OPENCLASSROOMS

Formation Openclassrooms de Data Analyst

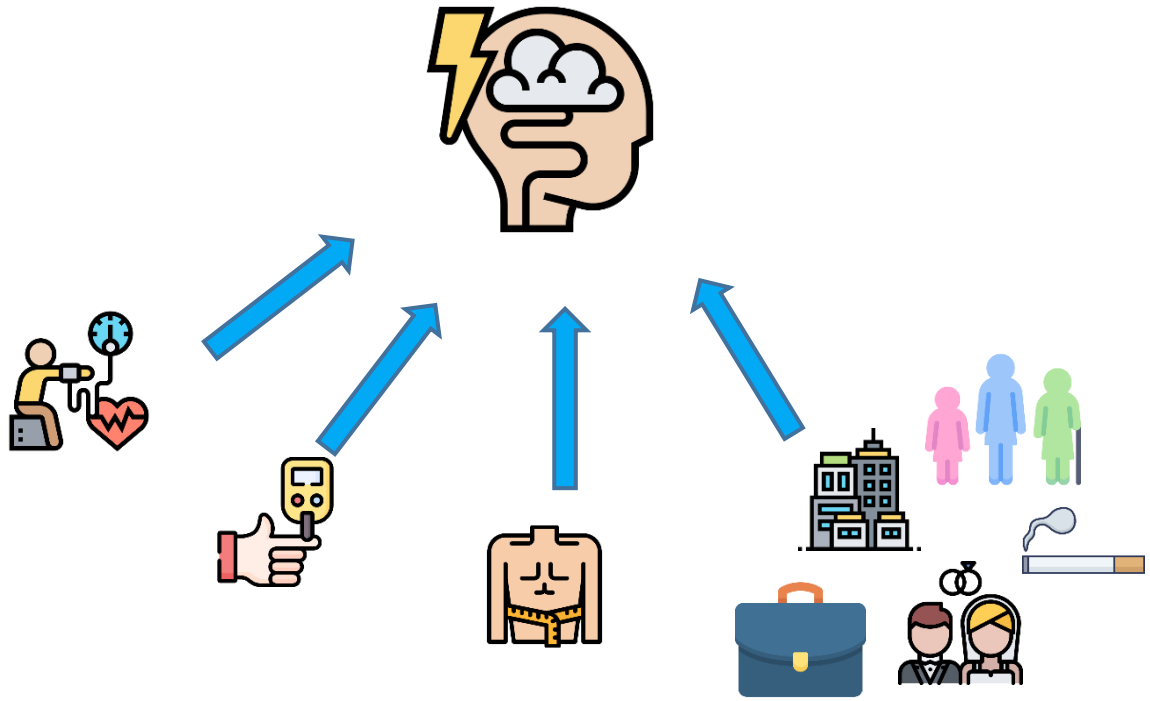
Projet n°8 : Communiquez vos résultats

RAPPORT PRESENTE PAR :

LENA VERBOOM

Détection des profils susceptibles d'être atteints par un AVC

(Accident Vasculaire Cérébral)

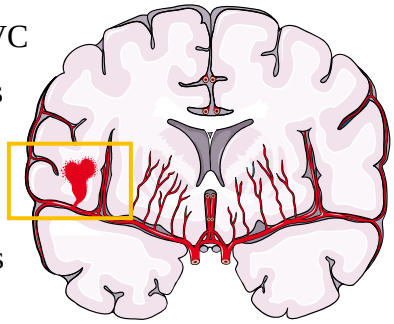


Août 2020

Table des matières

I. Introduction

L'objectif de mon projet est la détection de profils susceptibles d'être atteints par un AVC (Accident Vasculaire Cérébral). Un AVC correspond à une rupture ou à une obstruction d'un vaisseau sanguin au niveau du cerveau. En effet, en France les AVC touchent environ 140 000 personnes par an, soit 1 AVC toutes les 4 minutes. Les conséquences d'un AVC varient d'un cas à l'autre, certains décèdent et d'autres gardent des séquelles neurologiques plus ou moins graves. Les séquelles les plus fréquentes sont les troubles de mémoire, d'équilibre et visuel.



L'AVC étant la première cause de handicap et la deuxième cause de décès dans le monde, le but de mon projet est de réduire au maximum le risque d'une personne d'être atteint par un AVC. Ceci va être réalisé en identifiant les caractéristiques qui sont susceptibles de favoriser l'apparition d'AVC telle que l'âge, le style de vie ou les problèmes de santé des personnes. Avec ces caractéristiques un modèle de classification va permettre de prédire quelles sont les personnes à risques et ces personnes vont pouvoir être prises en consultation médicale. Évidemment notre modèle de classification va favoriser la prédiction de personnes atteintes d'un AVC, afin de privilégier les faux positifs par rapport aux faux négatifs.

Les données utilisées pour ce projet ont été téléchargés sur Kaggel, une plateforme web qui propose des milliers de datasets pour des compétitions en science des données. Elles contiennent des informations sur 5110 individus tels que l'âge, le sexe, le type de travail ou le type de résidence (fig.1)

	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

Fig.1 : *Visualisation des différentes caractéristiques des individus (5110 lignes x 12 colonnes).*

Parmi ces individus choisis aléatoirement, 249 ont subi un AVC et 4861 n'ont pas eu d'AVC. Afin d'être dans la capacité à étudier le risque d'être atteint d'un AVC, la première étape du projet est d'effectuer un nettoyage des données.

II. Nettoyage des données

A. Replacement des valeurs NaN

Premièrement les valeurs non numériques (NaN) ont été recherchées afin de ne pas avoir des difficultés dans l'analyse des données par la suite. Cette recherche a permis de montrer 201 valeurs NaN pour la variable Bmi (figure 2).

```
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
```

Fig.2 : *Visualisation des valeurs NaN.*

Ces valeurs NaN ont été remplacées par la moyenne des Bmi (Body Mass Index) dans la dataframe, c'est-à-dire par 28.89.

B. Création de la colonnes « diabète »

Par la suite une nouvelle colonne « diabète » a été créée afin de visualiser les personnes atteintes par cette maladie (figure 3). En effet pour une personne souffrant de diabète le taux de glucose dans le sang dépasse 126 mg/ld. Lorsque cette valeur est dépassée dans la colonne 'avf_glucose_level', on attribue à la personne la valeur 1 et 0 si la valeur est en dessous.

avg_glucose_level	bmi	smoking_status	stroke	diabete
228.69	36.600000	formerly smoked	1	1
202.21	28.893237	never smoked	1	1
105.92	32.500000	never smoked	1	0

Fig.3 : *Visualisation de la nouvelle colonne 'diabète'.*

C. Préparation des données pour l'équilibration

La prochaine étape est la préparation de données afin de pouvoir les équilibrer et par la suite les utiliser pour les modèles de classification. En effet il y a plus de personnes qui n'ont pas eu d'AVC (4861) que de personnes qui ont eu un AVC (249), ce qui favoriserait la détection de personne sans AVC avec le modèle et biaiserai l'utilisation du modèle pour la prédiction d'AVC. Pour cela les données de textes ont été transformées en binaire par la création de

dictionnaire. Pour le sexe, la situation de mariage et le type de résidence les différents dictionnaires suivants ont été créés (figure 4).

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

Fig.4 : *Création des 3 dictionnaires différents*.

Pour transformer la variable type de travail, la méthode 'get_dummies' de la librairie pandas a été utilisée. Cette méthode créée un DataFrame avec des colonnes de variables factices (figure 5).

	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
...	...	...	...	...	...
5105	0	0	1	0	0
5106	0	0	0	1	0
5107	0	0	0	1	0
5108	0	0	1	0	0
5109	1	0	0	0	0

Fig.5 : *Transformation de la variable type de travail en binaire*.

En ce qui concerne la variable statut fumeur, celle-ci a été transformée en dictionnaire sur le notebook, cependant elle n'a pas été utilisée pour le reste des analyses. En effet cette variable contient 1544 personnes caractérisées comme 'unknown', si cette variable est ajoutée ceci faussera les résultats d'analyse de l'effet statut fumeur sur l'apparition d'AVC.

Après avoir ajouté toutes les nouvelles colonnes dans la Dataframe, les variables avf_glucose_level et le bmi ont été transformées en logarithme afin de minimiser les écarts entre les différentes données. Le tableau final qui va pouvoir être utilisé pour le modèle de classification est le tableau suivant (figure 6).

	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

Fig.6 : *Visualisation de la DataFrame finale (5109 lignes x 16 colonnes).*

III. Analyses univariée et bivariée des données

A. *Sexe des individus*

Dans le but de pouvoir visualiser les différences entre les individus, la DataFrame a été divisée en deux groupes : le groupe AVC et le groupe non-AV.

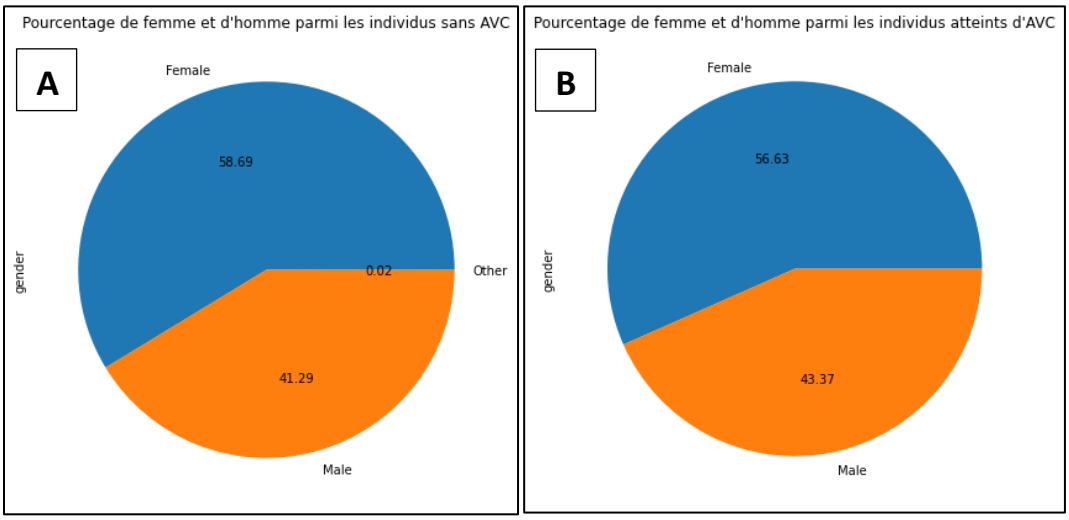
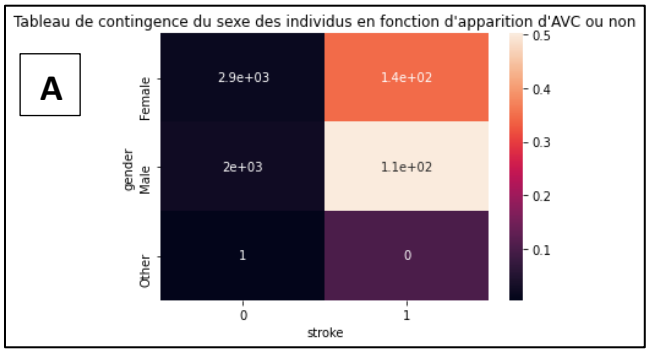


Fig.7 : *Représentation du pourcentage de femme et d'homme parmi les individus avec et sans AVC.*

En ce qui concerne la répartition des sexes entre les deux groupes, on observe 59 % d'hommes, 41% de femme et 0.02 % d'enfant pour le groupe sans AVC (figure 7A) et 57 % d'hommes et 43 % de femmes pour le groupe d'individus atteints par un AVC (figure 7B). Cette répartition de sexe est approximativement équivalente entre les deux groupes.

Afin de déterminer s'il y a une dépendance entre le sexe de l'individu et l'atteinte ou non d'un AVC, un tableau de contingence (figure 8A) et un test de Chi2 ont été réalisés (figure 8B).



B

Test Chi2 : Sexe	
P-value	0.78

Fig.8 : *Tableau de contingence du sexe des individus en fonction de l'apparition d'AVC et test Chi2 (H0 : Il n'y a pas de différence significative entre le sexe et l'apparition d'AVC, H1 : Il y a une différence significative entre le sexe et l'apparition d'AVC).*

Le tableau de contingence montre une différence dans le nombre total des femmes dans le groupe AVC. Le test de Chi2, qui permet d'analyser l'indépendance entre deux variables qualitatives, a montré une P-value de 0.78. Cette P-value est supérieure à 0.05 donc on accepte l'hypothèse nulle et on peut conclure qu'il n'y a pas de différence significative entre les femmes et les hommes et l'apparition d'AVC.

B. Age des individus

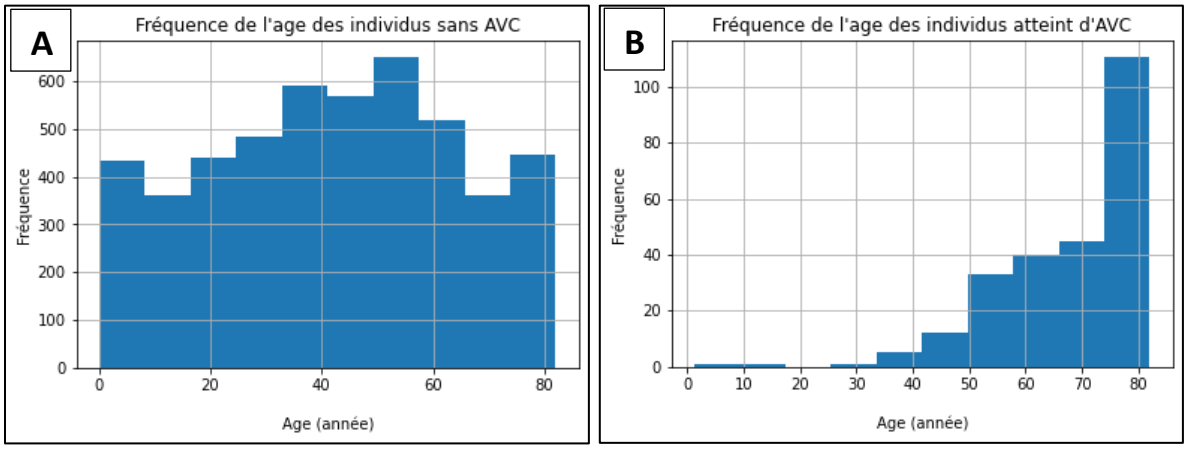


Fig.9 : *Histogramme représentant la fréquence d'âge parmi les individus avec et sans AVC.*

Lorsqu'on observe la fréquence d'âge des deux groupes, on remarque que pour les personnes non atteintes d'AVC l'âge varie entre 0 et 80 ans (figure 9A), alors que pour les personnes atteintes d'AVC la plupart sont âgée entre 50 et 80 ans (figure 9B).

