IMT Atlantique

UE Data Santé
Technopôle de Brest-Iroise - CS 83818
29238 Brest Cedex 3
www.imt-atlantique.fr

Rapport

Data Santé - Accident Vasculaire Cérébral

Benjamin DEMOLIN
FIP 3A - TAF HEALTH

Thomas PERRIN
FISE 2A - TAF HEALTH

Inès THAO
FISE 2A - TAF HEALTH

Date d'édition : 24 mars 2022

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Sommaire

Liste des figures

Liste des tableaux

1. Introduction

Un accident vasculaire cérébral (AVC) est un déficit neurologique soudain d'origine vasculaire causé par un infarctus ou une hémorragie au niveau du cerveau. Le terme « accident » souligne l'aspect soudain ou brutal des symptômes, mais dans la plupart des cas les causes de cet accident sont internes (liées à l'âge, l'alimentation ou l'hygiène de vie, notamment).

Les symptômes varient beaucoup d'un cas à l'autre selon la nature de l'AVC (ischémique ou hémorragique), l'endroit et la taille de la lésion cérébrale : aucun signe remarquable, perte de la motricité, perte de la sensibilité, trouble du langage, perte de la vue, perte de connaissance, décès.

D'après l'Organisation Mondiale de la Santé (OMS), l'AVC est la première cause de handicap physique de l'adulte et la deuxième cause de décès dans la plupart des pays occidentaux (Europe, États-Unis, etc.). Dans ces pays, un individu sur 200 est atteint d'un accident vasculaire cérébral chaque année. Par exemple, en France en 2019, on dénombre chaque année plus de 140 000 nouveaux cas d'accidents vasculaires cérébraux, soit un toutes les quatre minutes selon l'INSERM (Institut national de la santé et de la recherche médicale).

Nous avons cherché, à travers ce projet, à connaître quelles variables entraient en jeu et étaient corrélées aux cas d'accidents vasculaires cérébraux ; quelles étaient les variables les plus "importantes" à surveiller et correspondaient à des facteurs à risques ; et quelle était la meilleure méthode de prédiction et donc présentait le moins d'erreurs dans la prédiction de cas d'AVC.

Pour cela, nous avons utilisé la base de données en accès libre "Stroke Prediction Dataset" disponible sur Kaggle : https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset.

Nous noterons par la suite :

- "stroke" pour AVC ;
- "avcP" pour les cas positifs ;
- "avcN" pour les cas négatifs.

2. Description du jeu de données

Notre jeu de données est composé de dix variables et une variable d'intérêt : "stroke" (en bleu). Cf. Figure 1.

Nom	Type	Niveau
stroke	Factor	avcN / avcP
gender	Factor	Female / Male
age	Numérical	-
hypertension	Factor	htN / htP
heart_disease	Factor	hdN / hdP
ever_married	Factor	No / Yes
work_type	Factor	never_worked / govt_job / private / self_employed
Residence_type	Factor	Rural / Urban
avg_glucose_level	Numérical	-
bmi	Numérical	-
smoking_status	Factor	formerly / never / smokes

FIGURE 1 – Jeu de données

Les données sont segmentables en deux catégories : données quantitatives (Numerical) et données qualitatives (Factor).

NB :

- Une seule personne n'avait pas indiqué de ***"gender"***, nous l'avons retirée du dataset ;

- Nous avons regroupé les *enfants* dans *"never_worked"* de ***"work_type"*** ;
- ***"avg_glucose_level"*** signifie taux moyen de glycémie ;
- ***"bmi"*** signifie IMC : Indice de Masse Corporelle.

3. Variables quantitatives

Étudions premièrement les variables de manière globale pour avoir une idée d'ensemble des variables quantitatives.

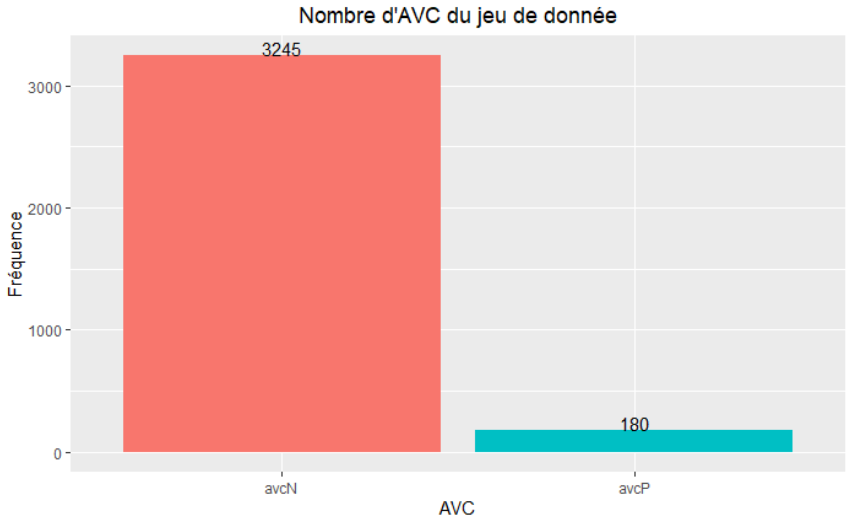


FIGURE 2 – Répartition des cas d'AVC du jeu de données

On remarque que le nombre de cas négatifs est bien plus important que le nombre de cas positifs. Ces données seront rééquilibrées avant les apprentissages et expliquées dans la partie 6.

3.1. Vision globale des variables quantitatives

3.1.1. AVC et Tranches d'âge

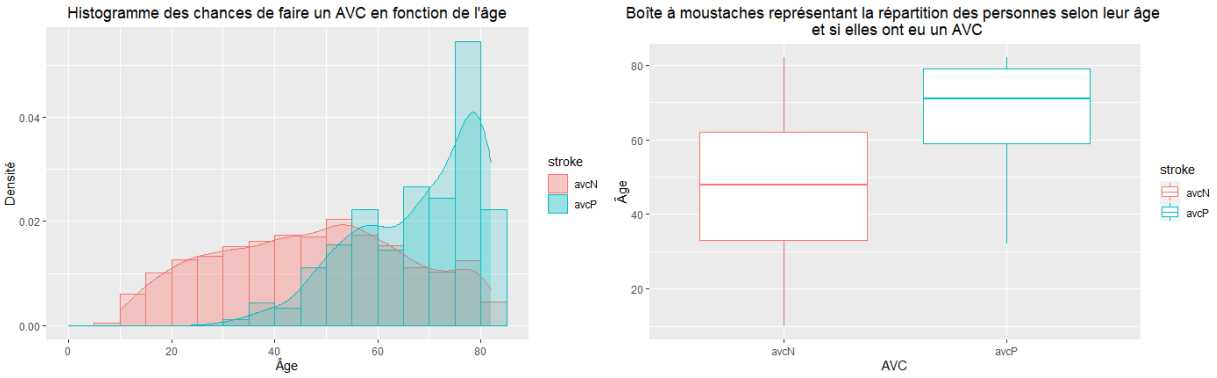


FIGURE 3 – Répartition des cas d'AVC en fonction de l'âge

Nous pouvons voir d'après la figure 3, que les cas d'AVC sont segmentables selon des tranches d'âge, les cas d'AVC se situant plutôt autour de 70 ans. Ainsi, nous avons décidé de segmenter de manière arbitraire selon la densité et la répartition des cas.

