ECONOMETRIE DES BIG DATA

ATALLAH Joanne & COUSTILLAC Célestine

Les maladies transmises sexuellement (MTS) sont un véritable enjeu de santé publique, à la fois sur le plan médical et social. Parmi elles, la gonorrhée se démarque par la facilité de son dépistage en laboratoire, mais aussi par sa fréquence élevée dans certains groupes de population.

Dans le cadre d'un programme de dépistage mené par des médecins en cabinet privé, des données ont été systématiquement recueillies lors des consultations. Cela permet de mieux comprendre les profils des personnes concernées, et surtout d'identifier celles qui sont les plus à risque.

Ce travail s'inscrit dans une approche concrète et appliquée. Il vise à fournir aux professionnels de santé des repères clairs pour **mieux cibler les publics prioritaires**.

I – TRAITEMENT DES DONNEES

I.I - Importation des données

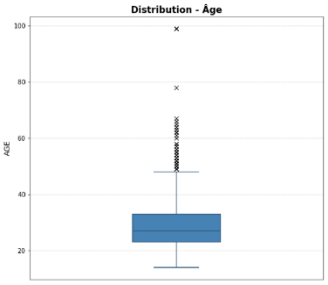
Les données analysées proviennent d'un **programme de dépistage de la gonorrhée** mené en médecine. Elles ont été importées sous format .csv et comprennent 3 144 observations décrivant des individus ayant consulté pour différents motifs. La base contient 13 variables à dominante catégorielle, incluant des informations socio-démographiques (sexe, âge, état-civil), comportementales (orientation sexuelle, nombre de partenaires, antécédents d'MTS), ainsi que le résultat du dépistage (DIAGN), qui constitue notre variable cible.

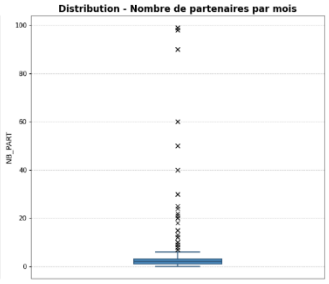
I.II - Nettoyage des données

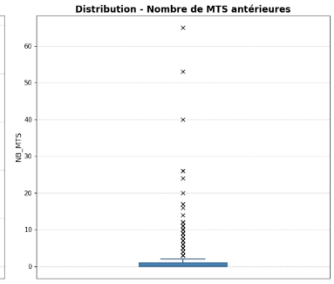
a. Détection des valeurs aberrantes

Dans un premier temps, afin de traiter les **valeurs extrêmes susceptibles de biaiser l'analyse**, nous avons considéré comme NA les observations déclarant plus de 30 infections sexuellement transmissibles antérieures ou 40 partenaires sexuels ou plus par mois. Cette neutralisation des extrêmes permet de préserver les individus dans l'échantillon, tout en différant leur traitement précis à une étape ultérieure dédiée à la gestion des valeurs manquantes. (Les données sont Skewed to the right)

	AGE	NB_PART	NB_MTS
count	3144.00	3143.00	3144.00
mean	28.58	5.21	1.21
std	8.41	15.02	2.56
min	14.00	0.00	0.00
25%	23.00	1.00	0.00
50%	27.00	2.00	1.00
75%	33.00	3.00	1.00
max	99.00	99.00	65.00







b. Traitement des valeurs manquantes

Par la suite, nous avons **identifié les valeurs devant être stipulée comme des valeurs manquantes**. Outres les cellules vides, certaines modalités spécifiques ont été considérées comme des valeurs manquantes : les valeurs 9 pour l'état civil, l'orientation sexuelle, MTS antérieur, la raison de la visite, l'histoire du contact, la culture et le diagnostic ainsi que les valeurs 99 pour l'âge et le nombre de partenaire.

Dans un second temps, nous avons procédé à la **correction des valeurs manquantes**. Les observations présentant plus de deux valeurs manquantes ont été supprimées car les observations avec un nombre important de valeurs manquantes apportent peu d'information exploitable, tout en risquant de fausser l'analyse si elles sont conservées. De même, la variable DIAGN étant notre variable cible, il n'était pas possible d'intégrer des individus sans information sur le diagnostic dans la phase de modélisation. Ainsi, nous avons également supprimer les observations pour lesquelles la variable DIAGN était absente.

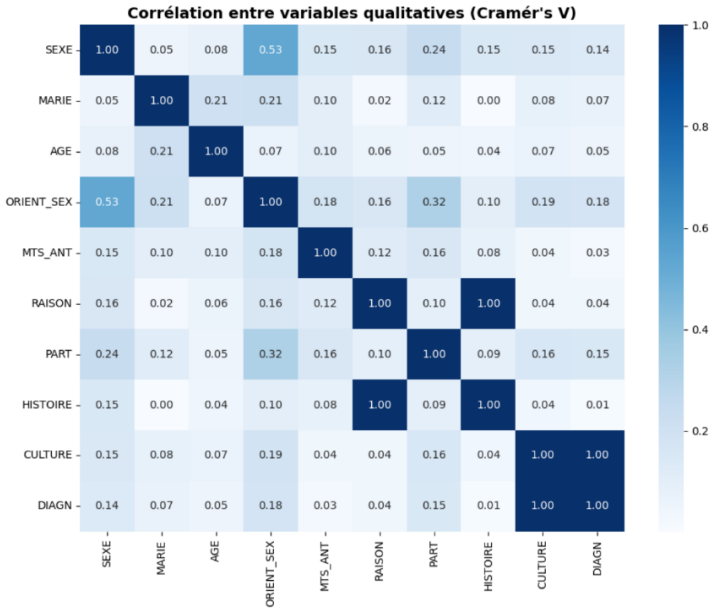
Enfin, les valeurs manquantes restantes ont été imputées par la médiane, une approche simple et robuste permettant de limiter l'influence des valeurs extrêmes, notamment dans des variables ordonnées.

I.III - Transformation des variables

Plusieurs transformations ont été appliquées aux variables afin de faciliter l'analyse statistique et d'assurer une interprétation cohérente dans les modèles.

Tout d'abord, l'âge des individus a été converti en variable binaire, distinguant les **personnes de moins de 30 ans de celles de 30 ans et plus**. Ce seuil permet d'analyser d'éventuelles différences de risque entre jeunes adultes et adultes plus âgés. Le nombre d'antécédents d'MTS a également été transformé en variable dichotomique : il s'agissait ici d'identifier simplement si la personne avait **déjà eu une MTS ou non**. De la même manière, le nombre de partenaires sexuels mensuel a été réduit à une variable binaire opposant les **personnes peu actives sexuellement (moins de trois partenaires) à celles considérées comme très actives (trois partenaires ou plus)**. Enfin, l'état civil a été recodé en une variable binaire MARIE, distinguant les **personnes mariées des non mariées**, incluant les individus célibataires, séparés/divorcés, veufs.

De plus, l'identifiant unique des individus **(variable ID) a été supprimé**, car il ne contenait aucune information utile à l'analyse statistique et ne servait qu'à l'identification administrative.

