Projet SAS

Classification d'étoiles

Dany FADEL

Université Paris-Dauphine

Overview

1. Présentation des données
2. Analyse univariée
3. Analyse Bivariée
4. Modélisation
5. Conclusion

Section 1

Présentation des données

Presentation des données

- Jeu de données disponible à l'adresse suivante: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset
- Le jeu de données contient 5110 observations de 12 variables, dont l'une est un id unique
- Mon but est de prédire si un patient risque d'avoir une crise cardiaque ou non, ce qui correspond à la variable binaire stroke dans le jeu de données
- On a pour cela accès à 10 variables descriptives, dont 7 sont catégorielles et 3 sont numériques.
- Les variables catégorielles sont: gender, heart_disease, hypertension, ever_married, work_type, Residence_type et smoking_status.
- Les variables quantitatives sont: age, avg_glucose_level et bmi,

Tableau des données

Obs.	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
1	9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
2	51676	Female	61	0	0	Yes	Self-employed	Rural	202.21	N/A	never smoked	1
3	31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
4	60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
5	1665	Female	79	1	0	Yes	Self-employed	Rural	174.12	24	never smoked	1
6	56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly smoked	1
7	53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
8	10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never smoked	1
9	27419	Female	59	0	0	Yes	Private	Rural	76.15	N/A	Unknown	1
10	60491	Female	78	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1

Section 2

Analyse univariée

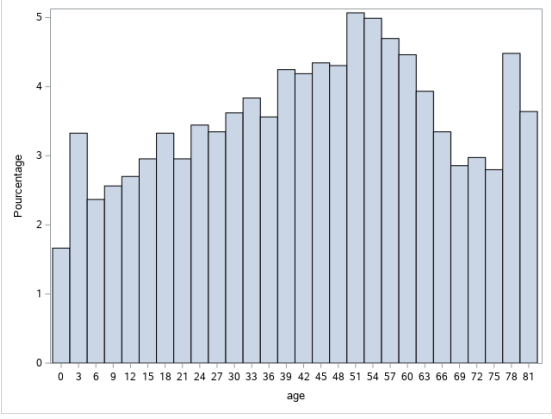
Variables Quantitatives

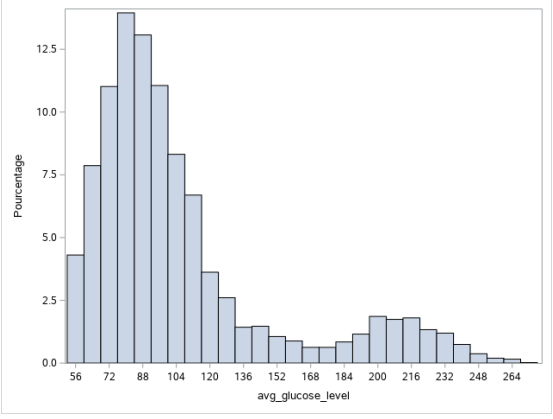
Après avoir remplacé les valeurs manquantes de bmi par la valeur médiane, on a:

La procédure MEANS

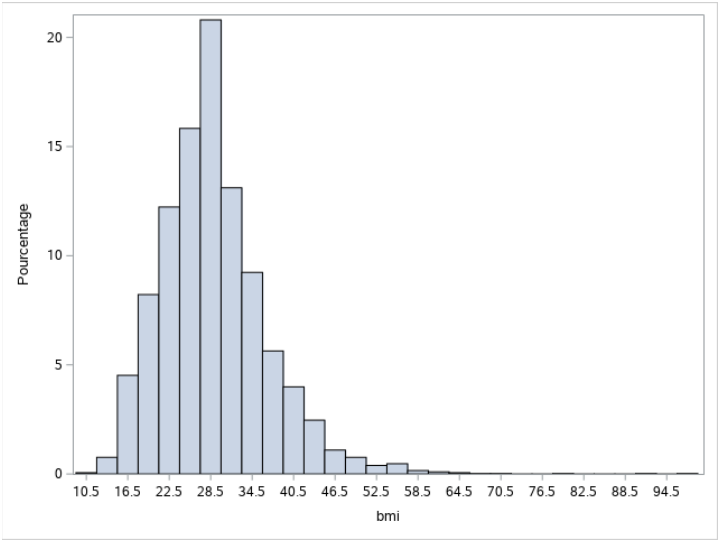
Variable	N	Nbre manquant	Minimum	Quartile inférieur	Moyenne	Quartile supérieur	Maximum
age	5110	0	0.0800000	25.0000000	43.2266145	61.0000000	82.0000000
avg_glucose_level	5110	0	55.1200000	77.2400000	106.1476771	114.0900000	271.7400000
bmi	5110	0	10.3000000	23.8000000	28.8620352	32.8000000	97.6000000