Nous obtenons ainsi 4 tranches d'âge :

- **Moins de 30 ans :** 0_30 ;
- **Entre 31 et 50 ans :** 31_50 ;
- **Entre 51 et 65 ans :** 51_65 ;
- **66 ans et plus :** 66_*plus* ;

Nous obtenons donc la répartition suivante :

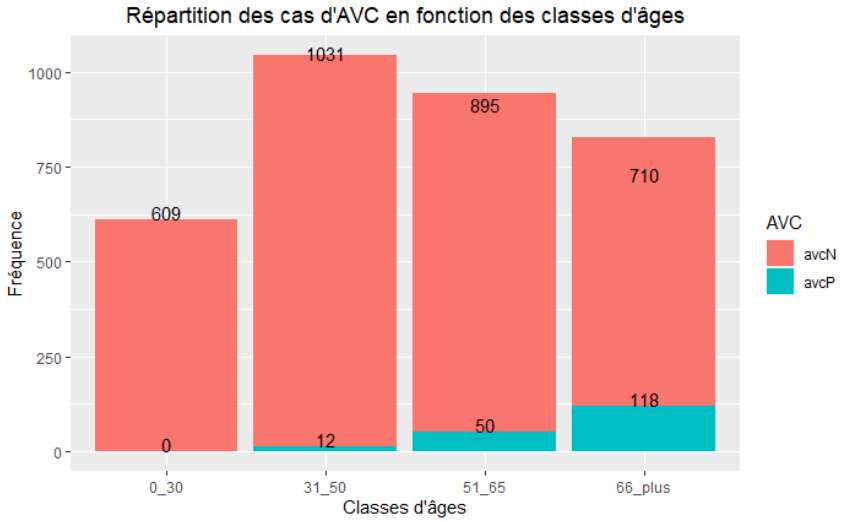


FIGURE 4 – Répartition des cas d'AVC en fonction des classes d'âges

3.1.2. AVC et Classes de niveaux de glucose

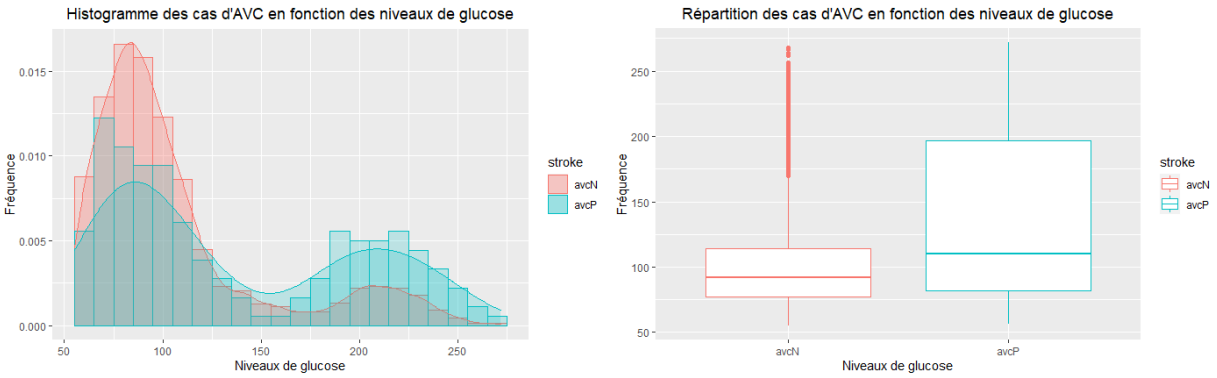


FIGURE 5 – Répartition des cas d'AVC selon leur niveau de glucose

Nous pouvons voir d'après la figure 5, que les cas d'AVC sont segmentables selon des classes de niveaux de glucose. Nous avons décidé de segmenter selon les niveaux de glucoses estimés par les médecins. Pour cela, nous nous sommes appuyés sur les données du site de la Fédération des Diabétiques : https://www.federationdesdiabetiques.org/information/glycemie

Nous obtenons ainsi 4 classes de niveaux de glucose :

- **Moins de 70 g/L :** 0_69 ;
- **Entre 70 et 100 g/L :** 70_100 ;
- **Entre 101 et 126 g/L :** 101_126 ;
- **127 g/L et plus :** 127_*plus* ;

Nous obtenons donc la répartition suivante :

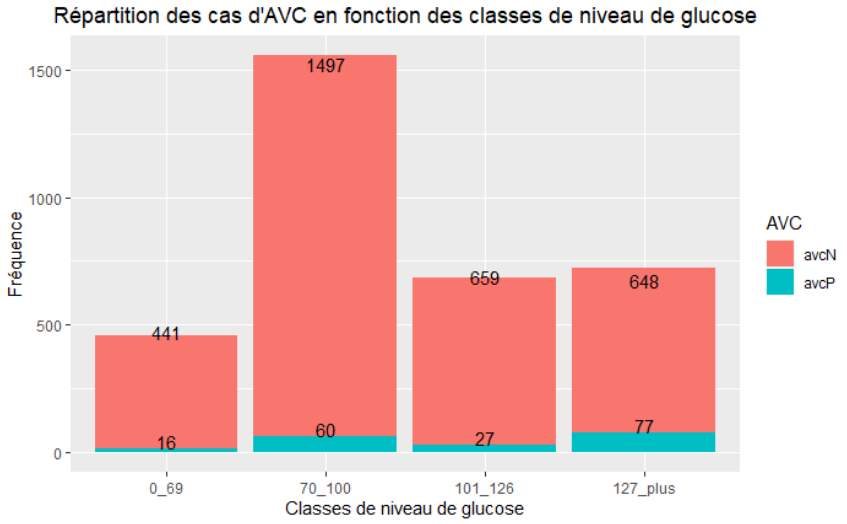


FIGURE 6 – Répartition des cas d'AVC en fonction des classes de niveaux de glucose

3.1.3. AVC et Classes d'IMC

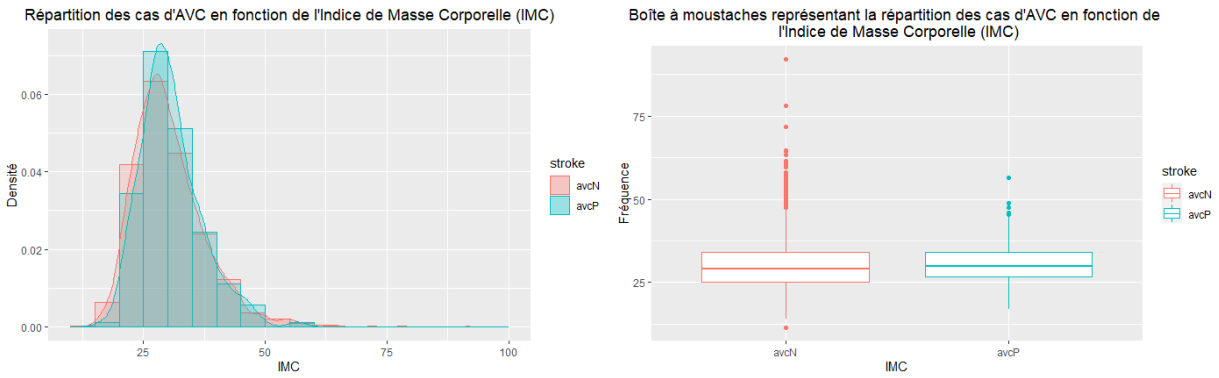


FIGURE 7 – Répartition des cas d'AVC en fonction de l'IMC

Nous pouvons voir d'après la figure 7, que les cas d'AVC sont segmentables selon des classes d'Indice de Masse Corporelle (IMC). Nous avons décidé de segmenter selon la table d'IMC donnée par la Haute Autorité de Santé (HAS) :
https://www.has-sante.fr/upload/docs/application/pdf/2009-09/table_imc_230909.pdf