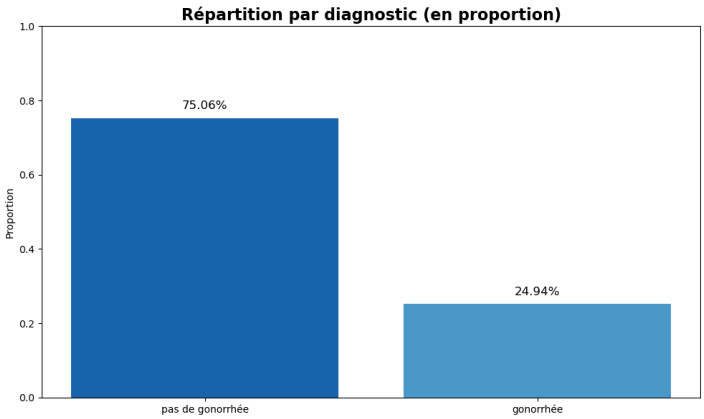


Enfin, à partir des V de Cramer, une mesure statistique utilisée pour évaluer la force de l'association entre deux variables nominales. Nous remarquons que deux couples de variables apparaissent comme fortement redondants, CULTURE et DIAGN, qui mesurent tous deux le résultat du dépistage de la gonorrhée, ainsi que RAISON et HISTOIRE, où la déclaration d'un partenaire contaminé est systématiquement associée à un motif de consultation pour contact. Ainsi, afin d'éviter la multi-colinéarité, nous avons décidé de **supprimer les variables CULTURE et RAISON** qui recode de manière quasiment équivalente DIAGN et HISTOIRE.

Ainsi, après traitement notre base est composée de **3 071 individus**.

I.IV - Équilibre de la variable cible

Enfin, nous avons examiné la distribution de la variable DIAGN. Celle-ci indique que **25 % des individus** présentent une gonorrhée (766 cas sur 3 071), contre 75 % de cas négatifs, cette répartition ne présente donc **pas de fort déséquilibre**.




```
DIAGN
0    2305
1     766
```

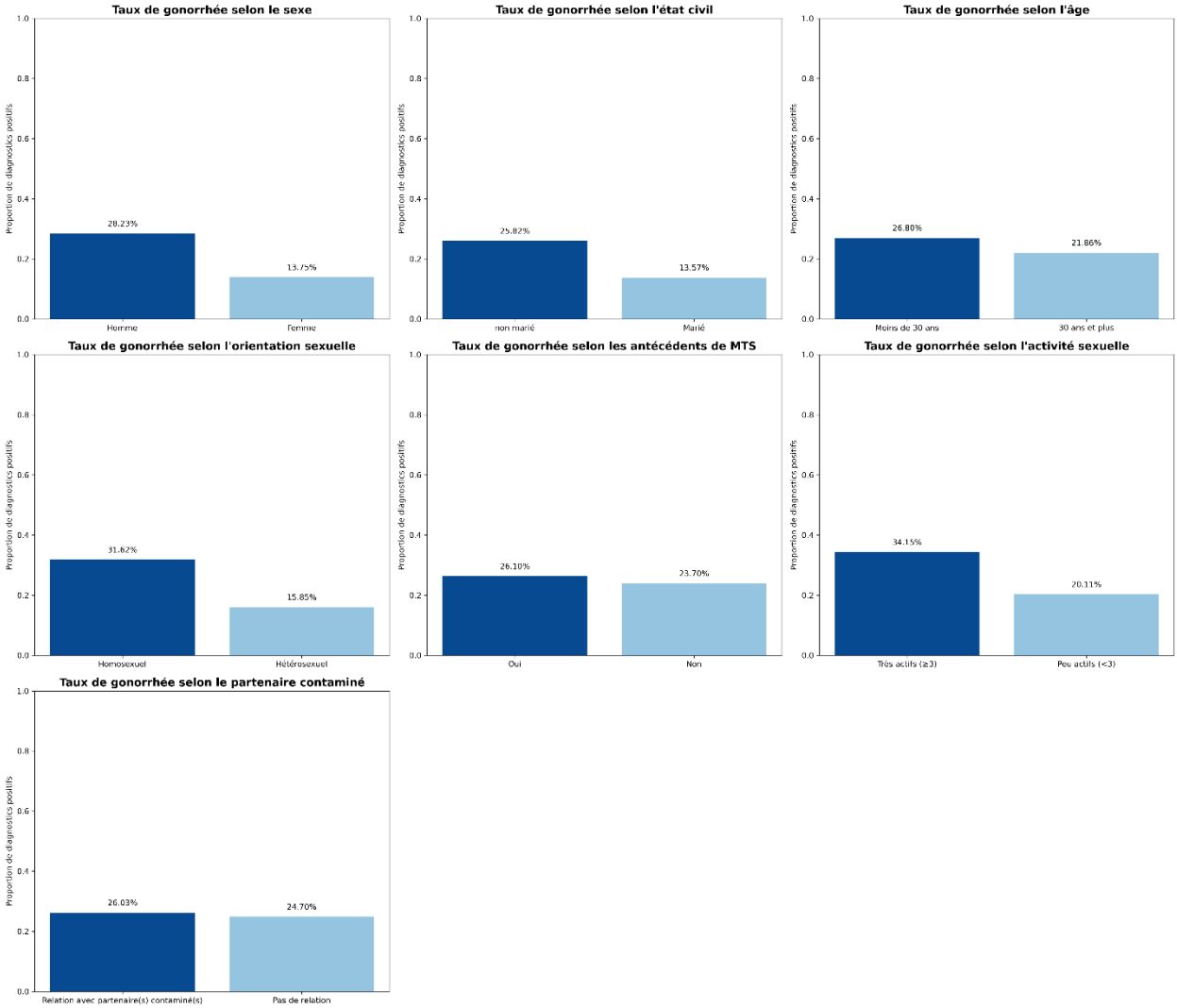
À l'issue de ces étapes de traitement, notre base de données ne contient plus de valeurs aberrantes ni de valeurs manquantes non traitées. L'échantillon final est ainsi constitué d'observations complètes, nettoyées et prêtes pour l'analyse statistique et la modélisation.

Après traitement, la base finale est composée de **3 071 individus** et comporte **8 variables**. Le tableau ci-dessous présente les principales variables de notre base de données, accompagnées de leur signification :

Variables	Codage
SEXE	0 = Homme, 1 = Femme
MARIE	0 = non marié, 1 = marié
AGE	0 = moins de 30 ans, 1 = 30 ans et plus
ORIENT_SEX	0 = homosexuel(le), 1 = hétérosexuel(le)
MTS_ANT	0 = aucune infection antérieure déclarée, 1 = au moins une infection antérieure
PART	0 = moins de 3 partenaires par mois , 1 = 3 partenaires ou plus par mois
HISTOIRE	0 = pas de relation, 1 = relation avec partenaire(s) contaminé(s) par MTS
DIAGN	0 = pas de gonorrhée, 1 = gonorrhée

II - ANALYSE DESCRIPTIVE DES DONNEES

Le graphique ci-dessous présente les **taux de diagnostic positif à la gonorrhée (DIAGN = 1) pour chaque modalité** des variables explicatives retenues. Ces taux sont calculés au sein de chaque sous-groupe, et permettent ainsi d'identifier les profils les plus exposés à la maladie.



