RÉPUBLIQUE DU SÉNÉGAL

UNIVERSITÉ GASTON BERGER DE SAINT-LOUIS

U.F.R. DE SCIENCES APPLIQUÉES ET TECHNOLOGIE

Département de Mathématiques Appliquées

Mémoire présenté pour l'obtention du diplôme de Master

en Mathématiques Appliquées

Parcours : Sciences des Données et Applications (SDA)

Option : Statistique

présenté par

ADAMA SALL

Modélisation des risques d'AVC à l'aide de la Statistique des extrêmes et des techniques de machine learning

Soutenu le 24 Juillet 2025 devant le jury composé de :

El hadj Déme	Président	Professeur titulaire
Aliou DIOP	Encadrant	Professeur titulaire
Aba DIOP	Co-encadrant	Professeur titulaire
Sadibou AIDARA	Examinateur	Maitre assistant

Encadré et supervisé par **Pr. Aliou Diop** et **Pr. Aba Diop**

Année Universitaire : 2023-2024

DÉDICACES

Je dédie ce mémoire :

À la personne la plus précieuse de ma vie, ma mère ; ma confidente **Awa Dieng**, une femme exemplaire, dévouée, forte, courageuse, qui s'est battue sans relâche pour la réussite de son fils.

À mon père, **Baba Sall**, un modèle de courage, d'humilité et d'éducation. Un homme intègre, humble, toujours prêt à guider et à soutenir. Merci d'avoir été un repère, un ami, un guerrier et un véritable compagnon de route.

Ce mémoire vous est dédié pour tout ce que vous avez fait de moi. Vous m'avez transmis des richesses immatérielles, des valeurs, des principes et une force intérieure que je ne pourrai jamais vous rendre à leur juste valeur.

Je dédie également ce mémoire :

À ma jumelle, **Awa Sall**, avec qui je partage un lien unique, une présence constante, un soutien indéfectible et des encouragements silencieux mais puissants.
À mon frère, **Babacar Sall**, ma soeur **Seynabou Sall**, mon neveu **Moustapha Sy** et également **Ndeye Diama Gaye** et **Macoumba Sall** pour votre présence et votre bienveillance, toujours là dans les moments importants.
À toute ma famille, pour l'amour, la force et les valeurs que vous m'avez transmis, et pour n'avoir jamais cessé de croire en moi.

À toutes celles et ceux qui, un jour, m'ont demandé avec un sourire : « Alors, c'est pour quand la soutenance ? » Eh bien, réjouissez-vous : le moment est enfin arrivé ! Je soutiens aujourd'hui.

REMERCIEMENTS

*Avant toute chose, je rends grâce à **Allah, le Tout-Puissant**, pour m'avoir accordé la santé, la patience, la persévérance et la force nécessaires à la réalisation de ce mémoire.*

Je tiens à exprimer ma profonde gratitude au **Professeur Aliou Diop**, pour son accompagnement rigoureux, ses conseils avisés et sa disponibilté constante tout au long de ce mémoire. Ses remarques pertinentes ont grandement enrichi la qualité de ce travail.

Mes remerciements vont également au **Professeur Aba Diop**, pour son encadrement bienveillant, sa disponibilité, ses orientations précieuses et son soutien continu. Son expertise et ses suggestions m'ont été d'une grande aide à chaque étape de ce projet.

Messieurs les Professeurs, *par votre rigueur intellectuelle, votre engagement et votre bienveillance, vous êtes pour moi des références. Votre présence attentive et constante, tant en cours que lors du suivi individuel, a été une véritable source d'inspiration.*

Je souhaite également remercier **l'Université Alioune Diop de Bambey**, qui m'a offert des bases solides durant ma licence. Ce socle de connaissances a constitué un appui fondamental dans la poursuite de mon parcours universitaire. Je lui suis profondément reconnaissant pour la qualité de la formation reçue.

Je souhaite exprimer ma profonde gratitude à mes amis, à mes camarades de classe et à tous ceux qui, de près ou de loin, m'ont encouragé, soutenu, écouté, conseillé ou simplement cru en moi.Votre bienveillance, vos mots d'encouragement et votre présence à mes côtés, dans les moments de doute comme dans les moments de joie, ont été précieux et m'ont permis d'aller jusqu'au bout de ce travail. Merci à chacun d'entre vous d'avoir cru en moi et de m'avoir donné la force de persévérer jusqu'à la réussite.

« *Je vous recommande trois éthiques qui constituent votre viatique pour l'atteinte des objectifs : Je vous exhorte à la discipline dans la quête des sciences, à la réflexion et à l'approfondissement dans les recherches à l'exemple des devanciers* »
Serigne Cheikh Ahmed Tidjan Sy Al Maktoum

Résumé

Ce mémoire porte sur **la modélisation des risques d'AVC à l'aide de la Statistique des extrêmes et des techniques de machine learning**. Face à l'importance croissante des AVC dans la santé publique, l'objectif est d'identifier les facteurs de risque associés à la survenue d'un AVC à partir des variables socio-démographiques puis de quantifier la probabilité individuelle et d'évaluer la performance des différents modèles prédictifs. Ce travail commence par une revue des méthodes statistiques classique notamment la régression logistique puis introduit la Statistique des extrêmes (GEV) pour mieux analyser les cas rares et graves. Plusieurs algorithme de machine learning (*Random Forest, SVM, KNN, Gradient Boosting, Réseau de neurones*) sont ensuite appliqués et comparés pour leur capacité à prédire le risque d'AVC.
L'étude montre l'apport combiné des méthodes statistiques avancées et de techniques de machine learning dans la modélisation des risques d'accident vasculaire cérébral (AVC).

Les études confirment l'influence significative de plusieurs facteurs de risque, notamment l'âge avancé, l'hypertension, les maladies cardiaques et un taux de glucose élevé. Les méthodes traditionnelles, comme la régression logistique se sont révélées pertinentes pour identifier et quantifier ces facteurs de risque, confirmant leur pertinence pour l'interprétation clinique, bien qu'elles présentent certaines limites en termes de performance prédictive. La Statistique des extêmes (GEV) a permis quant à elle, d'identifier des seuils critiques pour les variables continues (âge, glycémie, IMC), mettant en évidence les profils extrêmes les plus vulnérables..

En parallèle, les techniques de machine learning, bien que souvent considérées comme une boîte noire en raison de leur faible interprétabilité ont démontré une supériorité en termes de précision prédictive. Elles ont également mis en lumière l'importance d'autres variables explicatives telles que le statut tabagique, le type de travail ou encore le statut matrimonial. Parmi ces méthodes, le Gradient Boosting s'est distingué par ses excellentes performances, offrant une bonne capacité discriminante pour identifier les patients à risque. Quant au réseau de neurones, il a affiché un rappel exceptionnel, détectant la quasi-totalité des cas d'AVC, ce qui en fait un outil adapté pour le dépistage.
En résumé, ce travail met en lumière la complémentarité entre les approches statistiques et les méthodes d'apprentissage automatique pour une modélisation plus fine et efficace. La première reste essentielle pour l'analyse explicative des facteurs de risque tandis que la deuxième apporte une puissance prédictive, ouvrant la voie à une prévention et une prise en charge mieux ciblées des populations.

Mots clé : Accident Vasculaire Cérébral, Modélisation, Statistique des extrêmes, Loi généralisée des valeurs extrêmes, Machine Learning, Facteurs de risque cardiovasculaire,

Abstract

This thesis focuses the modeling of stroke (AVC) risk using extreme value statistics and machine learning techniques. Given the growing public health impact of stroke, the main objective is to identify risk factors associated with stroke occurrence from socio-demographic variables, quantify individual risk probabilities, and evaluate the performance of various predictive models. The work begins with a review of classical statistical methods, notably logistic regression, and introduces extreme value statistics (GEV) to better analyze rare and severe cases. Several machine learning algorithms (Random Forest, SVM, KNN, Gradient Boosting, and Neural Networks) are then applied and compared for their ability to predict stroke risk.

The study highlights the combined value of advanced statistical methods and machine learning in stroke risk modeling. The results confirm the significant influence of several risk factors, especially advanced age, hypertension, heart disease, and elevated glucose levels. Traditional methods like logistic regression prove useful for identifying and quantifying these factors, supporting their clinical interpretability, though with some limitations in predictive performance. Extreme value statistics (GEV) enable the identification of critical thresholds for continuous variables (age, glucose, BMI), revealing the most vulnerable extreme profiles.

In parallel, machine learning techniques, although often seen as black-box models due to limited interpretability, demonstrated superior predictive accuracy.They also uncovered the importance of additional variables such as smoking status, type of occupation, and marital status. Among these methods, Gradient Boosting exhibited excellent discriminative power, while neural networks achieved outstanding recall, detecting nearly all stroke cases—making them highly suitable for screening..
In summary, this work underscores the complementarity between statistical and machine learning approaches. Traditional methods remain essential for explanatory analysis, while machine learning offers powerful predictive capabilities. Together, they pave the way for more targeted prevention and improved care for at-risk populations.

Keywords : Stroke, Modeling, Extreme value statistics, Generalized Extreme Value (GEV) distribution, Machine Learning, Cardiovascular risk factors

Liste des abréviations

Abréviation	Signification
AVC	Accident Vasculaire Cérébral
OMS	Organisation Mondiale de la Santé
ML	Machine Learning
GEV	Loi Généralisée des Valeurs Extrêmes
MLE	Estimation du maximum de vraisemblance (Maximum Likelihood Estimation)
AIC	Akaike information criterion (critère d'information d'Akaike)
SVM	Support Vector Machine
RF	Random Forest
GB	Gradient Boosting
NN	Réseau de Neurones (Neural Network)
CNN	Réseau de neurones convolutif (Convolutifal Neural Network)
AUC	Aire sous la courbe (Area Under the Curve)
ROC	Receiver Operating Characteristic
IMC	Indice de Masse Corporelle
VP	Vrai Positif
VN	Vrai Négatif
FP	Faux Positif
FN	Faux Négatif
TPR	Taux de Vrais Positifs
FPR	Taux de Faux Positifs
KNN	K-Nearest Neighbors
CSV	Comma-Separated Values

Table des matières

Introduction générale

Motivations et contextes

L'accident vasculaire cérébral (AVC) constitue l'une des problématiques de santé publique à l'échelle mondiale. C'est une pathologie grave, une urgence neurologique fréquente impliquant à la fois un pronostic vital et fonctionnel [6]. L'Accident Vasculaire Cérébral est une maladie lourde de conséquences causant les principaux décès et d'handicap dans le monde. Selon les statistiques de l'organisation mondiale de la santé (OMS), environ 15 millions de personnes souffrent d'un AVC chaque année dans le monde. Parmi elles 5 millions décèdent et 5 d'autres millions conservent des séquelles permenantes ce qui fait de lui la deuxième cause de décès dans le monde, juste après les maladies cardiovasculaires. L'impact de cette maladie ne se limite pas à la mortalité, elle est principale cause d'incapacité physique chez les adultes et reprénsente un défis majeur pour les systèmes de santé (Selon l'OMS [8]).

En 2005 l'OMS estimait jusqu'à 16 millions le nombre de personnes pouvant être victimes d'un AVC à travers le monde et à 6,2 millions le nombre de décès attribué à cette maladie . Ce nombre pourra atteindre 23 millions avec 7,8 millions de décès en 2030 par le simple fait du vieillissement de la population.

En Afrique subsaharienne (ASS), les maladies non transmissibles tels que les AVC étaient considérées pendant longtemps comme rares. Aujourd'hui, l'Afrique pourrait avoir des taux d'incidence et de prévalence d'AVC jusqu'à 2 à 3 fois plus élevés que ceux de l'Europe occidentale et des États-Unis du fait de l'allongement de l'espérance de vie et de l'exposition aux nombreux facteurs de risque vasculaires. L'OMS estime que 80% des AVC se produiront dans les pays en voie de développement en 2030.

D'aprés les progrès de la science, un certains nombres de facteur liés à la survenue de l'AVC ont été mis en évidence . Il s'agit des facteurs de risque modifiables et les facteurs non modifiables (selon O'Donnell et all).

Les systémes de santé peinent à trouver les meilleurs solutions pour l'amélioration de la qualité, accroître l'éfficacité et la réduction des coûts de soins. Face à ce défi sanitaire, la construction et l'utilisation des modèles d'analyse prédictive apparaissent comme une solution prometteuse pouvant modéliser et anticiper les risques chez un patient ou de prédire ses chances de compilcation ou de récidive après un évènement.

Objectif du mémoire

Ce mémoire vise à explorer et appliquer des méthodes statistiques avancées et de machine learning pour la modélisation des risques d'AVC. L'objectif principal est d'identifier les facteurs de risque associés à la survenue d'un accident vasculaire cérébral à partir des variables cliniques et démographiques en utilsant les modèles traditionnelles et des techniques de machine learning. Ensuite quantifier la probabilité individuelle de la survenue d'AVC et d'évaluer la performance des modèles pour la prediction du risque AVC à l'aide des critères objectifs , l'objectif est aussi de de caractériser le profil des patients à haut risque et d'étudier la distribution des variables extrêmes chez ces patients.

En résumé, ce travail démontre l'apport de la statistique des extrêmes et du machine learning pour la prédiction et la prévention des AVC, en mettant en avant la capacité de ces approches à mieux cibler les individus les plus exposés à un risque grave et rare

Questions de recherche

Ce mémoire s'inscrit dans une double perspective :

— D'une part, identifier les facteurs cliniques et démographiques les plus significativement associés à la survenue d'un AVC.

— D'autre part, évaluer la pertinence des outils statistiques avancés notamment la statistique des valeurs extrêmes (GEV) et les techniques de machine learning pour modéliser ce risque à l'échelle individuelle.

Ces approches visent à dépasser les limites des modèles traditionnels, souvent centrés sur des moyennes et moins sensibles aux comportements extrêmes, c'est-à-dire les profils de patients exposés à des valeurs inhabituelles (âge très avancé, glucose très élevé, etc.), qui sont pourtant souvent les plus à risque.

Dans le contexte actuel où les maladies cardiovasculaires représentent l'une des premières causes de mortalité dans le monde, l'accident vasculaire cérébral (AVC) constitue un enjeu majeur de santé publique. Une meilleure compréhension de ses facteurs de risque ainsi qu'une capacité prédictive efficace sont essentielles pour améliorer la prévention et la prise en charge des populations vulnérables. Ainsi, la question centrale qui guide cette étude est la suivante :

Quels sont les facteurs cliniques et démographiques les plus significativement associés à la survenue d'un accident vasculaire cérébral (AVC), et dans quelle mesure les approches statistiques avancées notamment la statistique des extrêmes (GEV) et les techniques de machine learning permettent-elles de prédire individuellement ce risque, d'identifier les profils de patients les plus exposés à des événements rares et graves, et d'améliorer la performance par rapport aux modèles classiques ?

Chapitre 1

REVUE BIBLIOGRAPHIQUE ET FONDEMENT THÉORIQUE

1 Introduction aux AVC

En raison de leur fréquence, de leur gravité et de leur coût, les accidents vasculaires cérébraux (AVC) constituent un des problemes de santé publique les plus préoccupant actuellement et pour les prochaines années. Ils sont définis par l'Organisation mondiale de la santé (OMS) comme « **le développement rapide de signes cliniques localisés ou globaux de dysfonction cérébrale avec des symptômes durant plus de vingt quatre heures pouvant entraîner la mort, sans autre cause apparente qu'une origine vasculaire** ». Cette perte soudaine de la fonction cérébrale due à un infarctus ou à une hémorragie, distinguant ainsi deux entité pathologiques principales : les accidents ischémiques ou infarctus cérébraux, représentant 80% des AVC et les accidents hémorragiques ou hématomes intra cérébraux représentant 20% des AVC

L'AVC ischémique, aussi appelé infarctus cérébral ou ramollissement du cerveau se présente comme l'un des principaux handicaps acquis dans le monde impactant de manière significative l'autonomie et la qualité de vie des individus touchés, Il est le plus souvent le résultat d'un thrombus (un caillot qui se forme dans une artère), d'une embole (un corps étranger, le plus souvent un caillot, qui, porté par la circulation, va obstruer l'artère en aval), ou d'un rétrécissement de l'artère causé par l'athérosclérose (épaississement de la paroi interne de la paroi artérielle). L'athérosclérose est la première cause d'accident vasculaire cérébral ischémique (50 à 60% des cas)

Les AVC hémorragiques sont causés par un épanchement de sang dans le tissu cérébral. Leur cause est généralement une hypertension artérielle, ou beaucoup plus rarement, des formations maliques vasculaires (angiome, anévrisme), des troubles de la coagulation ou des complications d'un traitement anticoagulant. Les AVC hémorragiques représentent 10 à 15 % de l'ensemble des AVC, soit 10 à 20 cas par 100000 habitants. [7]

1.2 Importance de la modélisation

La modélisation joue un rôle crucial dans le domaine de la medecine. Face à des prédictions alarmantes de 23 millions de cas annuel d'ici 2030, modéliser les risques liés à cette maladie devient une priorité. Sur le plan médical que socio-économique, la modélisation aide à identifier précocement les patients à haut risque facilitant la mise en place des stratégies préventives

ciblées, d'optimiser les soins médicaux permettant de prioriser les patients nécèssitant une prise en charge urgente et enfin personnaliser les soins pour pouvoir les adapter aux besoins spécifiques de chaque patient.

2 Modèles traditionnels pour la modélisation des risques d'AVC

2.1 Présentation de la régression logistique

La régression logistique est un modèle linéaire de classification statistique probabilité permettant d'étudier la relation entre un ensemble de variable quantitave $X = (X_1, X_2, \ldots, X_p)$ et une variable qualitative Y binaire. Comme c'est un modèle linéaire on utilise une fonction de score $S(X) = a_1X_1 + a_2X_2 + ... + a_pX_p$. L'objectif est de déterminer les coefficients $a_1, ..., a_p$ tels que le score $S(x)$ soit positif lorsque les probabilité d'appartenance au groupe 1 est élevé et négatif dans le cas contraire.

Pour la probabilité $P_1(x)$ d'appartenance aux groupes, on applique une fonction logistique (**sigmoïde**) au score qui a la forme indiquée dans la figure 3.4. Sa forme explicite est définie par :

$$P(x) = \frac{1}{1 + e^{-S}},$$

Cette fonction prend en entrée une valeur réelle $-\infty \leq S \leq \infty$ et renvoie une valeur dans l'intervalle $[0, 1]$

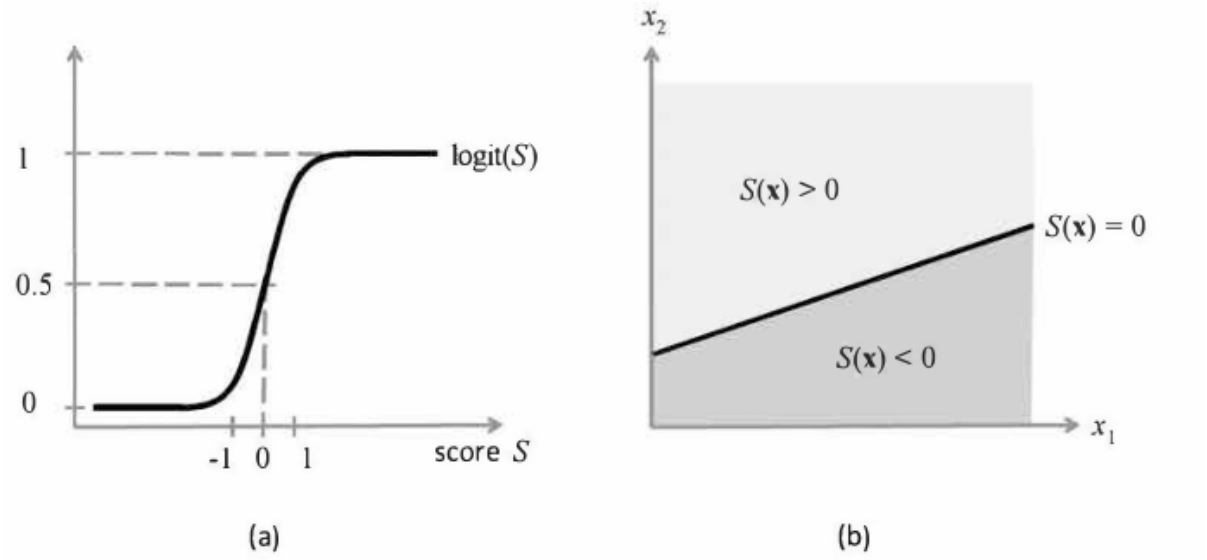


FIGURE 1.1 – Fonction sigmoïde

3. Statistique des extrêmes

3.1 Définition et objectif

Historiquement la Statistique des extrêmes ou la théorie des valeurs extrêmes est l'une des branche de la Statistique qui s'intéresse spécifiquement à l'étude des phénomenes rares, c'est-à-dire ceux dont la probabilité d'apparition est faible mais qui peuvent causer d'importants dégâts (ex : catastrophes naturelles, crues, épidémies, etc.). La Statistique des extrêmes analyse le comportement des queues de la distribution limite des maxima (ou minima) observés dans un ensemble de données, là où se trouve les événements rares mais potentiellement catastrophiques. Dans ce contexte nous pouvons considérer que la théorie des valeurs extrêmes comme le contrepartie de la statistique classique, qui est principalemnet basée sur l'étude des valeurs autour de leur moyenne. Dans le cadre de la statistique classique les valeurs aberrantes ne sont pas prise en compte. Par contre elle font l'objet de toute les attentions en théorie des valeurs extrêmes. Cette théorie est apparue entre 1920 et 1940, grâce aux travaux de **Fisher-Tippet**, de **Gumbel** et de **Gnedenko** qui a pour but de fournir des méthodes pour estimer les probabilités d'événements extrêmes, les quantiles extrêmes et les risques associés.

Plaçons nous dans le cadre statistique et considérons une suite de variables aléatoire $X_1, \ldots, X_n$ indépendante et identiquement distribuée (i.i.d) de fonction de répartition F. On note la statistique d'ordre de l'echantillon le réarrangement croissant suivant :

$$X_{1,n} \leq X_{2,n} \leq \cdots \leq X_{n,n}$$

Dans la théorie des valeurs extrêmes deux statistiques d'ordre sont particulièrement intéressants pour l'étude des évenements extrêmes, le minimum de l'échantillon ($Min = X_{1,n}$) et le maximun ($Max = X_{n,n}$). On notera qu'il existe une relation pour passer de l'un à l'autre :

$$Min = -Max(-X_1, \ldots, -X_n)$$

3.2 Concepts fondamentaux (étude de quelques distributions)

3.2.1 Distribution de probabilité de Fréchet

La distribution dite de Fréchet a été introduite par le mathématicien **François Maurice Réné Fréchet**. Il a developpé cette distribution comme loi des valeurs extrêmes. Les travaux de Fréchet ont inspiré les recherches de *Fisher-Tippett* en 1928 et ceux de *Gumbel* en 1958. La distribution de Fréchet est l'une des distributions les plus populaires dans de nombreux domaines, par exemple en hydrologie, la distribution de Fréchet est utilisé pour des evenements extrêmes.