Nous obtenons ainsi 4 classes d'IMC en kg/m^2 :

- **Moins de 25 :** 0_25 ;
- **Entre 25 et 30 :** 25_30 ;
- **Entre 31 et 35 :** 30_35 ;
- **35 et plus :** 35_plus ;

Nous obtenons donc la répartition suivante :

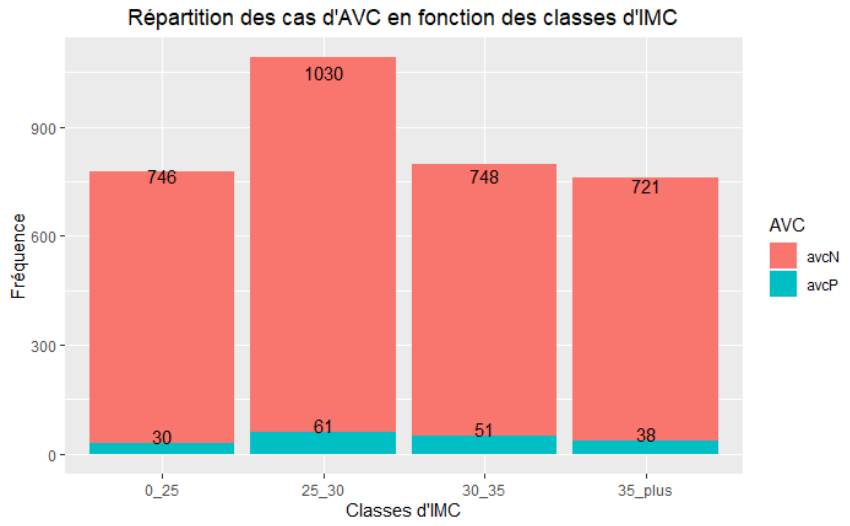


FIGURE 8 – Répartition des cas d'AVC en fonction des classes d'IMC

3.2. Tests statistiques des variables quantitatives

Nous voulons vérifier s'il y a une différence significative entre les cas avérés et non avérés dans les groupes d'âges, de niveaux de glucose et d'IMC. Pour cela, nos variables étant quantitatives, nous effectuons des tests de Student avec un intervalle de confiance de 95%. Nous regardons donc la valeur de $p - value$: si p-value est inférieure au seuil de 5%, on rejette l'hypothèse nulle en faveur de l'hypothèse alternative, et le résultat du test est déclaré « statistiquement significatif ».

Les résultats des tests de Student sont indiqués pour chacune des variables quantitatives dans le tableau Table 1 suivant :

	Âge	Niveau de glucose	IMC
p-value	$< 2.2e-16$	$< 2.2e-16$	0.4973
Significatif	Oui	Oui	Non

TABLE 1 – Résultats des tests de Student des variables quantitatives

On peut ainsi conclure de la significativité des variables "âge" et "niveau de glucose", mais pas de "IMC".

3.3. Corrélation entre les variables quantitatives

Pour effectuer une analyse épidémiologique, il nous faut évaluer la corrélation entre nos variables d'intérêt, notamment les variables quantitatives que sont l'âge, le niveau de glucose ainsi que l'IMC.

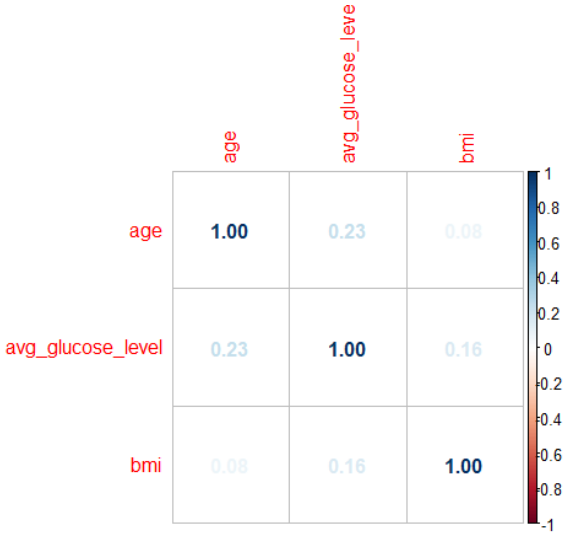


FIGURE 9 – Taux de corrélation entre les variables

On observe par ailleurs que les corrélations ne sont pas très marquées, la plus haute étant celle reliant l'âge et le niveau de glucose : 0.23. (Fig. 9).

4. Variables qualitatives

À présent, étudions les variables qualitatives. Nous avons pour cela également transformé les variables quantitatives précédentes en variables qualitatives en les discrétisant.

4.1. Vision globale des variables qualitatives

	Attribut	Pourcentage AVC	p-value
Genre	Femme	5,03	0,517
	Homme	**5,6**	
Hypertension	Sans	4,08	1,16e-16
	Avec	**13,97***	
Maladie Cardiaque	Sans	4,47	1,91e-15
	Avec	**17,48***	
Statut Matrimonial	Jamais marié	2,42	4,11e-5
	A déjà été marié(e)	**6,16***	
Type de travail	Gouvernemental	0	5,31e-3[1]
	Pas de travail	4,47	
	Secteur privé	4,95	
	Indépendant	**7,63**	
Type de résidence	Zone rurale	5,12	0,783
	Zone urbaine	**5,39**	
Statut fumeur	**Ancien fumeur**	**6,82***	0,049
	Non-fumeur	4,54	
	Fumeur	5,29	
Classe d'âge	0-30	0	6,87e-44
	31-50	1,15	
	51-65	5,29	
	66 et plus	**14,25***	
Classe de glucose	0-69	3,5	1,60E-11
	70-100	3,85	
	101-126	3,94	
	127 et plus	**10,62***	
Classe d'IMC	0-25	3,87	0,145
	25-30	5,59	
	30-35	**6,38**	
	35 et plus	5,01	

FIGURE 10 – Variables qualitatives et chances d'avoir un AVC

Légende :

- [1] *L'approximation du khi-carré peut être incorrecte.*
- ***Variable ayant le plus de chances d'être associées à la survenue d'un AVC.***
- ** Résultat significatif avec une erreur de 5%.*

4.2. Tests statistiques et corrélation des variables qualitatives

La significativité des résultats obtenus ainsi que la dépendance entre les différentes variables qualitatives de la figure 10 ont été déduites des résultats de tests du χ^2.