Afin d'observer la différence d'âge en fonction des deux groupes un graphique en boîte à moustache a été réalisé (figure 10).

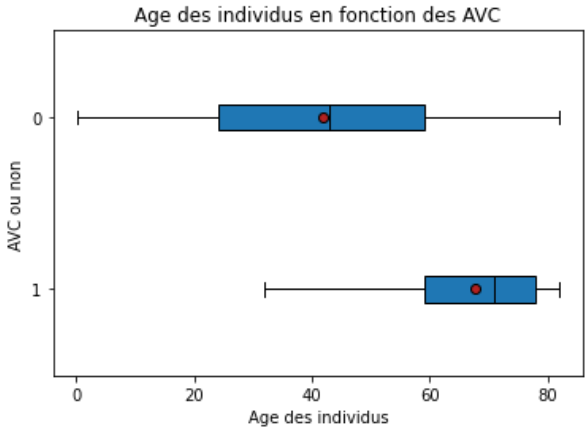


Fig.10 : *Boîtes à moustaches représentant l'âge des individus en fonction de l'apparition d'AVC ou non.*

Sur cette représentation on remarque que l'âge des individus atteints d'AVC a tendance à être situé entre 60 et 80 ans et serait donc plus âgé.

Afin de déterminer s'il y a une dépendance entre l'âge et l'apparition ou non d'un AVC, un test d'ANOVA Welch peut être réalisé. Cependant avant d'effectuer ce test il faut que les conditions d'homoscédasticité et la normalité des variables soient respectées. Pour vérifier l'homoscédasticité des variables, un test de Levene a été effectué (figure 11A) et pour la normalité un test de Jarque – Bera et un qq-plot ont été effectués (figure 11B-C).

A

Test Levene : Age	
P-value	1

B

Test Jarque-Bera : Age	
P-value	1.2e-49

C

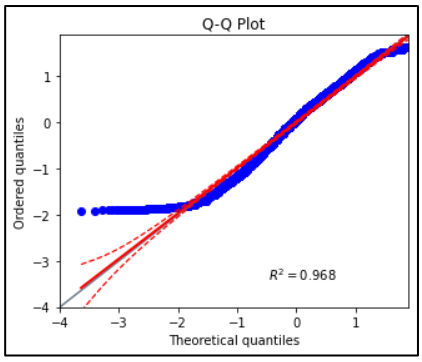


Fig.11 : *Test Levene (H0 : les variances des deux groupes sont égales, H1 : au moins une des variances est différente), Test Jarque-Bera (H0 : les échantillons suivent une distribution normale, H1 : Les échantillons ne suivent pas une distribution normale) et Q-Q plot de la variable âge.*

Le test de Levene a montré une p-value supérieure à 0.05 donc on accepte l'hypothèse nulle et on peut conclure que les variances des différents groupes sont équivalentes. Le test de Jarque-Bera a montré une p-value inférieure à 0.05 donc l'hypothèse 0 est rejetée et on peut conclure que la variable ne suit pas une distribution normale. Ceci est également vérifié à l'aide du Q-Q plot où l'on observe une courbe bleue qui ne se trouve pas dans les limites en pointillés rouge. Vu que la condition d'homoscédasticité est respectée, un test d'ANOVA welch est réalisé et également un test de kruskal-wallis, qui est un test non paramétrique (figure 12).

A

Test ANOVA welch : Age	
P-value	2.11e-95

B

Test Kruskal-Wallis : Age	
P-value	3.72e-71

Fig.12 : *Test ANOVA welch (H0 : Les deux moyennes ne sont pas égales, H1 : les deux moyennes sont égales), Test Kruskal-Wallis (H0 : Les médianes des populations sont égales, H1 : les moyennes des populations sont différentes).*

Les tests d'ANOVA welch et de Kruskal-Wallis ont montré une p-value inférieure à 0.05 donc on accepte l'hypothèse H0 et on peut conclure que les moyennes et les médianes entre les deux groupes sont significativement différentes. Il y a une dépendance entre l'âge et l'atteinte ou non d'un AVC.

C. Hypertension artérielle chez les individus

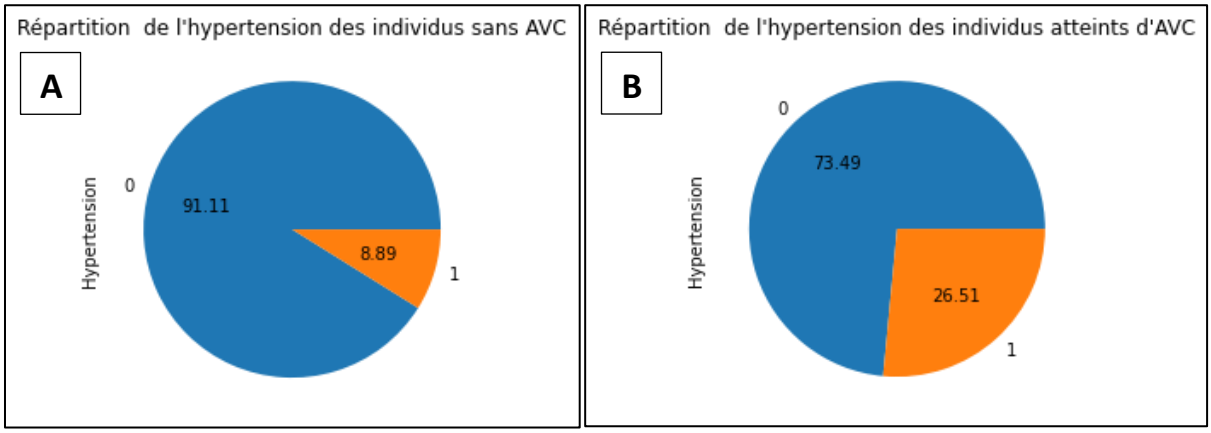
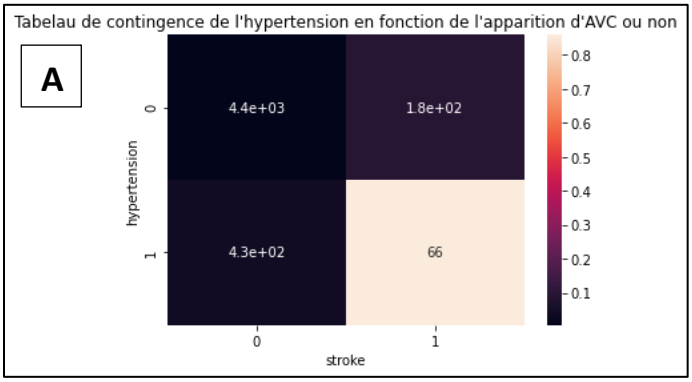


Fig.13 : *Représentation du pourcentage d'hypertension parmi les individus avec et sans AVC.*

En ce qui concerne la répartition de personnes atteintes d'hypertension entre les deux groupes, on observe 9 % de personnes atteintes par l'hypertension pour le groupe sans AVC (figure 13A) et 27 % pour le groupe d'individus atteints par un AVC (figure 7B). Les personnes atteintes d'hypertension ont tendance à être plus touchées par les AVC.

Afin de déterminer s'il y a une dépendance entre l'hypertension et l'atteinte d'une personne d'un AVC ou non, un tableau de contingence (figure 14A) et un test de Chi2 ont été réalisés (figure 14B).



B

Test Chi2 : hypertension	
P-value	1.66e-19

Fig.14 : *Tableau de contingence de l'hypertension des individus en fonction de l'apparition d'AVC et test Chi2 (H0 : Il n'y a pas de différence significative entre l'hypertension et l'apparition d'AVC, H1 : Il y a une différence significative entre l'hypertension et l'apparition d'AVC).*

Le tableau de contingence montre une différence dans le nombre total de personnes atteintes d'hypertension dans le groupe AVC. Le test de Chi2 a montré une P-value inférieure

à 0.05 donc on rejette l'hypothèse nulle et on peut conclure qu'il y a une différence significative entre les personnes atteintes d'hypertension et l'apparition d'AVC.

D. Maladie cardiaque chez les individus

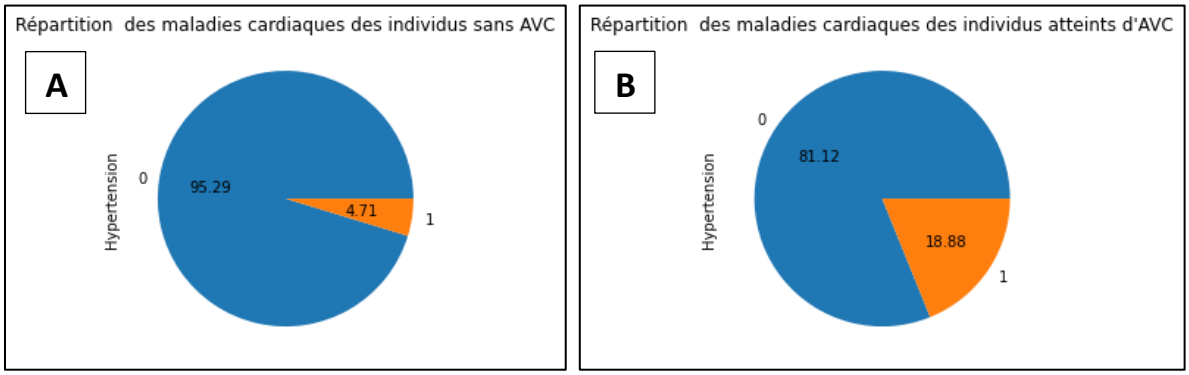
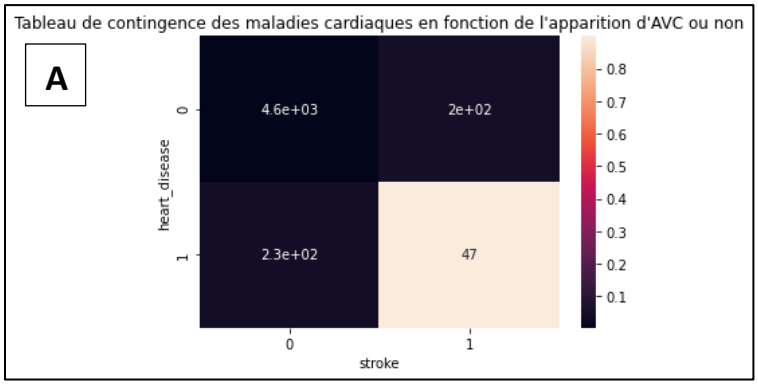


Fig.14 : *Représentation du pourcentage des maladies cardiaques parmi les individus avec et sans AVC.*

En ce qui concerne la répartition de personnes atteintes de maladie cardiaque entre les deux groupes, on observe 5 % de personnes atteintes par une maladie cardiaque pour le groupe sans AVC (figure 14A) et 19 % pour le groupe d'individus atteints par un AVC (figure 14B). Les personnes atteintes par les maladies cardiaques ont tendance à être plus touché par les AVC.

Afin de déterminer s'il y a une dépendance entre les maladies cardiaques et l'atteinte d'une personne d'un AVC ou non, un tableau de contingence (figure 15A) et un test de Chi2 ont été réalisé (figure 15B).



B

Test Chi2 : maladie cardiaque	
P-value	2.08e-21

Fig.15 : *Tableau de contingence des maladies cardiaques des individus en fonction de l'apparition d'AVC et test Chi2 (H0 : Il n'y a pas de différence significative entre les maladies cardiaques et l'apparition d'AVC, H1 : Il y a une différence significative entre les maladies cardiaques et l'apparition d'AVC).*

Le tableau de contingence montre une différence dans le nombre total de personnes atteintes de maladie cardiaque dans le groupe AVC. Le test de Chi2 a montré une p-value inférieure à 0.05 donc on rejette l'hypothèse nulle et on peut conclure qu'il y a une différence significative entre les personnes atteintes de maladie cardiaques et l'apparition d'AVC.

E. Statuts mariés des individus

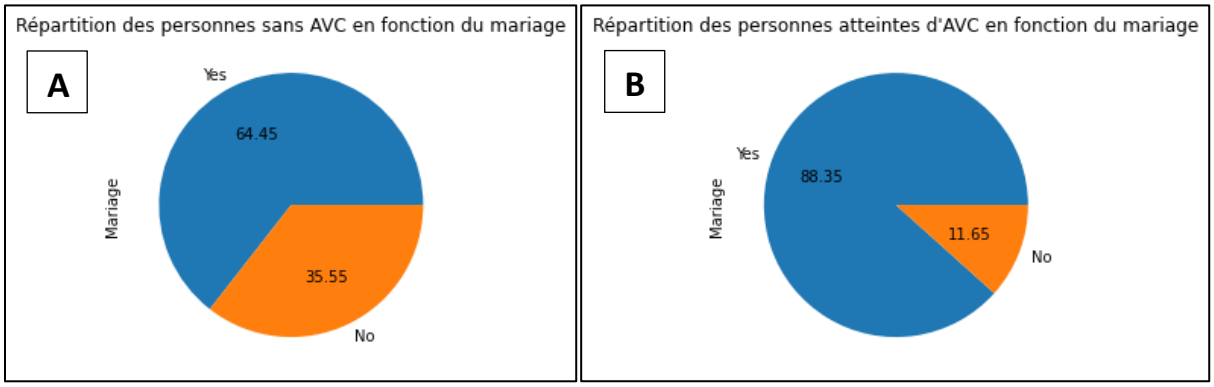
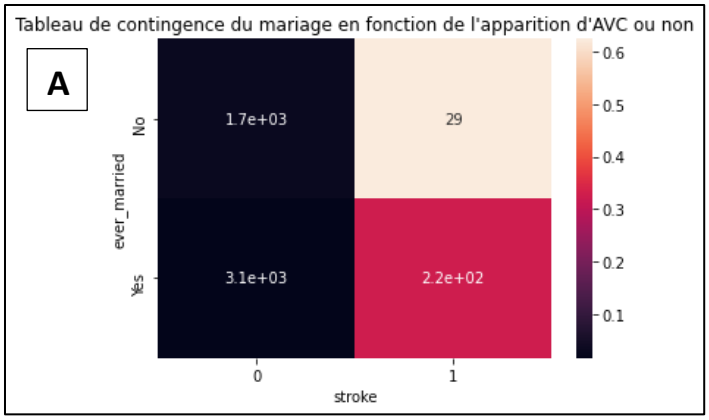


Fig.16 : *Représentation du pourcentage de statut mariée parmi les individus avec et sans AVC.*

En ce qui concerne la répartition de personnes mariées entre les deux groupes, on observe 64 % de personnes mariées pour le groupe sans AVC (figure 16A) et 88 % pour le groupe d'individus atteints par un AVC (figure 16B). Les personnes mariées ont tendance à être plus touché par les AVC.