Définition 3.2.1 : On dit qu'une variable aléatoie X est de distribution de Fréchet de paramètre $\mu, \sigma > 0$ et $\gamma > 0$ si sa densité de probabilité peut s'écrit telle que :

$$h_{\gamma,\sigma,\mu}(x) = \begin{cases} 0 & six < \frac{1}{\gamma} \\ \frac{\gamma}{\sigma}\left(\frac{x-\mu}{\sigma}\right)^{-1-\gamma} \exp\left(-\left(\frac{x-\mu}{\sigma}\right)^{-\gamma}\right) & six \geq \frac{1}{\gamma}, \end{cases} \tag{2.1}$$

où γ est le paramètre de forme, σ est le paramètre d'échelle et μ est le paramètre de position. Dans la figure 1.2 nous allons visualiser l'effet de μ et de σ sur la densité de distribution de Fréchet.

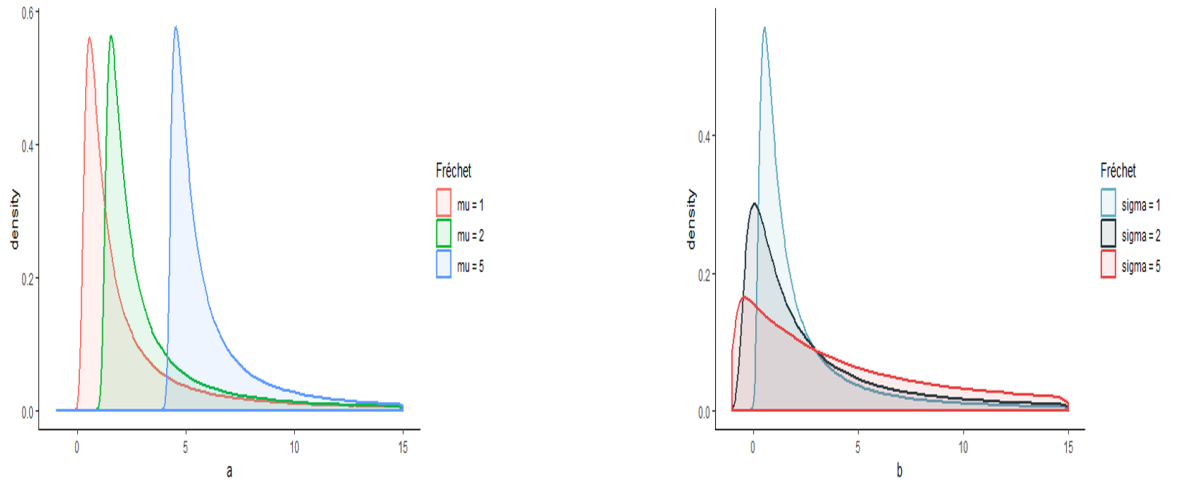


FIGURE 1.2 – Fonction de densité de Fréchet : (a) pour différentes valeurs de μ, (b) pour différentes valeurs de σ

Il est clair sur la figure 1.2(a) que la moyenne μ joue un rôle de translation. Par contre sur la figure 1.2(b) le paramètre d'echelle σ controle l'aplatissement de la distribution.

Remarque 3.2.1 : On constate suur la figure 1.2 que la fonction de densité de Fréchet est est asymétrique à droite et bornée à gauche. Cette asymétrie implique que la moyenne est supérieure à la médiane, ce qui est typique des distributions à longue queue droite.

3.2.2 Distribution de probabilité de Weibull

La distribution dite de Fréchet a été introduite par le mathématicien suédois Waloddi Weibull. De nos jour cette distribution est couramment utilisé dans l'analyse de survie et en fiabilité des systemes. La distribution de Weibull peut également être utilisée dans de nombreux domaine notamment la biologie, l'économie, l'hydrologie etc [28]

Définition 3.2.1 : On dit qu'une variable aléatoie X est de distribution de Weibull de paramètre $\mu, \sigma > 0$ et $\gamma < 0$ si sa densité de probabilité peut s'écrit telle que :

$$h_{\mu,\sigma,\gamma}(x) = \begin{cases} 0 & \text{si } x > \frac{1}{\gamma} \\ \frac{\gamma}{\sigma}\left(\frac{x-\mu}{\sigma}\right)^{\gamma-1} \exp\left(-\left(\frac{x-\mu}{\sigma}\right)^{\gamma}\right) & \text{si } x \leq \frac{1}{\gamma} \end{cases} \tag{2.4}$$

où γ est le paramètre de forme, σ le paramètre d'échelle et μ le paramètre de position. Dans la figure 1.3 nous allons visualiser l'effet de μ et de σ sur la densité de distribution de Weibull.

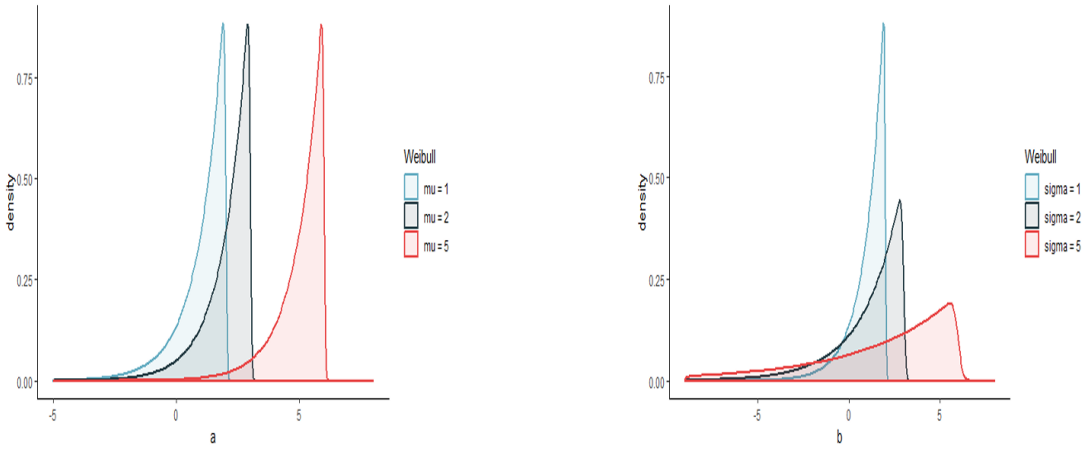


FIGURE 1.3 – Fonction de densité de Weibull : (a) pour différentes valeurs de μ, (b) pour différentes valeurs de σ

Il est clair sur la figure 1.3(a) que la moyenne μ joue un rôle de translation. Par contre sur la figure 1.3(b) le paramètre d'echelle σ controle l'aplatissement de la distribution.

Remarque 3.2.1 : On constate sur la figure 1.3 que la fonction de densité de Weibull est asymétrique à gauche et bornée à droite. Cette asymétrie implique que la moyenne est inférieure à la médiane, ce qui est typique des distributions à longue queue gauche.

3.2.3 Distribution de Gumbel

La distribution dite de Gumbel a été introduite par le mathématicien américain Emil Julius Gumbel. Il a développé cette distribution comme loi des valeurs extrêmes. La distribution de Gumbel est l'une des distributions populaires dans de nombreux domaines grâce à sa fléxibilité.

Définition 3.2.1 : On dit qu'une variable aléatoie X est de distribution de Fréchet de paramètre μ et $\sigma > 0$ si sa densité de probabilité peut s'écrit telle que :

$$f(x) = \exp\left(-\left(\frac{x-\mu}{\sigma}\right)\right) \cdot \left(\frac{x-\mu}{\sigma}\right)^{\gamma}$$

où σ est le paramètre d'échelle et μ le paramètre de position. Dans la figure 1.4 nous allons visualiser l'effet de μ et de σ sur la densité de distribution de Gumbel.

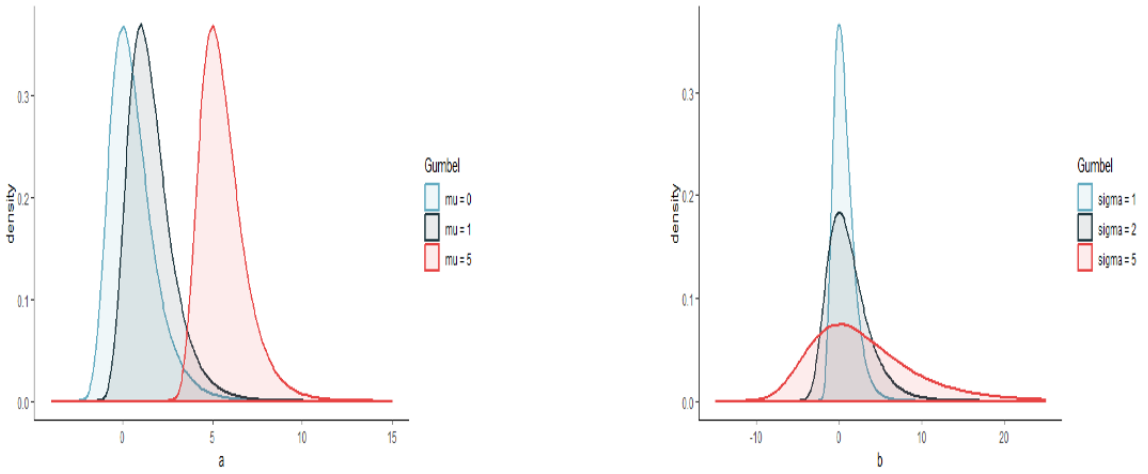


FIGURE 1.4 – Fonction de densité de Gumbel : (a) pour différentes valeurs de μ, (b) pour différentes valeurs de σ

Il est clair sur la figure 1.4(a) que la moyenne μ joue un rôle de translation. Par contre sur la figure 1.3(b) le paramètre d'echelle σ controle l'aplatissement de la distribution.

3.3 Modèles extrêmes utilisés

3.3.1 Loi généralisée des valeurs extrêmes (GEV)

La loi GEV (Generalized Extrem Value) est une famille de lois de probabilté qui modélise la distribution limte des maxima ou minima d'échantillon de variables aléatoires indépendantes et identiquement distribuées (i.i.d).Considérons $(X_1, \ldots, X_n)$ une suite de variable aléatoire i.i.d et de fonction de répartion F. Pour étudier le comportement extrême des événements, on considére la variable aléatoire $M_n = \max(X_1, \ldots, X_n)$, le maximum de l'échantillon de taille n telque :

$$\begin{aligned} F_{M_n}(x) &= P(M_n \leq x) \\ &= P(X_1 \leq x, \ldots, X_n \leq x) \\ &= P(X_1 \leq x) \cdots P(X_n \leq x) \\ &= (F(x))^n. \end{aligned}$$

On remarque que cette équation est dégénérée c'est à dire elle converge rapidement vers 0 ou 1. La fonction de répartion F est inconnue en pratique, ce qui limite son intérêt dans l'étude des queues de distribution. Alors Fisher et Tippet ont propose une solution a ce probleme en énoncant un théorème qui fut l'un des fondements de la théorie des valeurs extrêmes.

Théorème 3.4.1 : (Fisher-Tippett-Gnedenko)
S'il existe deux suites $a_n > 0$ et b_n et une loi non-dégénérée G telles que

$$\mathbb{P}\left\{\frac{M_{(n)} - b_n}{a_n} \leq x\right\} = F^n(a_n x + b_n) \to G(x)$$

alors G est nécessairement de loi $GEV(\mu, \sigma, \gamma)$

$$G_{\mu,\sigma,\gamma}(x) = \exp\left(-\left(1+\gamma\frac{x-\mu}{\sigma}\right)_+^{-1/\gamma}\right),\ x \in \mathbb{R}$$

- $\mu \in \mathbb{R}$ = paramètre de localisation
- $\sigma > 0$ = paramètre d'échelle
- $\gamma \in \mathbb{R}$ = paramètre de forme

Le théoreme de Fisher-Tippett-Gnedenko indique si le nombre d'observation extrême est élevé, la distribution des valeurs extrêmes converge vers la distribution généralisé des valeurs extrêmes (GEV). Si F verifie les hypothèses ci-dessus, on dit que F appartient au domaine d'attraction de G et on note $F \in DA(G)$. La loi $GEV(\mu, \sigma, \gamma)$ généralise trois ensembles de lois qui sont caractérisés par le paramètre de forme γ. Ce paramétre permet de contrôler la lourdeur de la queue de distribution et répartir les distributions limites en trois domaine d'attraction, **Fréchet, Weibull, Gumbel**.

Type I : Fréchet ($\alpha > 0$) : domaine des lois à queue lourde

$$\Phi_\alpha(x) = \begin{cases} 0 & \text{si } x \leq 0, \\ e^{-x^{-\alpha}} & \text{si } x > 0, \end{cases}$$

Type II : Weibull :($\alpha < 0$) : domaine des lois à queue bornée à droite

$$\Psi_\alpha(x) = \begin{cases} e^{-(-x)^{\alpha}} & \text{si } x < 0, \\ 1 & \text{si } x \geq 0. \end{cases}$$

Type III : Gumbel ($\alpha = 0$) : domaine des lois à queue légère

$$\Lambda(x) = e^{-e^{-x}}, \quad x \in \mathbb{R}.$$

L'expression la plus générale est la loi limite $G(x)$ des valeurs extrêmes développée par Gnedenko (1943), Von Mise (1954) et Jenkison (1955) est definie par :

$$G_{\gamma,\mu,\sigma}(x) = \begin{cases} \exp\left[-\left(1+\gamma\frac{x-\mu}{\sigma}\right)^{-1/\gamma}\right] & \text{si } \gamma \neq 0 \\ \exp\left[-\exp\left(-\frac{x-\mu}{\sigma}\right)\right] & \text{si } \gamma = 0 \end{cases}$$

3.4 Estimation du modèle

3.4.1 Maximum de Vraisemblance (MLE)

L'estimation par maximum de vraisemblance est une méthode paramétrique proposée par **Ronald Alymer Fisher** en 1912. C'est l'un des estimateurs les plus utilisés en Statistique. Elle consiste à maximiser la fonction de vraisemblance associée à une loi prédéfinie pour obtenir les estimateurs des paramètres. Pour un échantillon $x = \{x_1, \ldots, x_n\}$, où les x_i sont des réalisations indépendantes d'une variable aléatoire X ayant comme distribution la fonction $f_X(x)$ de paramètre θ (θ peut être un vecteur de paramètre) alors la fonction de vraisemblance pour cet échantillon est donnée par :

$$\mathcal{L}(\theta|x) = \prod_{i=1}^{n} f_X(x_i|\theta).$$

Cette fonction peut s'interpréter comme la probabilité d'obtenir l'échantillon x si X avait comme ditribution la fonction $f_X(x)$ de paramètre θ. Le but de l'estimation de maximum de

vraisemblance est de trouver les valeurs des paramètres $\hat{\theta}$ qui maximisent cette fonction de vraisemblance. c'est-à-dire $\hat{\theta} = \arg\max_{\theta} L(\theta \mid x)$.

En général, on travaille souvent avec la log-vraisemblance pour simplifier les calculs :

$$\ell(\theta \mid x) = \ln \mathcal{L}(\theta \mid x) = \sum_{i=1}^{n} \ln f_X(x_i \mid \theta) \tag{1.1}$$

— **Pour la loi GEV**

Dans le cas d'une distribution GEV, pour un échantillon donné, si $\gamma \neq 0$ la vraisemblance est définie par :

$$\mathcal{L}(\mu, \sigma, \gamma) = \prod_{i=1}^{n} \frac{1}{\sigma} \left(1 + \gamma \frac{x_i - \mu}{\sigma}\right)^{-1/\gamma - 1} \exp\left\{-\left(1 + \gamma \frac{x_i - \mu}{\sigma}\right)^{-1/\gamma}\right\} \tag{1.2}$$

Pour simplifier les calculs, nous utilisons la log-vraisemblance :

$$\ell(\mu, \sigma, \gamma) = -n \ln \sigma - \left(1 + \frac{1}{\gamma}\right) \sum_{l=1}^{n} \ln\left(1 + \gamma \frac{x_l - \mu}{\sigma}\right) - \sum_{l=1}^{n} \left(1 + \gamma \frac{x_l - \mu}{\sigma}\right)^{-1/\gamma} \tag{1.3}$$

La log-vraisemblance pour le type Gumbel ($\gamma = 0$) est donnée par :

$$\ell(\mu, \sigma) = -n \ln \sigma - \sum_{l=1}^{n} \left(\frac{x_l - \mu}{\sigma}\right) - \sum_{l=1}^{n} \exp\left\{-\left(\frac{x_i - \mu}{\sigma}\right)\right\} \tag{1.4}$$

Pour trouver les estimateurs des paramètres μ, σ et γ, nous maximisons la log-vraisemblance en résolvant les équations suivantes :

Pour $\gamma = 0$:

$$\frac{\partial \log \mathcal{L}(\mu, \sigma)}{\partial \mu} = 0 = n - \left[\sum \exp\left(-\left(\frac{x_i - \mu}{\sigma}\right)\right)\right] \tag{1.5}$$

$$\frac{\partial \log \mathcal{L}(\mu, \sigma)}{\partial \sigma} = 0 = n + \sum \frac{x_i - \mu}{\sigma} \left[\exp\left(-\left(\frac{x_i - \mu}{\sigma}\right)\right) - 1\right] \tag{1.6}$$

Pour $\gamma \neq 0$:

$$\frac{\partial \log \mathcal{L}(\mu, \sigma, \gamma; X)}{\partial \mu} = \left(\frac{1}{\gamma} + 1\right) \sum_{i=1}^{k} \frac{\gamma}{\sigma + \gamma(x_i - \mu)} \qquad -\frac{1}{\sigma} \sum_{i=1}^{k} \left(1 + \gamma \frac{x_i - \mu}{\sigma}\right)^{-\frac{1}{\gamma} - 1} = 0$$

$$\begin{aligned} \frac{\partial \log \mathcal{L}(\mu, \sigma, \gamma; X)}{\partial \sigma} = -\frac{k}{\sigma} \quad &+ \left(\frac{1}{\gamma} + 1\right) \sum_{i=1}^{k} \frac{\gamma(x_i - \mu)}{\sigma^2 + \gamma\sigma(x_i - \mu)} \\ &- \sum_{i=1}^{k} \left(\frac{x_i - \mu}{\sigma^2}\right) \left(1 + \gamma \frac{x_i - \mu}{\sigma}\right)^{-\frac{1}{\gamma} - 1} = 0 \end{aligned} \tag{1.7}$$

$$\frac{\partial \log \mathcal{L}(\mu,\sigma,\gamma;X)}{\partial\gamma} = -\left(1-\frac{1}{\gamma^2}\right)\sum_{i=1}^{k}\log\left(1+\gamma\frac{x_i-\mu}{\sigma}\right)$$
$$-\left(1+\frac{1}{\gamma}\right)\sum_{i=1}^{k}\frac{x_i-\mu}{\sigma+\gamma\sigma(x_i-\mu)} \quad +\sum_{i=1}^{k}\left(1+\gamma\frac{x_i-\mu}{\sigma}\right)^{-\frac{1}{\gamma}}$$
$$\times\left\{\frac{x_i-\mu}{\sigma\gamma+\sigma\gamma^2(x_i-\mu)} \quad -\frac{1}{\gamma^2}\log\left(1+\gamma\frac{x_i-\mu}{\sigma}\right)\right\}=0$$

La résolution de ces systèmes est relativement difficle, il n'admet pas généralement de solutions explicites, dans ce cas on fait appel à des méthodes d'optimisation numérique

3.4.2 Méthode des L-moments

Les moments, tels que la moyenne, la variance, l'asymétrie et la kurtosis, sont traditionnellement utilisés pour décrire les caractéristiques d'une distribution univariée. ***Hosking*** a introduit une approche alternative utilisant les L-moments qui sont des combinaisons linéaires des statistiques d'ordres. La méthode des estimateurs des L-moments est un ensemble de statistiques qui fournissent une méthode robuste pour résumer les caractéristiques d'une distribution de probabilité [10].

Définition 3.4.2 Soit X une variable aléatoire à valeur réelles avec fonction de distribution cumulative $F(x)$ et fonction quantile $x(F)$, et soit $X_{1,n} \leq X_{2,n} \leq \cdots \leq X_{n,n}$ les statistique d'ordre d'un échantillon aléatoire de taille n tiré de la distribution X. Le n-ième moment L définie par **Hosking** [10] est la quantité donnée par :

$$\lambda_r = r^{-1}\sum_{k=0}^{r-1}(-1)^k\binom{r-1}{k}\mathbb{E}\left[X_{r-k,r}\right] \quad r=1,2,\ldots$$

Le L dans le "L-moment" souligne que λ_r est une fonction linéaire des statistiques d'ordre attendu. L'espérance de la statistique d'ordre est donnée par :

$$E(X_{i,r}) = \frac{r!}{(i-1)!(r-i)!}\int_0^1 x(F)F^{i-1}(1-F)^{r-i}\,dF$$

Lorsqu'on compare les L-moments aux moments conventionnels, les L-moments présentent certains avantages, notamment leurs existences, leur unicité et leur robustesse (car ils sont des combinaisons linéaires). La formule ci-dessus peut être réécrire sous une forme qui est particulièrement utile pour les calculs des L-moments d'une distribution de probabilité donnée.

$$\lambda_r = \int_0^1 x(F)P^*_{r-1}(F)\,dF, \quad r=1,2,\ldots,$$

$$P^*_r(t) = \sum_{i=0}^{r}(-1)^{r-i}\binom{r}{i}\binom{r+i}{i}t^i$$

où :

— $x(F)$ est la fonction quantile
— F est la fonction de distribution cumulative
— $P_r^*(t)$ est le polynôme de legendre decalé d'ordre r

Les deux premiers L-moments λ_1 et λ_2, appelés L-localisation et L-échelle, étant des mesures de localisation et d'échelle, Les coefficients de variation τ et les troisième et quatrième ratios de L-moment τ_3 et τ_4, appelés L-asymétrie et L-kurtosis, étant des mesures d'asymétrie et de kurtosis :

$$\tau = \frac{\lambda_2}{\lambda_1}, \quad \tau_3 = \frac{\lambda_3}{\lambda_2}, \quad \tau_4 = \frac{\lambda_4}{\lambda_2}$$

3.4.3 Méthode des Moindres carrés (MMC)

La méthode des moindres carrés est une technique fondamentale en statistique pour ajuster un modèle à un ensemble de données, en minimisant la somme des carrés des écarts entre les valeurs observées et les valeurs prédites par le modèle.