Variables Quantitatives





Variables Quantitatives



Variables Qualitatives

La procédure FREQ

gender	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
Female	2994	58.59	2994	58.59
Male	2115	41.39	5109	99.98
Other	1	0.02	5110	100.00

On supprime la ligne qui contient la valeur 'Other':

La procédure FREQ

gender	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
Female	2994	58.59	2994	58.59
Male	2115	41.39	5109	99.98
Other	1	0.02	5110	100.00

Variables Qualitatives:tableaux de fréquences

La procédure FREQ

hypertension	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	4611	90.25	4611	90.25
1	498	9.75	5109	100.00

La procédure FREQ

heart_disease	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	4833	94.60	4833	94.60
1	276	5.40	5109	100.00

La procédure FREQ

ever_married	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
No	1756	34.37	1756	34.37
Yes	3353	65.63	5109	100.00

La procédure FREQ

Residence_type	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
Rural	2513	49.19	2513	49.19
Urban	2596	50.81	5109	100.00

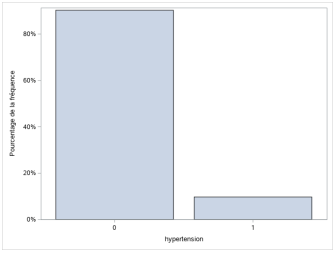
La procédure FREQ

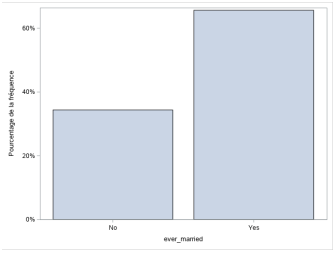
work_type	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
Govt_job	657	12.86	657	12.86
Never_worked	22	0.43	679	13.29
Private	2924	57.23	3603	70.52
Self-employed	819	16.03	4422	86.55
children	687	13.45	5109	100.00

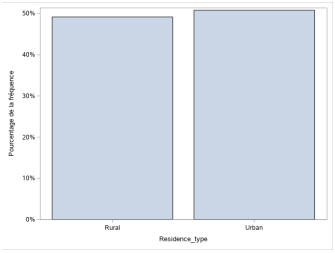
La procédure FREQ

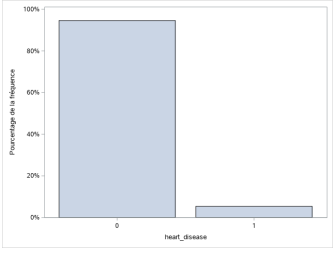
smoking_status	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
Unknown	1544	30.22	1544	30.22
formerly smoked	884	17.30	2428	47.52
never smoked	1892	37.03	4320	84.56
smokes	789	15.44	5109	100.00