Afin de déterminer s'il y a une dépendance entre le mariage et l'atteinte d'une personne d'un AVC ou non, un tableau de contingence (figure 17A) et un test de Chi2 ont été réalisés (figure 17B).



B

Test Chi2 : mariage	
P-value	1.63e-14

Fig.17 : *Tableau de contingence du statut mariée des individus en fonction de l'apparition d'AVC et test Chi2 (H0 : Il n'y a pas de différence significative entre le mariage et l'apparition d'AVC, H1 : Il y a une différence significative entre le mariage et l'apparition d'AVC).*

Le tableau de contingence montre une différence dans le nombre total de personnes mariées dans le groupe AVC. Le test de Chi2 a montré une p-value inférieure à 0.05 donc on rejette l'hypothèse nulle et on peut conclure qu'il y a une différence significative entre les personnes mariées et l'apparition d'AVC.

F. Type de travail des individus

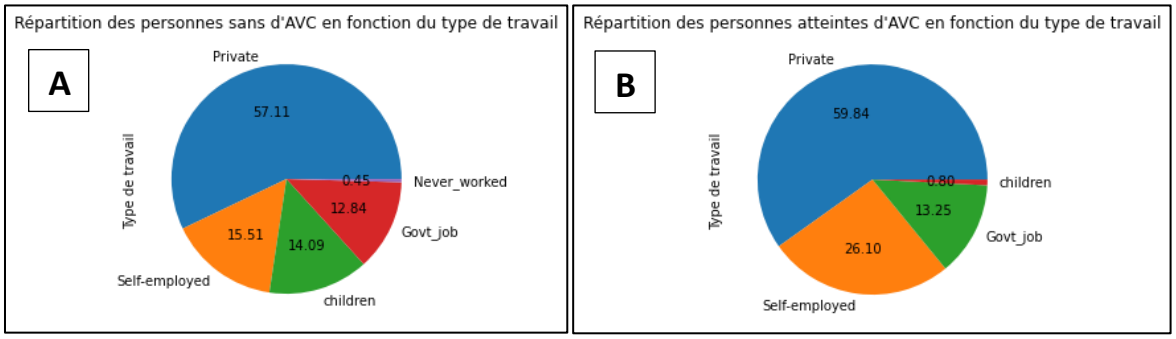
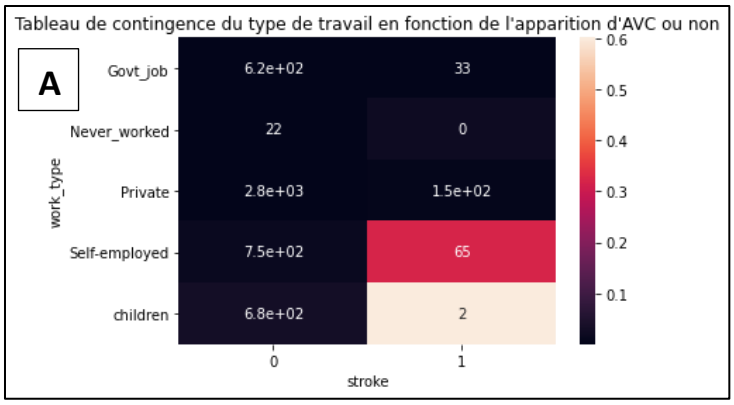


Fig.18 : *Représentation du pourcentage de type de travail parmi les individus avec et sans AVC.*

En ce qui concerne la répartition de type de travail des individus entre les deux groupes, on observe 57 % de personnes qui travaillent dans le privé, 16% comme travailleur indépendant pour le groupe sans AVC (figure 18A) et pour le groupe sans AVC 60 % de personnes qui travaillent dans le privé, 26% comme travailleur indépendant (figure 18B).

Afin de déterminer s'il y a une dépendance entre le type de travail et l'atteinte d'une personne d'un AVC ou non, un tableau de contingence (figure 19A) et un test de Chi2 ont été réalisés (figure 19B).



B

Test Chi2 : travail	
P-value	5.39e-10

Fig.19 : *Tableau de contingence du type de travail des individus en fonction de l'apparition d'AVC et test Chi2 (H0 : Il n'y a pas de différence significative entre le type de travail et l'apparition d'AVC, H1 : Il y a une différence significative entre le type de travail et l'apparition d'AVC).*

Le tableau de contingence montre une différence dans le nombre total de personnes qui sont travailleur indépendant dans le groupe AVC. Le test de Chi2 a montré une p-value inférieure à 0.05 donc on rejette l'hypothèse nulle et on peut conclure qu'il y a une différence significative entre le type de travail des personnes et l'apparition d'AVC.

G. Type de résidence des individus

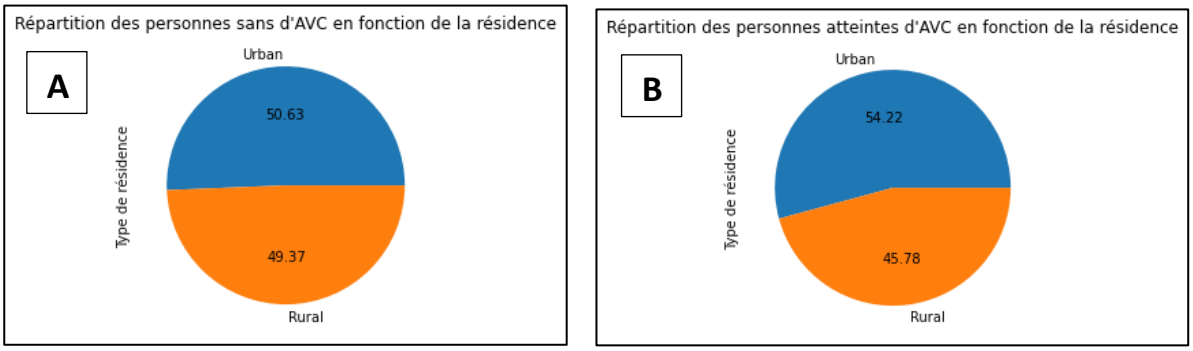
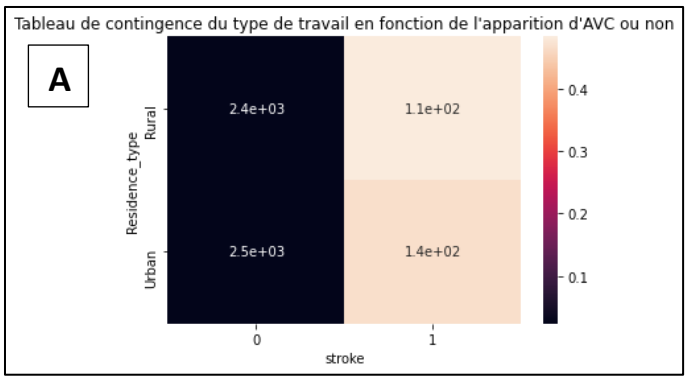


Fig.20 : *Représentation du pourcentage de type de résidence parmi les individus avec et sans AVC.*

En ce qui concerne la répartition du type de résidence des personnes entre les deux groupes, on observe 51 % des personnes qui habitent dans une résidence urbaine pour le groupe sans AVC (figure 20A) et 54 % pour le groupe d'individus atteints par un AVC (figure 20B). Cette répartition de résidence est approximativement équivalente entre les deux groupes

Afin de déterminer s'il y a une dépendance entre le type de résidence et l'atteinte d'une personne d'un AVC ou non, un tableau de contingence (figure 21A) et un test de Chi2 ont été réalisés (figure 21B).



B

Test Chi2 : résidence	
P-value	0.298

Fig.21 : *Tableau de contingence du type de résidence des individus en fonction de l'apparition d'AVC et test Chi2 (H0 : Il n'y a pas de différence significative entre le type de résidence et l'apparition d'AVC, H1 : Il y a une différence significative entre le type de résidence et l'apparition d'AVC).*

Le tableau de contingence montre une différence dans le nombre total de résidence urbaine et rurale dans le groupe AVC. Le test de Chi2 a montré une p-value supérieure à 0.05 donc on accepte l'hypothèse nulle et on peut conclure qu'il n'y a pas de différence significative le type de résidence et l'apparition d'AVC.

H. Statut fumeur des individus

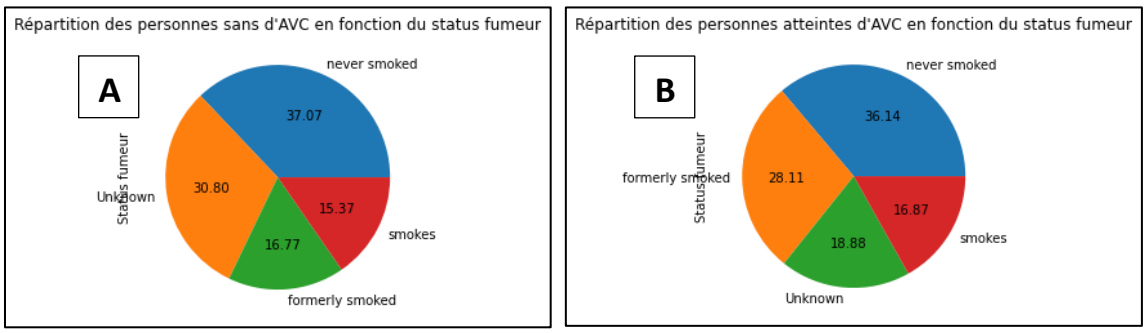
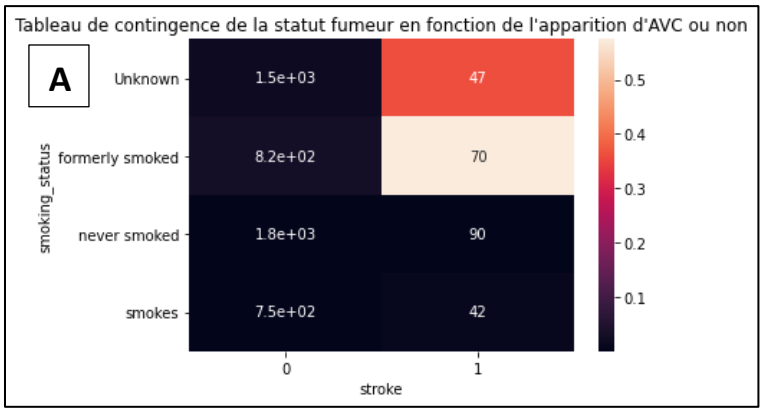


Fig.22 : *Représentation du pourcentage des statuts fumeur parmi les individus avec et sans AVC.*

En ce qui concerne la répartition du statut fumeur des individus entre les deux groupes, on observe 37 % de personnes qui n'ont jamais fumées, 30 % pour qui on ne connaît pas le statut pour le groupe sans AVC (figure 22A) et pour le groupe sans AVC 36 % de personnes qui n'ont jamais fumées, 28% pour qui on ne connaît pas la statue (figure 22B). Cette répartition du statut fumeur est approximativement équivalente entre les deux groupes.

Afin de déterminer s'il y a une dépendance entre le statut fumeur et l'atteinte d'une personne d'un AVC ou non, un tableau de contingence (figure 23A) et un test de Chi2 ont été réalisés (figure 23B).



B

Test Chi2 : fumeur	
P-value	0.00

Fig.21 : *Tableau de contingence du statut fumeur des individus en fonction de l'apparition d'AVC et test Chi2 (H0 : Il n'y a pas de différence significative entre le statut et l'apparition d'AVC, H1 : Il y a une différence significative entre le statut fumeur et l'apparition d'AVC).*

Le tableau de contingence montre une différence dans le nombre total dont on ne connaît pas le statut fumeur dans le groupe AVC. Le test de Chi2 a montré une p-value inférieure à 0.05 donc on rejette l'hypothèse nulle et on peut conclure qu'il y a une différence significative entre le statut fumeur des personnes et l'apparition d'AVC. Cependant cette conclusion n'est pas prise en compte dans mes analyses parce que le nombre de personnes pour qui on ne connaît pas le statut fumeur biaise le test statistique car il le considère comme une réelle condition. Pour cela cette variable ne va pas être utilisée pour le reste de l'étude.

I. Indice de masse corporelle des individus

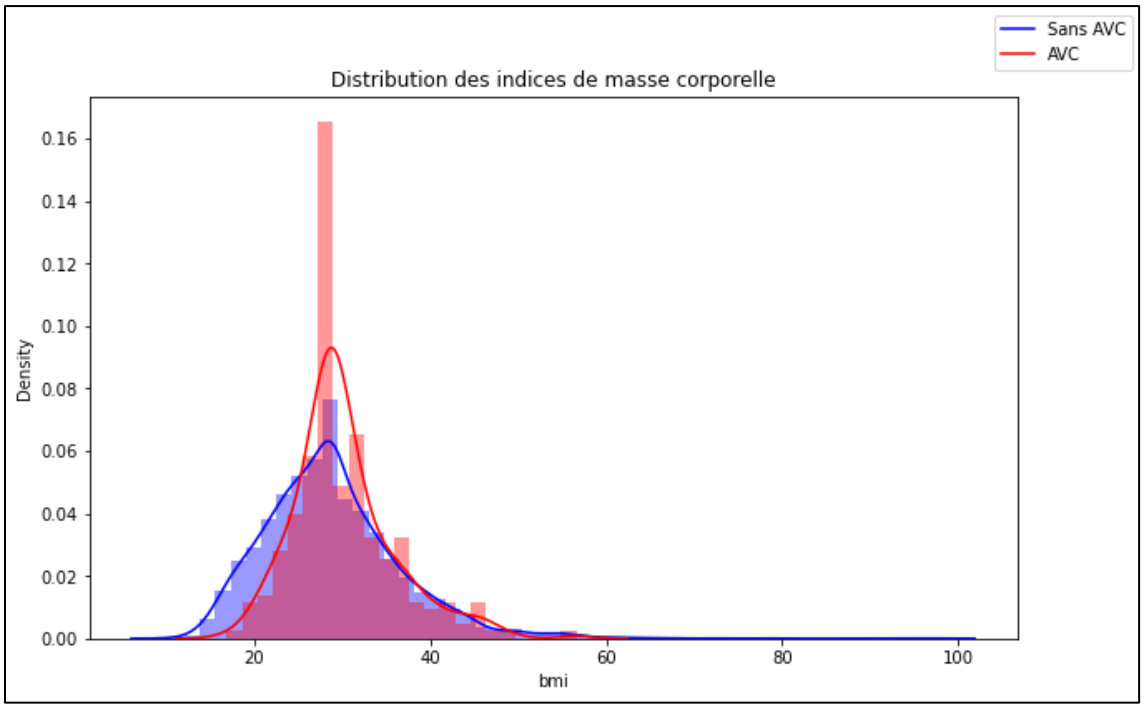


Fig.22 : *Histogramme représentant l'indice de masse corporelle des individus avec et sans AVC.*

Lorsqu'on observe la fréquence de l'indice de masse corporelle des deux groupes, on remarque que pour les personnes atteintes et non atteintes d'AVC les fréquences de masse corporelle sont situées entre 20 et 40 (figure 22). Cependant pour les personnes atteintes d'AVC on observe plus de personnes qui ont un indice de masse corporelle de 30 (courbe rouge).