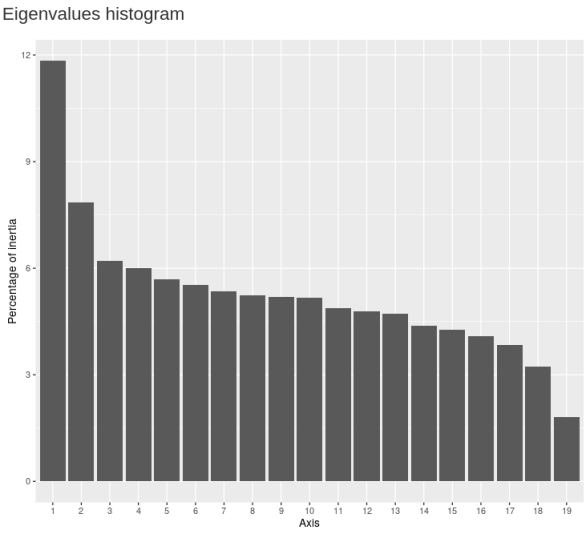
Pour cela, nous avons déterminé les matrices résiduelles afin d'examiner les différences entre les dénombrements attendus et observés pour déterminer les variables ayant l'impact le plus important ; ainsi que les p-value afin de déterminer si l'association entre les variables est statistiquement significative.

Nous avons arbitrairement choisi un seuil de signification égal à 0.05 tel que :

- Si $p \leq 0.05$: les variables présentent une association statistiquement significative ;
- Si $p > 0.05$: impossible de conclure que les variables sont associées.

5. Analyse des Correspondances Multiples (ACM)

On peut réaliser une analyse multivariée des variables pour identifier les effets combinés des variables explicatives sur la variable cible. On réalise une analyse factorielle multiple des correspondances (AFCM) en sélectionnant uniquement les variables qualitatives, dont les variables qualitatives obtenues en discrétisant les variables quantitatives (âge, glucose et IMC). La première ACM permet d'aboutir à des cartes de représentation sur lesquelles on peut visuellement observer les proximités entre les catégories des variables qualitatives et les observations.



(a) Histogramme des valeurs propres

Eigenvalues table

Axis	%	Cum. %
1	11.8	11.8
2	7.9	19.7
3	6.2	25.9
4	6.0	31.9
5	5.7	37.6
6	5.5	43.1
7	5.3	48.5
8	5.2	53.7
9	5.2	58.9
10	5.2	64.1
11	4.9	68.9
12	4.8	73.7
13	4.7	78.4
14	4.4	82.8
15	4.3	87.0
16	4.1	91.1
17	3.8	95.0
18	3.2	98.2
19	1.8	100.0

(b) Tableau des proportions des valeurs propres

FIGURE 11 – Valeurs propres de l'ACM

Sur le graphique des variables (Figure 12), les modalités partagées par les mêmes individus sont représentées par des points qui tendent à se regrouper ; la dissemblance produit, au contraire, de la distance. On remarque dans le cadran droit supérieur que les modalités d'âge supérieur à 66 ans, de maladie cardiaque et d'hypertension semblent fortement liées entre elle d'une part ainsi qu'aux deux grands principes de structuration des données. La modalité d'âge inférieur à 30 ans est également fortement liée aux deux premiers axes.

Parmi les autres résultats remarquables, on peut voir que le type de résidence ne semble pas avoir d'importance dans les risques de faire un AVC. Avoir un IMC inférieur à 35 s'oppose à un IMC supérieur à 35 selon le deuxième axe. Avoir un niveau de glucose inférieur à 127 s'oppose à avoir un niveau de glucose supérieur à 127.

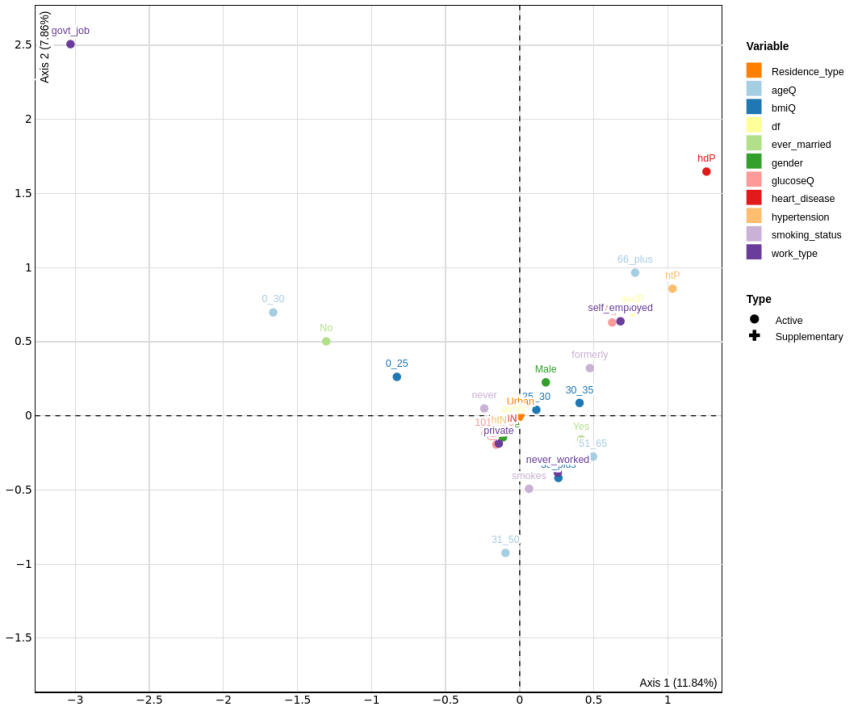


FIGURE 12 – Graphique des variables pour la première ACM

L'analyse exploratoire descriptive a montré que le risque d'avoir un AVC augmente avec l'âge, avec le fait d'avoir déjà été marié, d'avoir de l'hypertension, d'avoir une maladie cardiaque, et avec un niveau de glucose élevé. On cherche donc à étudier l'interaction entre ces variables.

On commence par étudier l'interaction entre l'âge et l'hypertension. Pour mieux mettre en évidence la liaison avec la pathologie, on construit par croisement une variable Age :Hypertension à 8 modalités.

Avec un test de dépendance avec la variable cible, on constate que les personnes les plus âgées (66 ans et plus) ont plus de chance de faire un AVC que les autres tranches d'âge et que, quelque soit la tranche d'âge, le fait d'avoir de l'hypertension augmente le risque de faire un AVC. On ajoute cette nouvelle variable interaction au dataframe et on lance une deuxième ACM.

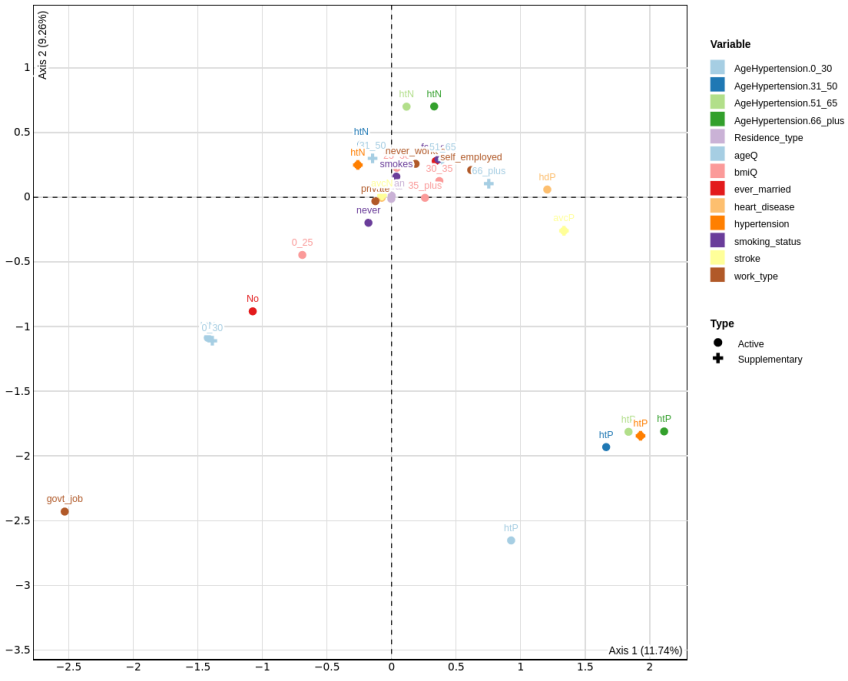


FIGURE 13 – Graphique des variables pour la deuxième ACM avec Age :Hypertension

On observe nettement deux groupes de modalité suivant la présence d'hypertension (en bas à droite) ou non (en haut au milieu), avec un risque plus élevé d'AVC si la personne a de l'hypertension. De plus, dans les deux groupes, un âge plus élevé décale la modalité selon l'axe 1 vers un plus fort risque d'AVC.