On observe notamment des taux de gonorrhée plus élevés chez les **hommes** (28,23 % contre 13,75 % chez les femmes), les **homosexuels** (31,62 % contre 15,85 % chez les hétérosexuels), les individus **très actifs sexuellement** (34,15 % contre 20,11 % chez les peu actifs), ainsi que chez les personnes **de moins de 30 ans** (26,80 % contre 21,86 % chez les 30 ans et plus). Le statut marital joue également un rôle, avec un taux plus élevé chez les **non mariés** (25,82 %) comparé aux mariés (13,57 %). De même, les individus ayant déclaré des **antécédents de MTS** présentent un taux légèrement plus élevé (26,10 %) que ceux sans antécédents (23,70 %). Enfin, un taux un peu plus important est également observé chez ceux **ayant eu un partenaire contaminé** (26,03 %) comparé à ceux sans relation connue à risque (24,70 %).

Ces résultats confirment l'importance de certains facteurs comportementaux et sociodémographiques dans la probabilité de contracter la gonorrhée. En particulier, le sexe, l'orientation sexuelle, la fréquence des rapports, l'âge et la situation maritale semblent jouer un rôle significatif. Cette analyse descriptive fournit ainsi une base solide pour la construction d'un modèle prédictif, en mettant en évidence les variables les plus discriminantes.

III – ANALYSE PREDICTIVE

III.I – Choix des variables

Afin d'optimiser la performance de notre modèle prédictif tout en limitant la complexité, nous avons procédé à une **sélection rigoureuse des variables explicatives**.

a. Test du Chi²

	Variable	Chi2	p-value	ddl
0	SEXE	59.640	0.0000	1
1	ORIENT_SEX	98.802	0.0000	1
2	PART	72.281	0.0000	1
3	MARIE	15.791	0.0001	1
4	AGE	9.135	0.0025	1
5	MTS_ANT	2.234	0.1350	1
6	HISTOIRE	0.363	0.5468	1

Cette analyse met en évidence plusieurs associations statistiquement significatives (p-value < 0,05), suggérant que certaines variables sont fortement liées à la présence de la gonorrhée. C'est notamment le cas du **sexe**, de l'**orientation sexuelle**, l'**activité sexuelle** dans le mois précédent, le **statut marital** ainsi que de l'**âge**. Ces facteurs apparaissent ainsi comme des éléments potentiellement déterminants dans l'identification des groupes à risque.

À l'inverse, d'autres variables, comme antécédents de MTS ou encore le contact ou non avec un partenaire contaminé, ne présentent pas de lien statistiquement significatif avec le diagnostic. Leur p-value > 0,05, ces variables n'ont pas, dans notre échantillon, d'influence directe sur l'issue du test.

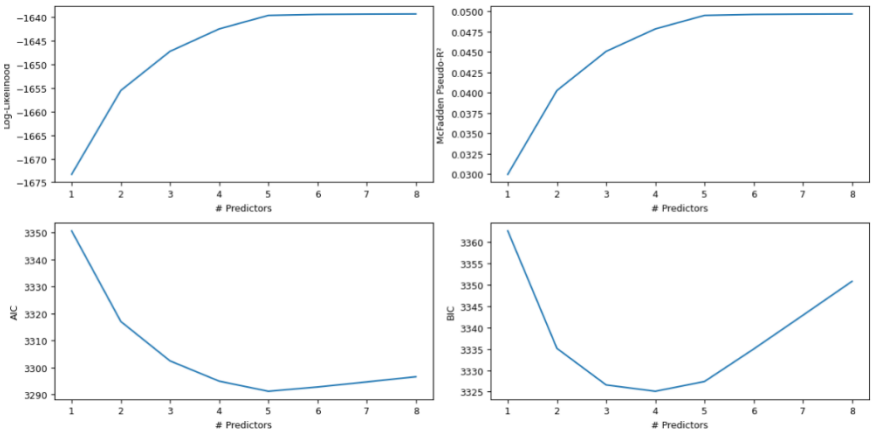
b. Best model

Afin d'identifier le meilleur sous-ensemble de variables explicatives, nous avons appliqué une **méthode de sélection fondée sur la log-vraisemblance dans le cadre d'un modèle de régression logistique**. Étant donné que notre variable cible (DIAGN) est binaire, la régression logistique a été privilégiée car contrairement à la régression linéaire, inadaptée aux variables qualitatives, la régression logistique permet de modéliser directement la probabilité d'un événement (ici, un diagnostic positif), tout en garantissant des prédictions comprises entre 0 et 1. Notre approche a ainsi consisté à tester toutes les combinaisons possibles de k variables, et à conserver, pour chaque valeur de k, le modèle présentant la log-vraisemblance maximale, c'est-à-dire le meilleur ajustement possible aux données.

Cette méthode a permis de comparer objectivement des centaines de modèles (2^p), tout en prenant en compte des indicateurs d'ajustement complémentaires comme le pseudo R^2, BIC et AIC.

Le meilleur compromis a été trouvé pour 5 variables, incluant notamment le **sexe**, l'**âge**, l'**orientation sexuelle**, l'**activité sexuelle** (PART) et le fait d'**avoir eu une relation avec un partenaire contaminé** (HISTOIRE).

Le modèle ainsi sélectionné présente un LLR p-value très significatif (< 0.001) et un pseudo R^2 de 0.0495 et le minimum AIC.



	Log-Likelihood
1	-1673.304659
2	-1655.519195
3	-1647.225163
4	-1642.467383
5	-1639.595859
6	-1639.388065
7	-1639.323587
8	-1639.286995

Les deux ensembles de variables proposés par les méthodes de sélection, le Chi² et le best model par log-vraisemblance, étaient quasi identiques, ne différant que par une seule variable : MARIE dans le premier cas, et HISTOIRE dans le second. Pour déterminer lequel de ces deux ensembles était le plus pertinent, nous les avons chacun testés au sein d'une régression logistique sur l'échantillon d'apprentissage.

Logit Regression Results

Dep. Variable:		DIAGN	No. Observations:			2149
Model:		Logit	Df Residuals:			2143
Method:		MLE	Df Model:			5
Date:	Tue, 08 Apr 2025		Pseudo R-squ.:			0.04076
Time:	01:24:03		Log-Likelihood:			-1154.7
converged:	True		LL-Null:			-1203.8
Covariance Type:	nonrobust		LLR p-value:			1.306e-19
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0540	0.313	-0.173	0.863	-0.667	0.559
AGE	-0.3297	0.114	-2.893	0.004	-0.553	-0.106
SEXE	-0.3496	0.164	-2.137	0.033	-0.670	-0.029
ORIENT_SEX	-0.5519	0.135	-4.099	0.000	-0.816	-0.288
PART	0.4823	0.109	4.435	0.000	0.269	0.695
MARIE	0.0620	0.187	0.332	0.740	-0.304	0.428

Dans l'ensemble issu du Chi² (SEXE, AGE, ORIENT_SEX, PART, MARIE), montrait une non-significativité de la variable MARIE (p-value > 0.05).