Afin d'observer la différence de l'indice de masse corporelle en fonction des deux groupes un graphique en boîte à moustache a été réalisé (figure 23).

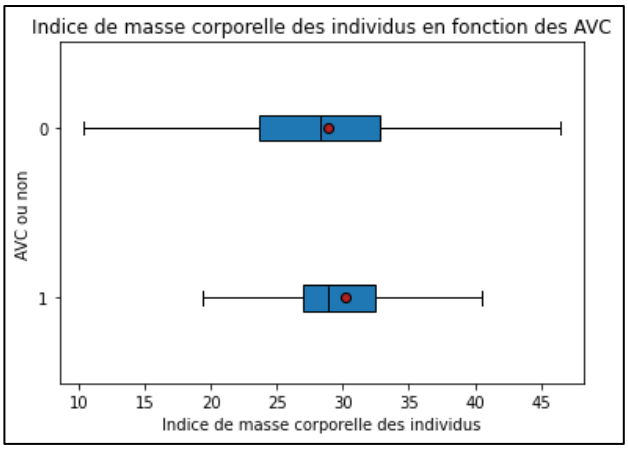


Fig.23 : *Boîtes à moustaches représentant l'indice de masse corporelle des individus en fonction de l'apparition d'AVC ou non.*

Sur cette représentation on remarque que l'indice de masse corporelle des individus atteints d'AVC à tendance à être entre 27 et 32 et serait donc un peu plus élevé.

Pour vérifier la normalité un test de Jarque – Bera et un qq-plot ont été effectués (figure 24A-B).

A

Test Jarque-Bera : Age	
P-value	0

B

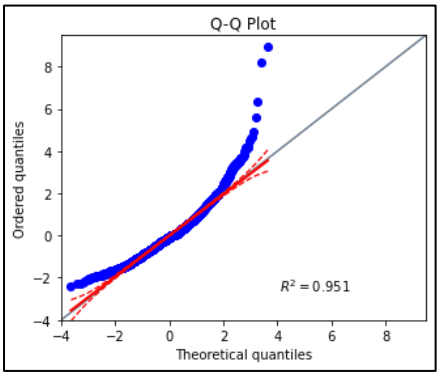


Fig.24 : *Test Jarque-Bera (H0 : les échantillons suivent une distribution normale, H1 : Les échantillons ne suivent pas une distribution normale) et Q-Q plot de la variable Bmi.*

Le test de Jarque-Bera a montré une p-value inférieure à 0.05 donc l'hypothèse nulle est rejetée et on peut conclure que la variable ne suit pas une distribution normale. Ceci est également vérifié à l'aide du Q-Q plot où l'on observe une courbe bleue qui ne suit pas la droite. Vu que la condition d'homoscédasticité est respectée, un test d'ANOVA welch est réalisé et également un test de kruskal-wallis (figure 25).

A

Test ANOVA welch : bmi	
P-value	0.00

B

Test Kruskal-Wallis : bmi	
P-value	0.00

Fig.25 : *Test ANOVA welch (H0 : Les deux moyennes ne sont pas égales, H1 : les deux moyennes sont égales), Test Kruskal-Wallis (H0 : Les médianes des populations sont égales, H1 : les moyennes des populations sont différentes).*

Les tests d'ANOVA welch et de Kruskal-Wallis ont montrés une p-value inférieure à 0.05 donc on accepte l'hypothèse H0 et on peut conclure que les moyennes et les médianes entre les deux groupes sont significativement différentes. Il y a une dépendance entre l'indice de masse corporelle et l'atteinte ou non d'un AVC.

J. Taux de glucose dans le sang

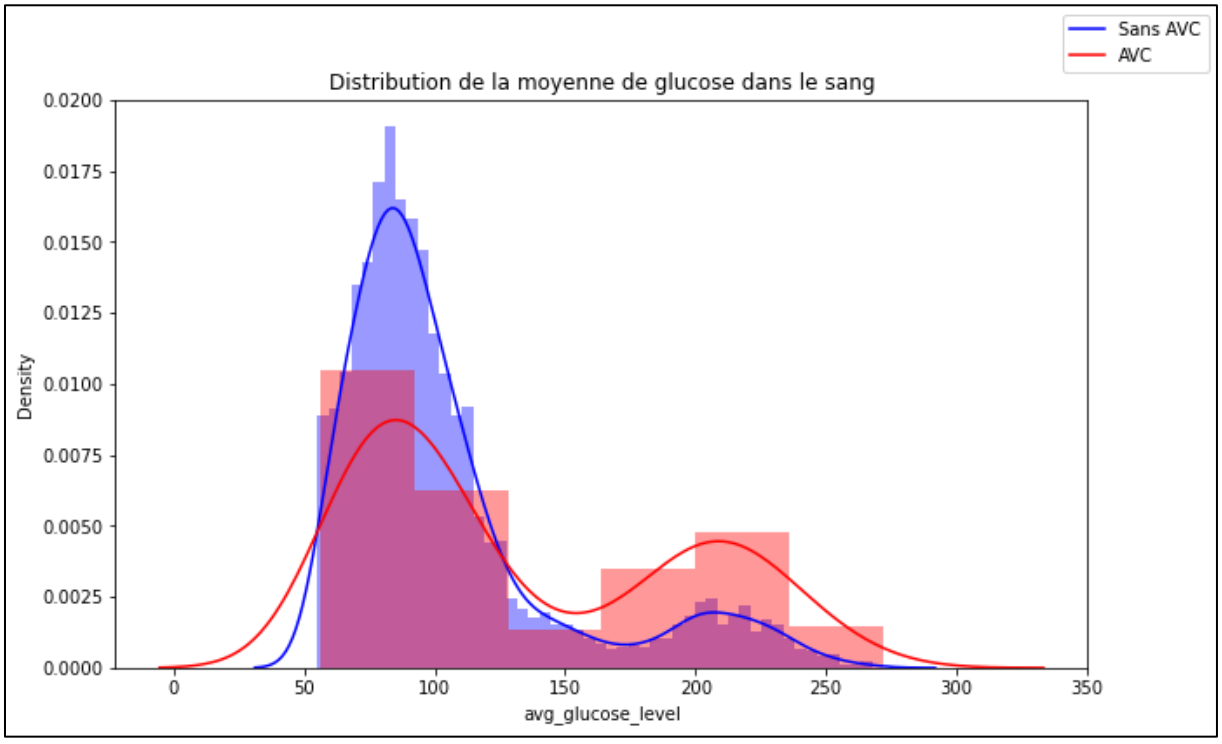


Fig.26 : *Histogramme représentant le taux de glucose dans le sang des individus avec et sans AVC.*

Lorsqu'on observe la fréquence du taux de glucose dans le sang des deux groupes, on remarque que pour les personnes atteintes et non atteintes d'AVC les fréquences de masse corporelle sont situées entre 50 et 250 (figure 26). Cependant pour les personnes atteintes d'AVC on observe plus de personnes qui ont un taux de glucose élevé entre 170 et 250 (courbe rouge).

Afin d'observer la différence de glucose dans le sang en fonction des deux groupes un graphique en boîte à moustache a été réalisé (figure 27).

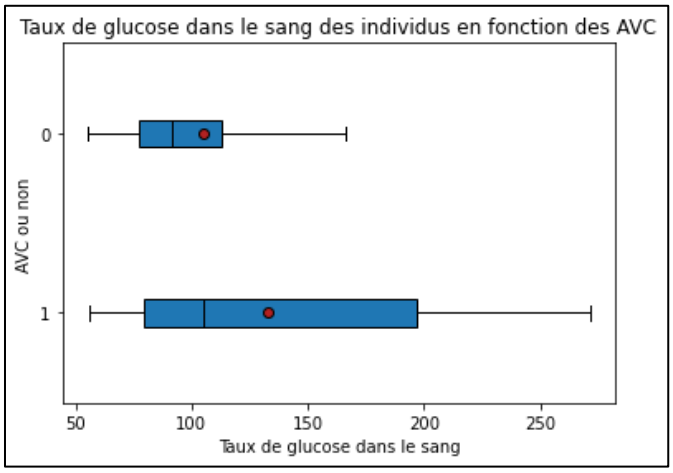


Fig.27 : *Boîtes à moustaches représentant le taux de glucose dans le sang des individus en fonction de l'apparition d'AVC ou non.*

Sur cette représentation on remarque que le taux de glucose dans le sang des individus atteints d'AVC a tendance à être entre 60 et 200 et serait plus élevé.

Pour vérifier la normalité un test de Jarque – Bera et un qq-plot ont été effectués (figure 28A-B).

A

Test Jarque-Bera : Age	
P-value	0

B

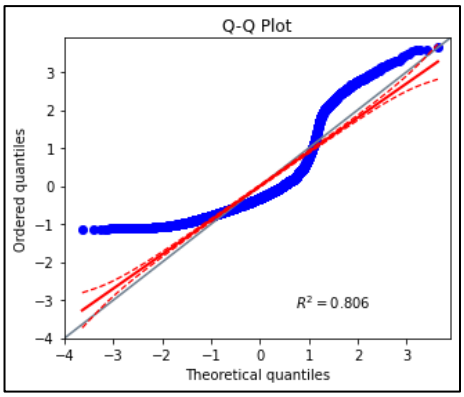


Fig.28 : *Test Jarque-Bera (H0 : les échantillons suivent une distribution normale, H1 : Les échantillons ne suivent pas une distribution normale) et Q-Q plot de la variable taux de glucose dans le sang.*

Le test de Jarque-Bera a montré une p-value inférieure à 0.05 donc l'hypothèse nulle est rejetée et on peut conclure que la variable ne suit pas une distribution normale. Ceci est également vérifié à l'aide du Q-Q plot où l'on observe une courbe bleue qui ne suit pas la droite. Vu que la condition d'homoscédasticité est respectée, un test d'ANOVA welch est réalisé et également un test de kruskal-wallis (figure 29).

A

Test ANOVA welch : glucose	
P-value	2.40e-11

B

Test Kruskal-Wallis : glucose	
P-value	3.6e-9

Fig.29 : *Test ANOVA welch (H0 : Les deux moyennes ne sont pas égales, H1 : les deux moyennes sont égales), Test Kruskal-Wallis (H0 : Les médianes des populations sont égales, H1 : les moyennes des populations sont différentes).*

Les tests d'ANOVA welch et de Kruskal-Wallis ont montrés une p-value inférieure à 0.05 donc on accepte l'hypothèse H0 et on peut conclure que les moyennes et les médianes entre les deux groupes sont significativement différentes. Il y a une dépendance entre le taux de glucose dans le sang et l'atteinte ou non d'un AVC.

K. Diabète chez les individus

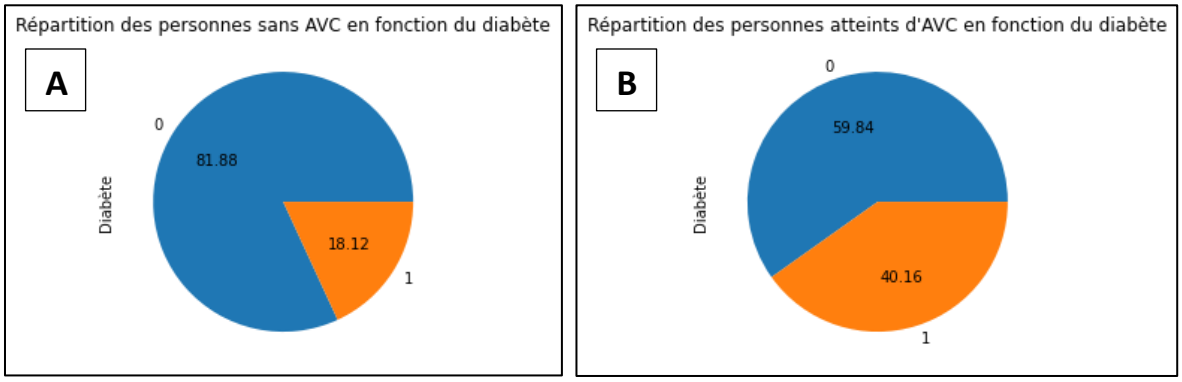
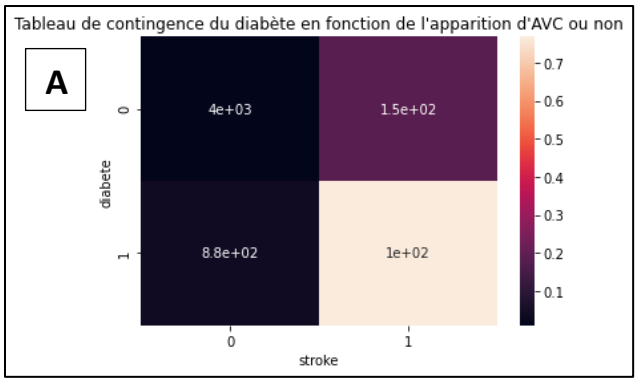


Fig.30 : *Représentation du pourcentage de personnes atteintes du diabète parmi les individus avec et sans AVC.*

En ce qui concerne la répartition de personnes atteintes de diabète entre les deux groupes, on observe 18 % de personnes atteintes par le diabète pour le groupe sans AVC (figure 30A) et 40 % pour le groupe d'individus atteints par un AVC (figure 30B). Les personnes atteintes par le diabète ont tendance à être plus touché par les AVC.

Afin de déterminer s'il y a une dépendance entre le diabète et l'atteinte d'une personne d'un AVC ou non, un tableau de contingence (figure 31A) et un test de Chi2 ont été réalisés (figure 31B).



B

Test Chi2 : diabète	
P-value	1.48e-17

Fig.31 : *Tableau de contingence du diabète des individus en fonction de l'apparition d'AVC et test Chi2 (H0 : Il n'y a pas de différence significative entre le diabète et l'apparition d'AVC, H1 : Il y a une différence significative entre le diabète et l'apparition d'AVC).*

Le tableau de contingence montre une différence dans le nombre total de personnes atteintes de diabète dans le groupe AVC. Le test de Chi2 a montré une p-value inférieure à 0.05 donc on rejette l'hypothèse nulle et on peut conclure qu'il y a une différence significative entre les personnes atteintes de diabète et l'apparition d'AVC.

Afin d'avoir une meilleure visualisation des variables significatives dans l'apparition d'AVC, un tableau récapitulatif a été créé.

Variables	*P-value*	*Apparition AVC*
Sexe	0.78	Indépendant
Age	2.11e-95	Dépendant
Hypertension artérielle	1.66e-19	Dépendant
Maladie cardiaque	2.09e-21	Dépendant
Mariage	1.63e-14	Dépendant
Type de travail	5.4e-10	Dépendant
Type de résidence	0.3	Indépendant
Statut fumeur	0.0	Dépendant mais pas pris en compte
Indice de masse corporelle	0.0	Dépendant
Taux de glucose dans le sang	2.4e-11	Dépendant
Diabète	1.48e-17	Dépendant

Fig.32 : *Tableau récapitulatif de significativité des différentes variables.*

En parallèle de ces analyses sur python, un dashboard à l'aide du logiciel Tableau a été créé (figure 33).

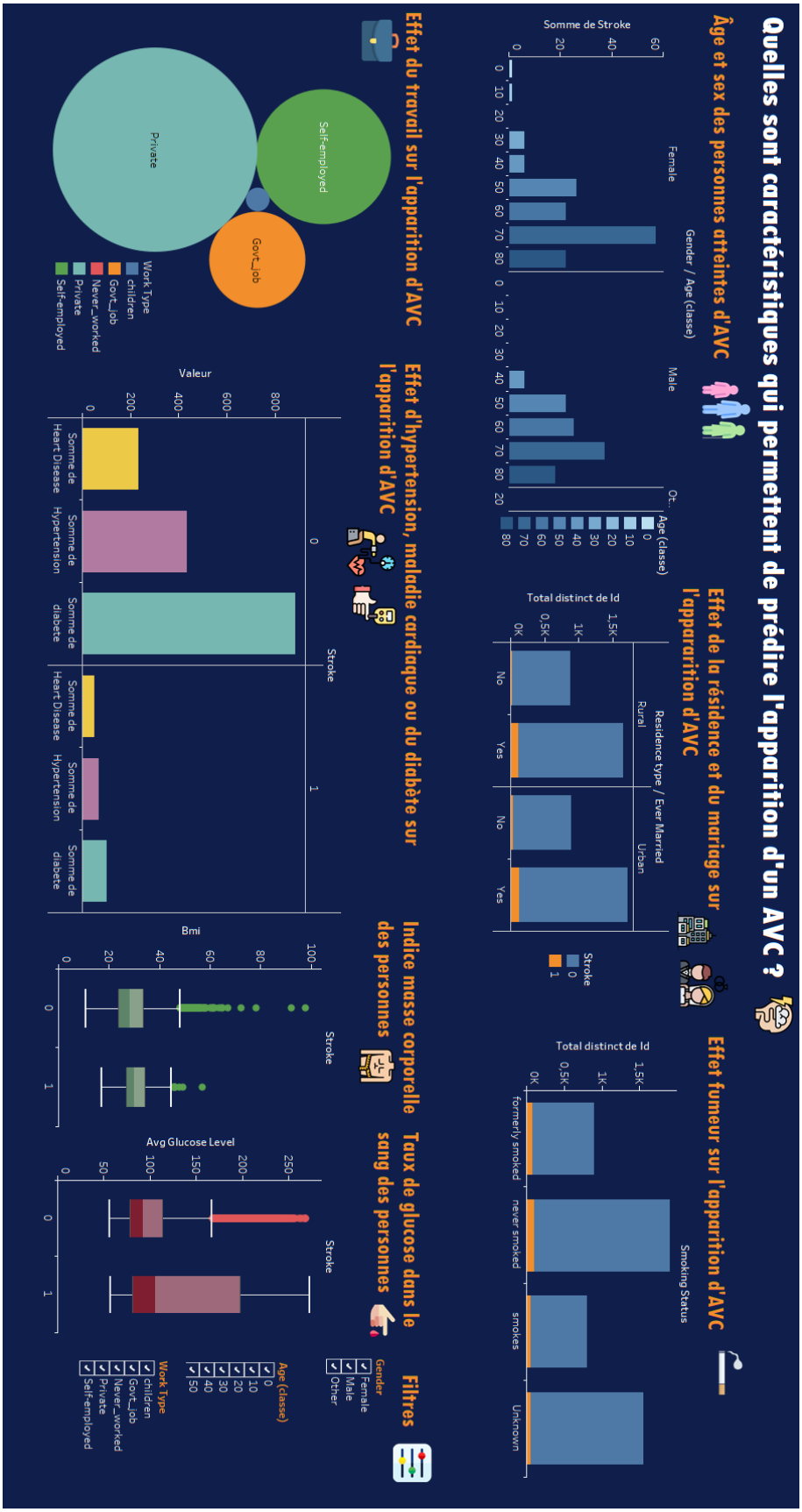


Fig.33 : *Dashboard des données à l'aide du logiciel Tableau.*

Sur ce dashboard on observe les différentes caractéristiques des personnes atteintes ou non d'AVC, ainsi que des filtres qui permettent de sélectionner, le sexe, la tranche d'âge ou le type de travail qu'on veut observer.

Afin de mieux visualiser le jeu de données, une analyse en composantes principales (ACP) a été réalisée. L'APC permet de résumer l'information des données dans un certain nombre de composantes principales dans le but de représenter les données sur un plan simple.

IV. Analyse en composante principale

Dans le but de déterminer le nombre de composantes principales une courbe de variance accumulée a été réalisée (figure 34).

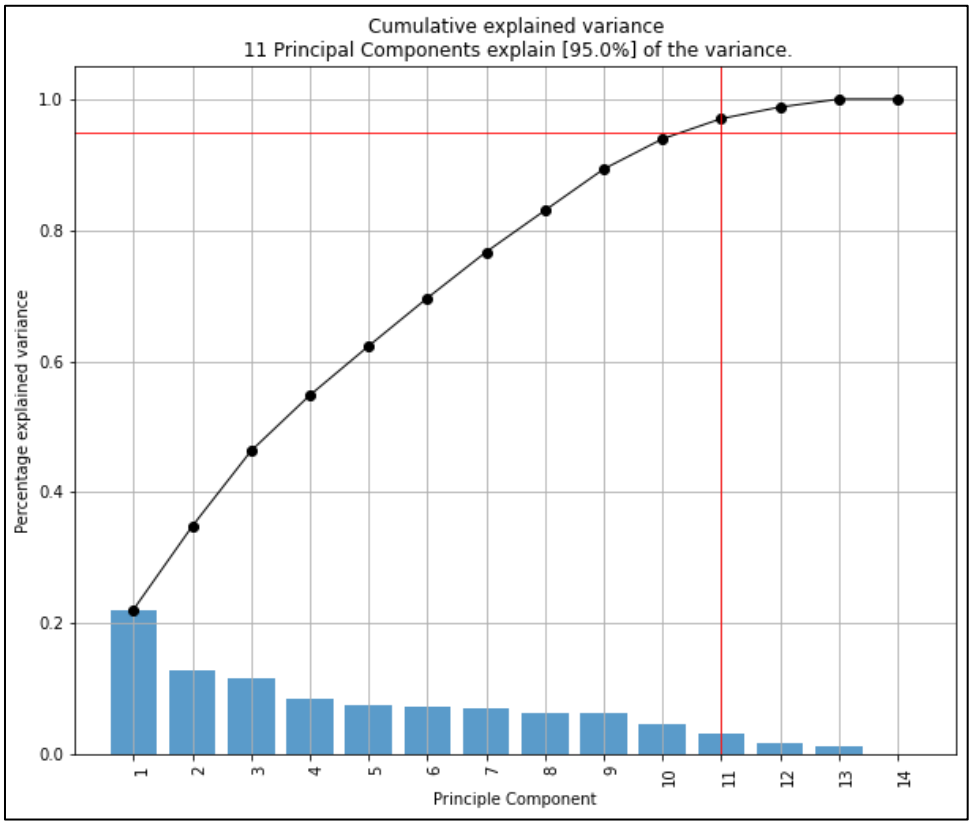


Fig.34 : *Courbe de variance accumulée pour la détermination du nombre de composante.*

On remarque que 95% de la variance est expliqué à 11 composantes principales et pour 3 composantes principales uniquement 45% de la variance est expliqué. Lorsqu'on représente les composantes dans un graphique 3D on observe la figure suivante (figure 35).

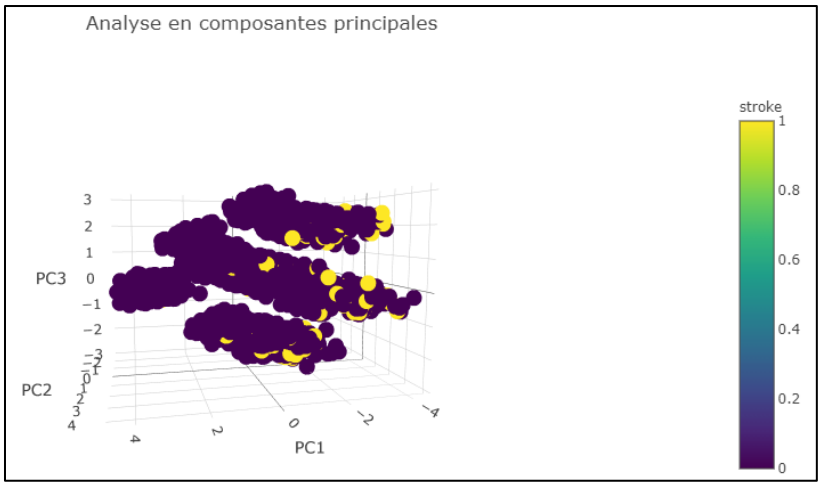


Fig.35 : *Graphique 3D de l'analyse en composante principale*

On remarque qu'avec uniquement 3 composantes principales il n'y a pas suffisamment d'information afin de séparer les deux groupes AVC et non AVC. Afin de visualiser la corrélation entre les différentes variables un cercle de corrélation a été réalisé (figure 36).

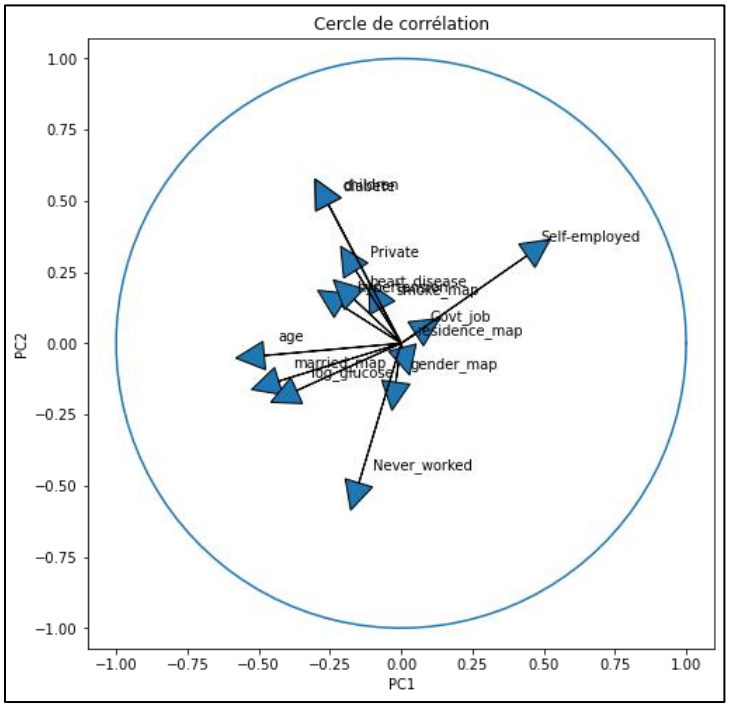


Fig.36 : *Cercle de corrélation des composantes principales 1 et 2.*

Sur ce cercle de corrélation on n'arrive pas réellement à avoir des informations sur la corrélation entre les différentes variables, cela est dû au fait qu'ils sont trop nombreux et qu'il n'y a pas assez de variance qui puisse expliquer les différentes variables.

Afin d'avoir une visualisation plus correcte de nos données, la technique de t-SNE a été utilisée (figure 37), ceci est une autre technique non-linéaire de réduction de dimension adaptée à la visualisation de données de grande dimension. L'algorithme calcule la distribution de probabilité des paires de points (les points proches ont une forte probabilité et les éloignés une faible) afin de les comparer et minimiser la divergence entre les deux distributions.

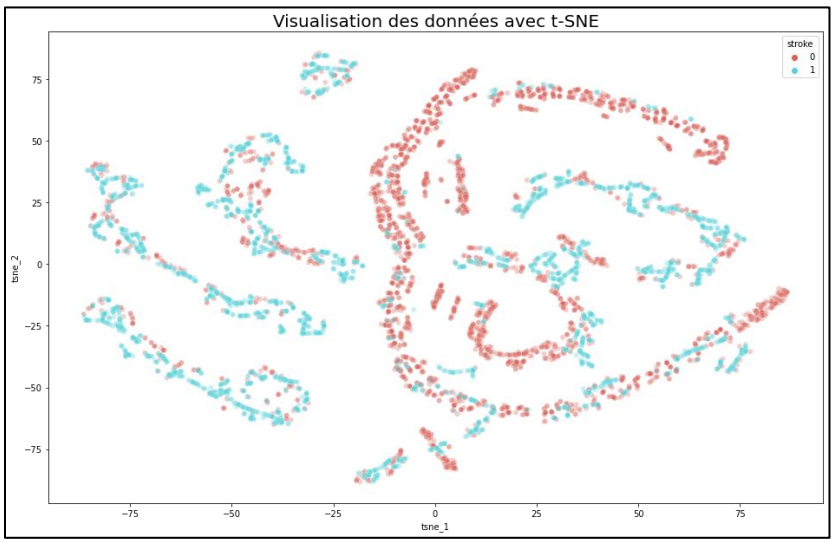


Fig.37 : *Visualisation des données à l'aide de la technique t-SNE.*

Dans cette représentation des données avec des dimensions réduites on observe deux groupes assez distincts, même si certains des points se retrouvent mélangés aux autres groupes.