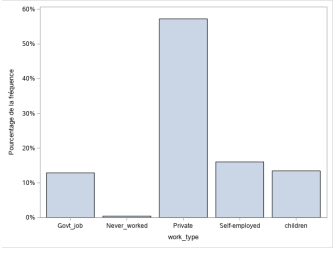
Variables Qualitatives: barplots

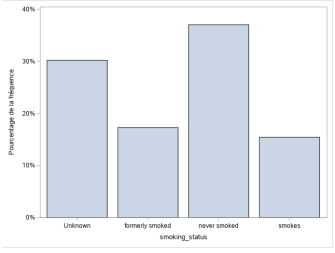










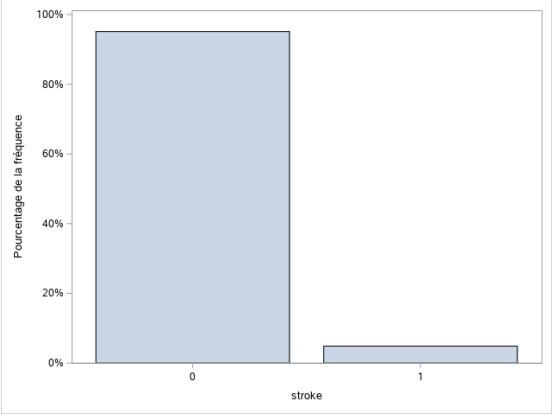


Variable cible

Stroke prend la valeur 1 si l'individu a eu une crise cardiaque, 0 sinon. On remarque que l'événement qui nous intéresse (valeur égale à 1) est rare dans notre jeu de données, ce qui pourrait poser problème lors de notre analyse.

La procédure FREQ

stroke	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	4860	95.13	4860	95.13
1	249	4.87	5109	100.00



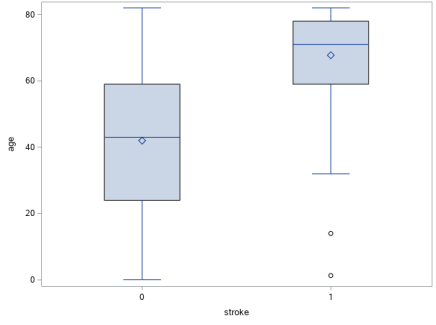
Section 3

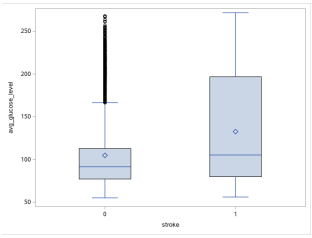
Analyse Bivariée

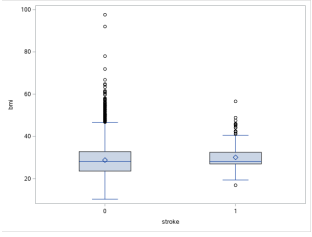
Croisement avec les variables quantitatives

La procédure MEANS

stroke	N obs	Variable	Minimum	Quartile inférieur	Moyenne	Médiane	Quartile supérieur	Maximum
0	4860	age	0.0800000	24.0000000	41.9748313	43.0000000	59.0000000	82.0000000
		avg_glucose_level	55.1200000	77.1200000	104.7875844	91.4650000	112.8100000	267.7600000
		bmi	10.3000000	23.6000000	28.8004321	28.1000000	32.8000000	97.6000000
1	249	age	1.3200000	59.0000000	67.7281928	71.0000000	78.0000000	82.0000000
		avg_glucose_level	56.1100000	79.7900000	132.5447390	105.2200000	196.7100000	271.7400000
		bmi	16.9000000	27.0000000	30.0903614	28.1000000	32.5000000	56.6000000







Sans surprise, les personnes ayant eu une crise cardiaque sont en moyennes plus vieux, et ont un niveau de glucose et un indice de masse corporelle plus élevés en moyenne.

Croisement avec les variables qualitatives

La procédure FREQ

Fréquence
Pct de col.

Table de stroke par gender

stroke	gender		
	Female	Male	Total
0	2853 95.29	2007 94.89	4860
1	141 4.71	108 5.11	249
Total	2994	2115	5109

La procédure FREQ

Fréquence
Pct de col.

Table de stroke par hypertension

stroke	hypertension		
	0	1	Total
0	4428 96.03	432 86.75	4860
1	183 3.97	66 13.25	249
Total	4611	498	5109

La procédure FREQ

Fréquence
Pct de col.