Principe général de la MMC La méthode des moindres carrés consiste à estimer les paramètres d'un modèle $f(x, \theta)$ en minimisant la somme des carrés des écarts entre les observations y_i et les valeurs théoriques $f(x, \theta)$:

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} (y_i - f(x_i; \theta))^2$$

Dans le cas de la GEV, l'idée est d'estimer les paramètres (μ, σ, γ) en ajustant les quantiles empiriques de l'échantillon aux quantiles théoriques issue de la distribution GEV, c'est à dire :

$$Q^{\text{GEV}}(p; \mu, \sigma, \gamma) = \begin{cases} \mu + \frac{\sigma}{\gamma} \left[(-\log p)^{-\gamma} - 1 \right], & \gamma \neq 0 \\ \mu - \sigma \log(-\log p), & \gamma = 0 \end{cases}$$

On cherche alors à minimiser l'erreur quadratique :

$$\min_{\mu,\sigma,\gamma} \sum_{i=1}^{k} \left(Q_i^{\text{emp}} - Q^{\text{GEV}}(p_i; \mu, \sigma, \gamma) \right)^2$$

où Q_i^{emp} sont des quantiles empiriques associés aux probabilités p_i et Q^{GEV} est les quantiles de la GEV.

5. Machine Learning

5.1 Définition et types

Le machine learning (ML) ou l'apprentissage automatique représente aujourd'hui l'une des révolutions technologiques les plus marquantes du 21 ieme siècle. C'est une branche de l'intelligence artificielle (IA) qui consiste à donner vie aux données en utilisant des notions mathématiques pour créer des modèles d'IA capables d'apprendre de manière autonome. Ce terme fut prononcé en 1959 par **Arthur Samuel**, il définit le

Machine Learning comme **«un domaine d'étude qui donne aux ordinateurs la capacité d'apprendre sans être explicitement programmés».** [38] L'apprentissage automatique est la branche de l'informatique qui utilise l'expérience passée pour apprendre et utiliser ses connaissances afin de prendre des décisions futures. L'apprentissage automatique se situe à l'intersection de l'informatique, de l'ingénierie et des statistiques. L'objectif de l'apprentissage automatique est de généraliser un modèle détectable ou de créer une règle inconnue à partir d'exemples données. [41]

Dans le domaine du Machine learning, on distingue deux principaux techniques d'apprentissage : **supervisé** et **non-supervisé** qui se distingue principalement sur la présence ou l'absence de donnée étiquetées.

5.2 Les méthodes supervisées

L'apprentissage supervisé, également connu sous le nom de machine learning supervisé, se définit par l'utilisation de jeux de données étiquetées pour entraîner des algorithmes à classer des données ou à prédire des résultats. L'apprentissage supervisé est divisé en deux partie : la **classification** (prédiction de catégorie discrete) et de la **régression** (prédiction de valeur continue). Cette méthode repose sur la modélisation d'une relation fonctionnelle entre une variable X dite explicative et une variable Y dite dépendante, formalisé par

$$Y = f(X)$$

où f est une fonction paramétrique ou non paramétrique apprise à partie des données. L'objectif principal est d'estimer une fonction g, qui pour une nouvelle entrée x, fournit une sortie $g(x)$ proche de la valeur réelle y.

Dans les problèmes de régression, on cherche à prédire la valeur d'une variable continue, c'est-à-dire une variable qui peut prendre une infinité de valeurs.

Dans un problème de classification, on cherche à classer un objet dans différentes classes, c'est-à-dire que l'on cherche à prédire la valeur d'une variable discrète (qui ne prend qu'un nombre fini de valeurs).

4.2.1 Forêt aléatoire (RF) :

Le forêts aléatoire (Random Forest) représente l'une des méthodes de ML supervisé les plus robutse et polyvalentes, capable de résoudre les problémes de classification comme de régression. Developpé par **Leo Breiman** en 2001, Les forêts aléatoires sont une combinaison de prédicteurs d'arbres, de sorte que chaque arbre dépend des valeurs d'un vecteur aléatoire échantillonné de manière indépendante et avec la même distribution pour tous les arbres de la forêt dans le but de former un modèle à la fois précis et stable [37]. L'algorithme de Random Forest repose sur trois idées principales :

- **Le Boostrap (Bagging)** : À partir d'un échantillon initial de N observations $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}$, dont chacune est décrite au moyen de p variables prédictives, on crée « artificiellement » B nouveaux échantillons de même taille N par tirage avec remise. Chaque arbre est entraîné sur un de ces échantillon bootsrap, ce qui introduit de la diversité entre les arbres.
- **Sélection aléatoire des variables (max-feature)** : Parmi les p variables prédictives disponibles pour effectuer la segmentation associée au nœud d'un arbre, on n'en utilise qu'un nombre $m < p$ choisies « au hasard ». Celles-ci sont alors utilisées pour effectuer la meilleure segmentation possible. Cette randomisation supplémentaire réduit la corrélation entre les arbres et améliore la performance globale.

- **Aggrégation par vote (méthode ensembliste)** : L'algorithme combine plusieurs algorithmes « faibles », en l'occurrence les B arbres de décisions, pour en constituer un plus puissant en procédant par vote. Concrètement, lorsqu'il s'agit de classer une nouvelle observation **x**, on la fait passer par les B arbres et l'on sélectionne la classe majoritaire parmi les B prédictions (le mode), et en regression la moyenne des sortie des arbres

Cette approche permet de réduire la variance des prédicteurs par rapport en un arbre unique, tout en conservant une faible erreur de biais. De plus les forêts aléatoire sont réputés par leur robutesse au bruit et leur capacité à gérer des données de grandes dimensions

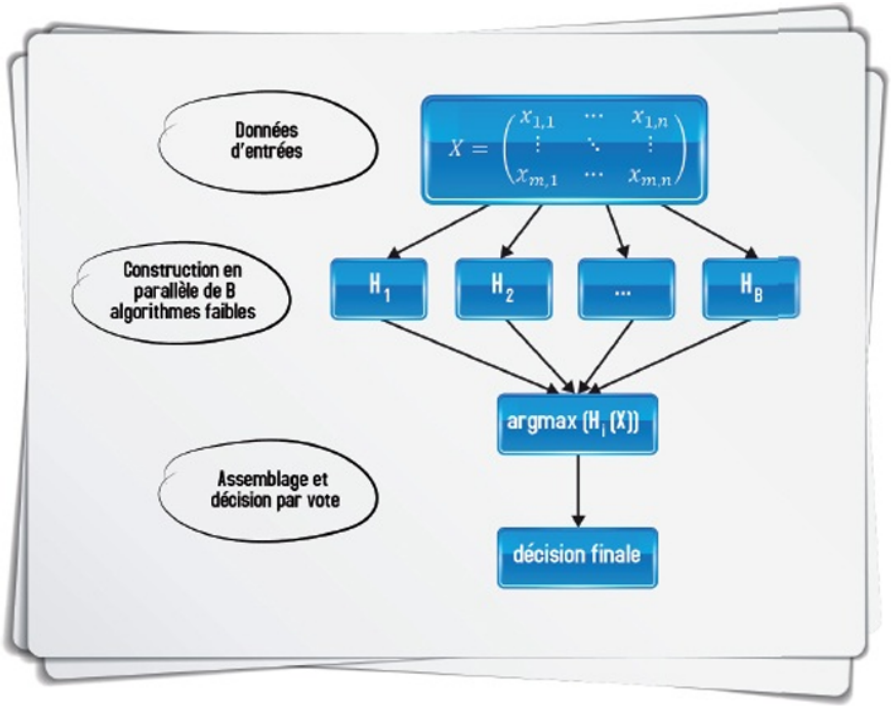


FIGURE 1.5 – Random forest : processus de construction et d'assemblage

4.2.2 Support Vector Machine (SVM) :

Support Vector Machine ou séparateur à vaste marge (SVM) est une technique d'apprentissage automatique non linéaire utilisée pour des tâches de classification et de régression. Cette méthode consiste à trouver un hyperplan optimal séparant les données en classe distinct en maximisant la marge entre elles. Cette méthode se faire en deux approche selon la linéarité des données. Pour un ensemble d'apprentissage

- Si les données sont linéairement séparable, le principe est de trouver un hyperplan linéaire $f(x)$ dont la forme est la suivante :

$$f(x) = \sum_{i=1}^{n} w_i x_i + b = \langle w, x \rangle + b$$

où w est le vecteur orthogonal à l'hyperplan et b est le déplacement par rapport à l'origine,$\langle ., . \rangle$ est le produit scalaire.si x_s est une vecteur de support(point le plus

proche de l'hyperplan) et $H = \{x \mid w^\top x + b = 0\}$ alors la marge est donnée par :

$$Marge = d(x_s, H) = \frac{2|w^T x_s + b|}{\|w\|}$$

.

L'objectif est de maximiser cette marge, ce qui revient à miniser $\|w\|$ sous la contrainte de normalisation $|w^\top x_s + b| = 1$. Alors on obtient le problème d'optimisation et la marge suivante

$$Marge = \frac{2}{\|w\|} et \begin{cases} \min\limits_{w,b} & \frac{1}{2}\|w\|^2 \\ \text{s.c.} & y_i(w \cdot x_i + b) \geq 1, \quad \forall i \in \{1, \ldots, n\} \end{cases}$$

- si les données sont non linéairement séparable, les SVM utilisent une fonction noyau φ (kernel) non linéaire pour projetter les données $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}$ de l'espace original à p dimension (p étant le nombre de variables prédictives) vers des données $\varphi(\mathbf{x}^{(1)}), \ldots, \varphi(\mathbf{x}^{(N)})$ dans un espace de dimension supérieur à p où ils seront « plus faciles à séparer », et dans lequel il sera alors possible de trouver une bande séparatrice linéaire. Le problème d'optimisation devient :

$$\begin{cases} \min\limits_{w,b} \frac{1}{2}||w||^2 + C\sum\limits_{i=1}^{n} \xi_i \\ \text{s.c.} \quad y_i(w \cdot \varphi(x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad \forall i \end{cases}$$

avec $\xi_i = \max\big(0, 1 - y_i(w^T\varphi(x_i) + b)\big)$.

— Si $y_i(w^T\varphi(x_i) + b) \geq 1$, alors $\xi_i = 0$. On retrouve le probléme linéairement séparable traité plus tôts

— Si $y_i(w^T\varphi(x_i) + b) < 1$, alors $\xi_i = 1 - y_i(w^T\varphi(x_i) + b) > 0$, alors le point est dans la marge mais du bon coté.

— Si $y_i(w^T\varphi(x_i) + b) < 0$, alors $\xi_i > 1$, alors le point est mal classé.

4.2.3 K-plus proches voisins (KNN) :

L'algorithme de K-plus proches voisins (KNN pour K Nearest Neighbors) est une méthode d'apprentissage supervisée non paramétrique, utilisée pour la classification en se basant sur la similarité avec les données d'apprentissage.

Soit un ensemble d'apprentissage $E = \{(y_i, \mathbf{x}_i)\}_{i=1}^n$ où y_i est la classe de l'observation i et $\mathbf{x}_i \in \mathbb{R}^p$ avec $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ est le vecteur de caractéristiques de dimension p.

Pour chaque nouvelle donnée $\mathbf{x}_u$, l'algorithme selon une métrique de distance d : $\mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_+$:

1. calcul les distances entre $\mathbf{x}_u$ et chaque $\mathbf{x}_i$ $d_i = d(\mathbf{x}_u, \mathbf{x}_i)$ (par exemple, la distance euclidienne $\|\mathbf{x}_u - \mathbf{x}_i\|_2$).
2. Sélectionner les k observations de E les plus proches de $\mathbf{x}_u$.
3. Attribuer à $\mathbf{x}_u$ le label majoritaire parmi les k voisins :

$$\hat{y} = \arg\max_{c \in \mathcal{C}} \sum_{i=1}^{k} \mathbb{I}(y_{(i)} = c),$$

où $\mathbb{I}$ est la fonction indicatrice et c est l'ensemble des classes possibles

4.2.4 Gradient Boosting :

Le Gradient Boosting est une technique d'apprentissage automatique supervisée particulierement puissante, utilisée pour résoudre des problèmes de régression et de classification. Il appartient à la famile des méthodes Boosting, qui construisent séquentiellement un modèle prédictif fiable en agrégeant plusieurs apprenants(modèle) «faible», c'est à dire des estimateurs légèrement meilleurs que le hasard [36]. Le principe est d'entraîner à chaque itération un nouvel apprenant faible (souvent un arbre de décision peu profond) qui se concentre sur la correction des erreurs (résidus) commises par l'ensemble des modèles précédents. Le Gradient Boosting utilise une approche itérative basée sur la descente de gradient pour minimiser une fonction de perte (loss function). Cette construction itérative permet d'obtenir un modèle global, appelé strong learner, avec une précision nettement améliorée. Cette méthode a été popularisée par l'algorithme AdaBoost (**Freund et Schapire, 1996)** et reste à la base de nombreuses méthodes modernes telles que XGBoost ou LightGBM (Fellous, 2019). ***La spécifité du Gradient Boosting introduite par*** Friedman (2002), est d'interpreter ce processus comme une descente de gradient dans un espace fonctionnel. Plus précisément, à chaque étape l'algorithme ajuste à nouveau le modèle dans la direction opposée au gradient de la fonction de perte, ce qui permet de minimiser efficacement l'erreur en espérance. Cette approche offre une grande flexibilité car elle peut s'appliquer à différente fonction de perte comme l'erreur quadratique pour la régression et la fonction logit pour la classification.

L'objectif central selon Friedman [34] est d'estimer $F^*(\mathbf{x})$ qui prédit une variable de sortie y à partir des variables d'entrée $\mathbf{x} = (x_1, \ldots, x_p)$, en minimisant l'espérance d'une **fonction de perte** sur la distribution conjointe des $(y, \mathbf{x})$. Formellement :

$$F^* = \arg\min_F \mathbb{E}_{y,\mathbf{x}}[L(y, F(\mathbf{x}))] = \arg\min_F \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_y(L(y, F(\mathbf{x}))) \mid \mathbf{x}\right] \tag{1.8}$$

Cette approche suppose une connaissance parfaite de la distribution des données, ce qui est rarement disponible en pratique. Dans ce cas, l'espérance théorique est remplacée par le risque empirique calculé sur l'échantillon d'entraînement :

$$R_n(F) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, F(\mathbf{x}_i)) \tag{1.9}$$

où $l(y, f(x))$ mesure l'erreur entre la prédiction $f(x)$ et l'observation y. La fonction de perte l doit être différentiable et convexe pour permettre l'application de la descente de gradient. Selon la nature du problème, différentes fonctions de perte sont utilisées :

— AdaBoost : pour la classification binaire, avec $l(y, f(x)) = \exp(-yf(x)), y \in \{-1, 1\}$
— LogitBoost : pour la classification binaire, avec $l(y, f(x)) = \log(1 + \exp(-yf(x)))$
— L2-Boosting : pour la régression, avec $l(y, f(x)) = (y - f(x))^2, y \in \mathbb{R}$

Algorithme de boosting L'algorithme produit une suite d'estimateurs de manière récursive :

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \lambda h_m(\mathbf{x}) \tag{1.10}$$

où :

— $h_m(\mathbf{x})$ est un nouvel estimateur faible ajusté sur les gradients négatifs (ou résidus) de la fonction de perte
— $\lambda \in [0, 1]$ est le taux d'apprentissage (paramètre de régularisation)

En pratique, la procédure est itérée de $m = 1$ jusqu'à une valeur M déterminée par validation croisée ou autre critère d'arrêt.

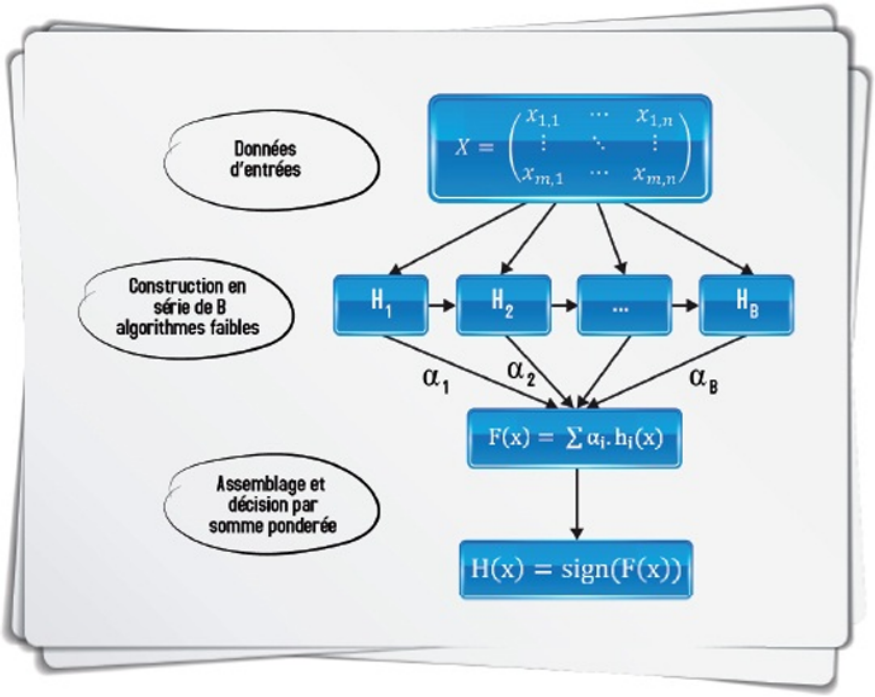


FIGURE 1.6 – Boosting :Processus de construction et d'assemblage

4.2.5 Réseau de neurones

Un réseau neuronal artificiel (ou simplement réseau de neurones) est un modèle prédictif qui reproduit le fonctionnement du cerveau humain. Ce dernier peut être vu comme un réseau complexe de neurones interconnectés, où chaque neurone reçoit des signaux d'entrée et effectue une somme pondérée de ces signaux, puis applique une fonction d'activation pour produire une sortie [29]. Cette approche cherche à reproduire la manière dont les neurones biologiques transmettent l'information dans le système nerveux. (voir figure 1.7)

L'une des premières formalisations des réseaux de neurones artificiels est proposée par **McCulloch et Pitts** dans les années 1940-1950. Ils ont modélisé ce fonctionnement en considérant les neurones comme une fonction de transfert f qui prend un certains nombres de signaux en entrée et renvoie une sortie binaire y

— le neurone reçoit plusieurs signaux d'entrée : $x_1, x_2, ..., x_n$.
— On calcule la somme pondérée $f = \sum_{i=1}^{n} w_i x_i$.
— Une fonction de seuil (step function) détermine la sortie :

$$y = \begin{cases} 1 & \text{si } f \geq 0 \\ 0 & \text{sinon} \end{cases}$$

Pour mieux visualiser ce principe, la figure 1.7 illustre la structure du perceptron, qui généralise le modèle originel de McCulloch et Pitts en intégrant la notion de biais et

une fonction d'activation. Cette représentation artificielle s'inspire du fonctionnement biologique du neurone :

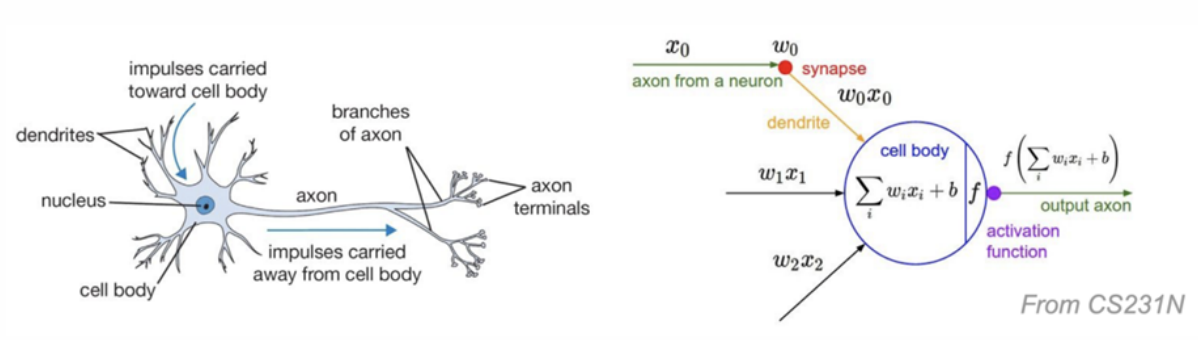


FIGURE 1.7 – Analogie entre neurone biologique et perceptron

La sortie du neurone est calculée comme suit :

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right)$$

Étant donné il existe plusieurs fonction d'activation dont nous allons citer quelques unes :

$$\text{Sigmoïde :} \quad \sigma(x) = \frac{1}{1+e^{-x}} \tag{1.35}$$

$$\text{Softmax :} \quad \sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad \text{avec } K \text{ classes} \tag{1.36}$$

$$\text{Tanh :} \quad \tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{1.37}$$

$$\text{ReLU :} \quad \text{ReLU}(x) = \max(0, x) \tag{1.38}$$

La capacité d'apprentissage automatique a été introduite par **Frank Rosenblatt** dans les années 1950, avec l'algorithme du perceptron. Celui-ci ajuste les poids en fonction de l'erreur observée entre la sortie attendue et la sortie prédite, selon une règle supervisée inspirée de la règle de Hebb.

Fonctionnement du perceptron simple

- **Structure et calcul** : Les perceptrons reçoivent plusieurs entrée $x_1, x_2, ..., x_n$ chacun associé à un poids $w_1, w_2, ..., w_n$, ainsi qu'un biais b. La sortie intermédiaire (somme pondérée) est calculée ainsi :

$$z = \sum_{i=1}^{n} w_i x_i + b$$

Cette valeur z est ensuite passée à travers une fonction d'activation généralement une foncton de seuil. Ainsi,le perceptron prend une décision binaire selon quel la somme pondérée des entrées dépasse ou non le seuil fixé par le biais.