Par la suite différents modèles de classifications sont réalisés dans le but de prédire l'apparition d'AVC. La régression logistique est une méthode statistique de prédiction de classes binaires qui provient d'une transformation de régression linéaire en logistique. La première étape de la régression est l'utilisation de Train test split qui permet de découper le dataset en données d'entrainement et données test. Par la suite le modèle est entrainé sur les données d'entrainement (X_train et Y_train) et les probabilités sont prédites. La performance du modèle de régression est évaluée à l'aide du classification report, qui donne des informations sur la précision et le recall.

V. Modèles de classification

A. *Classification à partir des données non équilibrées*

Afin d'observer l'utilité de l'équilibration des données, la première régression logistique a été réalisée avec les données déséquilibrées. La régression a été effectuée à l'aide de la librairie sklearn.linear_model et le classification report suivant a été obtenu (figure 38).

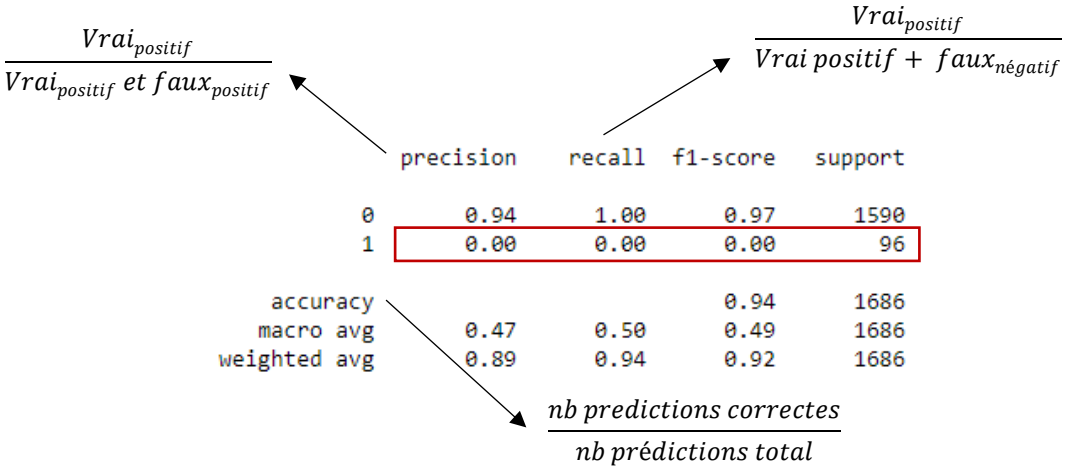


Fig.38 : *Classification report du modèle de régression à partir des données déséquilibrées.*

Le classification report donne plusieurs informations sur la performance du modèle, en effet la précision détermine le pourcentage des prédictions correctes donc l'exactitude des prévisions positives. Le recall détermine le pourcentage de cas positif correctement identifié. Le score F1 est une moyenne pondérée entre la précision et le recall (le meilleur score étant de 1 et le pire de 0). En ce qui concerne ce modèle, pour la variable stroke = 0 on observe 0.94 de précision et 1 de recall, ce qui veut dire il y a une bonne détection des personnes non atteintes d'AVC. Cependant pour la détection de personnes atteintes par un AVC, la précision et le recall se retrouvent à 0, ce qui veut dire qu'il y a une mauvaise détection des personnes atteintes d'AVC. Ceci est causé par les données déséquilibrées, en effet il n'y a que 249 personnes qui sont atteintes par des AVC dans notre dataset et le modèle favorise alors la prédiction de personnes non atteintes. Afin de remédier à ce problème, les données ont été équilibré à l'aide de la méthode SMOTE-Tomek.

B. *Classification à partir des données équilibrées avec SMOTE-Tomek*

SMOTE (Synthetic Minority Over-Sampling Technique) est une technique de suréchantillonnage qui permet de créer de nouveaux échantillons synthétiques de la classe

minoritaire (figure 39) et Tomek est une technique de sous-échantillonnage qui permet de supprimer les échantillons proches de la limite des deux classes afin d'augmenter la séparation des deux classes.

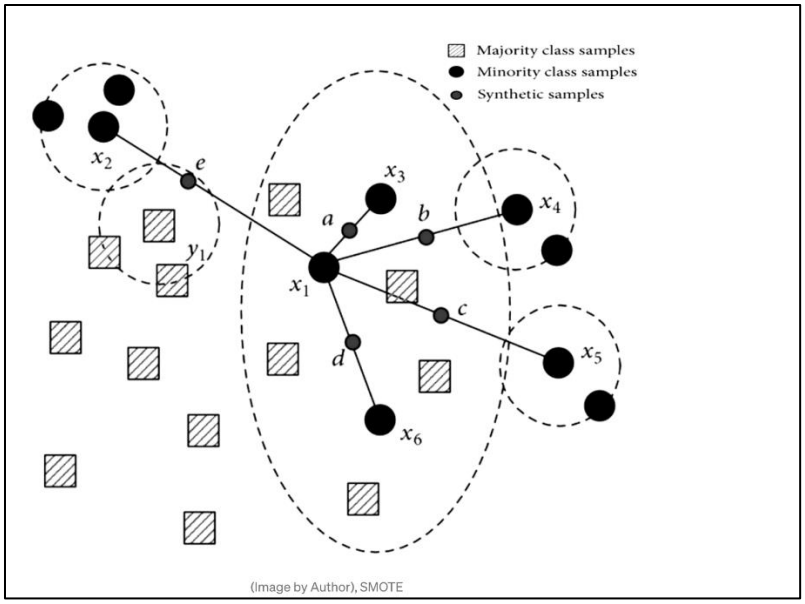


Fig.39 : *Principe de SMOTE pour générer de nouveaux échantillons dans la classe AVC.*

L'équilibrage des données s'est effectué sur les données X_train et y_train et a montré des valeurs équivalentes pour le groupe AVC et non AVC.

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

1) Classification avec LogisticRegression Classifier

La classification à l'aide de la librairie sklearn.linear_model a montré le classification report suivant (figure 40).

	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

Fig.40 : *Classification report du modèle de régression logistique sklearn à partir des données équilibrées.*

Pour la variable stroke = 0 on observe 0.97 de précision et 0.84 de recall, ce qui veut dire il y a une bonne détection des personnes non atteintes d'AVC. Pour la détection de personnes atteintes par un AVC, la précision se retrouve à 0.17 et le recall à 0.54, on remarque

une augmentation par rapport au modèle précèdent ce qui est dû à l'équilibrage des données. La matrice de confusion obtenue pour cette classification est la suivante (figure 41).

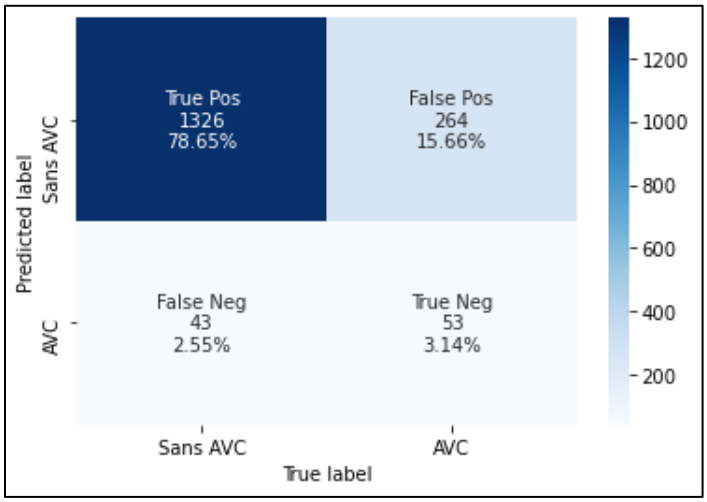


Fig.41 : *Matrice de confusion du modèle de la LogisticRegression.*

On observe que 78.65 % des non AVC sont prédit de manière correct, 3.14 % des AVC sont prédit de manière correct, 2.55 % des non AVC sont prédit comme étant des AVC et 15,66 % des AVC sont prédit comme des non AVC. Afin de mesurer la performance du classifieur une courbe de ROC a été tracé (figure 42).

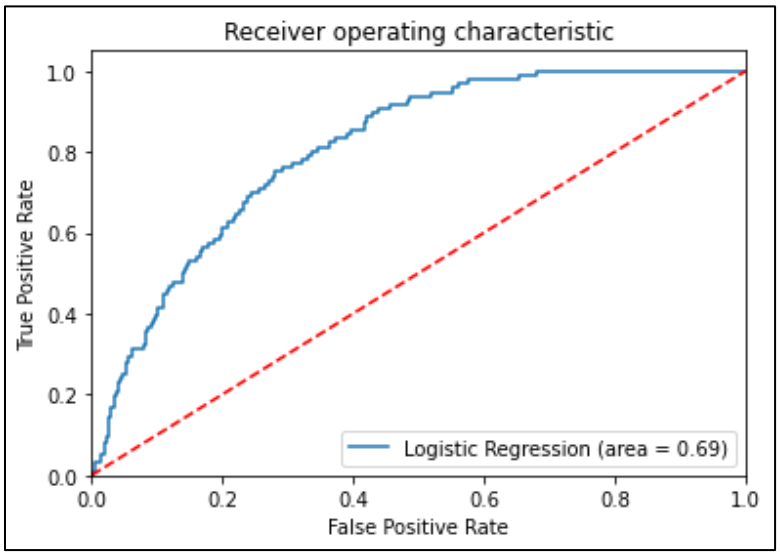


Fig.42 : *Courbe ROC du modèle de la LogisticRegression.*

Cette courbe représente le taux de vrai positif en fonction des faux positif et l'aire sous la courbe a permis d'obtenir un score de 0.69, ce qui est assez correct sachant qu'à 100% toutes les prédictions sont correctes. Afin d'évaluer le modèle plus précisément, la librairie SHAP a été utilisée et a permis d'obtenir les représentations suivantes (figure 43-44-45).

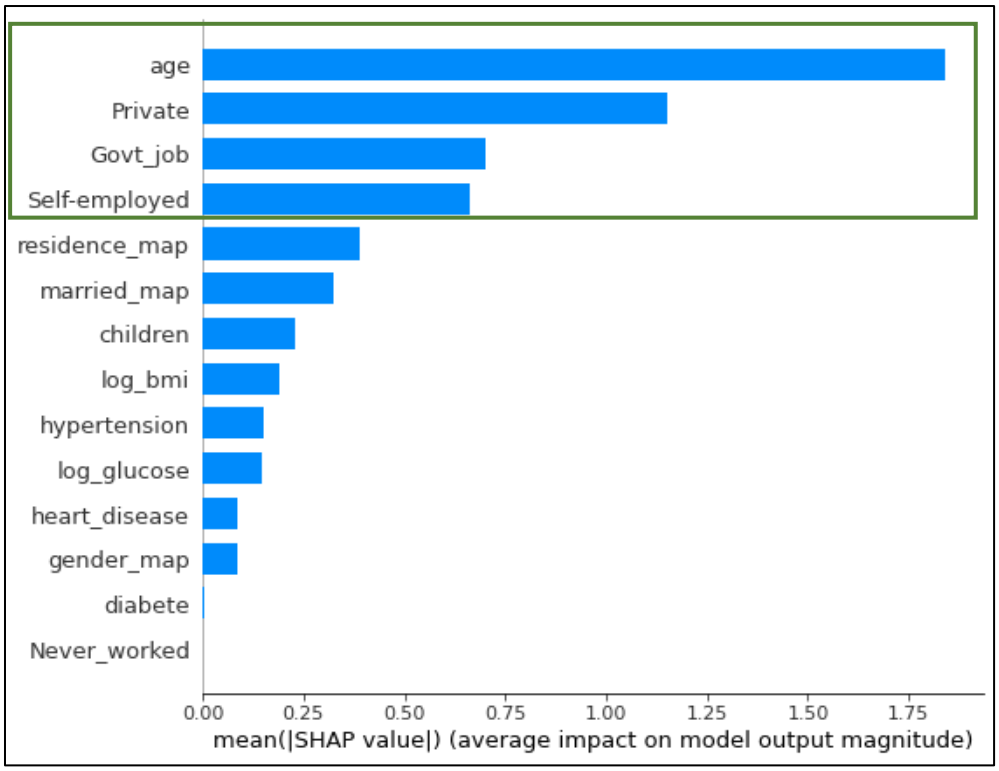


Fig.43 : *Interprétation des variables du modèle de la LogisticRegression avec SHAP.*

Ici on observe l'impact de chaque variable dans la prédiction, on remarque que les 4 variables qui ont le plus d'impact sont l'âge, et le type de travail qui est effectué par les individus. Dans le but d'observer l'effet négatif ou positifs des variables le graphique suivant a été réalisé (figure 44).

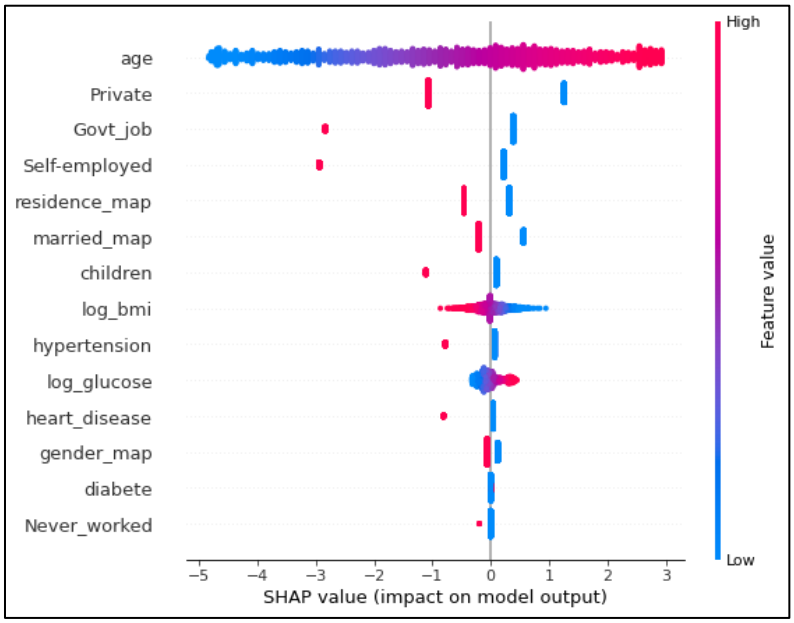


Fig.44 : *Interprétation des effet positif ou négatif des variables sur modèle de la LogisticRegression avec SHAP.*

On remarque que plus l'âge et le taux de glucose sont élevés plus les personnes ont tendances à être atteintes par un AVC et plus l'indice de masse corporelle est faible plus les personnes sont atteintes d'AVC. Afin de visualiser l'effet des variables dans la prédiction, les deux graphiques suivants ont été réalisés (figure 45A-B).

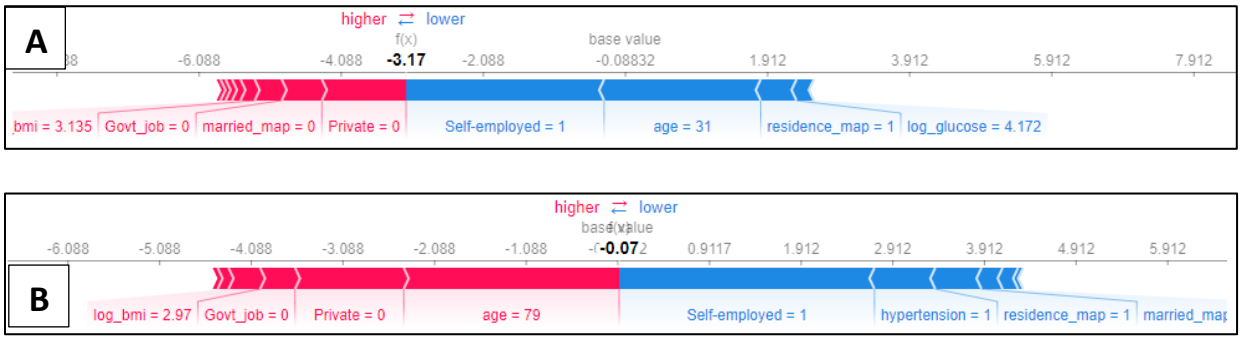


Fig.45 : *Interprétation des effets locaux des variables sur le modèle de la LogisticRegression avec SHAP.*

On observe dans la figure 45 A une personne non atteinte d'AVC, les variables travail indépendant, l'âge et le taux de glucose ont eu un effet important sur la prédiction. En ce qui concerne la personne prédite avec un AVC, les variables qui ont eu un effet sont l'âge et l'indice de masse corporelle. Par la suite, dans le but de comparer plusieurs modèles de classification, le Random Forest classifier a été utilisé.

2) Classification avec Random Forest Classifier

C'est un algorithme de classification qui calcule la moyenne des prévisions de plusieurs modèles d'arbre de décision pour réduire la variance des prévisions et donc l'erreur de prévision. La classification avec cet algorithme de forêts aléatoires a donné le classification report suivant (figure 46).

	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

Fig.46 : *Classification report du modèle de Random Forest à partir des données équilibrées.*

Cette classification a permis d'obtenir une précision de 0.96 et un recall de 0.87 pour la variable stroke = 0, et pour la détection de personnes atteintes d'AVC on observe une précision de 0.16 et un recall de 0.44. On remarque que par rapport à la classification précédente l'accuracy a augmentée jusqu'à 0.84.

3) Classification avec K Nearest Neighbors

Par la suite une classification à l'aide de l'algorithme K Nearest Neighbors a été réalisée. Le principe de cet algorithme est de choisir les k données les plus proches du point étudié dans le but d'en prédire sa valeur. Le modèle a donné le classification report suivant (figure 47).

	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

Fig.47 : *Classification report du modèle de K Nearest Neighbors à partir des données équilibrées.*

Cette classification a permis d'obtenir une précision de 0.96 et un recall de 0.84 pour la variable stroke = 0, et pour la détection de personnes atteintes d'AVC on observe une précision de 0.15 et un recall de 0.47. On remarque que l'accuracy du modèle est à 0.82.

En effet, ce qui est important dans le projet c'est la détection d'AVC et de favoriser la détection positive, pour cela le but est d'obtenir une valeur de recall pour la variable stroke = 1 la plus élevée. Afin de mieux visualiser les variables les plus importantes pour la prédiction d'AVC, l'algorithme RFE à été utilisé.

4) Sélection des variables les plus pertinentes avec la classification Random Forest

La RFE (Recursive Feature Elimination) est utilisée afin de sélectionner les variables du training dataset les plus pertinentes pour la prédiction (figure 48).

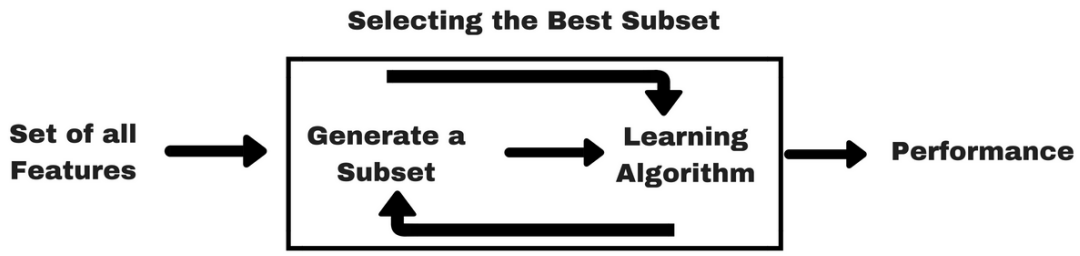


Fig.48 : *Schéma de la sélection d'attribut avec la RFE.*

Cette méthode supprime de manière récursive les attributs et créée un model sur les attributs restants. Ensuite elle utilise la précision du modèle dans le but d'identifier les combinaisons d'attributs qui contribuent le plus à la prédiction. La RFE à permis d'obtenir les résultats suivants (figure 49).

```
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
```

Fig.49 : *Visualisation des résultats de la RFE.*

Dans ces résultats de RFE on observe le nombre d'attribut correspondant qui vont de 2 à 16. Le recall de la classe positive est montré dans les résultats de la RFE. Le cross_val_score évalue le modèle de manière plus poussé avec une base de test différente de celle utilisée pour l'apprentissage. Pour la validation croisée, le dataset est découpé en 5 segments de façon aléatoire et on en utilise 4 pour l'apprentissage et 1 pour le test, ceci est répété 5 fois (figure 50).

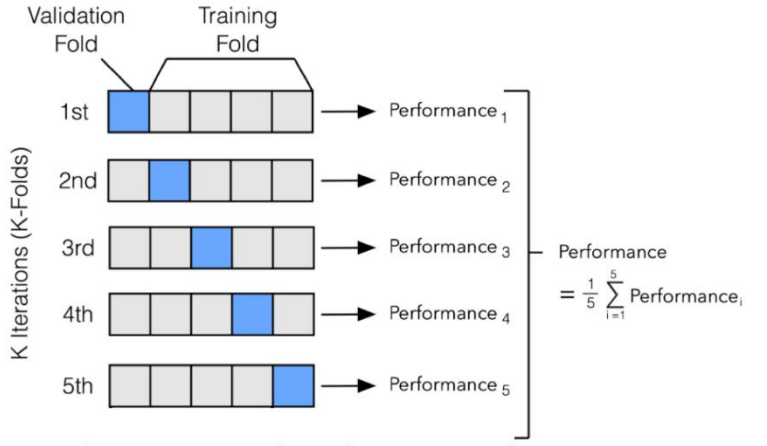
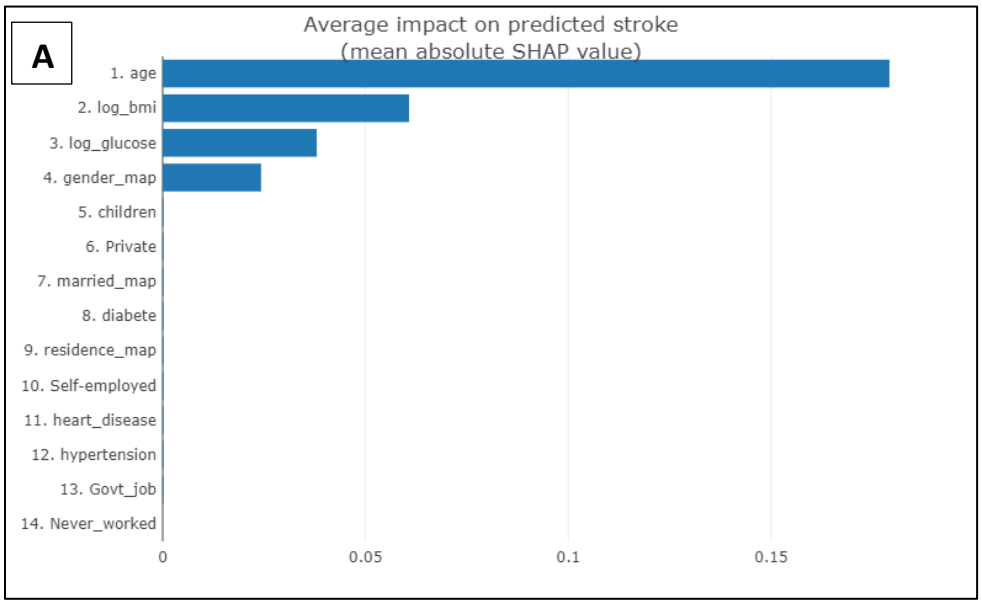
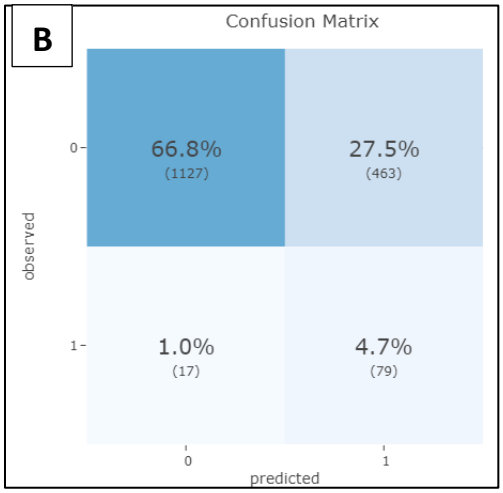
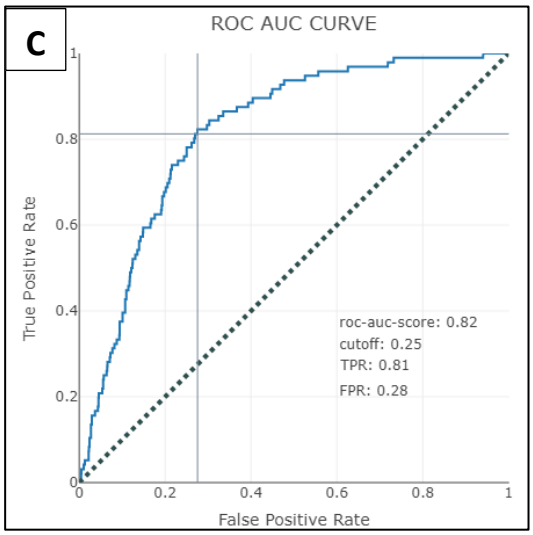


Fig.50 : *Schéma de la validation croisée.*

Dans les résultats de la RFE on remarque lorsque le nombre d'attribut augmente il y a une diminution du recall de la classe, ce qui veut dire que la détection d'AVC diminue. On remarque également que le score de la validation croisée augmente avec le nombre d'attribut. En comparant les différents scores et recall, le nombre d'attribut le plus pertinent serait au nombre de 4, avec un score de validation croisée de 0.91 et un recall de 0.60. Dans le but de visualiser la différence entre la classification avec tous les attributs et les 4 attributs sélectionnés, un dashboard a été créé et a montré les représentations suivantes (figure 51-54).





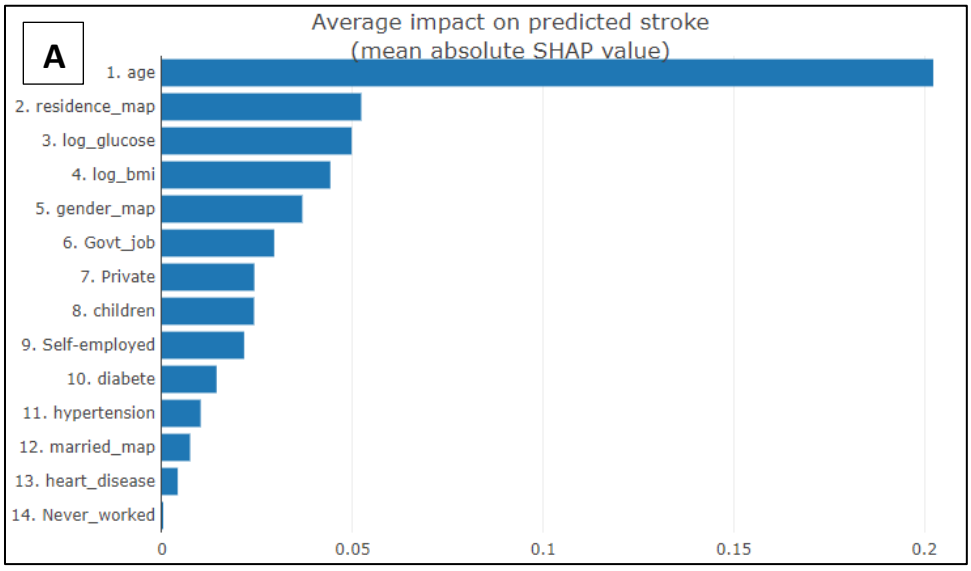


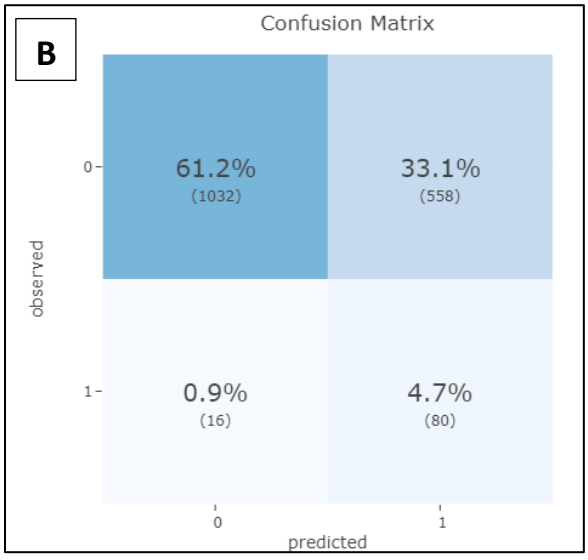
D

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

Fig.51 : *Représentation du dashboard avec la sélection des 4 attributs les plus pertinents.*

On observe que l'interprétation des variables les plus impactantes avec SHAP montre uniquement les 4 attributs du modèle, qui sont l'âge, l'indice de masse corporelle, le taux de glucose dans le sang et le sexe (figure 51A). La matrice de confusion ici possède un seuil de probabilité a 0.25, ce qui augmente le nombre de faux positif (non AVC qui sont prédit comme AVC (27.5%)), notre modèle va donc favoriser de prédire plus la présence d'AVC au lieu de prédire l'inverse induisant l'apparition de faux négatif. En effet, l'objectif du projet est d'effectuer de la prévention d'AVC, donc s'il a des personnes qui sont prédites à avoir un AVC au lieu de ne pas en avoir (faux positif) c'est moins gênant que les faux négatifs car certains critères peuvent faire qu'il y a tout de même un petit risque d'AVC, ainsi le mieux c'est d'effectuer une prévention pour ces personnes-là également. La courbe ROC montre un score de 0.82, ce qui est plus élevé par rapport au modèle précédent.