Table de stroke par heart_disease

stroke	heart_disease		
	0	1	Total
0	4631 95.82	229 82.97	4860
1	202 4.18	47 17.03	249
Total	4833	276	5109

La procédure FREQ

Fréquence
Pct de col.

Table de stroke par ever_married

stroke	ever_married		
	No	Yes	Total
0	1727 98.35	3133 93.44	4860
1	29 1.65	220 6.56	249
Total	1756	3353	5109

Croisement avec les variables qualitatives

La procédure FREQ

Fréquence
Pct de col.

Table de stroke par Residence_type

stroke	Residence_type		
	Rural	Urban	Total
0	2399 95.46	2461 94.80	4860
1	114 4.54	135 5.20	249
Total	2513	2596	5109

La procédure FREQ

Fréquence
Pct de col.

Table de stroke par smoking_status

stroke	smoking_status				
	Unknown	formerly smoked	never smoked	smokes	Total
0	1497 96.96	814 92.08	1802 95.24	747 94.68	4860
1	47 3.04	70 7.92	90 4.76	42 5.32	249
Total	1544	884	1892	789	5109

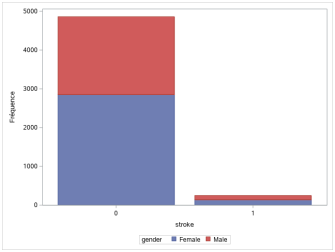
La procédure FREQ

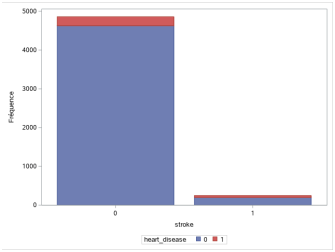
Fréquence
Pct de col.

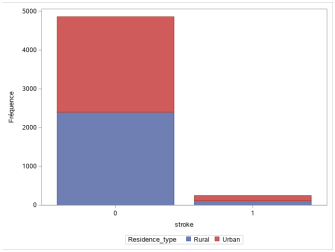
Table de stroke par work_type

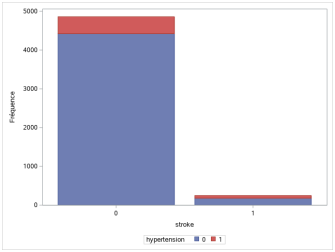
stroke	work_type					
	Govt_job	Never_worked	Private	Self-employed	children	Total
0	624 94.98	22 100.00	2775 94.90	754 92.06	685 99.71	4860
1	33 5.02	0 0.00	149 5.10	65 7.94	2 0.29	249
Total	657	22	2924	819	687	5109

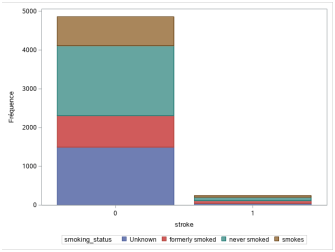
Croisement avec les variables qualitatives: graphes

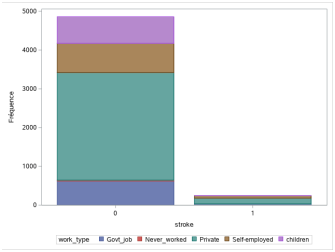




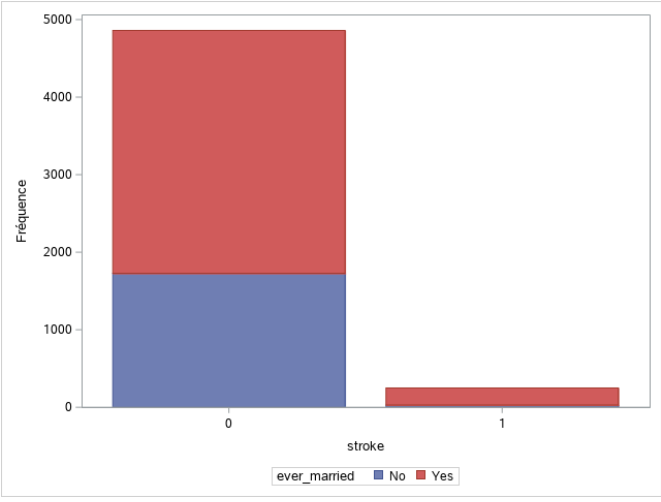








Évitez de vous marier, risque de crise cardiaque !