- **Apprentissage et ajustement des poids** : L'apprentissage du perceptron repose sur l'ajustement des poids et du biais afin de minimiser les erreurs de prédictions sur un ensemble d'entrainement, À chaque itération pour chaque exemple (X, y_{true}), le perceptron calcul la sortie prédite y et les points sont mis à jour selon la règle :

$$W = W + \alpha(y_{true} - \hat{y})X$$

où α est le taux d'apprentissage, y_{true} la sortie attendue, $\hat{y}$ la sortie prédite et X le vecteur d'entrée.

- **Fonction de coût et optimisation** : Pour évaluer la performance du perceptron, on utilise une fonction de coût, comme la fonction de log-loss pour la classification binaire :

$$L = -\frac{1}{m}\sum_{i=1}^{m}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

L'optimisation des paramètres (poids et bias) ce fait par descente de gradient, en calculant les dérrivés partielles de la fonction coût par rapport à chaque paramètre

5.3 Les méthodes non supervisées

L'apprentissage non supervisé, également connu sous le nom de machine learning non supervisé, utilise des algorithmes de machine learning pour analyser et regrouper des données non étiquetées. Dans l'apprentissage non supervisé, les algorithmes apprennent par eux-mêmes sans aucune supervision ni aucune variable cible fournie. Il s'agit de trouver des structures et des relations cachées dans les données données (clusters, anomalie).Contrairemenet au méthode supervisé, il ne nécessite pas de sortie prédefinie (Y). L'apprentissage non supervisé est divisé en deux partie : le regroupement(ou clustering) et l'association. Les catégories de l'apprentissage non supervisé sont les suivantes : Réduction de dimensionnalité, Clustering.
Cette partie ne sera pas développée dans ce memoire.

Chapitre 2

MÉTHODOLOGIE ET MODÉLISATION

1. Introduction

Ce chapitre présente la démarche méthodologique adaptée ainsi que les différentes étapes de modélisation mise en oeuvre dans le cadre de cette étude. Il constitue une phase essentielle car il permet de structure de manière rigoureuse l'approche analytique allant de la compréhension des données jusqu'a l'interprétation des résultats issus de la modélisation.
La première étape de cette démarche consiste en une description précise du jeu de donnée utilisé, incluant la nature des variables et leur distribution. Des visualisations graphiques et des analyses descriptives seront faites pour mettre en évidence les insights pertinents. Cette phase est suivie d'un traitement préalable des données afin de préparer les données pour les étapes de modélisation.

Une analyse exploiratoire approfondie est ensuite menée pour identifier les tendances, les relations entre variables et dégager les premiers indices quant aux facteurs associés au risque d'AVC. La suite du chapitre est consacrée à la modélisation qui repose sur une double approche : La théorie des valeurs êtremes et les techniques de machine learning.

2. Données et prétraitements

2.1. Source et description

Dans le cadre de ce mémoire, l'analyse et la modélisation prédictive de l'accident vasculaire cérébral (AVC) reposent sur l'utilisation d'un jeu de donnée ouvert, spécifiquement conçu pour la recherche et l'apprentissage automatique. Ce jeu de données a été publié en 2021 sur la plateforme Kaggle. Il rassemble des informations démographiques et cliniques telles que le sexe, l'âge, la présence d'hypertention ou de maladies cardiaques, le statut tabagisque ainsi que d'autres variables pertinentes pour l'évaluation du risque d'AVC (référence officielle [42]).

La base de donnée utilisée comprend **5110 patients**, chacun décrit selon **11 caracteristiques** dont 3 sont de type numérique et 8 de type catégoriel. La variable cible de notre étude est la variable binaire **AVC** indiquant si un individu a été victime ou non d'AVC . La répartition de cette variable dans l'échantillon est désequilibrée avec environ 4.87% de patient ayant subi un AVC et 95.13% n'ayant pas eu . Les tableaux

suivants donnent les descriptions detaillées des données.

TABLE 2.1 – Caractéristique des patients et leur définition

Variable	Définition
Âge	Âge du patient
Sexe	Sexe du patient (femme, homme, autre)
Hypertension	Augmentation anormale de la pression artérielle sur les parois des artères
Maladie Cardiaque	Insuffisance coronarienne
Situation Matrimoniale	Statut marital du patient (marié / non marié)
Type travail	Profession exercée par le patient
Résidence	Milieu de résidence du patient (urbain ou rural)
Taux moyen de glucose	Taux moyen de glucose sanguin chez le patient
IMC	Indice de masse corporelle du patient
Statut Fumeur	Statut tabagique du patient (jamais fumé, ancien fumeur, fumeur actuel)
AVC	Patient souffrant d'accident vasculaire cérébral (oui / non)

TABLE 2.2 – Répartition par type de travail

Type	**Effectif**	**Fréquence (%)**
Privé	2925	57.24
Indépendant	819	16.03
Enfants	687	13.44
Emploi public	657	12.86
Jamais travaillé	22	0.43

TABLE 2.3 – Répartition par statut tabagisque

Statut Fumeur	**Effectif**	**Fréquence (%)**
Jamais fumé	3436	67.24
Ancien fumeur	885	17.32
Fumeur actuel	789	15.44

TABLE 2.4 – Répartition par lieu de résidence

Résidence	**Effectif**	**Fréquence (%)**
Urban	2596	50.8
Rural	2514	49.2

TABLE 2.5 – Répartition des AVC

AVC	**Effectif**	**Fréquence (%)**
0 (Non)	4861	95.13
1 (Oui)	249	4.87

TABLE 2.6 – Répartition Hypertension

Hypertension	**Effectif**	**Fréquence (%)**
Non	4612	90.25
Oui	498	9.75

TABLE 2.7 – Répartition Maladie Cardiaque

Maladie Cardiaque	**Effectif**	**Fréquence (%)**
Non	4834	94.6
Oui	276	5.4

TABLE 2.8 – Répartition des comorbidités

TABLE 2.9 – Répartition démographique

Âge	$0-45$ ans	$46-60$ ans	$61-82$ ans
Effectif	1570	2236	1304
Fréquence (%)	30.72	43.76	25.52

Sexe	Femme	Homme	Autre
Effectif	2994	2115	1
Fréquence (%)	58.59	41.39	0.02

Situation Matrimoniale	Non marié	Marié
Effectif	1757	3353
Fréquence (%)	34.38	65.62

2.2. Analyse exploratoire des données

L'Analyse exploratoire des données (AED) est une étape fondamentale du processus d'analyse statistique. Elle permet de mieux comprendre la structure et les caractéristiques des données avant la modélisation en identifiant les relations, les anomalies et les tendances cachées sur les données. Dans cette partie nous explorons les données en s'appuyant sur les méthodes descriptives et graphiques.

L'Analyse exploiratoire des données vise principalement à résumer les principales caractèristiques d'un jeu de données, souvent à l'aide des mesures statistiques (moyenne, médiane, écart-type, etc). Ce travail d'exploration permet de détecter les erreurs, incohérences ou valeurs abberantes suscesptibles d'influencer les analyses, ensuite identifier la distribution des variables et repérer les éventuelles particularités dans les données et enfin examiner les liens potentiels entre les variables.

2.2.1 Analyse decriptive univariée

Dans le cadre de ce mémoire, l'Analyse exploratoire des données a été appliquée aux variables numériques principales du jeu de données telles que l'**âge**, le **taux de glucose moyen** et l'**indice de masse corporelle** (IMC). Comme illustré dans la figure 2.1, des outils graphiques tels que les histogrammes et les boxplots ont été utilisés pour visualiser la distribution de ces variables. Cette premiere analyse descriptive permet de dire que seule la variable IMC suit une distribution normale avec une confirmation de test de shapiro-wilk, de détecter la présence des valeurs extrêmes (voir figure 2.1).

Cette méthode a été appliquée aussi sur les variables catégorielles pour permettre de visualiser la répartition des patients selon les caractéristiques tels que le **genre**, **l'hypertension**, la **maladie cardiaque**, la **situation matrimoniale** le **type de travail**, la **résidence** et le **statut de fumeur**. La figure 2.2 présente la distribution de ces variables catégorielles et elle montre des désequilibres entre les groupes. Par exemple **95.1**% des patients n'ont pas présenté d'accident vasculaire cérébral tandis que **4.9**% sont atteints. De plus **90,3**% des patients ne souffrent pas d'hypertension et **94,6%** des cas étudiés présentent une maladie cardiaque. Les femmes sont beaucoup plus représentées que les hommes, de même que les mariés et les personnes qui n'ont jamais fumé. (voir figure 2.2)

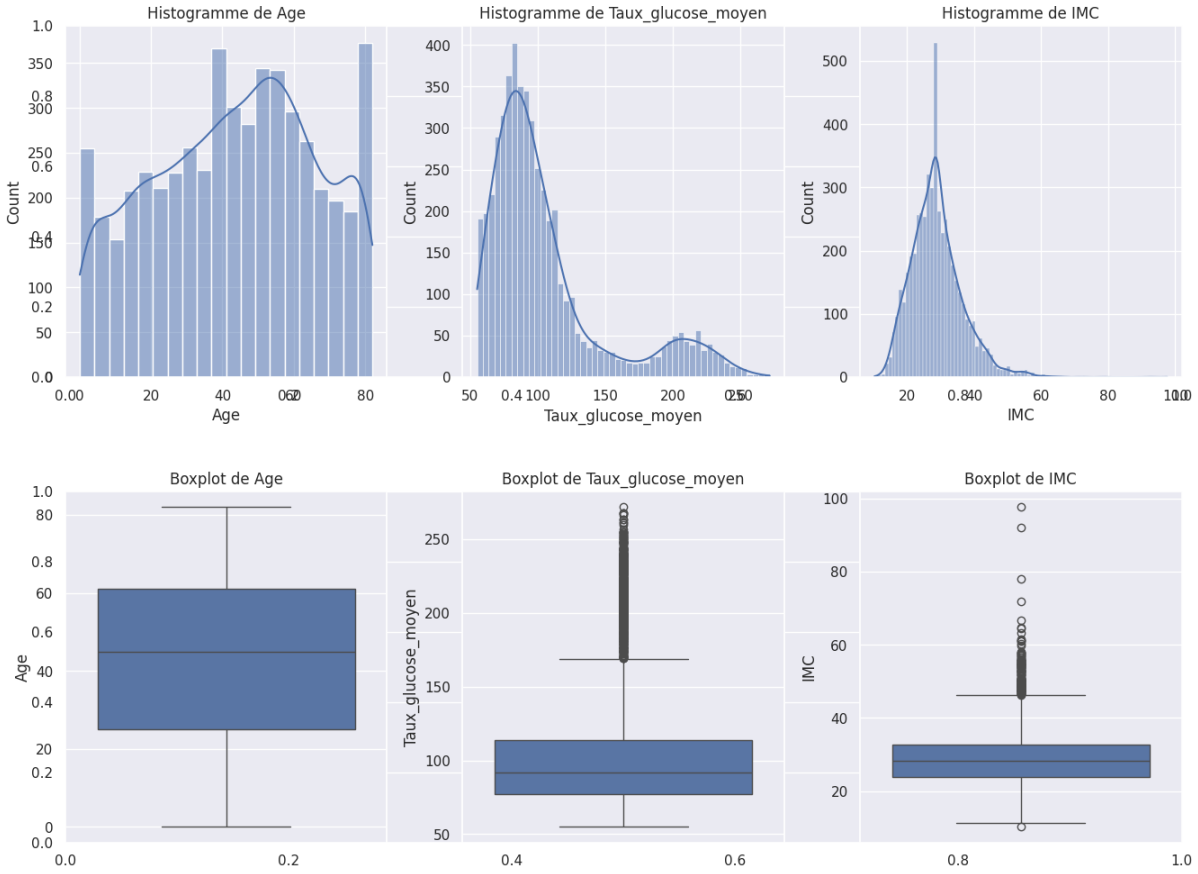


FIGURE 2.1 – Distribution des variables numériques

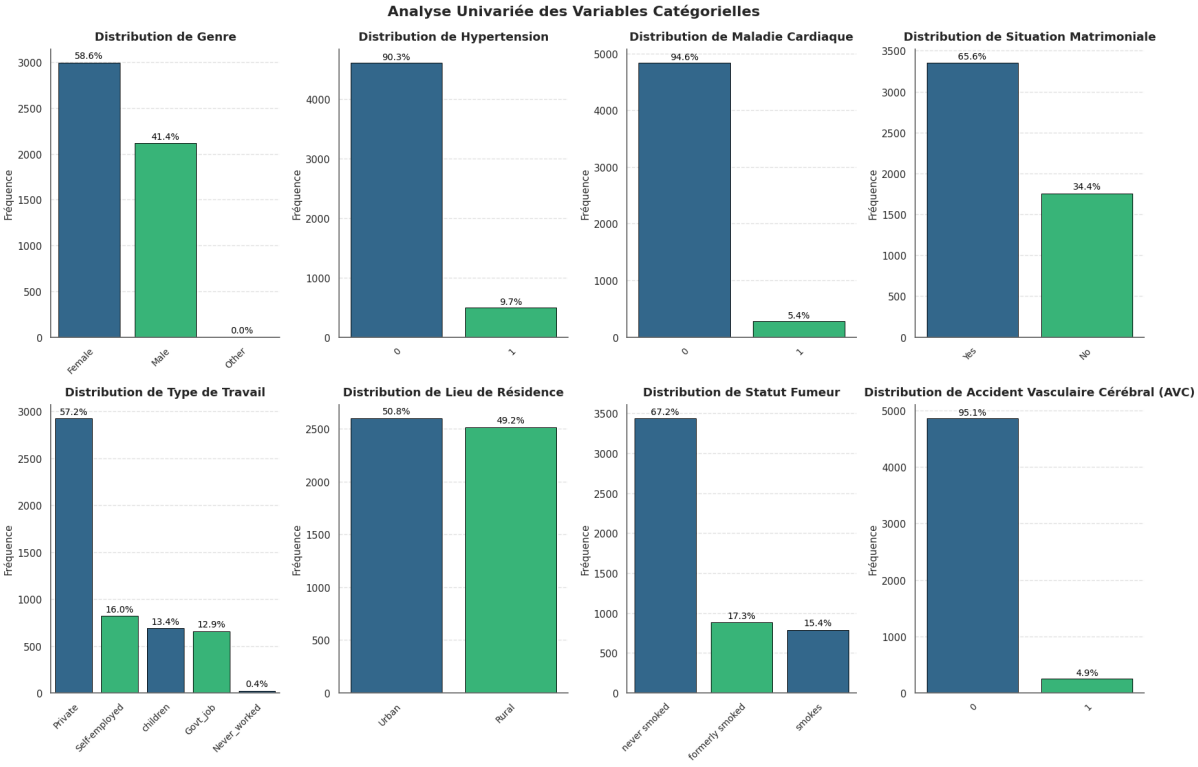


FIGURE 2.2 – Distribution des variables catégorielles

2.2.1 Analyse descriptive bivariée

L'analyse bivarée vise à etudier les relations entre deux variables afin de comprendre leurs interactions et leurs influences sur la variable cible.
L'examen des relations entre les variables ne révéle aucune une relation linéaire significative et que toutes les corrélations sont positives et statistiquement significatives avec une p-value inférieure à **0,05** (confirmé par un test de spearman). Les corrélations entre les variables restent faibles, seuls l'âge et l'IMC montrent une corrélation modérée.
Des tests de chi carré ont été réalisés pour évaluer les associations entre les variables catégorielles et la variable cible. Les résultats montrent que toutes les variables étudiées présentent une relation statistiquement significative avec l'AVC, à l'exception du **genre** et de la **résidence** pour lesquels aucune relation n'a été détectée.
L'analyse des résultats montre aussi que l'accident vasculaire cérebral est proportionnellement plus fréquent chez les femmes, les patients souffrant d'hypertension, ceux ayant une maladie cardiaque, les personnes mariées, les anciens fumeurs et fumeurs actuels et dans une moindre mesure, en milieu urbain. Ces résultats mettent en évidence certains profils à risque au sein de la population étudiée.
Par ailleurs, l'étude des variables quantitatives montrent que les patients ayant présenté un AVC sont en moyenne plus âgés et présentant un taux de glucose moyen plus élevé que ceux n'ayant pas eu d'AVC. De plus leur indice de masse corporelle (IMC) est légèrement supérieur à celui des autres patients. Ces différences suggèrent que l'âge avancé, l'hyperglycémie et IMC élevé pourraient constituer des facteurs de risque pour la survenue de l'AVC. Les graphes ci dessous montrent les resultats obtenus sur l'analyse bivariée.

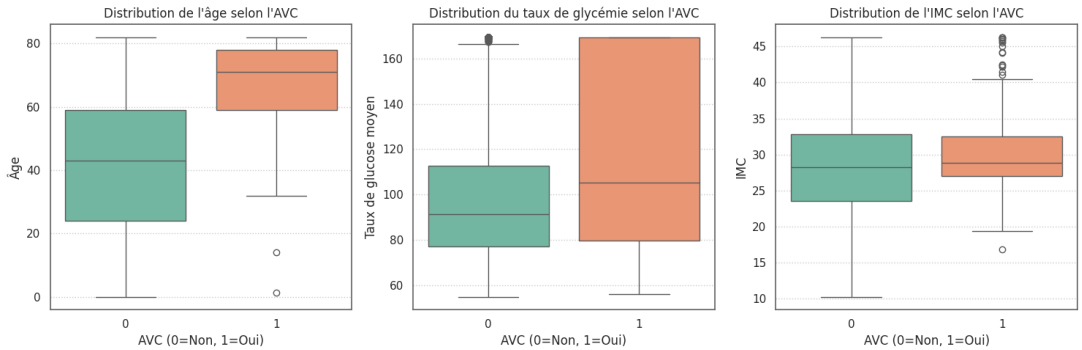


FIGURE 2.3 – Répartition des cas AVC selon les variables numerique

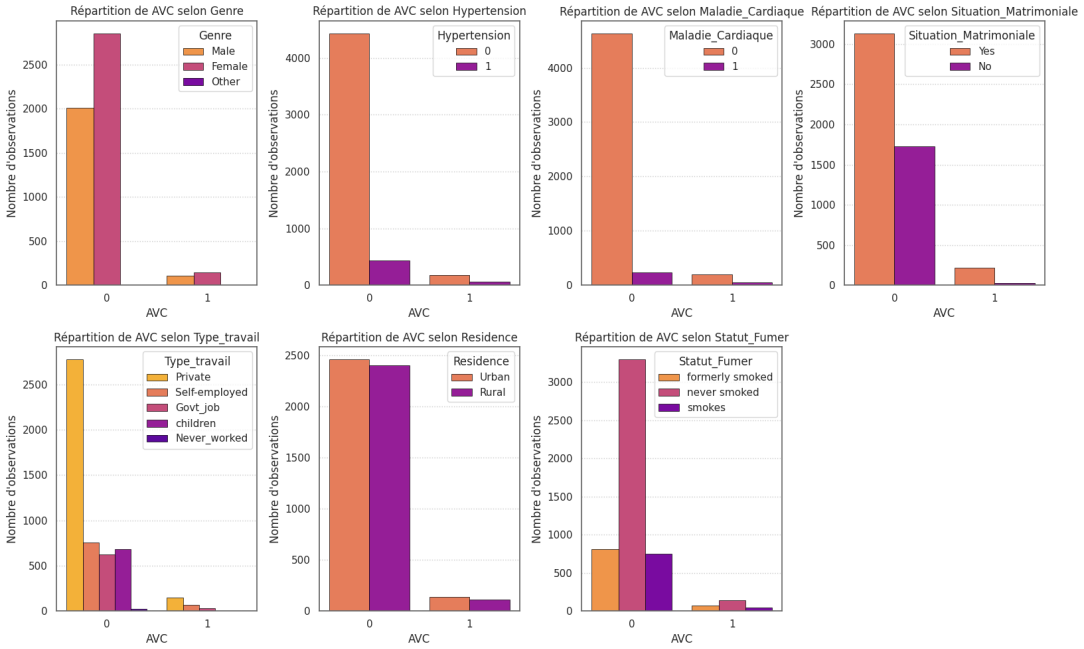


FIGURE 2.4 – Répartition des cas AVC selon les variables catégorielle

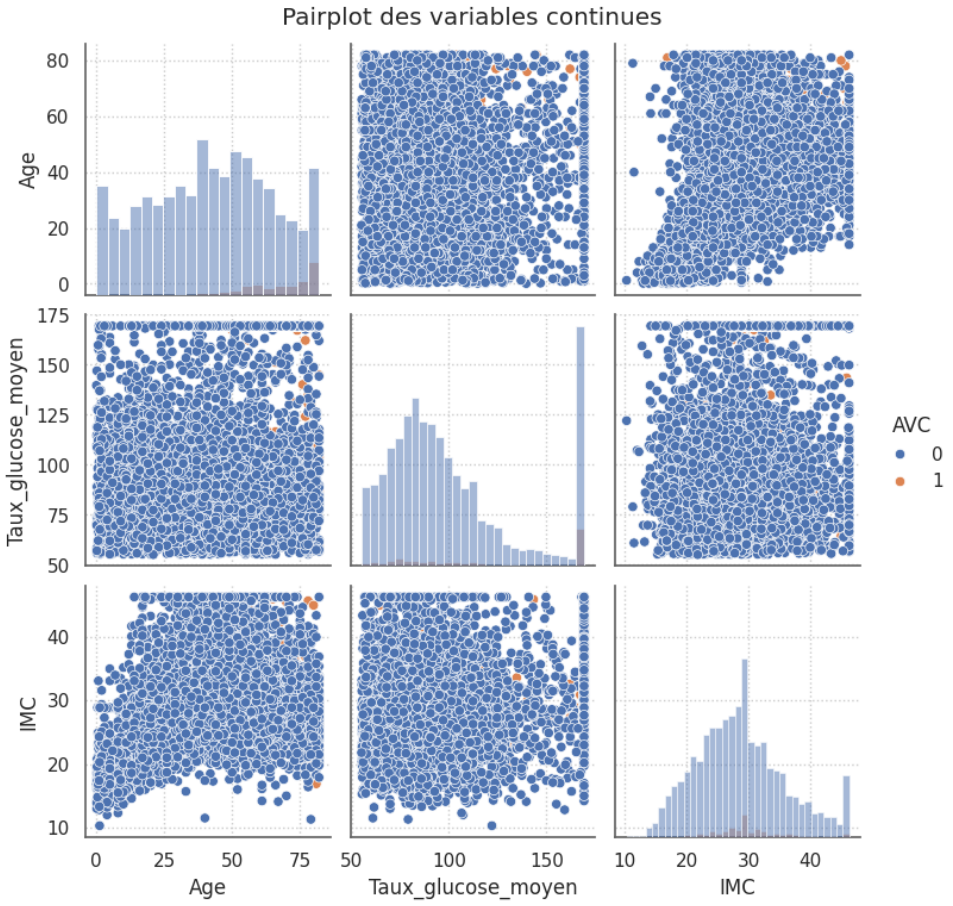


FIGURE 2.5 – Nuage de points des variables

2.3 Préparation des données

Cette partie constitue une étape importante dans tous projets d'analyse ou de modélisation. Elle englobe plusieurs processus dont le nettoyage des données, le traitement des valeurs manquantes, la détection des anomalies et la tranformation et structuration des données afin de les rendre exploitables. Dans le cadre de ce memoire nous concentrons sur les méthodes techniques.