Logit Regression Results

Dep. Variable:		DIAGN	No. Observations:			2149
Model:		Logit	Df Residuals:			2143
Method:		MLE	Df Model:			5
Date:	Tue, 08 Apr 2025		Pseudo R-squ.:			0.04098
Time:	01:12:18		Log-Likelihood:			-1154.4
converged:	True		LL-Null:			-1203.8
Covariance Type:	nonrobust		LLR p-value:			1.008e-19
	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0171	0.214	0.080	0.936	-0.402	0.437
AGE	-0.3403	0.108	-3.146	0.002	-0.552	-0.128
SEXE	-0.3672	0.165	-2.230	0.026	-0.690	-0.044
ORIENT_SEX	-0.5617	0.131	-4.283	0.000	-0.819	-0.305
PART	0.4885	0.109	4.485	0.000	0.275	0.702
HISTOIRE	0.1089	0.135	0.807	0.420	-0.156	0.374

Dans l'ensemble issu du best model (SEXE, AGE, ORIENT_SEX, PART, HISTOIRE), révélait cette fois aussi une non-significativité d'HISTOIRE.

Ce qui suggère qu'aucune variable n'apporte d'information discriminante suffisante dans ce contexte. Pour éviter le sur-ajustement et renforcer la parcimonie du modèle, nous avons donc choisi de **ne retenir que les 4 variables les plus robustes**, présentes dans les deux approches et significatives tout en conservant un bon compromis entre interprétabilité et performance, à savoir : **l'âge, le sexe, l'orientation sexuelle, ainsi que l'activité sexuelle (PART).**

III.II – Prédiction

Pour construire notre modèle prédictif, nous avons commencé par **découper la base de données en deux sous-échantillons : un échantillon d'apprentissage (*train*) et un échantillon de test**, afin d'évaluer la capacité de généralisation du modèle sur des données non vues. Ce découpage permet de calibrer le modèle sur un ensemble et de tester sa performance réelle sur un autre, en s'assurant qu'il ne s'agit pas d'un surajustement aux données d'entraînement.

a. Régression logistique

Pour prédire la probabilité d'un diagnostic positif, nous commençons par ajuster un **modèle de régression logistique** sur notre échantillon d'apprentissage, en utilisant les variables explicatives sélectionnées précédemment. Voici le modèle obtenu :

Logit Regression Results

Dep. Variable:	DIAGN	No. Observations:	2149
Model:	Logit	Df Residuals:	2144
Method:	MLE	Df Model:	4
Date:	Tue, 08 Apr 2025	Pseudo R-squ.:	0.04072
Time:	00:37:44	Log-Likelihood:	-1154.8
converged:	True	LL-Null:	-1203.8
Covariance Type:	nonrobust	LLR p-value:	2.590e-20

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0220	0.214	0.103	0.918	-0.397	0.441
AGE	-0.3417	0.108	-3.160	0.002	-0.554	-0.130
SEXE	-0.3520	0.163	-2.155	0.031	-0.672	-0.032
ORIENT_SEX	-0.5623	0.131	-4.291	0.000	-0.819	-0.305
PART	0.4834	0.109	4.446	0.000	0.270	0.696

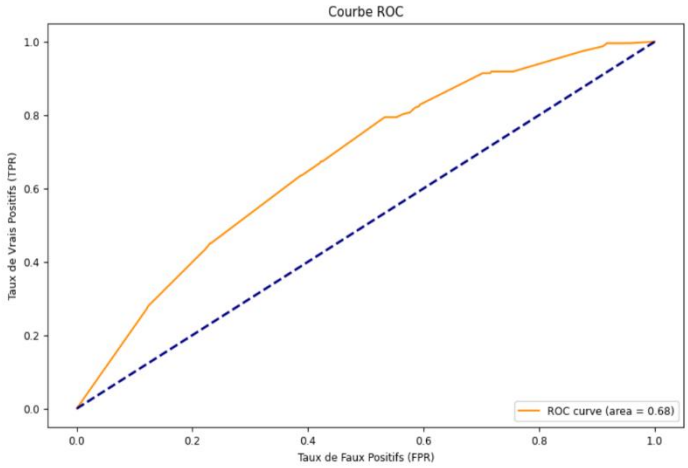
DataFrame contenant les seuils et les indicateurs :

	Threshold	Precision	Recall	F1 Score
0	0.20	0.302067	0.815451	0.440835
1	0.21	0.302067	0.815451	0.440835
2	0.22	0.302067	0.815451	0.440835
3	0.23	0.302067	0.815451	0.440835
4	0.24	0.302067	0.815451	0.440835
5	0.25	0.316804	0.493562	0.385906
6	0.26	0.316804	0.493562	0.385906
7	0.27	0.316804	0.493562	0.385906
8	0.28	0.316804	0.493562	0.385906
9	0.29	0.316804	0.493562	0.385906
10	0.30	0.316804	0.493562	0.385906
11	0.31	0.316804	0.493562	0.385906
12	0.32	0.316804	0.493562	0.385906
13	0.33	0.316804	0.493562	0.385906
14	0.34	0.431818	0.326180	0.371638
15	0.35	0.431818	0.326180	0.371638
16	0.36	0.431818	0.326180	0.371638
17	0.37	0.431818	0.326180	0.371638
18	0.38	0.431818	0.326180	0.371638
19	0.39	0.431818	0.326180	0.371638
20	0.40	0.431818	0.326180	0.371638
21	0.41	0.431818	0.326180	0.371638
22	0.42	0.433333	0.278970	0.339426
23	0.43	0.433333	0.278970	0.339426
24	0.44	0.433333	0.278970	0.339426
25	0.45	0.433333	0.278970	0.339426
26	0.46	0.433333	0.278970	0.339426
27	0.47	0.433333	0.278970	0.339426
28	0.48	0.433333	0.278970	0.339426
29	0.49	0.433333	0.278970	0.339426

Par défaut, la régression logistique considère qu'un individu est "positif" (malade) si la probabilité prédite dépasse 50 %. Cependant, dans notre contexte médical, ce seuil peut ne pas être optimal. En effet, il est parfois préférable de détecter davantage de cas positifs, même au risque de quelques faux positifs supplémentaires. Pour affiner cette décision, nous avons **testé différents seuils de classification** et, pour chacun, calculé trois indicateurs clés : la précision (*precision*) : la part des individus prédits positifs qui sont réellement positifs, le rappel (*recall*) : la part des vrais malades correctement identifiés par le modèle, ainsi que le F1-score qui représente la moyenne harmonique entre précision et rappel, qui mesure le compromis entre les deux.

Le F1-score étant particulièrement utile dans les contextes où l'on souhaite éviter à la fois les faux positifs et les faux négatifs, nous avons donc **retenu comme seuil optimal celui qui maximise le F1-score. Ce seuil s'est avéré être 22 %, ce qui signifie que nous considérons un individu comme malade dès que sa probabilité prédite dépasse 0,22**.