Section 4

Modélisation

Classification binaire

On s'intéresse à la prédiction des attaques cardiaques, donc à la prédiction de l'appartenance à la classe 1 de la variable stroke.

Le jeu de données est séparé en deux jeux train et test. La classe 1 de stroke étant sous-représentée, il est nécessaire d'effectuer la séparation sur chaque classe afin de s'assurer d'avoir la classe 1 dans les deux jeux de données. Cela nous permet également d egarder les même proportions entre les classes que le tableau d'origine.

Pour effectuer notre classification, on va utliser et comparer deux méthodes qui sont la régression logistique et les arbres de décision.

Régression Logistique

La première méthode que l'on va appliquer es tla régression logistique. On va pour cela utiliser la procédure *logistic*. On obtient les résultats suivants:

Régression logistique

La procédure LOGISTIC

Statistiques d'ajustement pour les données SCORE

Table	Fréquence totale	Log-vraisemblance	Taux d'erreur	AIC	AICC	BIC	SC	R-carré	R carré remis à l'échelle max.	AUC	Score de Brier
WORK.TEST	1533	-237.4	0.0489	482.8908	482.917	504.2308	504.2308	0.077704	0.240275	0.849447	0.042017

Matrice de confusion régression logsitique

La procédure FREQ

Fréquence
Pct de ligne

Table de F_stroke par I_stroke		
	I_stroke(Dans : stroke)	
F_stroke(De : stroke)	0	Total
0	1458 100.00	1458
1	75 100.00	75
Total	1533	1533

La régression prédit tout dans la classe 0. Ce modèle ne prend donc pas en compte la classe 1 qui est sous-représentée.

Arbre de décision

Essayons maintenant un arbre de décision. On utilise l'algorithme C4.5 avec l'indice de gini. Sur SAS, on va utiliser la procédure *hpsplit*. La matrice de confusion obtenue en évaluant le modèle sur l'échantillon test est:

Matrice de confusion arbre de décision

La procédure FREQ

Fréquence Pct de ligne	Table de Actual par Predicted	
	Predicted	
Actual	0	Total
0	1458 100.00	1458
1	75 100.00	75
Total	1533	1533

Encore une fois, seule la classe 0 est prédite. Il nous faut donc un stratégie pour prendre en compte la rareté de la classe 1. L'idée va être de rééquilibrer les classes.

Stratégies de sur-échantillonnage et de sous-échantillonnage

Une stratégie possible pour rééquilibrer les proportions des deux classes est d'effectuer un tirage avec remise des individus de la classe 1, que l'on ajoute ensuite à notre jeu de données. Cela permet de rééquilibrer les proportions artificiellement. C'est ce que l'on appelle le sur-échantillonnage (oversampling en anglais).

Une seconde stratégie est à l'inverse d'effectuer un tirage sans remise des individus de la classe 0, afin d'en réduire le nombre. L'inconvénient de cette méthode est qu'il peut y avoir une perte d'information. C'est le sous-échantillonnage, ou *undersampling* en anglais

Pour éviter des problèmes de sur-apprentissage (notamment pour le sur-échantillonnage) et pour pouvoir comparer les deux stratégies, on va garder l'échantillon test utilisé séparément et effectuer un rééchantillonnage sur le tableau train.