2.3.1 Nettoyage et traitement des valeurs manquantes

La gestion des valeurs manquantes est une étape essentielle pour assurer la qualité des données. Dans cette étude, différentes méthodes ont été appliquées en fonction de la nature des variables concernées :

- **IMC :** Les valeurs manquantes ont été imputée par la **moyenne**, une méthode justifiée par la distribution symétrique de cette variable (*Moyenne = Médiane = Mode*) et une dispersion faible des données.
- **Statut fumeur :** Les entrées marquées par "***Unknow***" ont d'abord été remplacer par ***NAN*** puis imputées par le mode (never smoke), étant donné une modalité de cette variable.

2.3.3 Traitement des valeurs aberrantes

Pour détecter et corriger les valeurs aberrantes la méthode des bornes (IQR) a été appliquée aux variables numériques. La procédure adaptée est la suivante :

1. Calcul des quartiles (Q_1, Q_3) et de l'intervalle interquartile (IQR).
2. Définition des seuils :
   - Toute valeur inférieure à $Q_1 - 1,5 \times IQR$ est remplacée par Q_1.
   - Toute valeur supérieure à $Q_3 + 1,5 \times IQR$ est remplacée par Q_3.

2.3.4 Normalisation et transformation des variables

Pour garantir que les variables numériques soient sur une même échelle et améliorer la performance des algorithmes sans que certaines variables n'influencent pas les modèles, une normalisation par standarisation a été appliquée sur les données. Cette méthode consiste à centrer et réduire les données en transformant chaque variable pour qu'il ait de moyenne 0 et un ecart-type 1. La transformation est donnée par la formule suivante :

$$X_{\text{norm}} = \frac{X - \mu}{\sigma} \tag{2.1}$$

où :

- X est la valeur originale de la variable
- μ est la moyenne de la variable
- σ est l'écart-type de la variable

2.3.5 Equilibrage des données

Dans le jeu de données la répartition de patient ayant eu AVC est 4.87% contre 95.13%, cette déséquilibre pourrait affecter les résultats en favorisant la classe majortaire. Pour résoudre ce problème de déséquilibre, plusieurs méthodes de réechantillonnage sont courament utilisées afin d'équilibre la base :

— **Suréchantillonnage (Oversampling)** : Cette technique consiste à augmenter la classe minoritaire en créant des échantillons synthétiques. Parmi les méthodes nous avons apliquée sur nos données le **SMOTE (Synthetic Minitory Over-sampling Technique)**, qui se base sur une sélection pour chaque observation x dans la classe minoritaire, de ses plus proches voisins x_{zi} dans le même classe et crée de nouveaux points par interpolations linéaire entre x et ses voisins. La génération d'échantillons synthétiques dans SMOTE est donnée par la formule suivante :

$$x_{\text{new}} = x + \delta \times (x_{zi} - x) \tag{2.2}$$

où :
— x_i est un échantillon de la classe minoritaire
— x_{zi} est son plus proche voisin (k-plus proche voisins)
— $\delta \in [0, 1]$ est un coefficient aléatoire

— **Sous-échantillonnage (Undersampling)** : Cette approche réduit le nombre d'exemple de la classe majoritaire en supprimant certaines d'entre eux, ce qui permet d'équilibrer les classes mais entrainant une perte d'information.

2.4 Sélection des variables

La sélection des variables est une étape déterminante dans le processus de la modélisation. Elle a été réalisée afin d'identifier les facteurs les plus pertinents pour prédire le risque d'AVC. Pour cela une approche combinée a été adaptée. D'une part, nous nous somme appuyés sur la littérature médicale, en retenant des facteurs de risque reconnus. D'autre part, des analyses statistiques ont été menées : test du Chi 2, corrélation de spearman, test de comparaison de moyenne ainsi qu'une vérification de multicolinéarité. Cette démarche a été adaptée dans ce contexte pour améliorer la performance des modèles, faciliter leur interprétation et minimiser le bruit introduit par des variables non significatives.

3. Modélisation à l'aide des méthodes traditionnelles

3.1. Modèle de régression logistique

La régression logistique a été utilisée pour la prédiction du risque d'AVC en raison de sa pertinence pour les problèmes de classification binaire. Dans notre cas, l'objectif est d'identifier les risques et de modéliser la probabilité de survenue d'un accident vasculaire cérébral (AVC) en estimant directement **P(AVC = 1 | X)** à travers une fonction logistique, qui transforme une combinaison linéaire des variables explicatives en une probabilité comprise entre 0 et 1. Nous avons réalisé une analyse de régression logistique du modèle défini comme suit :

$$\begin{aligned}\log\left(\frac{P(\textbf{AVC}=1)}{1-P(\textbf{AVC}=1)}\right) = \beta_0 &+ \beta_1 \times \textbf{Genre} + \beta_2 \times \textbf{Âge} + \beta_3 \times \textbf{Hypertension} \\ &+ \beta_4 \times \textbf{Maladie\_Cardiaque} + \beta_5 \times \textbf{Situation\_Matrimoniale} \\ &+ \beta_6 \times \textbf{Type\_travail} + \beta_7 \times \textbf{Statut\_Fumer} \\ &+ \beta_8 \times \textbf{Taux\_glucose\_moyen} + \beta_9 \times \textbf{IMC} + \beta_{10} \times \textbf{Résidence}\end{aligned} \tag{2.3}$$

Nous utilisons la méthode de sélection de variables par étapes inversées (***Backward*** en utilisant le test de Wald au niveau de 10) pour choisir les covariables significatives à conserver dans le

modèle final. Les résultats finaux de cette procédure d'ajustement sont présentés dans le tableau suivant .

TABLE 2.10 – Coefficients du modèle de régression logistique pour la prédiction d'AVC

Variable	**Coefficient**	**Erreur Std.**	**p-value**
(Intercept)	-7.686	0.381	< 2e-16 ***
Âge	0.069	0.005	< 2e-16 ***
Hypertension	0.380	0.163	0.019 *
Maladie cardiaque	0.336	0.187	0.073 .
Taux glucose moyen	0.006	0.002	0.00047 ***

Dans une seconde approche, l'âge a été transformé en variable catégorielle à trois modalités, afin de capturer la non-linéarité de son effet sur le risque d'AVC. Les classes sont définies comme suit :

- Classe 1 : [25–45[ans
- Classe 2 : [45–65[ans
- Classe 3 : [65–82] ans

Le modèle logistique suivant a été réajusté et les résultats finaux de cette procédure d'ajustement sont présentés dans le tableau suivant

TABLE 2.11 – Coefficients du modèle de régression logistique avec classes d'âge

Variable	**Coefficient**	**Erreur Std.**	**Odds Ratio**	**p-value**
Intercept	-5.399	0.530	0.0045	< 2e-16 ***
Âge (classe 2)	1.745	0.326	5.72	8.31e-08 ***
Âge (classe 3)	2.860	0.352	17.46	< 2e-16 ***
Hypertension	0.416	0.161	1.52	0.009890 **
Maladie Cardiaque	0.416	0.192	1.52	0.025 *
Taux Glucose Moyen	0.006	0.002	1.01	0.00055 ***

La probabilité ajustée du pronostic vital pour un individu est donnée par :

$$P(\text{AVC} = 1) = \frac{1}{1 + \exp\left(-S(X)\right)} \tag{2.4}$$

avec

$$\begin{aligned} S(X) = &-5.399 + 1.745 \times \mathbb{I}_{[\textbf{Âge} \in \textbf{classe 2}]} + 2.860 \times \mathbb{I}_{[\textbf{Âge} \in \textbf{classe 3}]} \\ &+ 0.416 \times \textbf{Hypertension} + 0.416 \times \textbf{Maladie\_Cardiaque} \\ &+ 0.006 \times \textbf{Taux\_glucose\_moyen} \end{aligned} \tag{2.5}$$

3.1. Modèle GEV (Generalized Extreme Value)

La loi GEV est particulièrement adaptée pour les évenements extrêmes telsque les risques élevés d'AVC. Elle permet de prendre en compte la queue de distribution des risques élevés,

ce qui est crucial pour l'identification des individus à risque dans une population donnée. Elle généralise trois types de distributions (***Fréchet, Weibull*** et ***Gumbul***) et est définie par ses paramètres de localisation (μ), d'échelle(σ) et de forme(γ). Pour une variable de réponse binaire Y_i et le vecteur des variables explicatives X_i, notons $P(Y_i = 1 \mid X_i = x)$ la probabilité conditionnelle d'infection. Puisque nous considérons la classe des modèles linéaires généralisés, nous suggérons la fonction de distribution cumulative GEV proposée par **Calabrese** et **Osmetti (2013)** comme la courbe de réponse donnée par :

$$\begin{aligned} P(Y_i = 1 \mid \mathbf{X_i} = \mathbf{x}) &= 1 - \exp\left[-\left(1 + \tau\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\right)\right)^{-1/\tau}\right] \\ &= 1 - GEV(-\beta x_i; \tau) \end{aligned} \tag{2.6}$$

où $\beta = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ est un paramètre de regréssion inconnu mesurant l'association entre les prédicteur potentiels et le risque d'infection et GEV(x ; τ) représente la probabilité cumulée pour la distribution GEV avec paramètre de localisation = 0, un paramètre d'échelle = 1 et un paramètre de forme inconnu

Nous utilisons une regréssion GEV après avoir adapter une procédure de sélection backward pour identifier les variables significatives. Les résultats finaux de cette procédure d'ajustement sont présentés dans le tableau suivant :

TABLE 2.12 – Coefficients du modèle BGEVA pour la prédiction d'AVC

Variable	**Estimation**	**Erreur Std.**	**Odds Ratio**	**p-value**
intercept	-3.1188	0.1419	0.044	< 2e-16 ***
Âge	0.0285	0.0022	1.029	< 2e-16 ***
Taux glucose moyen	0.0019	0.0006	1.002	0.000981 ***
Hypertension	0.2025	0.0812	1.224	0.01265 *
Maladie cardiaque	0.1951	0.0988	1.215	0.04845 *

Le modèle final pour estimer la probabilité de survenue d'un AVC en fonction des variables explicatives s'écrit :

$$\begin{aligned} \mathbb{P}(\textbf{AVC} = 1 \mid \mathbf{X}) = 1 - GEV\Big(- \big(- 3.12 + 0.029 \times \textbf{Âge} + 0.0019 \times \textbf{Taux\_glucose\_moyen} \\ + 0.195 \times \textbf{Maladie\_cardiaque} + 0.203 \times \textbf{Hypertension}\big);\, \tau\Big) \end{aligned} \tag{2.7}$$

Quantile extrême Les quantiles extrêmes correspondent aux valeurs trés hautes (ou très basses) d'une distribution, souvent située dans la queue de la distribution. Leur estimation est cruciale en analyse des valeurs extrêmes, notamment pour évaluer des risques rares. Pour chaque variable nous calculons le quantile extrême ascocié à partir des paramètres estimés par la loi généralisée des valeurs extrêmes. Sa formule mathématique d'ordre p est estimée par :

$$x_p = \begin{cases} \mu - \frac{\sigma}{\gamma}\left[1 - (-\log(1-p))^{-\gamma}\right], & \text{si } \gamma \neq 0 \\ \mu - \sigma \log\left(-\log(1-p)\right), & \text{si } \gamma = 0 \end{cases}$$

où :

— μ est le paramètre de localisation

— $\sigma > 0$ est le paramètre d'échelle
— γ est le paramètre de forme

Exemple numérique : Quantile extrême pour l'âge d'ordre 95 % est donnée par :

$$x_{0.95} = 65.9557223 - \frac{9.8738854}{-0.5545276}\left[1 - (-\log(1-0.95))^{-(-0.5545276)}\right]$$
$$\approx 50.61 \text{ ans}$$

4. Modélisation avec les techniques de machine learning

4.1 Choix des algorithmes

Le choix de l'algorithme de machine learning est une étape essentielle dans le processus de la modélisation. Ce choix dépend principalement de la nature du problème à traiter (régression, classification, clustering), des caractéristiques des données disponibles (volume, qualité, dimentionnalité), ainsi que des contraintes liés aux ressources informatiques et aux objectifs d'interprétabilité et de performance attendues. Il n'existe pas d'algorithme universellement optimal ; La sélection est donc généralement guidée par une analyse approfondie des données.

Dans notre cas le choix de ces modèles non paramétrique pour la machine learning se justifie aprés l'analyse exploratoire qui a révélé que nos variables d'étude ne suivent pas une distribution normale.

4.1.1 Forêt aléatoire (RF)

Le Forêt aléatoire (ou Random forest) est une méthode de machine learning particulièrement efficace pour les taches de classification. Elle repose sur la construction d'un ensemble d'arbre de décision, chacun étant entrainé sur un sous-échantillon aléatoire de donnée, puis sur l'agrégation des prédictions de ces arbres pour obtenir une décision finale plus robuste et moins sensible au surapprentissage. Dans le cadre de ce travail, le Random forest a été utilisé pour la prédiction du risque d'AVC à partir des variables cliniques et démographiques.

Optimisation des hyperparamètres : Afin d'optimiser les hyperparamètres du modèle, une recherche par validation croisée a été réalisé à l'aide d'une grille d'hyperparamètre. Cette procédure a permis d'examiner plusieurs combinaisons de paramètre, notamment :

- **Le nombre d'arbre** dans le forêt (n_estimators)
- **La profondeur maximal** des arbres (max_deph)
- **Le nombre minimal d'observation** par feuille (min_sample_leaf)

L'évaluation a été fais par une validation croisée à 5 plis, garantissant une estimation fiable des performances sur des sous-échantillon indépendants.

Meilleurs paramètres obtenus : Aprés évaluation systématique de plusieurs configurations, les paramètres optimaux retenus sont :

- **Le nombre d'arbre** : 8
- **La profondeur maximal des arbres** : 10
- **Le nombre minimal d'observation par feuille** : 1

Cette configuration a permis d'obtenir un score de validation croisée avec un taux de 87% selon les plis testés. Le modèle optimal se distingue par une bonne capacité de généralisation tout en capturant les interactions complexes entre les variables explicatives.

4.1.2 Support Vector Machine (SVM) :

Dans cette étude le SVM a été utilisé comme une méthode supervisé pour prédire la survenu d'un AVC. Le SVM est particulièrement adapté à la classification binaire. Son principe repose sur une recherche d'un hyperplan optimal séparant au mieux les différentes classes en maximisant la marge entre les points les plus proches de la frontière de décision (appélés vecteur support)

Optimisation des hyperparamètres : Pour obtenir un modèle performant, une recherche par grille a été menée en explorant plusieurs combinaisons des principaux hyperparamètres du SVM avec une fonction de noyau de type radial ou gaussien (RBF), adaptée à la modélisation des relations non linéaires entre les variables explicatives et la variable cible.

- **Paramètre de régulation C**
- **Gamma** (coefficient du noyau RBF)
- **Kernel** "rbf" noyau radial

La recherche a été effectuée par une validation croisée par 5 plis, garantissant la robustesse de modèle face à la variabilité des données d'entrainement.

Meilleurs paramètres obtenus : Au terme de cette optimisation le meilleur modèle obtenu correspond au paramètres suivants :

- **C** = 100
- **Gamma** = 1
- **Kernel** = RBF (noyan radial de base)

Ce modèle a été retenu comme modèle final SVM pour les phases d'evaluation sur les données test.

4.1.3 K-plus proches voisins (KNN)

Le K-plus proche voisins (KNN) est une méthode de classification supervisée basée sur la proximité des observations dans l'espace des variables explicatives. Le principe du KNN consiste à attribuer à une nouvelle observation la classe majoritaire parmi ses k voisins le plus proches, déterminés selon une mesure de distance choisie, généralement la distance euclidienne.

Optimisation des hyperparamètres : Pour obtenir un modèle performant, une recherche des meilleurs hyperparamètres a été menée en explorant plusieurs combinaisons des principaux hyperparamètres du KNN dans le but d'optimiser les performances du modèle KNN.

- **Le nombre de voisin (n_neighbors)**
- **La pondération des voisins (weight)**
- **La métrique des distances** "metrics"

Meilleurs paramètres obtenus : Au terme de cette optimisation le modèle obtenu correspond au paramètres suivants :

- **Nombre de voisin** = 7
- **pondération** = uniforme (chaque voisin a le même poids dans la décision)
- **Métrique de distance** = euclidienne

Ce modèle a été retenu comme modèle final KNN pour les phases d'evaluation sur les données test.

4.1.4 Gradient Boosting

Le Gradient Boosting est une méthode d'apprentissage supervisé fondée sur une combinaison séquentielle d'arbres de décision. Chaque nouvel arbre est construit pour corriger les erreurs du précédent, ce qui permet d'améliorer progressivement les performances globales du modèle. Dans ce travail le gradient boosting a été utilisé pour modéliser le risque d'AVC à partir des variables cliniques et démographiques.

Optimisation des hyperparamètres : L'optimisation des hyperparamètres a été réalisé à l'aide d'une recherche par grille afin de déterminer la meilleure configuration du modèle. Les paramètres exploités sont les suivants :

- **Nombre d'estimateur** (n_estimators)
- **Profondeur maximal des arbres** (max_depth)
- **Taux d'apprentissage** (learning_rate)
- **Fonction de perte** (log_loss)
- **Sous-échantillonage** (subsample)

L'optimisation a été menée par une validation croisée par 5 plis, en utilisant une critère de sélection l'aire sous la courbe ROC, ce qui permet de maximiser la capacité discriminante du modèle .

Meilleurs paramètres obtenus : Au terme de cette optimisation le meilleur modèle obtenu correspond au paramètres suivants :

- **Nombre d'estimateur** = 200
- **Profondeur maxima des arbres** = 5
- **Taux d'apprentissage** = 0,2
- **Fonction de perte** = log_loss
- **Sous-échantillonage** = 0,8

4.1.5 Réseau de neurones

Dans le cadre de cette étude un réseau de neurone profond (deep neuronal network) a été construit afin de modéliser le risque d'AVC à partir des variables cliniques et démographiques.

Architecture du modèle : Le réseau est de type séquentiel et comprend plusieurs couche entièrement connectée (dense), intercalées de technique de régularisation pour améliorer la robustesse et eviter le surapprentissage. La structure retenue est la suivante :

- **Couche d'entrée** : une couche dense de 128 neurones, activation ***ReLu***
- **Régularisation** : chaque couche est suivi d'une normalisation batch (BatchNormalization) et d'un dropout fixé à 20 %
- **Couches intermédiaires** : quatres couches supplémentaire (64,32,32,32 neurones) avec activation ReLu suivies du même mécanisme de régularisation.
- **Couche de sortie** :Un neurone avec activation sigmoïde.

Paramètres d'entraînement : Le modèle a été entrainé avec les spécifications suivantes :

- **Fonction de perte** : Binary_crossentropy
- **Optimisateur** : Adam, avec un taux d'apprentissage fixée à 0,001
- **Nombre d'époques** : 100
- **Taille de batch** : 32

4.2 Métriques d'évaluation de performance

L'évaluation des modèles de classifications est une étape cruciale pour mesurer leur efficacité et comparer leurs performances. Dans cette section nous présentons les métriques statistiques utilisées dans le cadre de notre étude pour quantifier la qualité des prédictions.

Matrice de confusion : La matrice de confusion est un tableau qui permet de comparer les prédictions d'un modèle de classification aux valeurs réelles. Elle détaille le nombre de prédiction correcte et incorrecte répartie par classe. Pour un problème de classification binaire, elle se présente ainsi :

		Observations		
		+	-	Total
Prédictions	+	Vrais positifs (VP)	Faux positifs (FP)	Total des positifs prédits (VP + FP)
	-	Faux négatifs (FN)	Vrais négatifs (VN)	Total des négatifs prédits (FN + VN)
	Total	Total des vrais positifs observés (VP + FN)	Total des vrais négatifs observés (FP + VN)	Taille totale de l'échantillon (N)

FIGURE 2.6 – Matrice de confusion (source [39])

— VP : Vrais positifs (True Positives)
— VN : Vrais négatifs (True Negatives)
— FP : Faux positifs (False Positives)
— FN : Faux négatifs (False Negatives)

Accuracy (Exactitude) : Mesure la proportion totale de prédictions correctes par rapport à l'ensemble des prédictions effectuées. :

$$\text{Accuracy} = \frac{VP + VN}{VP + VN + FP + FN}$$

Précision : Mesure la fiabilité des prédictions positives en calculant proportion de positifs de la population parmi tous les positifs prédits. La précision est utilisée comme un indicateur de performance lorsque l'objectif est de limiter le nombre de faux positifs :

$$\text{Precision} = \frac{VP}{VP + FP}$$

Recall (Rappel ou Sensibilité) : Le rappel permet de mesurer la proportion de positifs prédits parmi tous les positifs de la population. Il est intérressant lorsque nous devons identifier tous les échantillons positifs ; c'est-à-dire lorsqu'il est important d'éviter les faux négatifs. :

$$\text{Recall} = \frac{VP}{VP + FN}$$

F1-Score : le F1 score est la moyenne harmonique de la précision et du rappel, elle peut être une meilleure mesure que la précision sur des jeux de données de classification binaire déséquilibrés [33] :

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

AUC-ROC) : Mesure la capacité du modèle à distinguer les classes à tous les seuils :

— Courbe ROC :) La courbe ROC est alors tracée dans un espace de deux dimensions définies par Taux de vrais positifs (TPR en ordonnée et taux de faux positifs (FPR) en abscisse. Ce graphique permet d'identifier un ensemble de zones et de points remarquables, indiqué dans la figure 3.6a

— L'aire sous la courbe AUC $\in [0, 1]$. Si elle est égale à 1, elle indique que la discrimination des classes est parfaites).

