class : 0.6666666666666666
number of features : 3 cross_val_score : 0.8918746849061021 recall of positive class : 0.6041666666666666
number of features : 4 cross_val_score : 0.9130129841192675 recall of positive class : 0.6041666666666666
number of features : 5 cross_val_score : 0.9307761212841426 recall of positive class : 0.5416666666666666
number of features : 6 cross_val_score : 0.9307766839518177 recall of positive class : 0.4479166666666667
number of features : 7 cross_val_score : 0.9375146293595492 recall of positive class : 0.5104166666666666
number of features : 8 cross_val_score : 0.9375149106933867 recall of positive class : 0.5104166666666666
number of features : 9 cross_val_score : 0.9408833207295775 recall of positive class : 0.4791666666666667
number of features : 10 cross_val_score : 0.9369018842615099 recall of positive class : 0.4895833333333333
number of features : 11 cross_val_score : 0.9372079754766921 recall of positive class : 0.5
number of features : 12 cross_val_score : 0.9390450854354598 recall of positive class : 0.5104166666666666
number of features : 13 cross_val_score : 0.938433184338933 recall of positive class : 0.4791666666666667
number of features : 14 cross_val_score : 0.9439453582167486 recall of positive class : 0.4791666666666667
number of features : 15 cross_val_score : 0.942413776805488 recall of positive class : 0.4791666666666667
number of features : 16 cross_val_score : 0.9460888407245359 recall of positive class : 0.4479166666666667
```

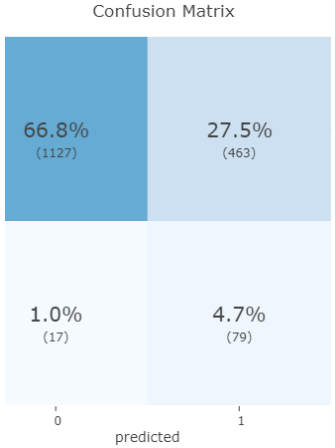
Validation croisée : base de test différent



4 Attributs : compromis entre un bon cross_val_score et le recall

Evaluation du modèle Random Forest Classifier avec 4 attributs :



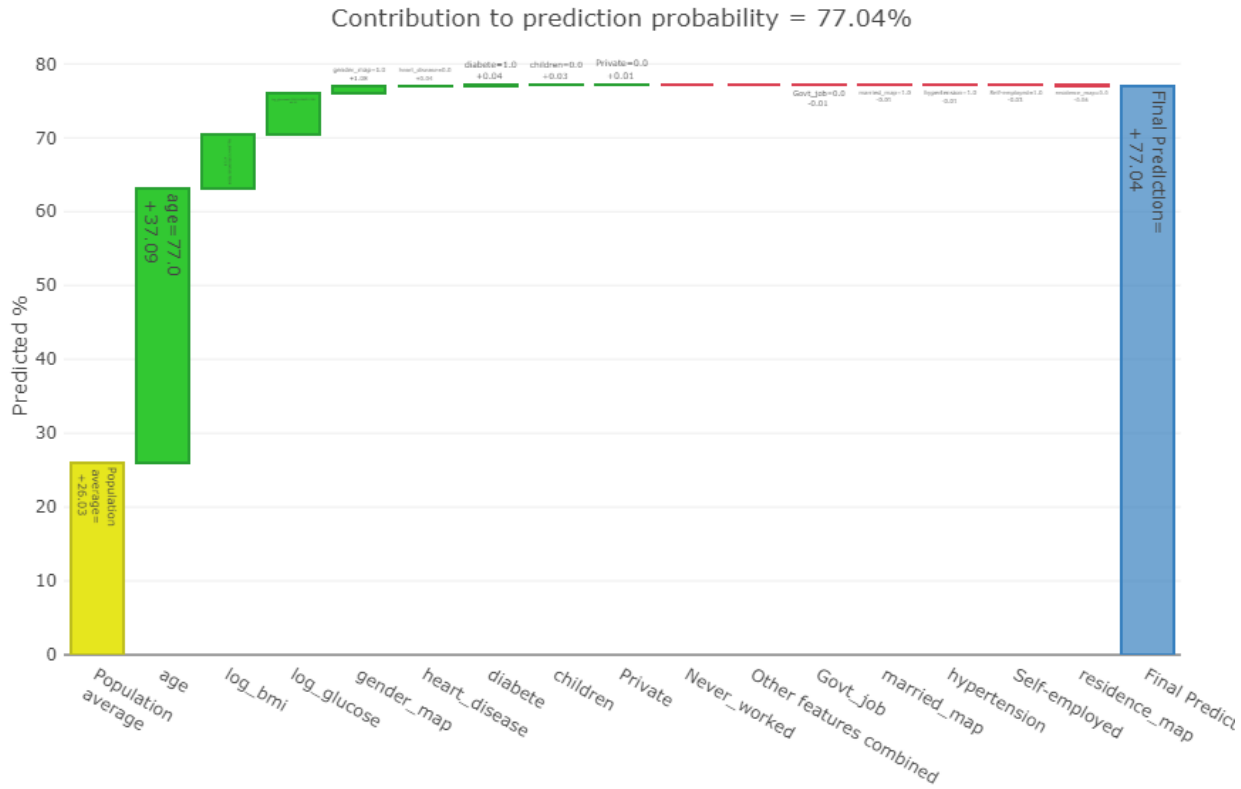


metric	Score
accuracy	0.731
precision	0.152
recall	0.812
f1	0.256
roc_auc_score	0.819

Seuil de proba : 0.25

4 Attributs : Age, imc, glucose dans le sang et sexe

Prédiction individuel :



Contribution : Age, imc, glucose dans le sang et sexe

- Les AVC touchent environ **16 millions** de personnes par an dans le monde

- Détecter les personnes à risques pour les envoyer en **consultation médicale**

- Différentes modèles de classifications: **accuracy et recall proche**

- RFE a permis d'**améliorer** le modèle avec **4 attributs** : un score ROC de **0.82**

- L'ensemble des variables ne permettent **pas de prédire à 100 %**
- **Amélioration** du modèle grâce à **d'autre facteurs de risque :**
  - Consommation d'alcool
  - Effet héréditaire
  - Cholestérol
  - L'alimentation
- Les personnes à **risque** sont pris en **consultation médicale** (pour alerter ou traitement)

Merci pour votre attention