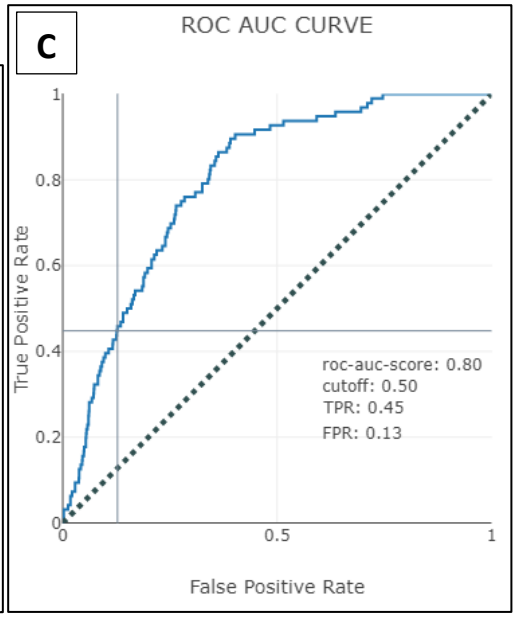
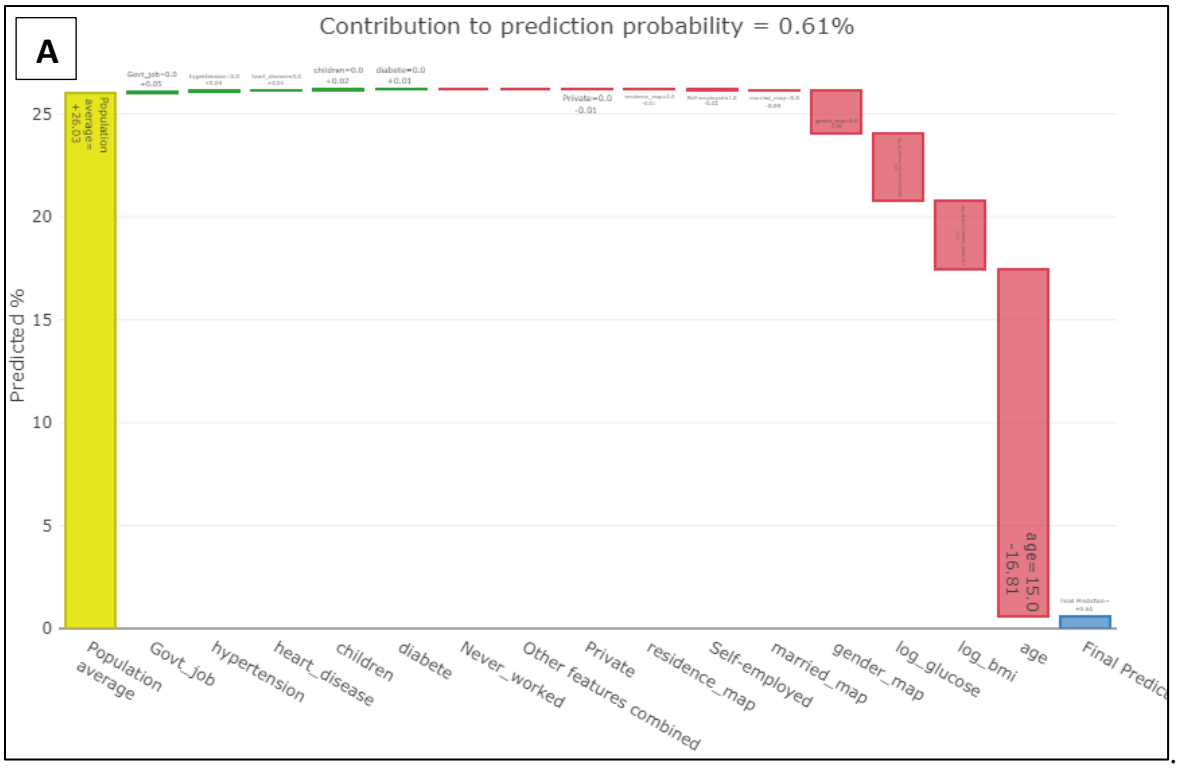


Fig.52 : *Représentation du dashboard avec la sélection de tous les attributs.*

On observe que l'interprétation des variables avec SHAP, des variables les plus impactantes montre l'âge, la résidence, l'indice de masse corporelle et le taux de glucose dans le sang (figure 52A). La matrice de confusion ici possède un seuil de probabilité a 0.25, ce qui augmente le nombre de la prédiction de faux positifs donc les non AVC qui sont prédits comme AVC (33.1%), notre modèle va donc favoriser de prédire plus la présence d'AVC au lieu de prédire l'inverse induisant l'apparition de faux négatif. La courbe ROC montre un score de 0.80, ce qui est moins élevé par rapport au modèle avec 4 attributs. Par la suite la contribution des variables a été observée pour deux individus, un AVC et un non AVC.



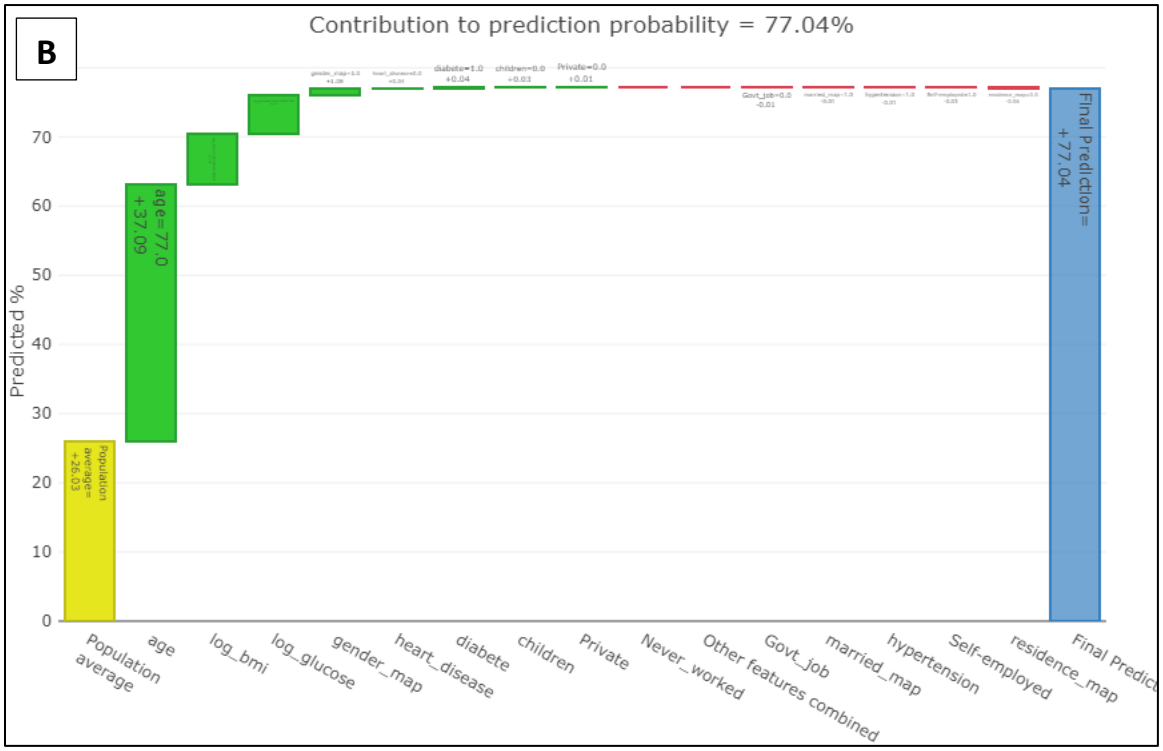


Fig.53 : *Représentation des variables pertinentes afin de prédire une personne non atteinte par un AVC (A) et atteintes (B) avec la classification avec 4 attributs.*

Pour une personne qui est prédite en non AVC (figure 53A), les variables qui jouent une importance sont l'âge (15 ans), l'indice de masse corporelle et le glucose dans le sang qui sont faibles et le sexe (homme). Pour une personne prédite en AVC (figure 53B), les variables qui

ont influencé cette prédiction sont l'âge (77 ans), l'indice de masse corporelle et le glucose dans le sang qui sont élevées et le sexe (femme).

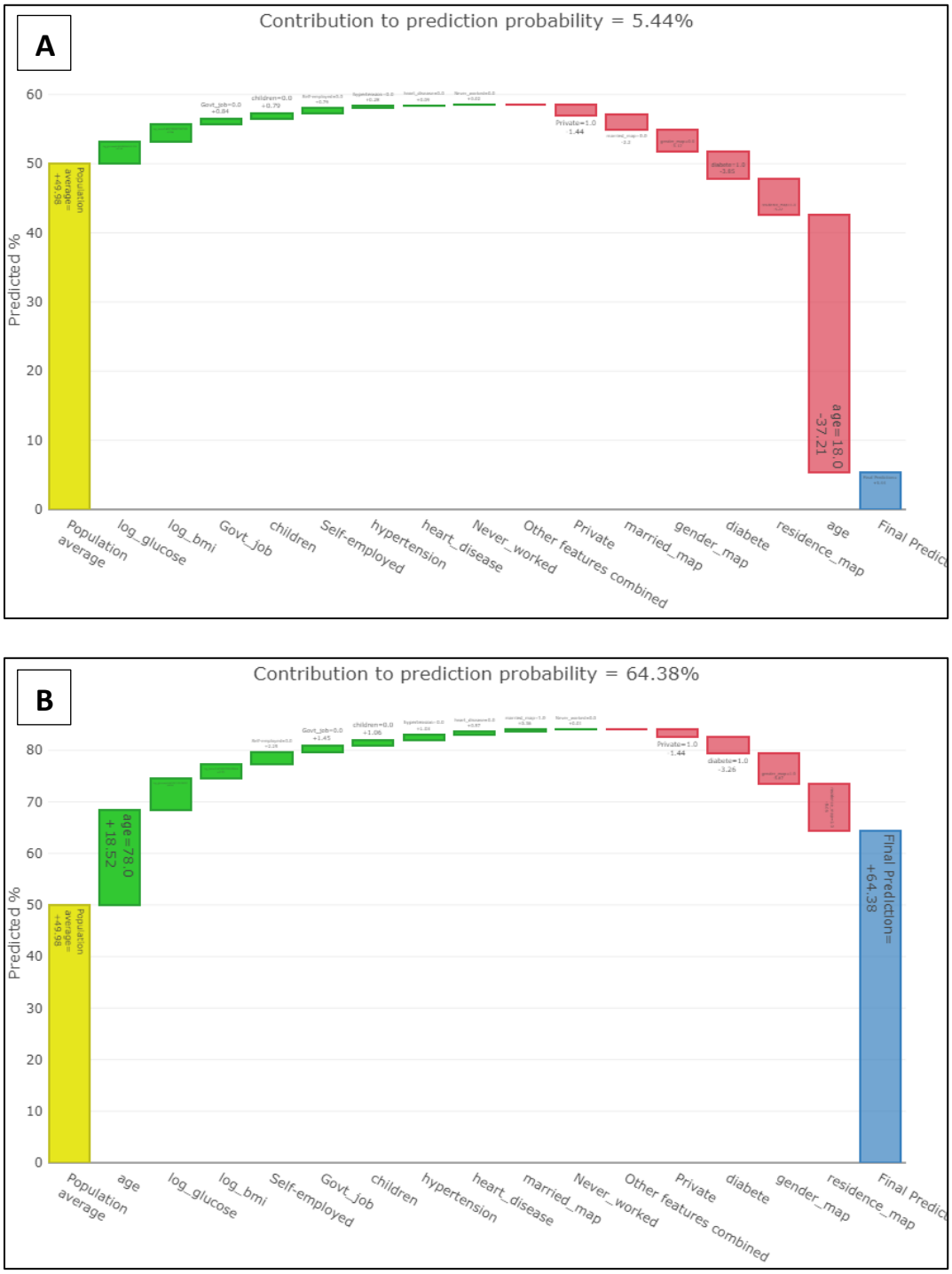


Fig.54 : *Représentation des variables pertinentes afin de prédire une personne non atteinte par un AVC (A) et atteintes (B) avec tous les attributs.*

Pour une personne qui est prédite en non AVC (figure 54A), les variables qui jouent une importance sont l'âge (18 ans), la résidence en zone rurale, le diabète et le sexe (homme). Pour

une personne prédite en AVC (figure 43B), les variables qui ont influencé cette prédiction sont l'âge (78 ans), l'indice de masse corporelle et le glucose dans le sang qui sont élevées et le sexe (femme).

VI. Conclusion

Pour conclure sur cette étude on a remarqué que les AVC sont de plus en plus fréquents dans le monde et touche environ 16 millions de personnes par an. L'AVC étant très dangereux pour les Hommes, le but de mon projet a été de détecter ces personnes qui sont susceptibles d'être atteintes par un AVC et d'effectuer une prévention. En effet ces personnes seraient prises en consultation médicale dans le but de réduire au maximum tous les risques d'apparition d'AVC.

Les différents modèles de classifications dans ce projet ont permis d'obtenir des valeurs d'accuracy et recall qui sont proches entre elles. Afin d'améliorer les prédictions des AVC une technique de RFE a été utilisée dans le but de sélectionner les variables les plus pertinentes. Cette technique a permis de définir que ces 4 variables sont l'âge, l'indice de masse corporelle, le taux de glucose dans le sang et le sexe. Cette dernière classification a permis d'obtenir un ROC score de 0.82 ce qui est un score assez convenable mais le réel but de projet c'est de détecter toutes les personnes à risque afin d'effectuer une prévention donc le score devrait être encore plus proche de 1. Ceci est du au fait qu'il n'y a certaines variables qui sont des facteurs à risque pour les AVC telles que l'alimentation, la consommation d'alcool, le cholestérol ou encore l'effet héréditaire, ne sont pas pris en compte dans ce modèle. Si d'autres variables avaient été à notre disposition la prédiction d'AVC serait encore plus précise et auraient donc améliorer le modèle de classification.