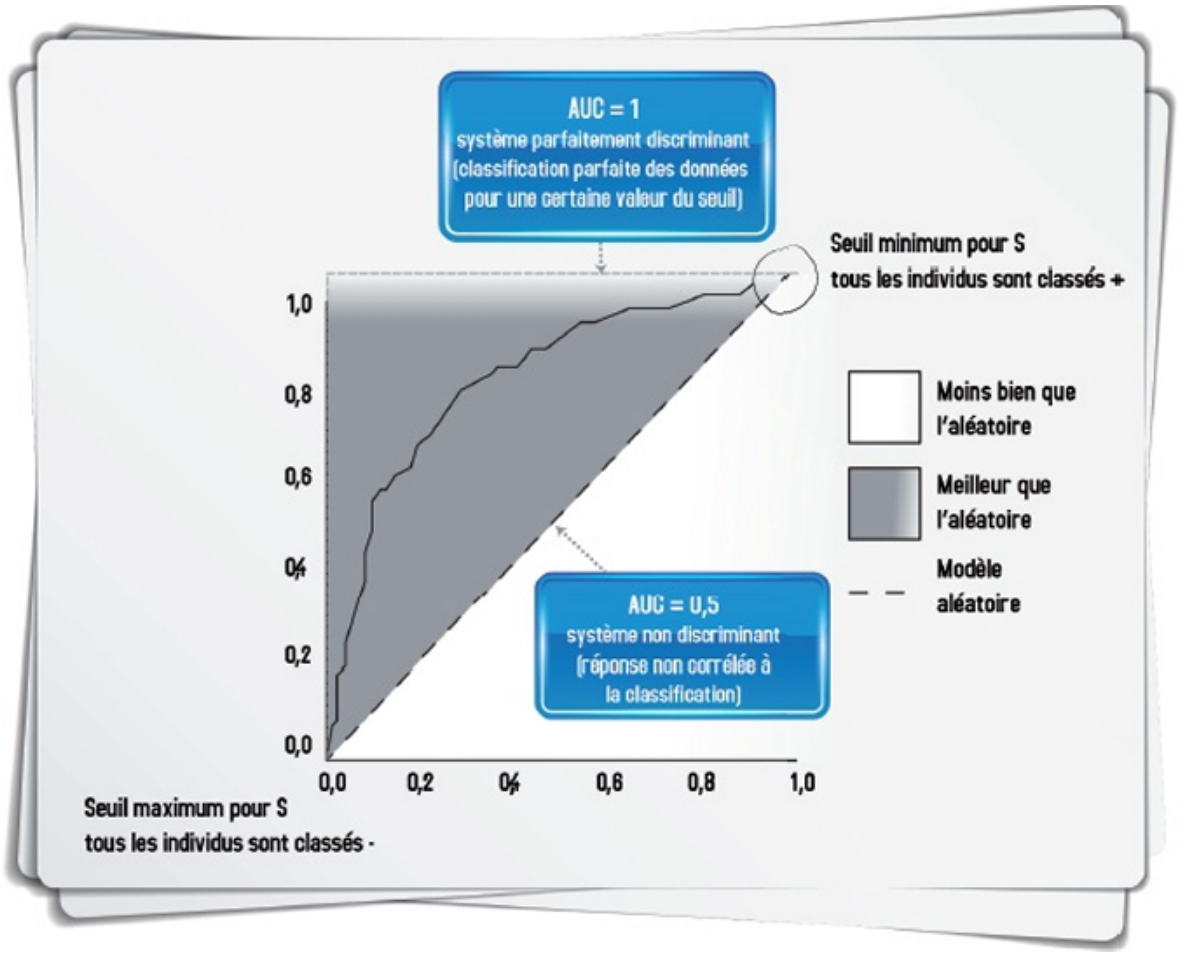


FIGURE 2.7 – Courbe ROC(source [39])

5. Conclusion

Le chapitre 2 a exposé la méthodologie suivie pour modéliser le risque d'AVC. Il a mis en avant l'importance d'une préparation rigoureuse des données (nettoyage, traitement des valeurs manquantes et aberrantes, normalisation, équilibrage des classes) pour garantir la qualité des analyses. L'analyse exploratoire a permis de repérer les tendances et les facteurs associés à l'AVC. Les variables les plus pertinentes ont été sélectionnées grâce à la littérature et à des analyses statistiques. Deux approches de modélisation ont été utilisées : la statistique des extrêmes (GEV) pour cibler les profils à risque élevé, et plusieurs algorithmes de machine learning pour maximiser la performance prédictive. Enfin, les modèles ont été évalués avec des métriques objectives, assurant la fiabilité des résultats. Ce chapitre établit ainsi une base solide pour l'analyse des résultats et la compréhension des risques d'AVC dans la population étudiée.

Chapitre 3

ANALYSE DES RÉSULTATS ET PERSPECTIVES

1. Introduction

Aprés avoir présenté le cadre théorique et les fondements méthodologiques pour la modélisation du risque de la survenue d'un accident vasculaire cérébral, ce chapitre est consacré à l'analyse détaillée des résultats obtenus. Il s'agit d'examiner la performance des différentes approches statistiques et algorithmes des apprentissages automatiques appliqués aux données cliniques et démographiques, ainsi que d'interprêter les facteurs de risque identifiés.

2. Résultats de la modélisation avec les méthodes traditionnelles

2.1. Résultats de la modélisation par la régression logistique

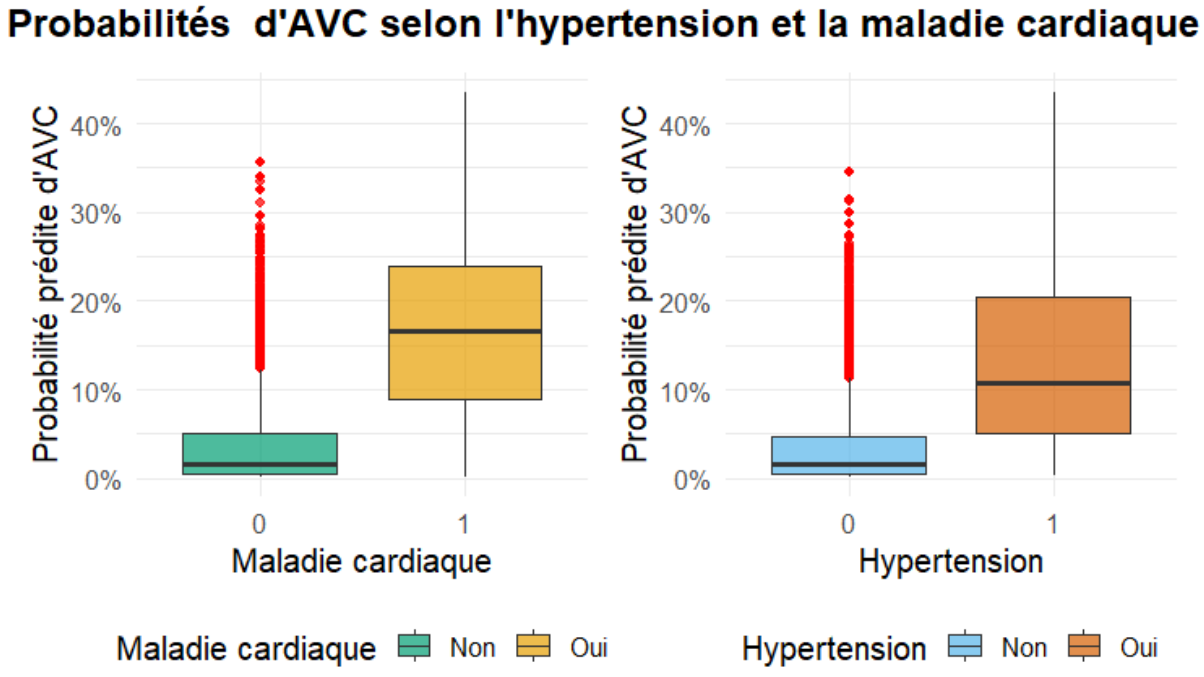
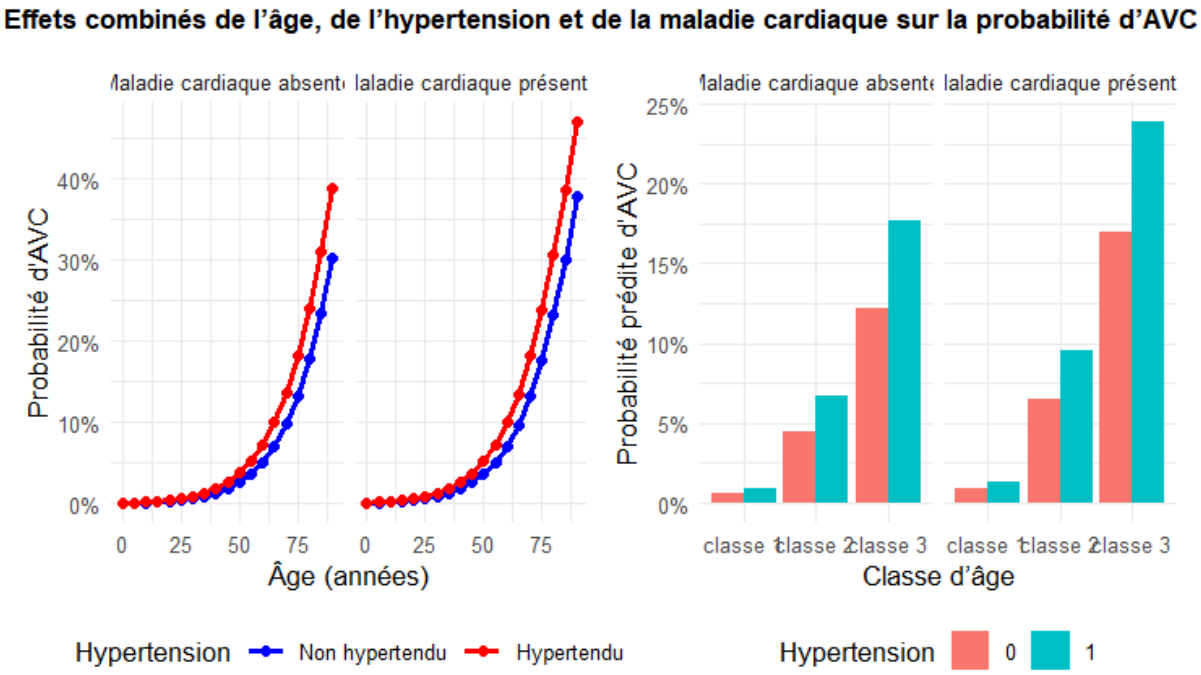
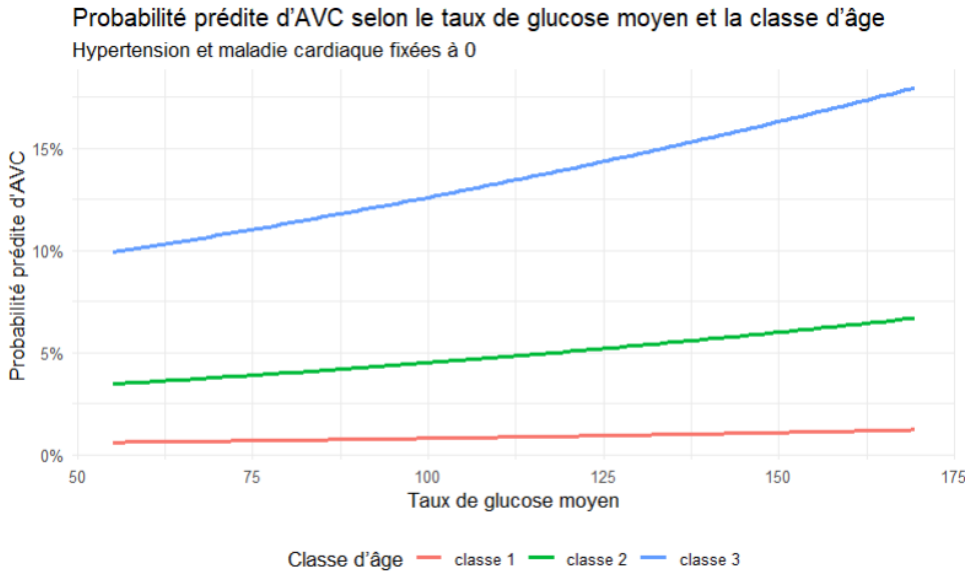


FIGURE 3.1 – Comparaison des risques d'AVC selon les facteurs cliniques



(a) Sans classe d'âge (b)Par classe d'âge

Glycémie moyenne fixée, analyse stratifiée par statut hypertensif



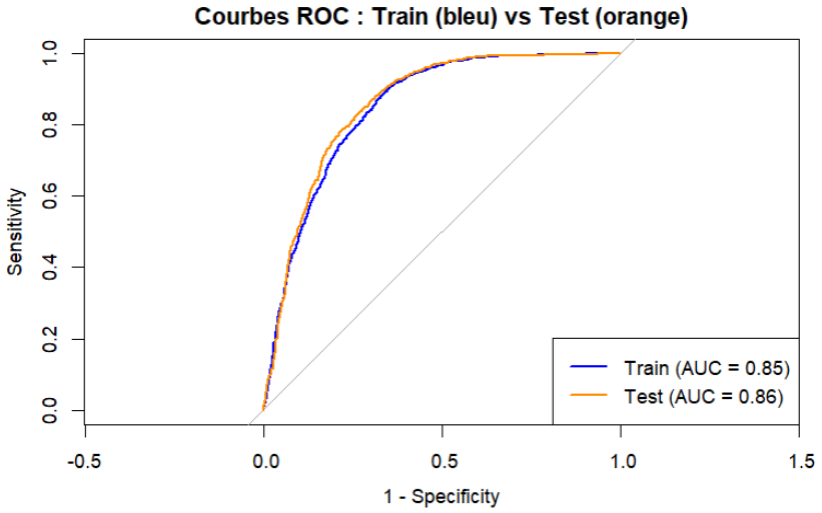
Maladie cardiaque et hypertension fixées à 0 ; analyse stratifiée par classe d'âge.

FIGURE 3.2 – Probabilité d'AVC selon l'âge et les facteurs de risque

FIGURE 3.3 – Performance du modèle de régression logistique

Métrique	Ensemble de test
Accuracy	77.23%
F1-Score	77.60%
Précision	75.60%
Recall	79.49%
Spécificité	75.03%
Taux de bon classement	77.23%
Taux de mauvais classement	22.77%

(a) Métriques de performance



(b) Courbe ROC (AUC = 0.85)

Identification et quantification du risque d'AVC par la régression logistique

La régression logistique a permis d'identifier les facteurs significatifs associés à la survenue d'un AVC dans la population étudiée et quantifier l'intensité de leur effet et de prédire la probabilité d'occurrence d'un AVC selon divers profils individuels. Dans l'étude considérée les facteurs retenus comme significatif sont ***l'âge, l'hypertension, la maladie cardiaque et le taux de glucose*** (voir le tableau 2.10). L'âge est le facteur le plus influent avec un odds ratio (OR) = 1,07 (IC95%[1,06 ;1,08]) indiquant qu'à chaque année supplémentaire, le risque d'AVC augmente de 7% . Cette relation linéaire se confirme également lorsqu'on regroupe les patients par classe d'âge : en effet, comparés la tranche des 25-45 ans (classe 1), les individus âgés de 45-65 ans (classe 2) présentent un risque d'AVC multiplié par 5,72 (IC95%[3,16 ;11,46]), tandis que ceux d'eux 65 + (classe 3) voient leur risque multiplié par 17,46 (IC95%[9,7 ;34,69]).

Ces résultats illustrent non seulement la progréssion régulière du risque avec l'âge, mais soulignent également **son accélaration chez les sujets âgés**, particulièrement au-delà de 65 ans. (voir tableau 2.11).

Outre l'âge, plusieurs autres facteurs de risque ont été identifiés. L'hypertension se distingue comme un facteur de risque majeur : les patients hypertendus présentent un risque d'AVC 52%

plus élevé que les non hypertendus avec un odds ratio (OR) de 1,52 et un intervalle de confiance à 95% allant de [1.1 ; 2.07] , ce qui témoigne d'un effet significatif au seuil de 5% . La maladie cardiaque est également associée à une augmentation du risque d'AVC bien que de manière marginalement significatif. Son odds ration est estimé à 1,52 (IC95% [1.04 ; 2.17]), indiquant une élévation de 52% du risque pour les patients cardiaques par rapport à ceux qui ne le sont pas. Concernant la glycémie , l'effet observé est modeste mais statistiquement significatif (voir tableau 2.11). Chaque unité supplémentaire du taux de glucose est associée à une augmentation de 1% du risque d'AVC (OR= 1,01, IC95%[1,00 ;1,01]).
Ces résultats montrent que l'hypertension et la maladie cardiaque agissent comme des amplificateurs importants du risque d'AVC, tandis que l'hyperglycémie, bien que l'effet modéré, contribue également au risque en particulier lorsqu'elle est combiné à d'autre facteurs comorbidités.

Estimation de la probabilité d'AVC selon les profils de risque

- **Influence de l'âge**
  L'âge exerce une influence marquée et progressive sur la probabilité de survenue d'un AVC. Les études montrent que le risque reste faible jusqu'à environ 50 ans, avec une probabilité généralement inférieure à 5 %. Ce pendant à partir de 60 ans, une augmentation rapide de la chance de survenu d'un AVC est observée (voir la figure 3.2a). À titre d'illustration, un patient âgé de 80 ans, souffrant de la maladie cardiaque, présente une probabilité d'AVC estimée à environ 45 % s'il est hypertendu, contre 37 % chez un non-hypertendu du même âge. Cette tendance se confirme quelle que soit la présence ou l'absence d'autres facteurs tels que la maladie cardiaque.(voir figure 3.2a)

  L'analyse par classes d'âge confirme cette progression. Dans la tranche des 25 à 45 ans (classe 1), le risque est très faible, inférieur à 2 %. Entre 45 et 65 ans (classe 2), le risque devient modéré, se situant généralement entre 5 % et 10 %. Enfin, pour les patients âgés de 65 à 82 ans (classe 3), le risque atteint des niveaux élevés, dépassant souvent 25%. (voir la figure 3.2a). À chaque classe d'âge, la probabilité predite sur les hypertendus est plus haute que celle prredite sur les non hypertendu et que l'écart entre les hypertendus et non hypertendus s'agrandit avec l'âge. L'ecart devient très marqué chez les patients les plus âgés (plus de 10% d'écart).
- **Influence du taux de glucose**
  Le taux de glucose moyen présente également une association significative avec le risque d'AVC, bien que son effet soit plus modéré que celui de l'âge. Les études montrent que pour toutes les classes d'âge, la probabilité de survenue d'un AVC augmente progressivement a mesure que le taux de glucose s'élève.
  — **Chez les patients les plus jeunes (classe 1, 25-45 ans)** la probabilté d'AVC reste trés faible, même pour des taux de glucose élevés, ne depassant pas 2%.
  — **Chez les patients agés de 45-65 ans**, la probabilité d'AVC augmente légèrement avec le taux de glucose : elle reste inférieure à 7% pour toutes valeurs du glucose
  — **Chez les personnes âgés de 65 et plus**, le risque prédictif est nettement élevé. Pour des taux de glucose élevés (autour de 175 mg/dl) , la probabilité de la survenur d'AVC peut dépasser 15%. (voir la figure **??**)
- **Influence de l'hypertension et maladie cardiaque**
  L'analyse des boxplots montre l'influence directe de l'hypertension et de la maladie cardiaque sur la probabilité prédite d'AVC. Ces graphes illustrent nettement que les patients ayant une maladie cardiaque ("Oui") présentent une probabilité d'AVC nettement

supérieure (médiane autour de 20 %) par rapport à ceux sans maladie cardiaque (médiane proche de 5 %). De même, l'hypertension est associée à une augmentation de la probabilité d'AVC (médiane autour de 10-12% chez les hypertendus contre moins de 5 % en l'absence d'hypertension). **L'hypertension et les maladies cardiaques sont deux facteurs cliniques majeurs qui influencent fortement le risque prédictif d'AVC.**

De plus la distribution des probabilités chez les hypertendus est plus étalée : 25% des individus de ce groupe ont une probabilité qui dépassent 20%, et certains cas atteignent ou dépassent 40%. La distribution est également plus étalée chez les patients cardiaques, avec une proportion plus importante d'individus ayant des probabilités élevées d'AVC. Ces observations confirment que l'hypertension et la maladie cardiaque augmentent toutes deux de manière indépendante la probabilité d'AVC. Plus encore, elles interagissent de façon cumulative, produisant un effet synergique : le risque maximal est atteint chez les patients âgés, hypertendus et porteurs d'une pathologie cardiaque. À l'opposé, les patients sans hypertension ni maladie cardiaque présentent majoritairement des probabilités prédictives faibles (<10 %), bien que quelques cas isolés dépassent ce seuil. Cela suggère que d'autres facteurs, tels que l'âge avancé ou une hyperglycémie élevée, peuvent à eux seuls conduire à un risque significatif, justifiant une approche multifactorielle dans l'évaluation du risque d'AVC. (voir la figure 3.1)

2.2. Résultats et interprétation de la modélisation avec GEV

En complément des approches classiques de régression, la loi des valeurs extrêmes généralisée (GEV) a été utilisée pour modéliser le comportement des variables continues dans les cas à haut risque. Cette méthode vise à identifier les seuils extrêmes au-delà desquels les valeurs observées deviennent cliniquement préoccupantes. Après ajustement du modèle BGEVA (*Binomial Generalized Extreme Value Additive Model*), les probabilités individuelles d'occurrence d'un AVC ont été estimées pour chaque patient dans le but de prédire le risque à partir des variables explicatives retenues. Afin de transformer ces probabilités en une classification binaire (patient à risque ou non), il a été nécessaire de déterminer un seuil optimal.

Détermination du seuil de classification par l'indice de Youden

Pour ce seuil, la courbe ROC du modèle a été tracée et l'indice de Youden a été utilisé. Cet indice permet d'identifier le point de la courbe qui maximise la somme de la sensibilité et de la spécificité, offrant ainsi le meilleur compromis entre détection de cas et limitation des faux positifs.

- Seuil optimal de probabilité (**indice de Youden**) : 0,188 (soit 18,8%)
- Spécificité associée à ce seuil : 0,693 (69.3%)
- Sensibilité associée à ce seuil : 0,851 (85,1%)

Cele signifie qu'un patient dont la probabilité prédite par le modèle est supérieure à **18,8 %** est classé comme à risque d'AVC ou même peut être un patient atteint. À ce seuil le modèle identifie correctement environ 85 % des cas d'AVC (sensibilité) tout en minimisant une spécificité correcte de 69 %

Résumé des patients identifiés à risque

Après application de ce seuil sur 5110 patients 33,4% ont été classés comme à risque. Les tableaux ci dessous résument leurs caractéristiques.