Nous avons ainsi évalué la performance globale de notre modèle en traçant la courbe ROC qui représente le compromis entre le taux de vrais positifs (sensibilité) et le taux de faux positifs. L'aire sous la courbe (AUC) constitue un indicateur synthétique de performance : plus elle est proche de 1, meilleur est le modèle.



Nous avons obtenu les résultats ci-contre. La courbe ROC obtenue présente une **aire sous la courbe (AUC) de 0.68**, ce qui indique une capacité de discrimination correcte mais perfectible. Autrement dit, le modèle parvient à distinguer les individus malades des non malades dans environ 68 % des cas. Ce résultat est meilleur que le hasard (AUC = 0.5), mais reste inférieur au seuil généralement recherché pour un bon modèle prédictif (souvent ≥ 0.8). Il suggère que le modèle capte une partie de l'information utile, mais pourrait être renforcé par un enrichissement des variables ou l'usage de méthodes plus complexes.

Afin d'évaluer la validité du modèle de régression logistique, nous avons appliqué le **test de Hosmer-Lemeshow** sur l'échantillon d'apprentissage. Ce test permet de comparer les valeurs observées aux valeurs prédites, en regroupant les individus par tranches de probabilité et en mesurant si les proportions de cas positifs prévues correspondent aux proportions effectivement observées. Il s'agit d'un outil standard pour tester la qualité de l'ajustement du modèle.

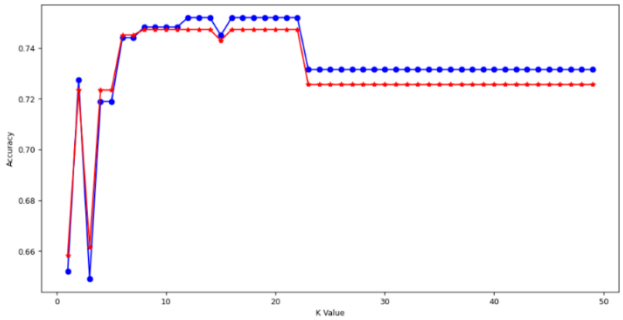
Dans notre cas, la statistique de test est de 6,360 avec une p-value associée de 0,3841. Cette p-value étant largement supérieure au seuil usuel de 5 % ($p > 0,05$), nous ne rejetons pas l'hypothèse nulle du test. Cela signifie que les prédictions du modèle sont cohérentes avec les observations, et que **le modèle présente un bon ajustement aux données**.

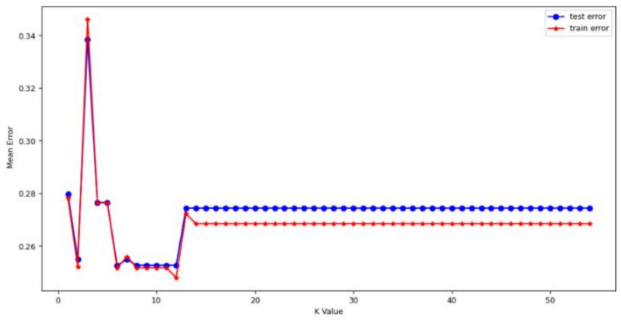
b. Modèle KNN

Par la suite, nous avons utilisé l'**algorithme des K plus proches voisins (KNN).** L'un des paramètres essentiels de ce modèle est le choix du nombre de voisins K, qui influence directement sa capacité de généralisation.

En effet, le KNN est un algorithme basé sur la proximité entre les observations : il attribue à un individu la classe majoritaire parmi ses k plus proches voisins dans l'espace des variables. Lorsque les variables sont binaires, les distances calculées entre individus sont simples, interprétables, et reflètent bien les différences de profils. Ce cadre est particulièrement adapté à des données de santé comportementale (sexe, âge, activité sexuelle, antécédents, etc.), où les groupes à risque peuvent être identifiés par similitude de réponses. Le KNN, ne supposant aucune forme fonctionnelle entre les variables et la cible, permet ainsi de détecter des structures locales ou non linéaires dans les données, ce qui peut être complémentaire à des approches plus paramétriques comme la régression logistique.

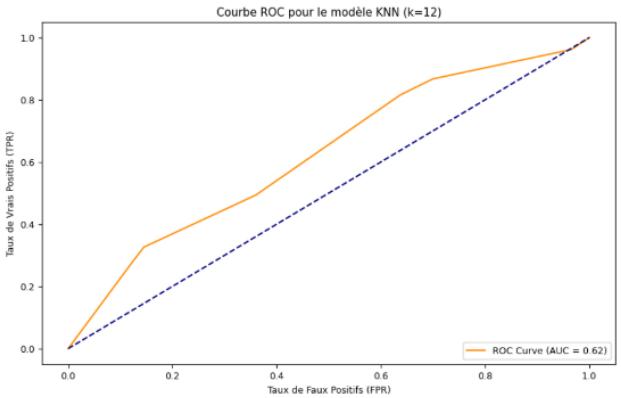
Pour évaluer les performances du modèle selon différentes valeurs de K comprises entre 1 et 55 (proche de la racine carrée de la taille de l'échantillon), nous avons analysé 2 indicateurs complémentaire : l'accuracy (proportion de bonne classification) et le taux d'erreur moyen. Ces indicateurs ont été calculées sur les ensembles d'apprentissage (en bleu) et de test (en rouge) afin de détecter d'éventuels cas surapprentissages (overfitting) ou de sous-apprentissage (underfitting).





Un bon compromis serait par exemple **K = 12**, qui assure une maximisation de l'accuracy tout en minimisant l'erreur.

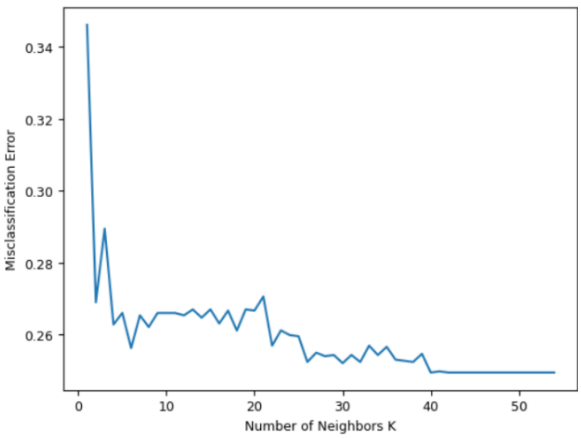
Nous obtenons ainsi la courbe ROC ci-contre avec le modèle KNN entraîné selon le découpage train/test classique, pour une valeur de K = 12. La **surface sous la courbe (AUC) est de 0.62**, ce qui indique une capacité de discrimination correcte, bien que légèrement inférieure à celle de la régression logistique (AUC = 0.68).