Sur-échantillonnage: régression logistique

Cette fois, le modèle prédit bien dans les deux classes, on a alors la matrice de confusion suivante:

Matrice de confusion régression logistique suréchantillonage

La procédure FREQ

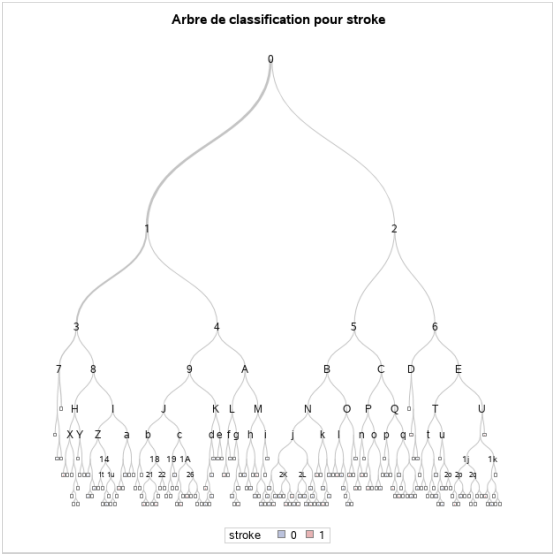
Fréquence
Pct de ligne

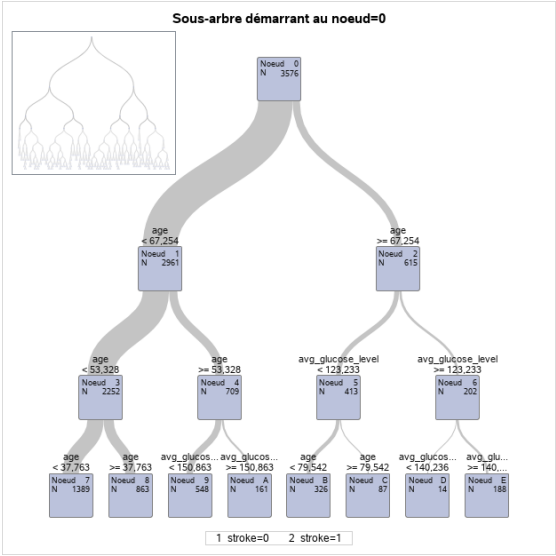
Table de F_stroke par I_stroke

F_stroke(De : stroke)	I_stroke(Dans : stroke) 0	1	Total
0	1181 81.00	277 19.00	1458
1	21 28.00	54 72.00	75
Total	1202	331	1533

On lit une sensibilité (classe positive = 1) de 72% et une spécificité de 81%. On a cependant un recall de 16%.

Sur-échantillonnage: arbre de décision





Sur-échantillonnage: arbre de décision

On a ici seulement 13% de vrais positifs.

Matrice de confusion arbre de décision suréchantillonnage

La procédure FREQ

Fréquence Pct de ligne

Table de Actual par Predicted

Actual	Predicted 0	Predicted 1	Total
0	1404 96.30	54 3.70	1458
1	65 86.67	10 13.33	75
Total	1469	64	1533

Sur-échantillonnage: forêt aléatoire

On utilise *hpforest* et *hp4score* pour entraîner une forêt aléatoire et l'évaluer sur léchantillon test. On passe alors à un taux de 74% de vrais positifs, mais une précision (recall) sur les positifs qui vaut 16%:

Matrice de Confusion forêt suréchantillonnage

La procédure FREQ

Fréquence Pct de ligne

Table de Actual par Predicted

Actual	Predicted 0	Predicted 1	Total
0	1153 79.08	305 20.92	1458
1	19 25.33	56 74.67	75
Total	1172	361	1533

Sous-échantillonnage: régression logistique

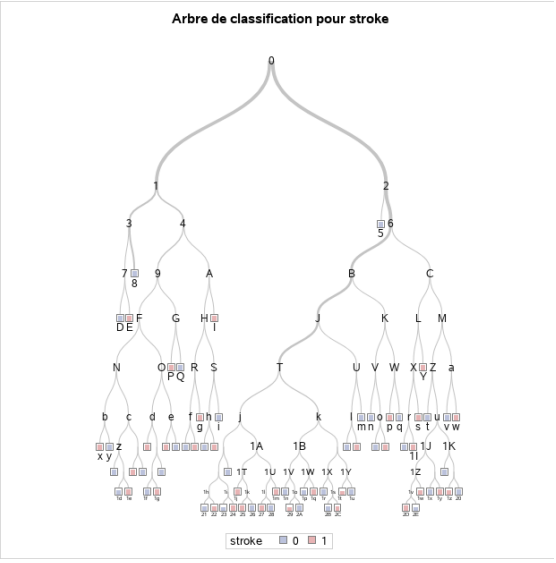
Matrice de confusion régression logistique sous-échantillonage

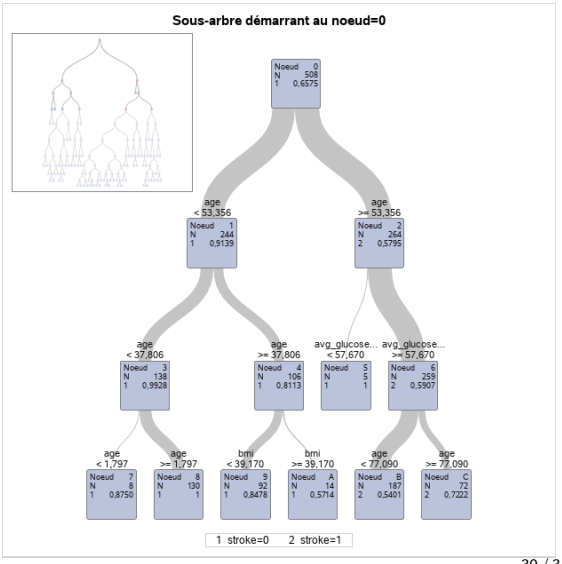
La procédure FREQ

Fréquence Pct de ligne	Table de F_stroke par I_stroke		
	I_stroke(Dans : stroke)		
F_stroke(De : stroke)	0	1	Total
0	1256 86.15	202 13.85	1458
1	26 34.67	49 65.33	75
Total	1282	251	1533

On lit une sensibilite de 65% et un recall de 20%.

Sous-échantillonnage: arbre de décision





Sous-échantillonnage: arbre de décision

On a ici 54% de vrais positifs mais 0.11 de prévision positive juste (recall).

Matrice de confusion arbre de décision sous-échantillonnage

La procédure FREQ

Fréquence Pct de ligne	Table de Actual par Predicted		
	Predicted		
Actual	0	1	Total
0	1142 78.33	316 21.67	1458
1	34 45.33	41 54.67	75
Total	1176	357	1533

Sous-échantillonnage: forêt aléatoire

On utilise *hpforest* et *hp4score* pour entraîner une forêt aléatoire et l'évaluer sur léchantillon test. On passe alors à un taux de 52% de vrais positifs, mais une précision (recall) sur les positifs qui vaut 12%:

Matrice de Confusion forêt sous-échantillonnage

La procédure FREQ

Fréquence Pct de ligne

Table de Actual par Predicted

Actual	Predicted 0	Predicted 1	Total
0	1179 80.86	279 19.14	1458
1	36 48.00	39 52.00	75
Total	1215	318	1533

Section 5

Conclusion

Conclusion

Les stratégies de sur-échantillonnage et de sous-échantillonnage nous permettent d'avoir des prédictions pour la classe minoritaire. La régression logistique semble donner un recall légèrement meilleur que les arbres de décision ou les forêts aléatoires, tout en ayant des résultats similaires sur d'autres métriques comme la sensibilité par exemple.

Cependant un recall de moins de 20% signifie que l'on a un faux positif dans 80% des crises cardiaques prédites. Ce n'est pas l'idéal...

Il existe d'autres méthodes qui permettent de prendre en compte des événements rares, qu'on ne traite pas ici. Dans les méthodes des rééchantillonnage il y a SMOTE, on aurait aussi pu essayer de pénaliser la régression logistique...