TABLE 3.1 – Répartition des variables catégorielles chez les patients à risque

Variable	Modalité	Proportion (%)
AVC		
	Non	87,6
	Oui	**12,4**
Genre		
	Femme	56,7
	Homme	43,3
Hypertension		
	Non	76,1
	Oui	23,9
Maladie Cardiaque		
	Non	84,8
	Oui	**15,2**
Situation Matrimoniale		
	Marié	**92,2**
	Non marié	7,8
Statut Fumer		
	Ancien Fumeur	**56,8**
	Fumeur actuel	27,4
	Non Fumeur	15,8
Type Travail		
	Indépendant / travailleur autonome	**54,4**
	Secteur privé	30,2
	Sans emploi / enfant / inactif	15,5
Résidence		
	Urbain	**51.6**
	Rural	48.4

Ce tableau présente les caractéristiques des patients identifiés à risque d'AVC. On observe que :

— 12% des patients ont effectivement présenté un AVC.
— La majorité des patients risqués sont des femmes et des personnes mariées. Cette surreprésentation peut s'expliquer par la structure démographique et social (espérance de vie plus élevée chez les femmes et majorité des personne sont ou ont été en couple)
— Une forte proportion d'ancien fumeur et d'actuels fumeurs, Cela confirme l'importance du tabac comme facteur de risque d'AVC
— Les patients hypertendus et cardiaque sont fréquents dans les patients à risque ou patient ayant subi un AVC.
— Les personnes à risque sont majoritairement actives, notamment dans le secteur autonome et la répartition entre mileu urbain et rural est équilibrée.

TABLE 3.2 – Statistiques descriptives des patients classés à risque d'AVC

Variable	Min	1er Quart.	Médiane	Moyenne	3e Quart.	Max
Âge	42	61	68	67,89	76	82
Taux glucose moyen	55,2	81,5	101,4	127,9	191,7	271,7
IMC	11,3	26,7	29,2	30,47	33,7	60,9
Probabilité prédite	0,1881	0,2359	0,3074	0,3209	0,3862	0,6550

Les patients à risque ou ayant subis un AVC sont majoritairement âgés avec une moyenne proche de 68 ans confirmant que l'âge avancée est un facteur dominant. La moyenne du taux de glucose est élevée (127,9 mg/dL traduisant la diabètique) et le 3e quartile très élevé (191,7 mg/dL) indiquent une hyperglycémie fréquente dans les patients à risque (25% de ce groupe dépassent le seuil de 191,7 mg/dl). Les patients à risque présentent en moyenne un indice de masse corporelle de 30,47, ce qui correspond à la limite du surpoids et du debut de l'obésité. La moyenne de la probabilité prédite dans ce groupe est de 32% avec des valeurs allant jusqu'a 65,5% (voir tableau 3.2).

TABLE 3.3 – Paramètres GEV estimés

Variable	Méthode	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\gamma}$	AIC
Âge	MLE	65.96	9.87	-0.55	***12187.09***
	L-moments	64.94	9.21	0.33	13756.30
	OLS	65.10	9.13	-0.36	12238.21
Glucose	MLE	92.33	33.55	0.42	***18204.66***
	L-moments	98.96	42.69	0.09	18353.16
	OLS	104.04	49.66	-0.11	18533.20
IMC	MLE	27.88	5.31	-0.08	***10898.16***
	L-moments	27.82	4.98	0.04	11026.73
	OLS	27.73	4.93	-0.02	10937.71

Trois méthodes d'estimation ont été appliquées pour ajuster la loi GEV sur les variables numériques des patients identifiés à risque. La méthode de maximum de vraisemblance (MLE), L-moments et les moindres carrées (OLS). Pour chaque variable, la méthode de maximum de vraisemblance (MLE) offre le meilleur ajustement selon le critère AIC (voir tableau 3.3). Pour l'ensemble des variables considérées, la méthode du maximun de vraisemblance offrent le **meilleur ajustement**, comme raison les valeurs plus faibles de l'AIC.

— Concernant l'**âge,** l'AIC obtenue avec MLE est nettement inférieur à celui des L-moments, confirmant la qualité de l'ajustement. **Le paramètre de forme $\hat{\gamma}$ est négatif** pour MLE et OLS, ce qui suggère une distribution de **type Weibull**, à **queue bornée**, indiquant l'existence d'un âge maximal au-delà duquel le risque d'AVC devient rare.

— Pour le **taux de glucose**, l'ajustement par MLE donne également la meilleure performance (AIC = 18 154,42), contre 18 300,95 pour les L-moments et 18 480 pour OLS. Le **paramètre de forme $\hat{\gamma}$ est positf** dans les cas MLE et L-moments, indiquant une **distribution de type fréchet**, à **queue lourde**, révélant la présence de valeurs extrêmes élevées de glycémie chez les patients à risque ou ayant eu un AVC.

Cela confirme la survenue fréquente d'hyperglycémies très élevées dans ce groupe, qui peuvent jouer un rôle importante dans le déclenchement de l'AVC.

— En ce qui concerne l'**IMC**, l'ajustement GEV confirme également la supériorité de la méthode MLE (AIC= 10 865,69), suivie de près par OLS (10905,15) tandis que les L-moments donnent un ajustement plus faible (AIC= 10992,62). Le **paramètres de forme est proche de zéro** pour toutes les méthodes, ce qui suggère une **distrisbution proche de type Gumbel**, avec une queue exponentielle. Cela signifique que des valeurs extrêmes d'IMC sont possibles dans les patients à risque ou ayant subis un AVC, mais reste moins fréquentes et limitées par rapport au glucose.

La méthode du maximum de vraisemblance (MLE) fournit systématiquement le meilleur ajustement pour l'ensemble des variables numériques étudiées (âge, taux de glucose, IMC), comme l'indique le plus faible AIC. L'analyse du paramètre de forme révèle une queue bornée pour l'âge (**type Weibull**), une queue lourde pour le taux de glucose (**type Fréchet**) et une distribution proche de **Gumbel** pour l'IMC. Ces résultats traduisent la nature des extrêmes dans cette population à risque : existence d'une limite supérieure pour l'âge, possibilité de valeurs très élevées pour la glycémie traduisant la présence des patients diabètiques, et des cas d'obésité mais limités.

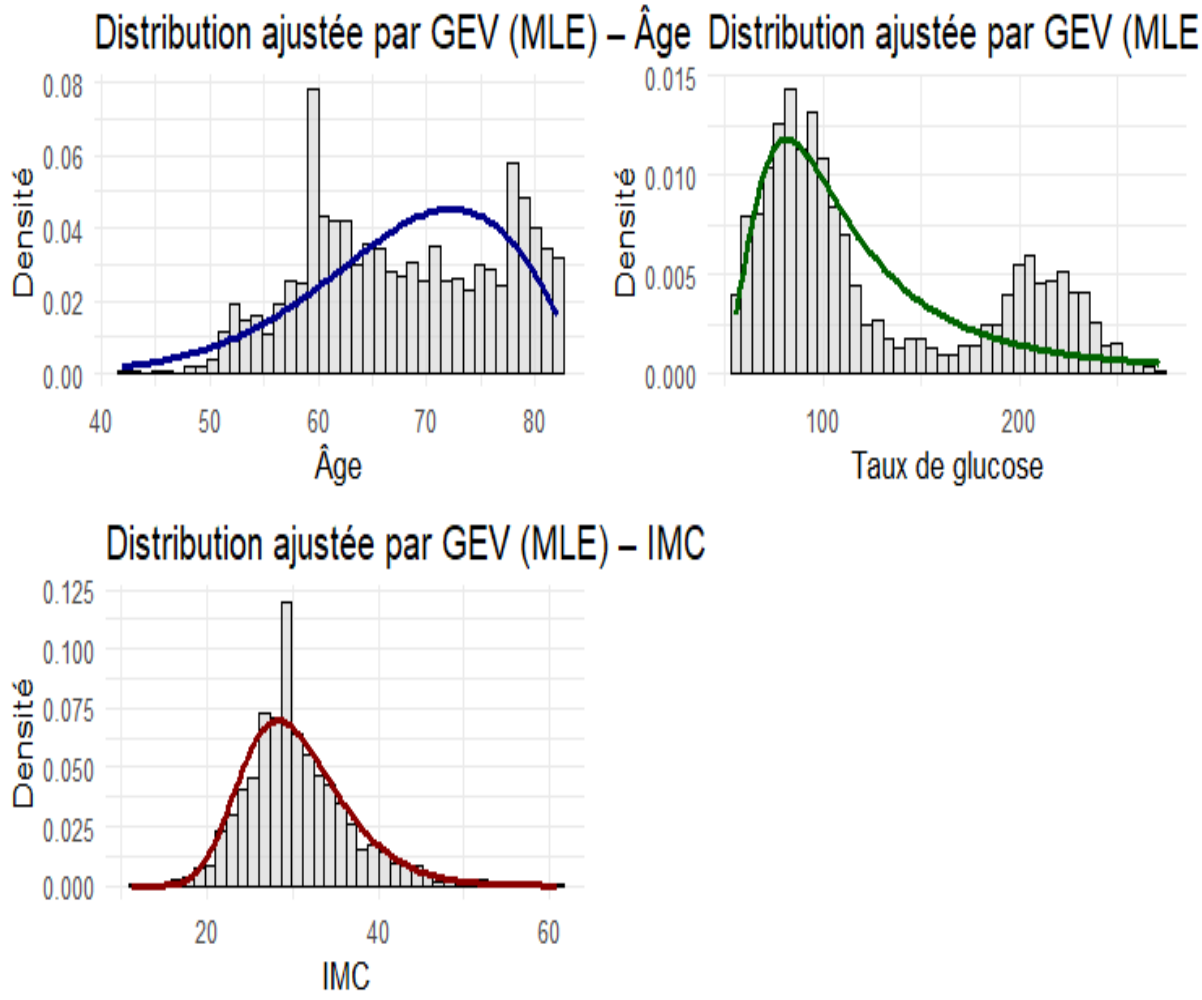


FIGURE 3.4 – Distribution des variables

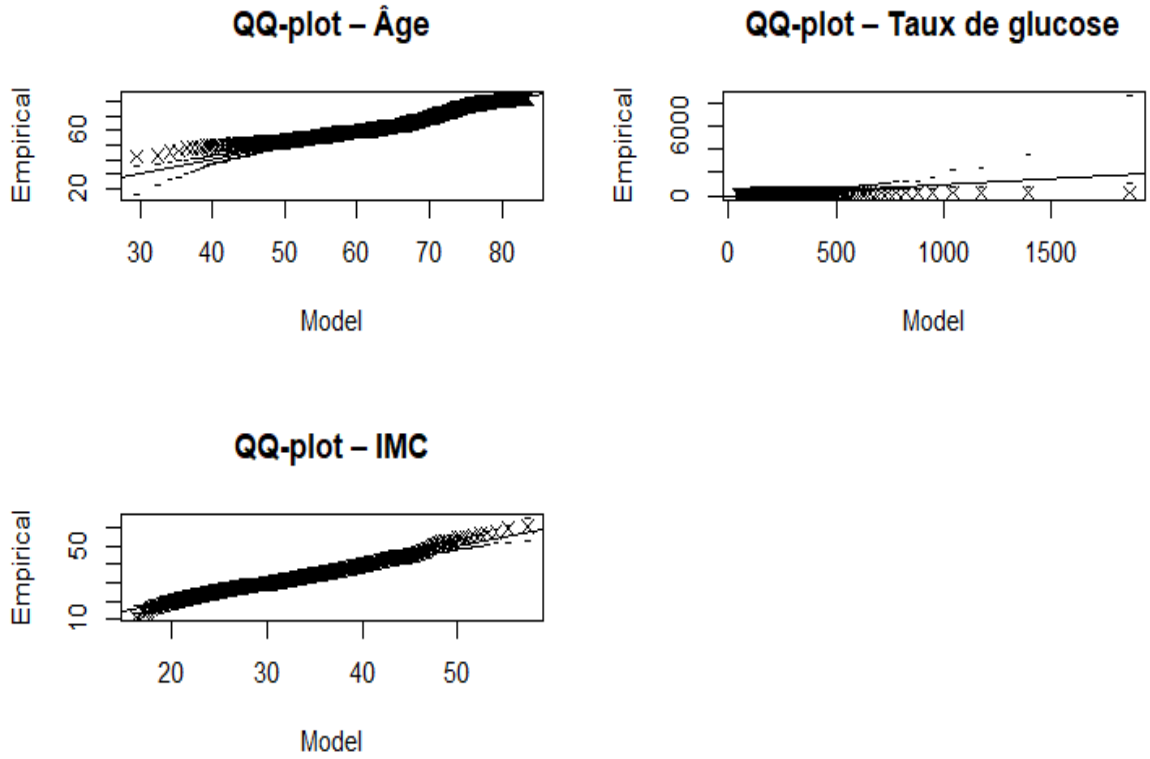


FIGURE 3.5 – Qualité d'ajustement des modèles

TABLE 3.4 – quantiles extrêmes

Quantile	Âge	Glucose	IMC
90%	78.70	218.24	38.77
95%	80.40	290.91	41.84
99%	82.47	564.68	48.16

Analyse des résultats des quantiles extrêmes

L'analyse des quantiles extrêmes issus de la modélisation par la loi GEV permet de mettre en évidence des profils particulièrement à risque dans la population étudiée. Pour la variable âge, 10 % des patients à risque ont plus de 78,7 ans, 5% dépassent 80,4 ans et 1% atteignent ou excèdent 82,5 ans, ce qui reflète une concentration notable de sujets très âgés parmi les cas à risque. Concernant le taux de glucose, les valeurs extrêmes sont particulièrement marquées : 10% dépassent 218mg/dL, 5% franchissent le seuil critique de 290,9 mg/dL, et 1% atteignent ou excèdent 564,7 mg/dL. Ces résultats confirment la prévalence élevée d'hyperglycémie sévères dans ce groupe. Enfin pour l'indice de masse corporelle (IMC), 10% des individus presentent un IMC supérieur à 38,8 kg/m2, 5% dépassent 41,8 kg/m2 et 1% atteignent ou excèdent 48,2 kg/m2, suggérant la présence de cas obésité massive.

Ces résultats indiquent que 5% des patients à risque cumulent des seuils élevés en âge (80,4 ans), glucose (290,9 mg/dL), ou IMC (41,8 kg/m2). tandis que 1% les plus extrêmes présentent des valeurs critiques atteignant ou dépassant les 564 mg/dL pour la glycémie ou 48 kg/m2 pour l'IMC.

2.3. Résultats de la modélisation avec les techniques de Machine Learning

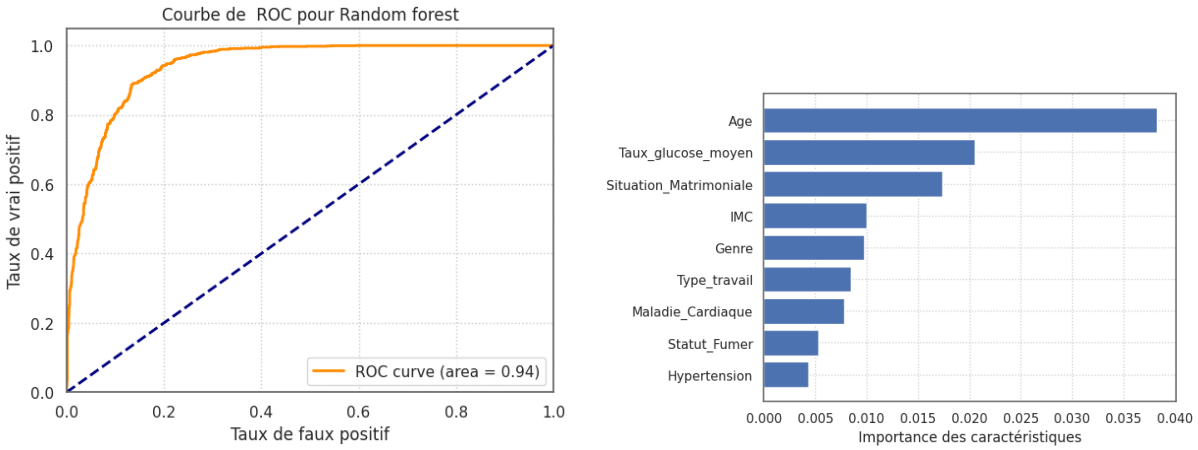
Random forêts

TABLE 3.5 – Performances du modèle Random Forest

Métrique	Train %	Test %
Accuracy	90.58	87.01
F1-Score	90.54	86.96
Précision	91.23	87.66
Recall (Sensibilité)	90.58	87.01

TABLE 3.6 – Taux de classification par classe pour le modèle Random forest

Classe	Bon classement (%)	Mauvais classement (%)
Sans AVC (0)	80.56	19.44
Avec AVC (1)	93.52	6.48



(a) Courbe de ROC

(b) Importance des variables

FIGURE 3.6 – Résultats du modèle Random Forest

Support Vector Machine (SVM)

TABLE 3.7 – Performances du modèle SVM

Métrique	Train %	Test %
Accuracy	96.58	91.77
Score F1	96.57	91.76
Précision	96.68	92.06
Recall	96.58	91.77

TABLE 3.8 – Taux de classification par classe pour le modèle SVM

Classe	Bon classement (%)	Mauvais classement (%)
Sans AVC (0)	87.72	12.28
Avec AVC (1)	95.86	4.14

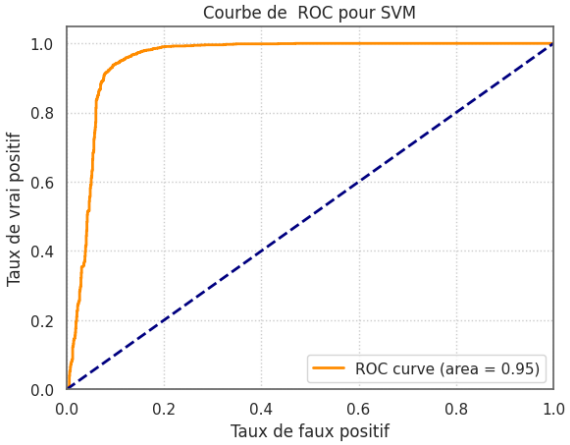


FIGURE 3.7 – Courbe ROC du modèle SVM (AUC = 0.92)

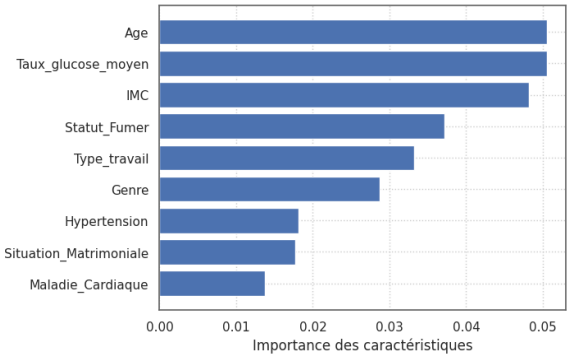


FIGURE 3.8 – Importance des variables dans le modèle SVM

FIGURE 3.9 – Résutats du modèle SVM

K-plus proche voisins (KNN)

TABLE 3.9 – Performances complètes du modèle KNN

Métrique	Train (%)	Test (%)
Accuracy	89.33	87.35
F1-Score	89.24	87.22
Précision	90.61	89.07
Recall	89.33	87.35

TABLE 3.10 – Taux de classification par classe pour le modèle KNN

Classe	Bon classement (%)	Mauvais classement (%)
Sans AVC (0)	87.72	12.28
Avec AVC (1)	95.86	4.14

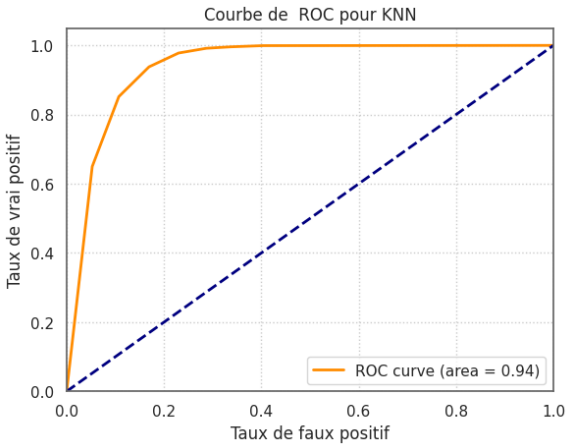


FIGURE 3.10 – Courbe ROC du modèle KNN (AUC = 0.94)

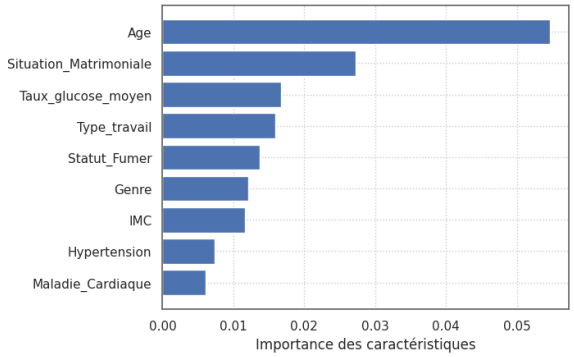


FIGURE 3.11 – Importance des variables dans le modèle KNN

FIGURE 3.12 – Performances du modèle KNN

Gradient Boosting

TABLE 3.11 – Performances du modèle Gradient Boosting

Métrique	Entraînement (%)	Test (%)
Accuracy	99.85	95.34
Score F1	99.85	95.34
Précision	99.85	95.39
Recall	99.85	95.34

TABLE 3.12 – Taux de classification par classe pour le modèle Gradient Boosting

Classe	Bon classement (%)	Mauvais classement (%)
Sans AVC (0)	87.72	12.28
Avec AVC (1)	95.86	4.14

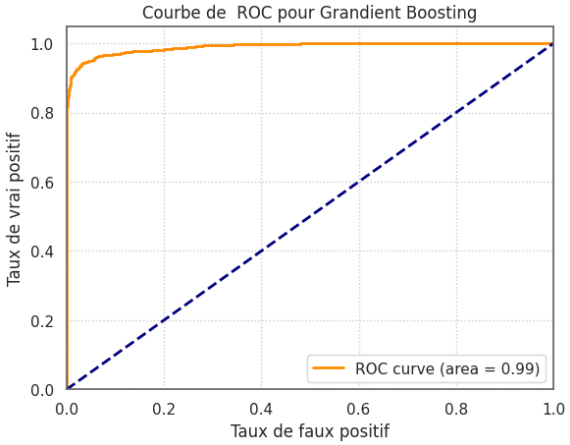


FIGURE 3.13 – Courbe ROC du Gradient Boosting (AUC = 0.99)