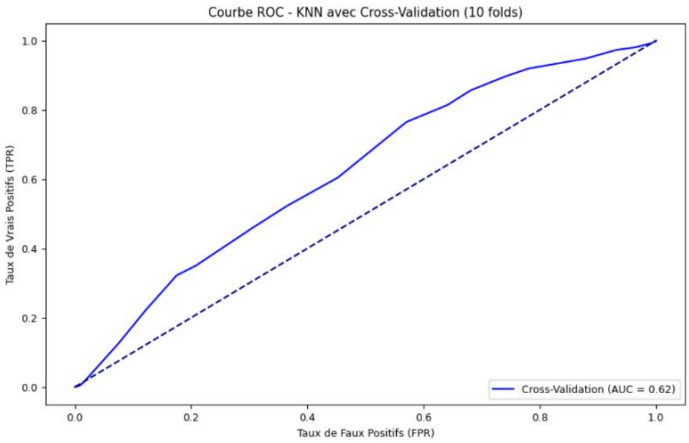


Pour nous assurer que nous avons bien retenu le meilleur modèle KNN, nous avons complété notre analyse en comparant les performances du modèle pour différentes valeurs de K, en utilisant des méthodes de validation croisée, telles que le **LOOCV** et la **10-fold cross-validation**.

Bien que la **validation croisée Leave-One-Out (LOOCV)** soit généralement recommandée pour les **petits échantillons**, il est tout à fait possible, et pertinent de l'appliquer à notre jeu de donnée étant donné que notre objectif est d'obtenir une estimation très précise de l'erreur de généralisation sans perdre trop de données dans les folds. LOOCV utilise n−1 observations pour l'entraînement à chaque itération, ce qui maximise l'utilisation des données disponibles et réduit le biais d'estimation. De son côté, la validation croisée 10-fold, bien que moins coûteuse en calcul, donne une estimation de l'erreur avec un peu plus de variance mais reste largement utilisée en pratique pour des raisons de performance et de stabilité. Dans les deux cas, avec 3071 observations, le volume est suffisant pour garantir une bonne représentativité, et le choix entre LOOCV ou 10-fold dépend alors davantage des ressources computationnelles disponibles et du niveau de précision souhaité plutôt que d'une contrainte liée à la taille de l'échantillon.

Dans le cas de la **10-fold cross-validation**, le choix de K a été guidé par la minimisation de l'erreur de mauvaise classification (*misclassification error*), comme illustré sur la courbe ci-dessous.



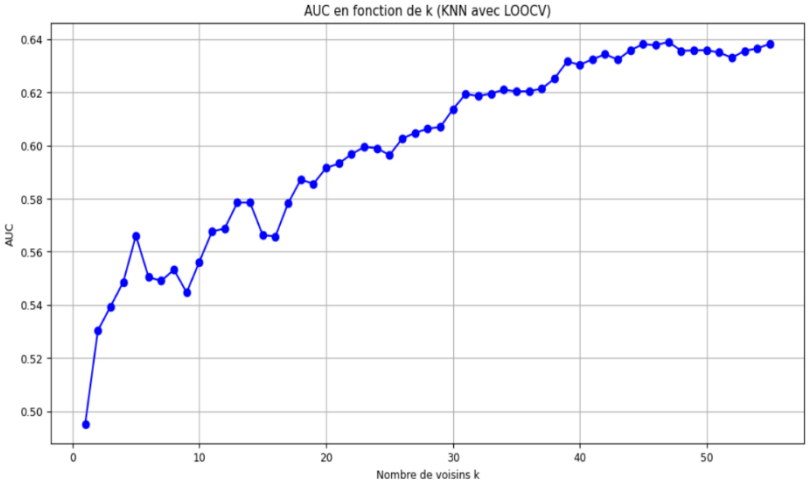


Ce processus nous a permis d'identifier K = 39 comme valeur optimale, pour laquelle nous avons ensuite analysé la courbe ROC afin d'évaluer plus finement sa capacité de discrimination.

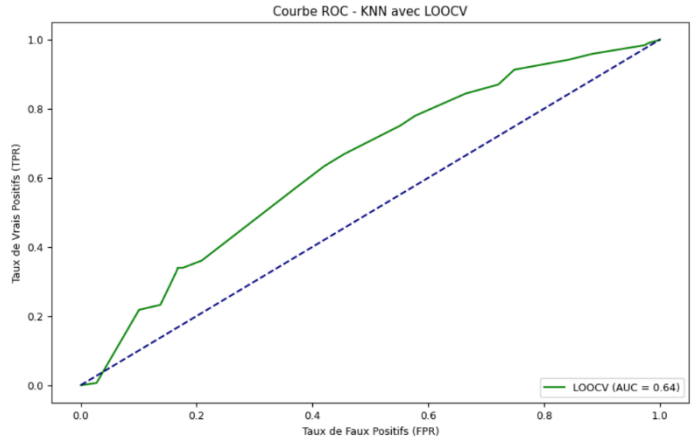
L'**aire sous la courbe (AUC) obtenue est de 0.62**, ce qui reflète une capacité de discrimination modérée. Cela signifie que, dans 62 % des cas, le modèle parvient à classer correctement un individu malade devant un individu non malade tiré au hasard. Cette performance reste équivalente à celle obtenue avec le découpage train/test classique.

Enfin, pour aller plus loin dans l'optimisation du modèle KNN, nous avons appliqué une **validation croisée Leave-One-Out (LOOCV)**, permettant d'évaluer la performance du modèle de façon très précise, en testant chaque observation à tour de rôle. Pour chaque valeur de K entre 1 et 55, nous avons calculé l'aire sous la courbe ROC (AUC), métrique de référence pour mesurer la capacité de discrimination d'un classifieur, indépendamment du seuil de classification choisi.

Les résultats, représentés dans le graphique à la page suivante, montrent une progression régulière de l'AUC avec l'augmentation de K, jusqu'à atteindre un maximum de 0.6389 pour K = 47.




```
k = 1, AUC = 0.4952
k = 2, AUC = 0.5304
k = 3, AUC = 0.5394
k = 4, AUC = 0.5486
k = 5, AUC = 0.5660
k = 6, AUC = 0.5504
k = 7, AUC = 0.5491
k = 8, AUC = 0.5532
k = 9, AUC = 0.5446
k = 10, AUC = 0.5562
k = 11, AUC = 0.5676
k = 12, AUC = 0.5687
k = 13, AUC = 0.5785
k = 14, AUC = 0.5785
k = 15, AUC = 0.5662
k = 16, AUC = 0.5658
k = 17, AUC = 0.5782
k = 18, AUC = 0.5871
k = 19, AUC = 0.5856
k = 20, AUC = 0.5915
k = 21, AUC = 0.5932
k = 22, AUC = 0.5967
k = 23, AUC = 0.5994
k = 24, AUC = 0.5989
k = 25, AUC = 0.5963
k = 26, AUC = 0.6026
k = 27, AUC = 0.6047
k = 28, AUC = 0.6063
k = 29, AUC = 0.6069
k = 30, AUC = 0.6136
k = 31, AUC = 0.6193
k = 32, AUC = 0.6186
k = 33, AUC = 0.6195
k = 34, AUC = 0.6209
k = 35, AUC = 0.6203
k = 36, AUC = 0.6204
k = 37, AUC = 0.6213
k = 38, AUC = 0.6251
k = 39, AUC = 0.6315
k = 40, AUC = 0.6303
k = 41, AUC = 0.6324
k = 42, AUC = 0.6343
k = 43, AUC = 0.6323
k = 44, AUC = 0.6358
k = 45, AUC = 0.6381
k = 46, AUC = 0.6377
k = 47, AUC = 0.6389
k = 48, AUC = 0.6355
k = 49, AUC = 0.6358
k = 50, AUC = 0.6358
k = 51, AUC = 0.6349
k = 52, AUC = 0.6330
k = 53, AUC = 0.6356
k = 54, AUC = 0.6365
k = 55, AUC = 0.6383
```



La courbe ROC ci-contre correspond au modèle KNN évalué par validation croisée LOOCV avec K = 47, valeur pour laquelle l'**AUC atteint 0.64**. Ce résultat confirme que cette configuration constitue la meilleure combinaison testée dans le cadre du KNN, en offrant la plus grande capacité de discrimination parmi l'ensemble des modèles évalués.

c. Arbre de décisions

Pour construire notre modèle d'arbre de décision, nous avons eu recours à une **optimisation des hyperparamètres via validation croisée (GridSearchCV)**. Cette méthode nous a permis d'identifier la combinaison la plus performante selon le critère de précision (accuracy). Les paramètres optimaux retenus sont les suivants : criterion = gini, max_depth = None, max_leaf_nodes = 5, min_samples_leaf = 1, et min_samples_split = 2. Afin de conserver un équilibre entre performance et lisibilité, nous avons néanmoins décidé de fixer manuellement la profondeur maximale de l'arbre à 4. Ce choix repose sur le constat que la précision du modèle reste stable à partir de ce niveau, tout en réduisant le risque de surapprentissage et en facilitant l'interprétation des résultats.

Pour évaluer la performance de l'arbre de décision entraîné avec les hyperparamètres optimaux, on peut analyser la classification report ci-dessus :

```
[[337 352]
 [ 53 180]]
              precision    recall  f1-score   support

           0       0.86      0.49      0.62       689
           1       0.34      0.77      0.47       233

    accuracy                           0.56       922
   macro avg       0.60      0.63      0.55       922
weighted avg       0.73      0.56      0.59       922

Accuracy : 0.5607
```

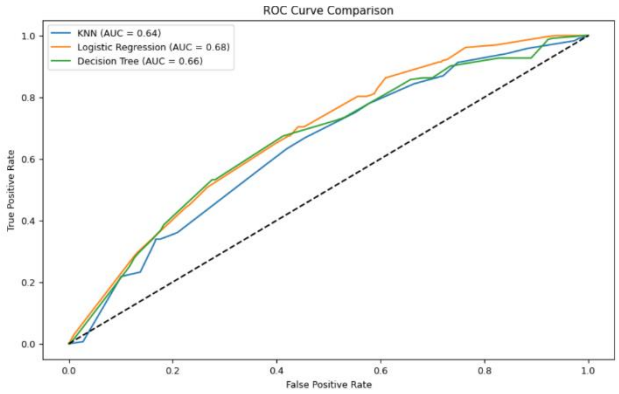
Le modèle atteint une **accuracy globale de 56%**, ce qui reste relativement modéré. Il prédit correctement les individus sans gonorrhée (classe 0) avec une **précision élevée de 86%**, mais présente un **rappel faible de 49%**, ce qui signifie qu'il ne détecte qu'environ la moitié des vrais cas négatifs. En revanche, pour les personnes atteintes de gonorrhée (classe 1), le **rappel est élevé (77%)**, ce qui montre que l'arbre parvient à identifier une grande partie des cas positifs, mais au prix d'une **précision faible (34%)**, traduisant un nombre important de faux positifs. Le **f1-score** pour la classe 1 est de **0.47**, ce qui reste acceptable dans un contexte de santé publique où **le rappel est souvent prioritaire afin de limiter les cas non détectés**. En somme, ce modèle semble favoriser la **sensibilité (rappel)** au détriment de la **spécificité**, ce qui peut être un **choix stratégique pertinent** si l'objectif est de ne pas rater de cas positifs, quitte à générer davantage de fausses alertes.

L'arbre obtenu permet une lecture intuitive des règles de classification, en identifiant des profils distincts selon l'âge, le sexe, l'orientation sexuelle, l'activité sexuelle ou encore l'historique de partenaires contaminés. La représentation graphique complète de cet arbre est présentée en annexe, afin d'en faciliter la consultation sans alourdir le contenu principal du rapport.

Son analyse nous a permis d'identifier plusieurs profils d'individus à risque variable de contracter la gonorrhée. L'orientation sexuelle apparaît comme le premier critère discriminant, les personnes **homosexuelles** étant globalement plus exposées. Au sein de ce groupe, le danger est particulièrement marqué chez les jeunes **hommes** très actifs sexuellement, c'est-à-dire de **moins de 30 ans et ayant trois partenaires ou plus par mois**. Ce sous-groupe présente une proportion importante de diagnostics positifs. À l'opposé, les femmes hétérosexuelles, âgées de plus de 30 ans, peu actives sexuellement, apparaissent comme les moins exposées, avec une majorité de cas négatifs. Cette segmentation claire permet de visualiser les règles de décision du modèle et de dégager des groupes cibles prioritaires pour les actions de prévention ou de dépistage.

III.III – Comparaison des différents modèles

Pour poursuivre notre analyse, il était essentiel de comparer les trois modèles testés (KNN, régression logistique, arbre de décision) afin de déterminer lequel offrait la meilleure capacité de discrimination. Le critère retenu est l'aire sous la courbe ROC (AUC). Plus l'AUC est élevée, plus le modèle est performant.



Comme l'illustre la courbe ci-contre, **le modèle de régression logistique présente la meilleure AUC (0.68)**, contre 0.66 pour l'arbre de décision et 0.64 pour le KNN. Nous avons donc retenu la régression logistique comme modèle final pour la suite de l'analyse, notamment pour la caractérisation des individus à risque à travers l'interprétation de ses coefficients.

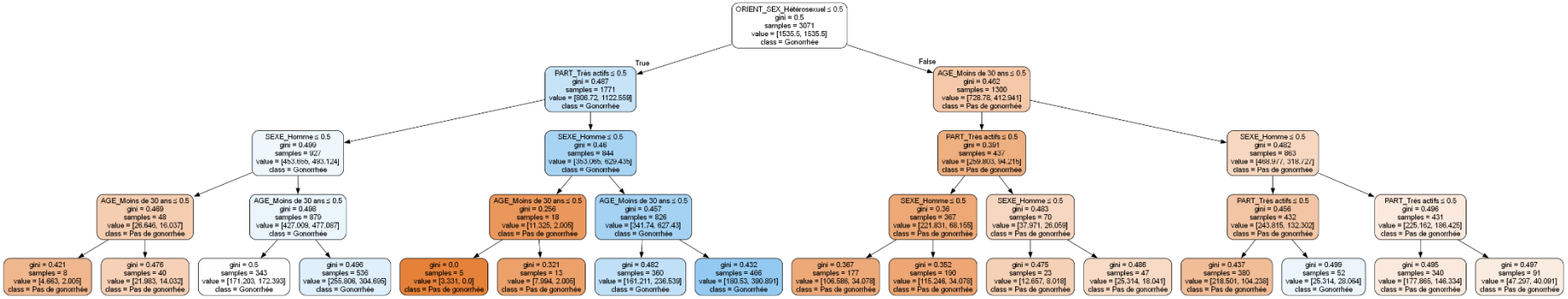
IV - CARACTERISATION DES GROUPES A RISQUE

L'analyse des résultats obtenus par la régression logistique permet d'**identifier les caractéristiques des individus les plus à risque en se basant sur les Odds Ratios** et les probabilités calculées pour chaque combinaison de variables.

Résultats pour chaque combinaison :

	AGE	SEXE	ORIENT_SEX	PART	Log-Odds	Odds Ratio	Probability
0	0	0	0	0	0.021981	1.022224	0.505495
1	0	0	0	1	0.505345	1.657557	0.623715
2	0	0	1	0	-0.540278	0.582586	0.368123
3	0	0	1	1	-0.056914	0.944676	0.485775
4	0	1	0	0	-0.329985	0.718934	0.418244
5	0	1	0	1	0.153380	1.165767	0.538270
6	0	1	1	0	-0.892244	0.409735	0.290647
7	0	1	1	1	-0.408879	0.664394	0.399181
8	1	0	0	0	-0.319711	0.726359	0.420746
9	1	0	0	1	0.163654	1.177807	0.540822
10	1	0	1	0	-0.881969	0.413967	0.292770
11	1	0	1	1	-0.398605	0.671256	0.401648
12	1	1	0	0	-0.671676	0.510852	0.338122
13	1	1	0	1	-0.188312	0.828356	0.453061
14	1	1	1	0	-1.233935	0.291145	0.225493
15	1	1	1	1	-0.750571	0.472097	0.320697

L'analyse des Odds Ratios montre que **les groupes présentant les risques les plus élevés sont principalement les jeunes hommes homosexuels actifs sexuellement** (AGE=0, SEXE=0, ORIENT_SEX=0, PART=1) avec un Odds Ratio de **1.66** et une probabilité associée de **62.4%**, ce qui signifie qu'ils ont environ **66% plus de chances** d'être diagnostiqués positifs par rapport à la catégorie de référence (peu actifs sexuellement). À l'inverse, les probabilités les plus faibles sont observées chez les femmes âgées hétérosexuelles peu actives sexuellement (AGE=1, SEXE=1, ORIENT_SEX=1, PART=0) avec un Odds Ratio de **0.29** et une probabilité de **22.5%**, indiquant un risque significativement réduit. De manière générale, une activité sexuelle élevée (PART = 1) augmente les Odds Ratios pour la majorité des combinaisons, suggérant que l'activité sexuelle est un facteur de risque majeur. Par exemple, pour un jeune homme homosexuel (AGE=0, SEXE=0, ORIENT_SEX=0), la probabilité passe de **50.5%** (PART=0) à **62.4%** (PART=1). L'orientation hétérosexuelle (ORIENT_SEX=1) semble être associée à un risque réduit par rapport aux homosexuels, sauf dans certaines configurations spécifiques. L'âge (AGE=1) agit globalement comme un facteur protecteur, réduisant systématiquement les probabilités de diagnostic positif par rapport aux jeunes (AGE=0).


ORIENT_SEX_Hétérosexuel ≤ 0.5
gini = 0.5
samples = 3071
value = [1535.5, 1535.5]
class = Gonorrhée
True
False
PART_Très actifs ≤ 0.5
gini = 0.487
samples = 1771
value = [806.72, 1122.559]
class = Gonorrhée
AGE_Moins de 30 ans ≤ 0.5
gini = 0.462
samples = 1300
value = [728.78, 412.941]
class = Pas de gonorrhée
SEXE_Homme ≤ 0.5
gini = 0.499
samples = 927
value = [453.665, 493.124]
class = Gonorrhée
SEXE_Homme ≤ 0.5
gini = 0.46
samples = 844
value = [353.065, 629.435]
class = Gonorrhée
PART_Très actifs ≤ 0.5
gini = 0.391
samples = 437
value = [259.803, 94.215]
class = Pas de gonorrhée
SEXE_Homme ≤ 0.5
gini = 0.482
samples = 863
value = [468.977, 318.727]
class = Pas de gonorrhée
AGE_Moins de 30 ans ≤ 0.5
gini = 0.469
samples = 48
value = [26.646, 16.037]
class = Pas de gonorrhée
AGE_Moins de 30 ans ≤ 0.5
gini = 0.498
samples = 879
value = [427.009, 477.087]
class = Gonorrhée
AGE_Moins de 30 ans ≤ 0.5
gini = 0.256
samples = 18
value = [11.325, 2.005]
class = Pas de gonorrhée
AGE_Moins de 30 ans ≤ 0.5
gini = 0.457
samples = 826
value = [341.74, 627.43]
class = Gonorrhée
SEXE_Homme ≤ 0.5
gini = 0.36
samples = 367
value = [221.831, 68.155]
class = Pas de gonorrhée
SEXE_Homme ≤ 0.5
gini = 0.483
samples = 70
value = [37.971, 26.059]
class = Pas de gonorrhée
PART_Très actifs ≤ 0.5
gini = 0.456
samples = 432
value = [243.815, 132.302]
class = Pas de gonorrhée
PART_Très actifs ≤ 0.5
gini = 0.496
samples = 431
value = [225.162, 186.425]
class = Pas de gonorrhée
gini = 0.421
samples = 8
value = [4.663, 2.005]
class = Pas de gonorrhée
gini = 0.476
samples = 40
value = [21.983, 14.032]
class = Pas de gonorrhée
gini = 0.5
samples = 343
value = [171.203, 172.393]
class = Gonorrhée
gini = 0.496
samples = 536
value = [255.806, 304.695]
class = Gonorrhée
gini = 0.0
samples = 5
value = [3.331, 0.0]
class = Pas de gonorrhée
gini = 0.321
samples = 13
value = [7.994, 2.005]
class = Pas de gonorrhée
gini = 0.482
samples = 360
value = [161.211, 236.539]
class = Gonorrhée
gini = 0.432
samples = 466
value = [180.53, 390.891]
class = Gonorrhée
gini = 0.367
samples = 177
value = [106.586, 34.078]
class = Pas de gonorrhée
gini = 0.352
samples = 190
value = [115.246, 34.078]
class = Pas de gonorrhée
gini = 0.475
samples = 23
value = [12.657, 8.018]
class = Pas de gonorrhée
gini = 0.486
samples = 47
value = [25.314, 18.041]
class = Pas de gonorrhée
gini = 0.437
samples = 380
value = [218.501, 104.238]
class = Pas de gonorrhée
gini = 0.499
samples = 52
value = [25.314, 28.064]
class = Gonorrhée
gini = 0.495
samples = 340
value = [177.865, 146.334]
class = Pas de gonorrhée
gini = 0.497
samples = 91
value = [47.297, 40.091]
class = Pas de gonorrhée