On choisit finalement de ne conserver que 4 modalités : les moins de 30 sans hypertension, les moins de 30 avec hypertension, les plus de 30 ans sans hypertension et les plus de 30 ans avec hypertension. On réalise une troisième ACM qui va nous permettre de conclure.

FIGURE 14 – Graphique des variables pour la troisième ACM avec Age :Hypertension

Finalement, une troisième ACM nous indique que les jeunes (moins de 30 ans) sans hypertension sont bien consistants avec un faible risque d'AVC. Mis à part pour la modalité a0_30htP, les cosinus carrés des modalités de la variable interaction sont élevés ce qui permet de s'assurer d'éviter des erreurs d'interprétation dues à des effets de projection.

Une deuxième étude par ACM nous permet d'observer l'interaction entre l'âge et l'hypertension. La modalité maladie cardiaque pour les plus de 30 ans, et dans une moindre mesure pour les moins de 30 ans, semble plus consistante avec le risque d'AVC que les autres modalités (pas de maladie cardiaque pour les moins de 30 ans et les plus de 30 ans). Cette première ACM nous permet de ne conserver que 3 modalités : les moins de 30 ans sans maladie cardiaque, les plus de 30 ans sans maladie cardiaque et ceux ayant une maladie cardiaque quelque soit l'âge. L'interaction entre maladie cardiaque et âge est très similaire à celle entre hypertension et âge. On peut en conclure que le fait d'avoir une maladie cardiaque est consistant avec le risque de faire un AVC et que les moins de 30 ans sans maladie cardiaque sont consistants avec un risque faible d'AVC.

Une troisième et dernière étude par ACM nous permet d'observer l'interaction entre l'hypertension et le niveau de glucose. Les modalités de niveau de glucose supérieur à 127 avec ou sans hypertension et l'hypertension qu'importe le niveau de glucose semblent plus consistantes avec le risque d'AVC que les autres modalités (absence de maladie cardiaque avec un niveau de glucose inférieur à 127). On choisit de ne conserver que 3 modalités glucose sous 127 sans hypertension, glucose supérieur à 127 sans hypertension et hypertension quelque soit le niveau de glucose. L'hypertension est plus consistante avec un fort risque d'AVC qu'un niveau élevé de glucose. On en conclut qu'un niveau de glucose inférieur à 127 sans hypertension est consistant avec un risque faible d'AVC.

6. Équilibrage du jeu de données

Tout d'abord, avant de se pencher sur les différents modèles d'apprentissages, il est intéressant d'examiner l'équilibrage du jeu de données.

6.1. Problématique

Comme illustré au début du rapport (Fig. 1), le jeu de données contient 3245 cas négatifs (avcN) et 180 cas positifs (avcP). Ainsi la classe "avcN" représente 94,7% du jeu de données, respectivement "avcP" représente 5,3% de celui-ci.

La problématique d'un tel déséquilibrage est que les modèles vont être surentraînés sur la classe "avcN". De ce fait, les modèles seront très peu performants sur la détection de la classe "avcP", or c'est surtout cette classe qui est importante car on souhaite prévenir les AVC.

Pour rectifier ce problème d'équilibrage, deux solutions possibles :

- L'oversampling
- L'undersampling

6.2. Oversampling

L'oversampling est une méthode qui va créer de nouveaux échantillons de la classe minoritaire. Il existe plusieurs méthodes d'oversampling, celle que nous avons choisie est la suivante :

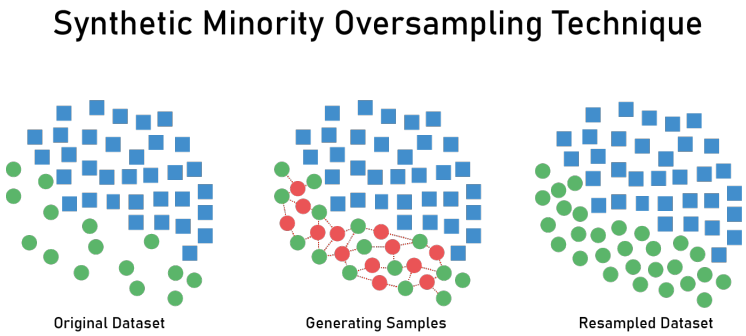


FIGURE 15 – Principe de l'oversampling

Le principe est d'utiliser un algorithme des plus proches voisins pour créer artificiellement de nouvelles données. Ce sont des données qui n'existent pas mais qui se rapprochent des données de la même classe. Pour utiliser facilement cette méthode, il existe le package "imbalance" qui met la méthode "oversample" à disposition.

6.3. Undersampling

Contrairement à l'oversampling qui va créer de nouveaux échantillons de la classe minoritaire, l'undersampling va supprimer des données de la classe majoritaire. Le principe de la méthode est de sélectionner aléatoirement des données de la classe majoritaire pour avoir le même nombre de données que dans la classe minoritaire.

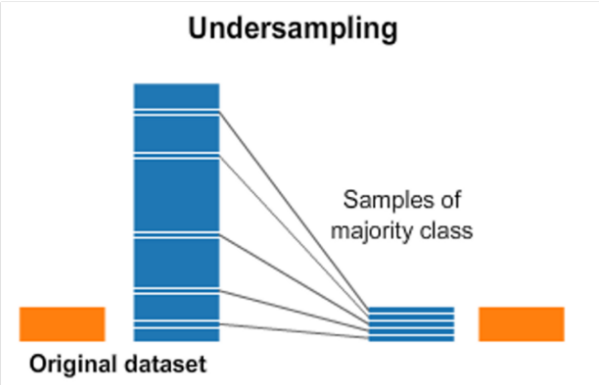


FIGURE 16 – Principe de l'undersampling

Le problème de cette méthode dans notre cas est que nous n'avons que 180 données dans la classe minoritaire. Un équilibrage avec cette méthode revient à supprimer 3000 données de la classe majoritaire et donc perdre $\backsim$ 90% du jeu de données.

6.4. Comparaison des différents équilibrages

Pour identifier quelle est la meilleure méthode d'équilibrage dans notre cas, nous avons décidé de faire quatre tests :

- Apprentissage avec jeu de base
- Apprentissage avec variables de base et oversampling
- Apprentissage avec variables quantitatives converties en qualitatives et oversampling
- Apprentissage avec variables de base et undersampling

Chaque équilibrage va être testé avec deux méthodes d'apprentissage (SVM et Random forest). Nous avons choisi ces méthodes car ce sont celles qui donnent les meilleurs résultats en général.

Pour comparer les différents tests quatre indicateurs sont utilisés :

- Erreur de classification : erreur de classification du modèle
- Kappa : permet de savoir si le modèle classe au hasard (0) ou non (1)
- Sensibilité : Pourcentage de cas négatifs bien classés
- Spécificité : Pourcentage de cas positifs bien classés

Méthode	Erreur	Kappa	Sensibilité	Specificité
svm unbalanced	5,26%	0	1	0
rf unbalanced	5,45%	0,13	0,99	0,09
svm oversample	15,42%	0,69	0,83	0,86
rf oversample	9,20%	0,82	0,88	0,94
svm oversample qualitative	18,65%	0,63	0,77	0,86
rf oversample qualitative	15,26%	0,69	0,82	0,87
svm undersample	20,37%	0,59	0,7	0,89
rf undersample	24,07%	0,52	0,72	0,8

FIGURE 17 – Principe de l'undersampling

Concrètement, les résultats des tests ressortent du tableau. Tout d'abord, le test avec le jeu de données non équilibré. Il est très performant en terme d'erreur de classification cependant grâce aux trois autres indicateurs on peut voir qu'il classe au hasard les données. Il va classer presque toutes les données en tant que cas négatif, le jeu étant composé à 94% de cette classe, l'erreur sera très faible. Ce jeu de données est inutilisable pour détecter les cas positifs.

Le second test qui ressort est le jeu de données avec les variables de base et l'oversampling. Les indicateurs sont tous très bons, ce qui signifie que le modèle classe aussi bien les cas positifs que négatifs. Les deux autres tests ne sont pas mauvais également mais moins performants tout de même. Dans la suite de l'étude, nous continuons avec le jeu de données avec les variables de base et l'oversampling pour ces raisons.

7. Apprentissages pour effectuer des prédictions

La dernière étape de l'étude est de tester différents modèles d'apprentissage pour examiner quel est le modèle le plus adéquat à notre problématique. Dans la suite, les résultats présentés sont les résultats des méthodes avec les paramètres optimisés. La métrique que nous avons décidé d'utiliser pour l'optimisation est le ROC. Cette mesure permet de comparer facilement deux modèles. Plus le ROC est élevé meilleur est le modèle.

Nous avons mis en paramètre 'preProcess = c("scale", "center")' lors de chaque train du fait que nous utilisons des variables quantitatives qui ne sont pas toutes au même niveau de grandeur.

7.1. Régression logistique binaire

La méthode qui est utilisée pour tuner la régression logistique binaire est celle de la selection forward par minimisation de l'AIC. Voici les paramètres finaux de la régression logistique tuned :

```
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -0.11771    0.04625  -2.545 0.010932 *
genderMale           -0.47953    0.04519 -10.612  < 2e-16 ***
age                   2.14094    0.06987  30.643  < 2e-16 ***
hypertensionhtP      -0.20762    0.04141  -5.013 5.35e-07 ***
heart_diseasehdP     -0.17411    0.04236  -4.110 3.95e-05 ***
ever_marriedYes      -0.51571    0.05421  -9.513  < 2e-16 ***
work_typegovt_job     0.64373    0.07522   8.558  < 2e-16 ***
work_typeprivate      0.24427    0.06077   4.019 5.84e-05 ***
work_typeself_employed -0.35116    0.05573  -6.301 2.95e-10 ***
Residence_typeUrban  -0.42652    0.04422  -9.645  < 2e-16 ***
avg_glucose_level     0.57740    0.04739  12.185  < 2e-16 ***
smoking_statusnever  -0.17942    0.04782  -3.752 0.000176 ***
smoking_statussmokes -0.34594    0.05154  -6.712 1.92e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

FIGURE 18 – Paramètre de la régression logistique binaire tuned

Le modèle final ne contient plus que neuf variables. La variable qui n'est plus présente dans le modèle est bmi.

7.2. K Nearest Neighbors (K-NN)

Pour optimiser le modèle K-NN, le paramètre sur lequel on peut facilement jouer est le nombre de K. Pour l'optimisation, k varie entre 1 et 50. En traçant le ROC en fonction de chaque k, il est simple de déterminer le nombre k optimal.

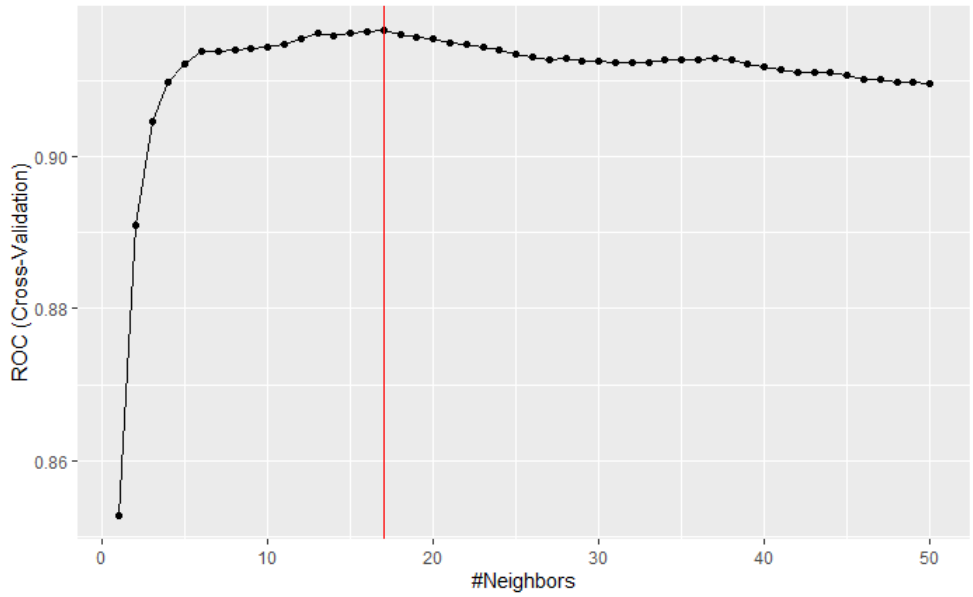


FIGURE 19 – Paramètre de K-NN tuned

Le trait rouge correspond à la valeur de k optimale et donc la valeur pour laquelle k maximise le ROC. La valeur k optimale est 17.

7.3. Support Vector Machines (SVM)

SVM est un modèle qui va chercher à maximiser ces marges. Le paramètre C de SVM permet de faire varier le coût. Le coût correspond au compromis entre mauvaise classification et simplicité du modèle. Le coût change également la valeur du ROC. Nous faisons varier C entre 0.01 et 1 avec un pas de 0.03 (pour accélérer le temps de calcul).

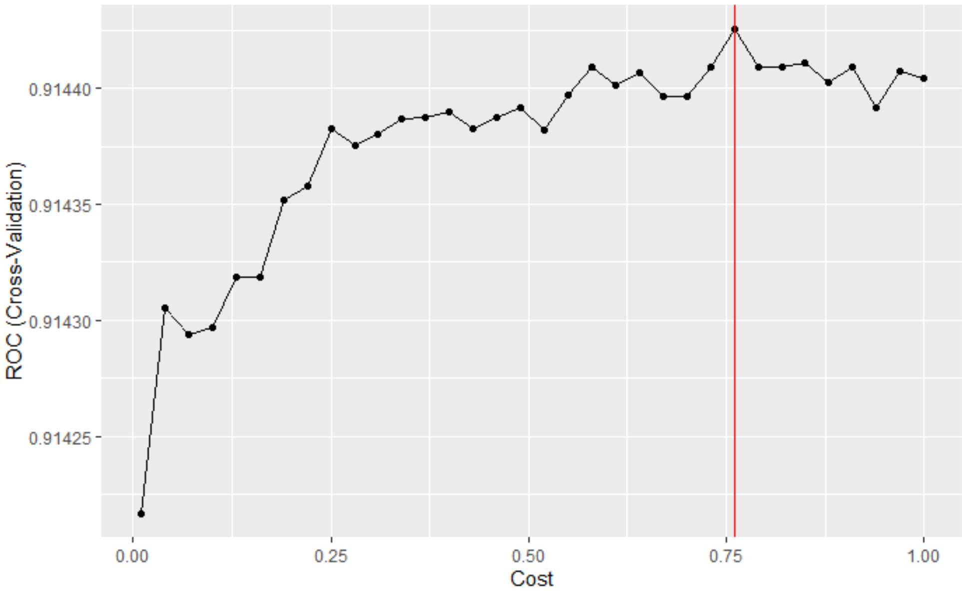


FIGURE 20 – Paramètre de SVM tuned

Le trait rouge correspond à la valeur C optimale et donc la valeur pour laquelle C maximise le ROC. La valeur de C optimale est 0,76.

7.4. Arbre de décision

Concernant l'arbre de décision, le paramètre le plus simple à modifier est sa profondeur. Nous obtenons d'abord un arbre facilement lisible mais avec des résultats pas très bons. De ce fait nous avons décidé d'augmenter un peu sa profondeur.

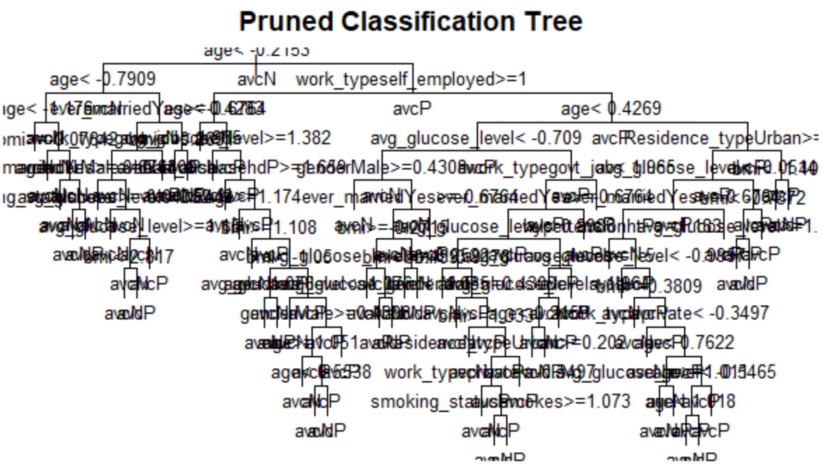


FIGURE 21 – Paramètre d'arbre de décision tuned

L'arbre n'est pas très lisible mais c'est un compromis entre un arbre trop entraîné sur le training set et un autre facilement lisible mais trop général et qui n'a pas de bonnes performances.

7.5. Forêt aléatoire

Pour optimiser l'arbre de décision, nous pouvons faire varier le nombre d'arbres de la forêt. Le modèle optimal est un modèle avec 500 arbres (valeur maximale que nous avons essayée). Il est intéressant de voir quelles sont les variables importantes dans la décision de la forêt.

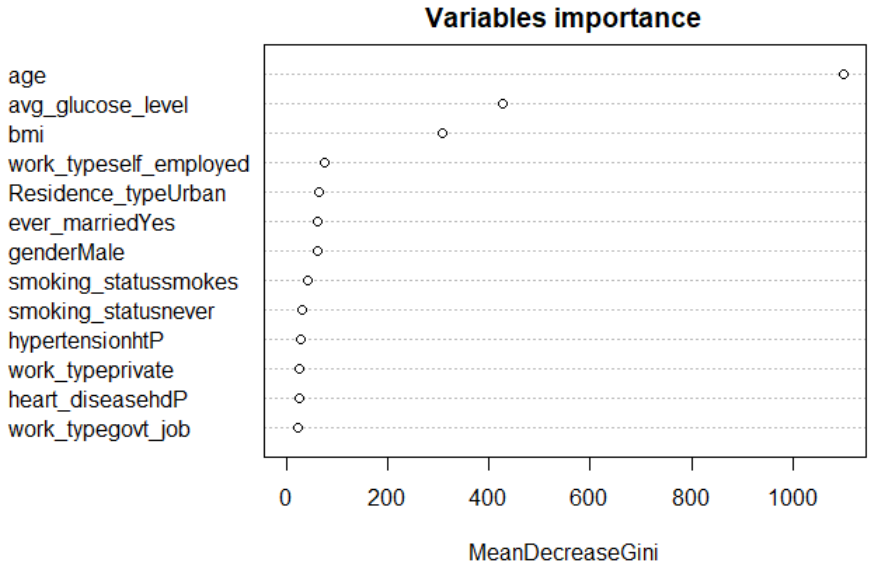


FIGURE 22 – Paramètre de la forêt aléatoire tuned

Globalement, dans la décision de la forêt trois variables ont un impact fort, ce sont les trois variables quantitatives :

- age
- avg_glucose_level
- bmi

7.6. Comparaison des différents modèles

Une fois les différents modèles entraînés, il est intéressant de comparer les résultats qu'ils donnent sur différents indicateurs pour déterminer le plus adéquat.

Methode	Erreur	Kappa	Sensibilité	Specificité	AUC
glm_tuned	15,52%	0,69	0,83	0,86	0,84
knn_tuned	15,78%	0,68	0,8	0,89	0,84
svm_tuned	15,31%	0,69	0,83	0,86	0,85
tree_tuned	14,23%	0,72	0,83	0,89	0,86
rf_tuned	8,99%	0,82	0,88	0,94	0,91

FIGURE 23 – Comparaison des résultats des modèles

À l'aide du tableau, on remarque tout de suite un modèle qui se démarque des autres. La forêt aléatoire a des performances supérieures à tous les autres modèles qui sont tous assez similaires en terme de performance.

Il est possible de regarder les indicateurs sous forme de boxplot pour voir la moyenne et l'écart-type du modèle.

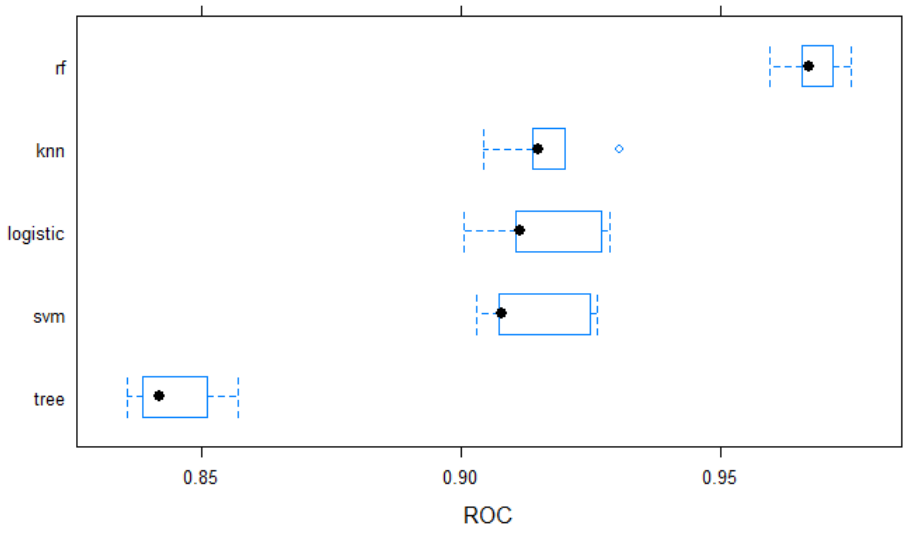


FIGURE 24 – Comparaison des ROC des modèles

En terme de ROC, la forêt aléatoire a globalement une faible variabilité comparée aux autres modèles en plus d'avoir le meilleur ROC. Le K-NN a un ROC assez intéressant. L'arbre de décision est de loin le moins bon des modèles en terme de ROC.

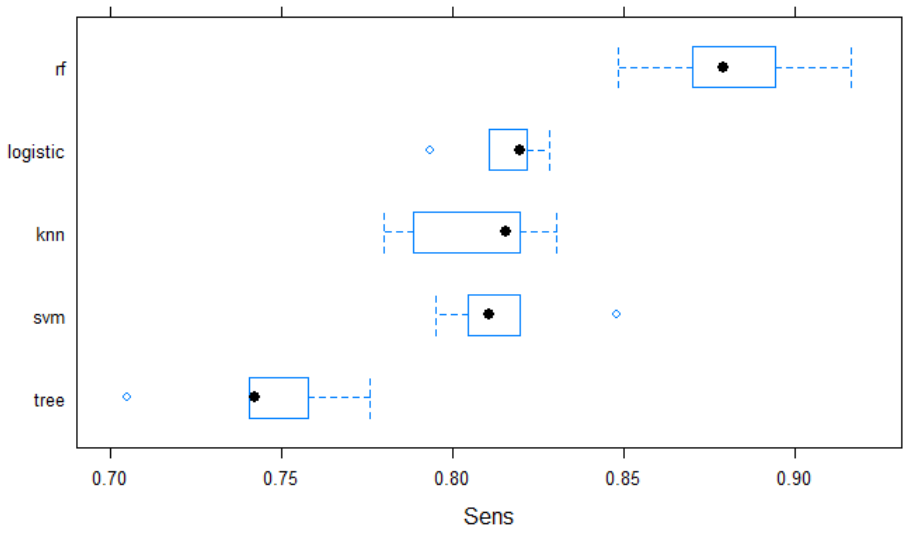


FIGURE 25 – Comparaison des sensibilités des modèles

En terme de sensibilité, la forêt aléatoire reste meilleure cependant elle a une forte variabilité. K-NN reste encore très bon sur ce paramètre. L'arbre de décision est quant à lui toujours le modèle le moins performant.

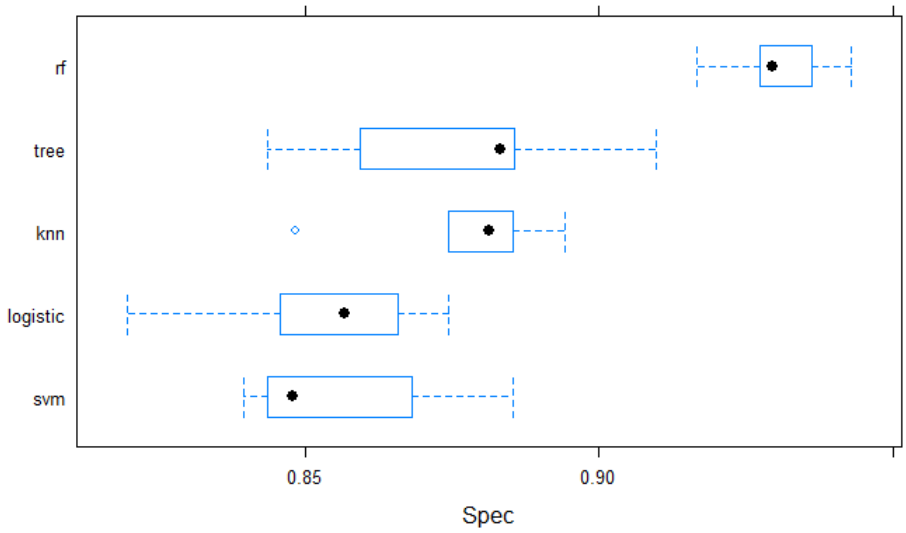


FIGURE 26 – Comparaison des spécificités des modèles

En terme de spécificité, la forêt aléatoire reste meilleure. K-NN reste très bon sur ce critère avec une faible variabilité. L'arbre de décision est meilleur sur ce paramètre mais il a une très grande variabilité.

Parmi les critères de comparaisons, la spécificité est certainement le critère le plus important. C'est le critère qui traduit à quel point notre modèle est performant dans la détection des cas positifs. Sachant qu'il est plus important de détecter un cas positif que négatif, c'est le critère qui pèse le plus lourd dans le choix de notre modèle.

Avec l'aide de tous ces indicateurs, on peut très clairement admettre que sur notre jeu de données la forêt aléatoire sera le modèle d'apprentissage le plus performant. Qui plus est, c'est un modèle que l'on peut interpréter. On sait que les variables quantitatives sont les variables qui ont le plus d'importance pour savoir si un patient a un risque d'AVC ou non. Ainsi, un médecin pourrait savoir sur quels paramètres se concentrer pour étudier le risque. Un modèle comme le SVM ne nous permettrait pas cela.

8. Conclusion

Le jeu de données que nous avons choisi est très intéressant du fait que les AVC peuvent toucher tout le monde et des fois sans prévenir. Les risques d'un AVC sont multiples, on peut aller de la paralysie à la mort du patient. C'est pour analyser ce genre de maladies et les prévenir que l'analyse de données est très intéressante.

Cependant, le gros problème du jeu de données est qu'il n'est pas équilibré. Il aurait été intéressant d'avoir autant de cas d'AVC positifs que de cas négatifs. Dans cette étude, la méthode qui a été mise en place a été d'équilibrer le jeu de données. Le choix entre l'oversampling et l'undersampling s'est posé. L'oversampling était plus performant car il a créé 3000 nouvelles données. L'undersampling pourrait être tout aussi performant si le déséquilibrage de nos classes n'était pas aussi grand. Utiliser l'undersampling dans notre jeu de données revient à supprimer 3000 données soit $\backsim$ 90% des données. C'est pourquoi dans notre cas, il est préférable d'utiliser l'oversampling.

Après l'étude des différents modèles, nous avons pu comparer les modèles avec plusieurs indicateurs. L'indicateur le plus important est la spécificité car il permet de donner la performance dans la détection des cas positifs. La forêt aléatoire est le modèle qui permet une très bonne interprétabilité des résultats en plus d'avoir les meilleures performances. On en conclut donc que c'est le modèle le plus adapté à la détection des futurs cas d'AVC.

Lien vers le code qui a servi à l'étude : https://github.com/BenjaminDemolin/AVC_DataSante