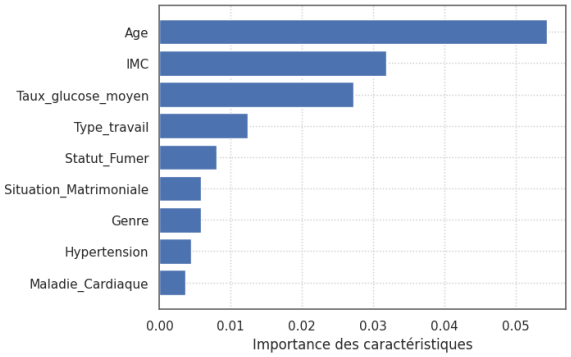


FIGURE 3.14 – Importance des variables (Gradient Boosting)

FIGURE 3.15 – Résultats du modèle Gradient Boosting

Réseau de neurone

TABLE 3.13 – Performances détaillées du réseau de neurones

Métrique	Entraînement (%)	Test (%)
Accuracy	90.90	89.03
Précision	85.28	83.56
Score F1	91.60	89.80
Recall	98.92	97.04

TABLE 3.14 – Taux de classification par classe pour le modèle Réseau de neurones

Classe	Bon classement (%)	Mauvais classement (%)
Sans AVC (0)	81.11	18.89
Avec AVC (1)	97.04	2.96

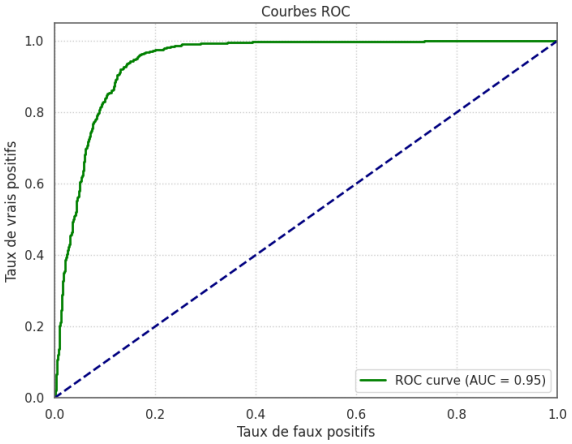


FIGURE 3.16 – Courbe ROC du réseau de neurones (AUC = 0.98)

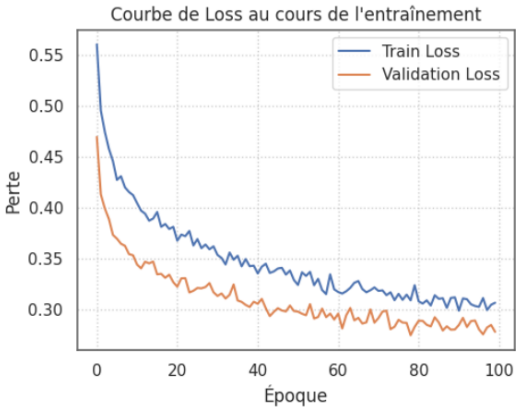


FIGURE 3.17 – Courbe d'apprentissage : Évolution de la fonction de perte

FIGURE 3.18 – Résultats du réseau de neurones

Interprétation des résultats de Machine learning

Plusieurs modèles d'apprentissage supervisé ont été testés pour prédire le risque d'AVC : *Random forêst, Support Vector Machine, K-Nearest Neighbors, Grandient Boosting, et les réseaux de neurones multicouches*. Ces modèles ont été évalués sur la base de leur capacité de classification (accuracy, précision, recall, f1-score) ainsi que le taux de bon et de mauvais classement aux classes avec AVC (classe 1) et sans AVC (classe 0)

Le tableau suivant montre les performance globales des modèles :

TABLE 3.15 – Comparaison des performances des modèles

Modèle	Accuracy(%)	Précision(%)	Recall(%)	F1-Score(%)
Réseau de Neurones (RN)	89.03	83.56	97.04	89.80
Gradient Boosting (GB)	95.34	95.39	95.34	95.34
SVM	91.77	92.06	91.77	91.76
KNN	87.35	89.07	87.35	87.22
Random Forest (RF)	87.01	87.66	87.01	86.96

TABLE 3.16 – Taux de classification par classe pour les différents modèles

Modèle	TBC C0 (%)	TMC C0 (%)	TBC C1 (%)	TMC C1 (%)
Random Forest	80.56	19.44	93.52	6.48
SVM	87.72	12.28	95.86	4.14
KNN	87.72	12.28	95.86	4.14
Gradient Boosting	87.72	12.28	95.86	4.14
Réseau de neurone	81.11	18.89	97.04	2.96

Note : Classe 0 = Sans AVC, Classe 1 = Avec AVC.

Performances globales des modèles

Les résultats obtenus montrent que le **Gradient Boosting** est le plus performant en terme d'**accuracy, précision, recall et F1-score**, avec une précision globale de **95%**. La courbe ROC montre une aire sous la courbe de très élevé traduisant une bonne capacité discriminante. Ces performances élevées traduisent une capacité remarquable à distinguer efficacement les individus à risque d'AVC tout en assurant un bon équilibre entre la détection des cas positifs (AVC) et négatifs (sans AVC).
Ce compromis entre la sensibilité et la spécificité est essentiel pour une application clinique fiable.

Le **réseau de neurones** obtient quant à lui une accuracy légèrement inférieure (**89,03%**), mais avec un rappel de **97,04%**, nettement supérieur à tous les autres modèles. Cela signifie que ce modèle est trés efficace pour détecter les patients à risque, parvenant à identifier la quasi-totalité des cas d'AVC (avec trés peu de faux négatifs), ce qui est crucial dans un contexte médical où manquer un cas positif peut avoir de lourde conséquences en terme de morbidité et de mortalité.

Le modèle **SVM** se positionne également comme une alternative robuste, avec un compromis entre la précision et le rappel bien équilibré (accuracy= **91,77%**, précision = **92,66 %**) adapté aux besoins cliniques où les faux positifs et faux négatifs sont également critiques.

Les modèles **KNN** et **Random Forest** présentent également de bonne performance métriques proches. Ces résultats indiquent que ces deux approches sont capables d'identifier de manière fiable les individus à risque d'AVC, bien qu'à un niveau légèrement inférieur au réseau de neurones, au Gradient Boosting ou au SVM.

Performances des modèles par classe

L'analyse plus fine par classe révèle que le Réseau de neurone excelle particulièrement dans la détection des patients ayant subi un AVC (classe 1), avec un taux de bon classement de **97,04%** et un taux d'erreur trés faible (2,96%). Cette caractéristique est importante dans le contexte médical, où il est primordial de minimiser les faux négatifs afin de ne pas manquer des cas critiques. En revanche ce modèle présente une précision plus faible (83,56%), ce qui signifie qu'il génère davantage de faux positifs (compromis souvent acceptable dans un cadre de depistage). Les modèles comme le **SVM**, **KNN** et le **Gradient Boosting** montrent des bonnes performances équilibrées avec des taux de bon classement proches de 87-88% pour les patients sans AVC (classe 0), et de 95-96% pour les patients avec AVC (classe 1). Ces résultats indiquent leur robustesse pour une classification fiable aussi bien dans la détection des cas positifs que négatifs.

4. Comparaison des méthodes

Dans cette section, nous comparons les différentes méthodes de classification utilisées pour la prédiction du risque d'AVC. Le choix du modèle dépend avant tout des objectifs cliniques visés : maximiser la détection des cas, trouver un compromis entre faux positifs et faux négatifs, ou encore permettre une interprétation claire des résultats.

Lorsque la priorité est donnée à la détection des AVC, le **réseau de neurones** s'impose comme la méthode la plus performante grâce à son excellent *recall* de **97,04 %**, indiquant

sa capacité à identifier la quasi-totalité des cas positifs. Cette propriété est cruciale dans une stratégie de dépistage où les faux négatifs peuvent avoir des conséquences cliniques graves.

En revanche, si l'objectif est de trouver un **équilibre entre précision, sensibilité**, le **Gradient Boosting** se démarque comme le modèle le plus performant globalement, avec une *précision* de **95,39 %** et un très bon compromis entre les deux classes. Il est suivi de près par le **SVM** et le **KNN**, qui offrent également de bons résultats. Si l'objectif est de garantir une interprétation claire des résultats les méthodes classiques comme la **régression logistique** et la **Statistique des extrêmes** restent meilleures bien que moins performantes, mais elles conservent un intérêt important. Elles permettent de modéliser la probabilité d'AVC en fonction des variables explicatives et d'interpréter directement l'effet de chaque facteur de risque via les coefficients associés. Elle reste ainsi un outil précieux dans un contexte épidémiologique où lorsque l'interprétabilité prime sur la performance.

Cette comparaison mets clairement en évidence la **supériorité des techniques de machine learning** face aux approches statistiques traditionnelles, notamment en matière de précision, de détection des cas à haut risque, et de robustesse prédictive. Contrairement aux modèles classiques comme la régression logistique ou ceux issus de la théorie des valeurs extrêmes (GEV), qui reposent sur des hypothèses fortes de linéarité et d'indépendance, les modèles d'apprentissage automatique présentent une **flexibilité bien supérieure** pour capter les **relations complexes et non linéaires** entre les facteurs de risque.

Cette flexibilité se traduit par trois avantages majeurs :

— Une meilleure capacité à modéliser les relations non linéaires entre les variables explicatives ;
— Une modélisation efficace des interactions complexes entre facteurs de risque ;
— Une adaptation plus fine aux caractéristiques des données médicales, souvent bruitées et déséquilibrées.

Cela étant, les méthodes statistiques classiques conservent leur pertinence dans certaines situations. La **théorie des valeurs extrêmes** (notamment via la loi GEV) est particulièrement utile pour **l'analyse des événements rares** et pour la détection de seuils critiques dans les queues de distribution. Quant à la régression logistique, elle demeure un modèle de référence pour **interpréter les relations entre variables** et pour identifier les facteurs de risque significatifs de manière transparente et communicable.

En somme, le **réseau de neurones** constitue la solution idéale pour des campagnes de dépistage ciblées, le **Gradient Boosting** offre un excellent compromis pour un usage clinique, tandis que les modèles statistiques conservent leur place dans l'analyse interprétative et la gestion des extrêmes.

5. Limites et axes d'amélioration

Dans cette étude bien que méthodologiquement solide, présente certaines limites. La taille de l'échantillon reste relativement restreinte, ce qui peut limiter la généralisation des résultats. Bien qu'un équilibrage des classes ait été réalisé, les données proviennent d'une source unique, sans validation externe, ce qui réduit leur représentativité. De plus, les variables utilisées pour la prédiction, bien que pertinentes (âge, glucose, IMC,etc.) ne couvrent pas l'ensemble des facteurs de risque potentiels. Par exemple des informations importantes comme les ***antécédants familliaux***, ***les troubles de conscience***, ***la déficience motrice, des signes de gravités tels que l'engagement cérébral ou l'hémorragie intraventriculaire ou encore les délais entre les premiers symptômes et la prise en charge***, auraient pu enrichir significativement l'analyse [1]

Les modèles de machine learning notamment le réseau de neurone et le Gradient Boosting

offrent d'excellentes performances mais restent peu interprétables, ce qui peut poser problème dans un contexte clinique. À l'inverse la régression logistique, bien qu'interprétable et souvent utilisée en épidémiologie, montre des performances limitées (accuracy de 78,03 %), notamment sur la détection des AVC. Ce résultats est attendu, car ce modèle suppose des relations linéaires entre les variables, ce qui est probablement trop restrictif pour modéliser des phénomènes complexes comme le risque d'AVC. De la même manière, l'utlisation de la loi GEV dans le cadre de la statistique des extrêmes repose sur des hypothèses fortes telles que la stationnarité et l'indépendance des observations, qui peuvent être difficile à vérifier dans ce contexte. Ces limites soulignent la nécessité d'un compromis entre performance, robustesse et explicabilité.

Conclusion générale, recommandations et perspectives

Le présent mémoire s'inscrit dans le champ de la modélisation des risques d'AVC en combinant la Statistique des extrêmes et les méthodes de machine learning. L'objectif principal était de développer, comparer et affiner des méthodes statistiques avancées et des techniques de machine learning afin d'anticiper la survenue d'événements rares et graves, et d'identifier les profils de patients les plus à risque.
Le premier chapitre a posé les bases théoriques à travers une revue approfondie des modèles prédictifs, en mettant l'accent sur les approches classiques telles que la régression logistique, mais aussi sur la Statistique des extrêmes, essentielle pour l'étude des cas rares. Les techniques de machine learning, notamment les algorithmes supervisés, ont également été présentés pour leur capacité à améliorer la détection des profils à haut risque.
Le deuxième chapitre a détaillé la méthodologie adoptée, depuis la préparation et l'exploration des données cliniques et démographiques jusqu'à la sélection des variables pertinentes et la mise en œuvre des différents modèles. Une attention particulière a été portée à la validation et à l'interprétation des résultats, afin de garantir la robustesse et la pertinence des prédictions.
Le troisième chapitre a présenté et discuté les résultats obtenus, mettant en évidence l'apport complémentaire des méthodes statistiques classiques et des techniques d'intelligence artificielle. Les principaux facteurs de risque d'AVC ont été identifiés : ***âge avancé, hypertension, antécédents cardiovasculaires, hyperglycémie, obésité***, ***tabac*** et ***type de travail***. L'étude a également soulignée l'importance de cibler les patients cumulant plusieurs de ces facteurs, car ils présentent une vulnérabilité accrue.

5.1. Recommandations

Les résultats de cette étude soulignent l'importance d'adopter des stratégies de prévention et d'intervention ciblées afin de réduire l'incidence et la gravité des accidents vasculaires cérébraux (AVC). Certaines populations présentent un rique nettement accru, justifiant une attention prioritaire. Il s'agit notamment des personnes âgées, particulièrement celles au-delà de 65 ans, ainsi que les patients souffrant d'hypertension artérielle, un facteur modifiable majeur. Les individus présentants des antécedants cardiovasculaires tels que la maladies cardiaque ainsi que les patients diabétiques ou présentant une hyperglycémie, doivent également bénéficier d'un suivi renforcé. Par ailleurs l'obésité constitue un facteur aggravant la vulnérabilité vasculaire justifant son inclusion dans les critères de ciblage.
Le risque est particulièrement aggravé chez les individus cumulant plusieurs de ces facteurs, par exemple l'âge avancé associé à l'hypertension ou l'hyperglycémie, Hypertension combinée à une maladie cardiaque, ou encore tabagisme avec d'autres comorbidités. Il est donc crucial de concentrer les effort de dépistage, de prévention et de suivi personnalisés sur ces populations vulnérables, afin de réduire la fréquence des AVC et d'améliorer la prise en charge des cas les

plus graves.

5.2. Perspectives de recherche future

Ce travail ouvre plusieurs pistes de prolongement pour approfonndir et élargir les résutats obtenus dans ce mémoire :

- **Approfondissement de la statistique des extrêmes avec la méthode POT**
  Aprés l'application de la loi généralisée des valeurs extrêmes (GEV) pour la modélisation des profils à très haut risque, il est pertinent d'explorer la méthode des excès au-dessus d'un seuil (Peak Over Threshold, POT), Cette approche permet d'analyser plus finement la distribution des valeurs en se concentrant sur les observations dépassant un certain seuil, offrant ainsi une meilleure estimation de probabilté d'événement rares et graves, comme les AVC survenant dans des conditions extrêmes d'âge, de glycémie ou d'IMC. L'utilisation conjointe de GEV et POT pourrait améliorer la détection et la caractérisation des patients les plus à risque.
- **Exploration des modèles de deep learning pour une précision accrue**
  Il serait intéressant de se tourner vers des modèles de deep learning (par exemple les réseaux de neurones profonds, des LSTM). Ces modèles, capables de traiter des volumes importants de données et de capturer des interactions complexes entre les variables, pourrait permettre d'améliorer encore la performance prédictive et d'identifier des patterns non detectés par les méthodes traditionnelles.
- **Intégration des méthodes d'imagerie médicale pour la détection précoce**
  Enfin, une perspective majeure réside dans l'intégration des données d'imagerie médicale (IRM, scanner cérébral) dans les modèles prédictifs. Les techniques de deep learning, notamment les réseaux de neurones convolutifs (CNN), sont particulièrement adaptées à l'analyse d'imagerie médicales et pourraient permettre une détection plus précoce et plus fiable des lésions cérébrales associées à l'AVC. Cette approche multimodale, combinant des données cliniques, démographique et imagerie, ouvrirait la voie à une modélisation plus complète et personnalisée du risque d'AVC.

En développant ces axes, la recherche future pourra renforcer la précision des modèles, améliorer la prévention et la prise en charge des patients à risque, et contribuer à une médecine prédictive et personnalisée dans le domaine des accidents vasculaires cérébraux.

Références

[1] Diop, A., Diop, A., Dème, E. H., & Sy, I. (2021). A case study of Stroke patients in Senegal : application of Generalized extreme value regression model. *African Journal of Applied Statistics*, 8(1), 1101–1110.

[2] Hbid, Y. (2021). *Modélisation et analyse prédictive des risques et des conséquences post accident vasculaire cérébral*. Thèse de doctorat, Sorbonne Université ; Université Cadi Ayyad (Maroc).

[3] Ozenne, B. (2015). *Modélisation statistique pour la prédiction du pronostic de patients atteints d'un AVC*. Thèse de doctorat, Université Claude Bernard-Lyon I.

[4] Léandre, C., & Com-Ruelle, L. (2019). *Repérer les facteurs de risque des patients hospitalisés pour un premier épisode d'AVC*. IRDES.

[5] Béjot, Y., Touzé, E., Jacquin, A., Giroud, M., & Mas, J. L. (2009). Épidémiologie des AVC. *Médecine/Sciences*, 25(8-9), 727–732.

[6] Biousse, V. (1994). Étiologies et mécanismes des accidents ischémiques cérébraux. *Annales de radiologie (Paris)*, 37(1-2), 11–16.

[7] Grillo, P., Velly, L., & Bruder, N. (2006). AVC hémorragique : nouveautés sur la prise en charge. *Annales françaises d'anesthésie et de réanimation*, 25(8), 868–873.

[8] World Health Organization. (2025). Accident vasculaire cérébral (AVC). ://www.emro.who.int/fr/health-topics/stroke-cerebrovascular-accident/index.html .

[9] Grimshaw, S. D. (1993). Computing maximum likelihood estimates for the generalized Pareto distribution. *Technometrics*, 35(2), 185–191.

[10] Hosking, J. R. M. (1990). L-moments : analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society. Series B*, 52(1), 105–124.

[11] Šimková, T. (2017). Statistical inference based on L-moments. *Statistika : Statistics and Economy Journal*, 97(1).

[12] Fisher, R. A., & Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24(2), 180–190.

[13] Gumbel, E. J. (1935). Les valeurs extrêmes des distributions statistiques. *Annales de l'institut Henri Poincaré*, 5(2), 115–158.

[14] Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of Mathematics*, 44(3), 423–453.

[15] Adlouni, S. E., & Ouarda, T. B. M. J. (2008). Comparaison des méthodes d'estimation des paramètres du modèle GEV non stationnaire. *Revue des sciences de l'eau*, 21(1), 35–50.

[16] Gilleland, E., & Katz, R. W. (2016). extRemes 2.0 : An extreme value analysis package in R. *Journal of Statistical Software*, 72, 1–39.

[17] Abdous, Y. C. F., Chebana, F., et al. (2019). Modélisation de pics sanitaires à l'aide de la théorie des valeurs extrêmes.

[18] Gardes, L. (2010). *Contributions à la théorie des valeurs extrêmes et à la réduction de dimension pour la régression*. Thèse de doctorat, Université Joseph-Fourier-Grenoble I.

[19] Chiu, Y. M. (2017). *Approches de modélisation des extrêmes dans l'étude des relations entre la santé et la météo*. Thèse de doctorat, INRS (Canada).

[20] Chiu, Y. M., Chebana, F., Abdous, B., et al. (2015). Modélisation des pics de mortalité et de morbidité hospitalière pour cause de maladies cardiovasculaires. *INRS, Centre Eau Terre Environnement*.

[21] Miranda, E. (2020). *Modélisation et caractérisation des risques extrêmes en fatigue des matériaux*. Thèse, Sorbonne Université.

[22] Oussama, B., & Meghlaoui, D. (2019). *Théorie des valeurs extrêmes : application au calcul de risques*. Thèse, Université des Frères Mentouri, Constantine.

[23] Doucet, E. (2014). Estimateurs à noyau et théorie des valeurs extrêmes. Université du Québec à Montréal.

[24] Tioguim, M. I., Delcaillau, M. D., et al. (2018). Modélisation d'extrêmes de séries temporelles : une étude empirique.

[25] Arbia, H., et al. *Modélisation statistique des valeurs extrêmes par la distribution GEV*. Thèse, University Kasdi Merbah Ouargla.

[26] Omrane, S. Estimation en théorie des valeurs extrêmes.

[27] Mbiakoup, O., & Nguyen, H. V. Approche multivariée de la théorie de la ruine : applications en assurance et risque alimentaire.

[28] Rinne, H. (2008). *The Weibull Distribution : A Handbook*. Chapman and Hall/CRC.

[29] Grus, J. (2020). *Data science par la pratique : fondamentaux avec Python*. Eyrolles.

[30] Lemberger, P., Batty, M., Morel, M., & Raffaëlli, J.-L. (2015). *Big Data et machine learning : Manuel du data scientist*. Dunod.

[31] Chollet, F. (2021). *Deep learning with Python*. Simon and Schuster.

[32] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.

[33] Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with Python*. O'Reilly Media, Inc.

[34] Friedman, J. H. (2001). Greedy function approximation : a gradient boosting machine. *Annals of Statistics*, 1189–1232.

[35] Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 7, 21.

[36] Ferraris, L., Liautaud, P., & Borel, É. (2022). Méthode de Gradient Boosting.

[37] Breiman, L. (2001). Random forests. *Machine learning*, 45, 5–32.

[38] Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, 3(3), 210–229.

[39] Morand, E. (2018). Biernat E. et Lutz M., 2017, Data science : fondamentaux et études de cas. *Population*, 73(2), 386–387.

[40] Marra, G., Calabrese, R., Osmetti, S. A. (2017). Package 'bgeva'.

[41] Dangeti, P. (2017). *Statistics for Machine Learning*. Packt Publishing Ltd.

[42] Soriano, F. (2021). Stroke Prediction Dataset. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset