Université de Bordeaux

Master Mathématiques Appliquées et Statistiques

Parcours Ingénierie des Risques Économiques et Financiers

Devoir maison

Analyse des Facteurs de Risque et Caractérisation des Groupes Associés à la Gonorrhée

---

Réalisé par : **Karama Fabarika Tiebele**

Douramane Moussa Barkiré

Encadrant : **Ursu Eugen**

Année académique 2024–2025

Introduction

Les maladies sexuellement transmissibles (MST) constituent un défi majeur en santé publique, tant par leur fréquence que par les complications graves qu'elles peuvent engendrer. Parmi ces infections, la gonorrhée se distingue non seulement par sa prévalence élevée, mais aussi par le fait qu'elle peut être diagnostiquée simplement et efficacement à l'aide d'un test de laboratoire. Toutefois, malgré la disponibilité de ces outils diagnostiques, plusieurs obstacles – notamment les préjugés sociaux et les comportements à risque – entravent le dépistage optimal de cette infection.

Dans le cadre d'un programme ciblé, des données issues de consultations effectuées par des médecins de pratique privée ont été recueillies afin d'explorer et de comprendre la dynamique de transmission de la gonorrhée. Ces données intègrent des informations démographiques (âge, sexe, orientation sexuelle), ainsi que des variables comportementales telles que le nombre de partenaires récents et l'historique des MST. Cette collecte vise à mettre en lumière les facteurs de risque les plus significatifs et à définir les groupes de population présentant une vulnérabilité accrue.

L'objectif principal de ce projet est triple :

1. **Identifier les facteurs de risque prédominants** en lien avec un diagnostic positif de gonorrhée.
2. **Caractériser les groupes à risque** afin d'orienter efficacement les stratégies de dépistage et de prévention.
3. **Formuler des recommandations opérationnelles** pour améliorer la prise en charge et la prévention de cette infection.

Pour atteindre ces objectifs, nous avons mis en œuvre une approche analytique reposant sur des techniques d'apprentissage automatique. Plus précisément, la régression logistique, les arbres de décision et le KNN (K-plus proches voisins) ont été exploités pour modéliser les relations entre les différentes variables et le diagnostic final. Ce rapport détaille l'ensemble des étapes suivies – depuis le prétraitement des données jusqu'à l'optimisation des modèles – et présente les résultats ainsi que les recommandations en vue d'une meilleure stratégie de dépistage ciblé.

I Description des données

Les données utilisées dans ce projet proviennent d'un programme de dépistage de la gonorrhée mené auprès de patients examinés par des médecins de pratique privée. Elles contiennent des informations démographiques, comportementales et médicales permettant d'identifier les facteurs de risque associés à un diagnostic positif. Le jeu de données comprend **3144 observations** et **12 variables** initiales, dont la description est détaillée ci-dessous.

I.1 Description des variables

- **ID** : Identifiant unique pour chaque patient (non utilisé dans l'analyse car il n'a pas de valeur explicative).
- **SEXE** : Sexe du patient (1 = Homme, 2 = Femme).
- **AGE** : Âge du patient (en années).
- **ORIENT_SEX** : Orientation sexuelle (1 = Hétérosexuel, 2 = Homosexuel).
- **MTS_ANT** : Antécédents de MST (0 = Non, 1 = Oui).
- **NB_MTS** : Nombre de MST antérieures.
- **RAISON** : Raison de la visite (variable catégorique).
- **NB_PART** : Nombre de partenaires.
- **HISTOIRE** : Historique médical pertinent (variable catégorique).
- **CULTURE** : Résultat de la culture.
- **DIAGN** : Diagnostic final de gonorrhée (Variable Cible : 1 = Positif, 0 = Négatif).

Les données comportent également des valeurs manquantes indiquées par certains codes spécifiques. En particulier, pour certaines variables, la valeur `9` est utilisée pour représenter une donnée manquante, tandis que pour d'autres, c'est la valeur `99`. Ces valeurs ont été traitées lors de l'étape de prétraitement pour garantir la qualité des analyses.

II Méthodologie

L'approche adoptée dans ce projet se décompose en plusieurs étapes afin d'assurer la rigueur et la robustesse de l'analyse. Les principales étapes sont le prétraitement des données, l'analyse exploratoire, la modélisation, la validation croisée, la sélection des variables
.

II.1 Prétraitement des données

- **Gestion des valeurs manquantes :** Les valeurs manquantes, indiquées par les codes `9` et `99` pour certaines variables, ont été remplacées par `NaN`. L'imputation a été réalisée par la moyenne pour les variables numériques et par le mode pour les variables catégoriques.
- **Feature Engineering :** Pour enrichir l'analyse, de nouvelles variables ont été créées :
  - `AGE_GROUP` : Catégorisation de l'âge en deux groupes (< 30 ans et ≥ 30 ans).
  - `MTS_ANT_GROUP` : Transformation de la variable `NB_MTS` pour distinguer entre les patients ayant déjà eu des MST et ceux qui n'en ont pas eu.
  - `PARTNER_GROUP` : Regroupement des patients en fonction du nombre de partenaires (≤ 1 partenaire pour les peu actifs et > 1 partenaire pour les très actifs).

— **Exclusion de variables :** Les variables `ID` et `CULTURE` ont été exclues de l'analyse, la seconde pour éviter le risque de fuite de données (data leakage) due à sa forte corrélation avec la variable cible.

Valeurs extrêmes : Pour traiter les valeurs aberrantes dans les variables numériques (`NB_PART` et `NB_MTS`), nous avons appliqué un écrêtage basé sur l'écart interquartile (IQR). Cette méthode consiste à identifier les valeurs situées en dehors des bornes définies par $Q_1 - 1.5 \times \text{IQR}$ et $Q_3 + 1.5 \times \text{IQR}$, puis à ramener ces valeurs aux seuils correspondants, ce qui permet de préserver la taille de l'échantillon. Pour la variable `NB_MTS`, les valeurs obtenues après écrêtage ont été arrondies à l'entier inférieur afin de respecter la nature discrète de cette variable.

II.2 Analyse exploratoire des données (EDA)

L'analyse exploratoire vise à comprendre la distribution des données et à identifier les relations potentielles entre les variables. Plusieurs actions ont été menées :

— **Calcul des statistiques descriptives (moyenne, min,max écart-type, etc.)** pour les variables numériques telles que `AGE`, `NB_PART` et `NB_MTS`.
— **Visualisations** à l'aide d'histogrammes, de diagrammes en barres et de boxplots pour examiner la distribution des variables et leur relation avec la variable cible `DIAGN`.

	AGE	NB_MTS	NB_PART
count	3137.000000	3137.000000	3137.000000
mean	28.449377	0.764106	2.444008
std	7.821566	0.819765	1.596259
min	14.000000	0.000000	0.000000
25%	23.000000	0.000000	1.000000
50%	27.000000	1.000000	2.000000
75%	32.000000	1.000000	3.000000
max	78.000000	2.000000	6.000000

FIGURE 1 – Statistiques descriptives pour `AGE`, `NB_MTS` et `NB_PART`.

Comme le montre la Figure 1, la variable `AGE` présente une moyenne d'environ 28 ans, tandis que `NB_MTS` (nombre de MST antérieures) et `NB_PART` (nombre de partenaires) ont également été évaluées. Ces informations offrent un premier aperçu de la tendance centrale

et de la dispersion des données, et constituent une base pour mieux cibler les analyses ultérieures.

II.3 Modélisation

Trois modèles d'apprentissage automatique ont été utilisés pour prédire le diagnostic de gonorrhée (`DIAGN`) :

1. **Régression logistique** : Un modèle linéaire qui offre une interprétabilité directe grâce aux coefficients, interprétés sous forme de log-odds.
2. **Arbre de décision** : Un modèle non linéaire qui segmente l'espace des caractéristiques à l'aide de règles conditionnelles et qui permet d'évaluer l'importance relative des variables.
3. **K-plus proches voisins (KNN)** : Un modèle non paramétrique qui prédit la classe d'une observation selon la majorité de ses voisins dans l'espace des caractéristiques, nécessitant une normalisation préalable des variables.

II.4 Validation croisée

Pour estimer de manière robuste la performance des modèles, une validation croisée à 5 plis ($k = 5$) a été mise en place. Cette méthode consiste à :

— Diviser le jeu de données en 5 sous-ensembles de taille équivalente.
— Entraîner le modèle sur 4 plis et le tester sur le 5ème, en répétant l'opération pour chaque pli.
— Calculer une mesure de performance moyenne (comme l'accuracy) pour obtenir une estimation fiable de la généralisation du modèle.

III Résultats

Cette section présente les résultats obtenus à partir des analyses exploratoires et des modèles prédictifs entraînés sur les données prétraitées. Les performances des modèles sont comparées, et les facteurs de risque les plus significatifs pour la gonorrhée sont identifiés.

III.1 Analyse des associations bivariées (Test du Khi-deux)

Avant de procéder à la modélisation multivariée, une analyse bivariée a été réalisée pour explorer l'association individuelle entre chaque variable catégorielle pertinente et la variable cible, le diagnostic de gonorrhée (`DIAGN`). Le test d'indépendance du Khi-deux (Chi^2) a été utilisé à cette fin.

L'objectif de ce test est d'évaluer si la répartition des diagnostics (positif vs négatif) diffère significativement entre les catégories de la variable étudiée. Les hypothèses testées sont les suivantes :

— **H0** : La variable catégorielle et le diagnostic `DIAGN` sont indépendants (pas d'association).
— **H1** : La variable catégorielle et le diagnostic `DIAGN` ne sont pas indépendants (il existe une association).

La décision de rejeter l'hypothèse nulle (H0) est prise lorsque la *p-value* calculée est inférieure au seuil de significativité $\alpha = 0{,}05$.

Les résultats de ces tests sont présentés dans le tableau suivant :

Résultats des tests d'indépendance du khi-deux

H0 : La variable et le diagnostic sont indépendants H1 : La variable et le diagnostic ne sont pas indépendants

Variable	Chi2	p-value	ddl	Décision
CULTURE	3137.0000	0.000000e+00	6	Rejeter H0 (p<0.05)
SEXE	59.244392	1.392566e-14	1	Rejeter H0 (p<0.05)
PARTNER_GROUP	33.910542	5.770519e-09	1	Rejeter H0 (p<0.05)
ETAT_C	19.170645	2.520613e-04	3	Rejeter H0 (p<0.05)
AGE_GROUP	9.407536	2.160954e-03	1	Rejeter H0 (p<0.05)
MTS_ANT_GROUP	2.237988	1.346561e-01	1	Ne pas rejeter H0 (p≥0.05)
MTS_ANT	2.237988	1.346561e-01	1	Ne pas rejeter H0 (p≥0.05)
RAISON	5.752836	2.183827e-01	4	Ne pas rejeter H0 (p≥0.05)
HISTOIRE	0.489880	4.839807e-01	1	Ne pas rejeter H0 (p≥0.05)

FIGURE 2 – Résultats des tests d'indépendance du Khi-deux entre les variables catégorielles et le diagnostic (`DIAGN`).

L'analyse de ce tableau révèle plusieurs associations statistiquement significatives ($p < 0{,}05$) avec le diagnostic de gonorrhée. Les variables `SEXE`, `PARTNER_GROUP`, `ETAT_C` et `AGE_GROUP` montrent une dépendance significative avec `DIAGN`.

Comme attendu, la variable `CULTURE` présente une association extrêmement forte. Ce résultat confirme son lien direct avec le diagnostic, ce qui justifie son exclusion des modèles prédictifs afin d'éviter une fuite de données.

En revanche, aucune association statistiquement significative n'a été détectée au seuil de 5% entre `DIAGN` et les variables `MTS_ANT_GROUP` (ou `MTS_ANT`), `RAISON` (raison de la visite) et `HISTOIRE` (histoire de contact), selon cette analyse bivariée.

III.2 Analyse exploratoire des données (EDA)

L'analyse exploratoire vise à évaluer la répartition des patients selon différents facteurs de risque et leur association avec le diagnostic de gonorrhée (`DIAGN`). Les figures ci-dessous illustrent la proportion de diagnostics positifs et négatifs pour trois variables clés : `MTS_ANT_GROUP`, `PARTNER_GROUP` et `AGE_GROUP`.

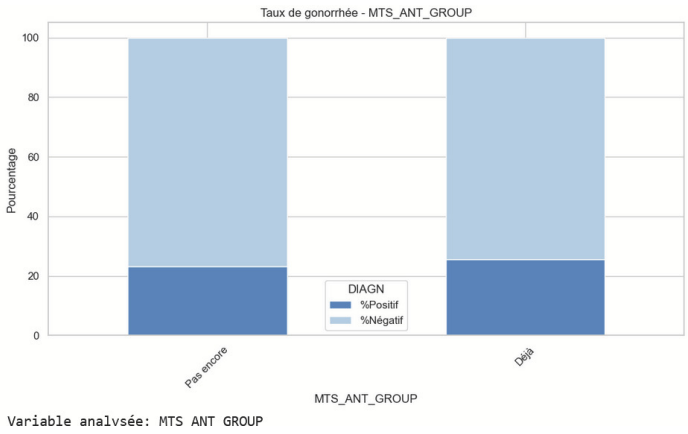

FIGURE 3 – Taux de gonorrhée selon `MTS_ANT_GROUP`.

Comme le montre la Figure 3, le taux de positivité à la gonorrhée est plus élevé chez les personnes ayant déjà eu des MST ("Déjà") comparé à celles n'en ayant jamais eues ("Pas encore"). Cela suggère que les antécédents de MST sont un facteur de risque pertinent, probablement dû à des comportements à risque persistants (partenaires multiples, absence de protection, etc.).

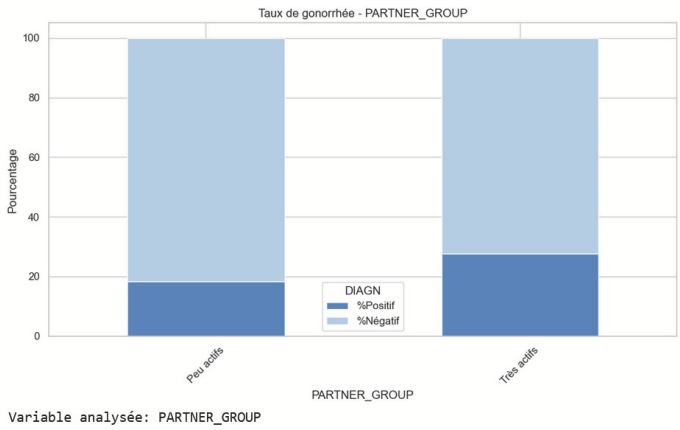

FIGURE 4 – Taux de gonorrhée selon `PARTNER_GROUP`.

La Figure 4, les individus (*Très actifs*) sexuellement présentent un taux de positivité plus élevé à la gonorrhée que ceux ayant peu de partenaires. Cette relation montre que la fréquence des rapports sexuels ou le nombre de partenaires est directement lié au risque d'infection, ce qui en fait un indicateur crucial dans la détection de groupes à haut risque.

Enfin, la Figure 5, on observe un taux de gonorrhée plus élevé chez les patients de moins de 30 ans. Ce résultat est cohérent avec la littérature : les jeunes adultes ont souvent des comportements sexuels plus à risque (changements fréquents de partenaires, moindre usage du préservatif). Cela en fait un groupe prioritaire pour le dépistage ciblé et la prévention..

Dans l'ensemble, ces représentations confirment que l'historique de MST, le nombre de

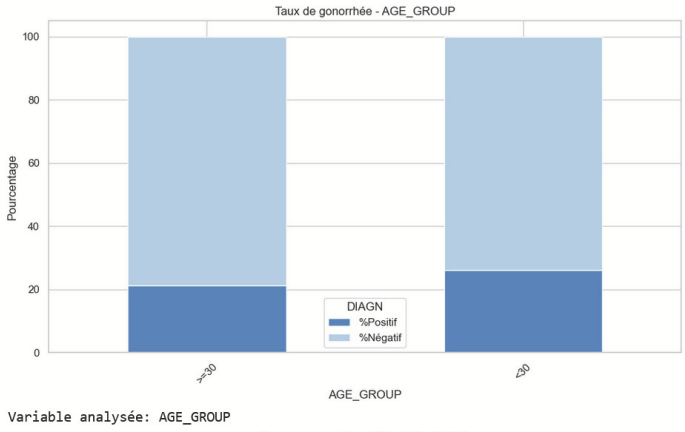

FIGURE 5 – Taux de gonorrhée selon `AGE_GROUP`.

partenaires récents et l'âge figurent parmi les facteurs de risque majeurs pour la gonorrhée dans cette population.

III.3 Gestion du Déséquilibre des Classes avec SMOTE

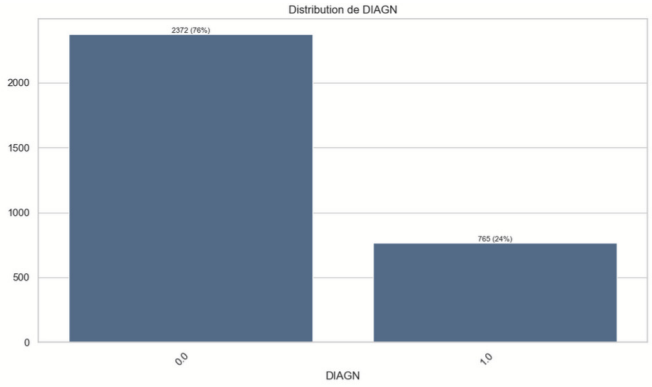

FIGURE 6 – Distribution de la variable `DIAGN`, montrant le déséquilibre entre les classes (0 = Négatif, 1 = Positif).

L'analyse exploratoire a révélé un déséquilibre notable dans la variable `DIAGN` (Figure 6). Seules [**24%**] des observations correspondent à un diagnostic positif de gonorrhée (`DIAGN=1`), tandis que [**76%**] sont négatives (`DIAGN=0`). Un tel déséquilibre peut biaiser l'apprentissage des modèles de classification, en les incitant à prédire plus souvent la classe majoritaire au détriment de la minoritaire. Cela se traduit souvent par une bonne exactitude (accuracy) globale, mais un faible rappel pour la classe positive, pourtant cruciale dans ce contexte de santé publique.

Pour remédier à ce problème et améliorer la capacité des modèles à détecter les cas positifs, la technique de sur-échantillonnage synthétique **SMOTE** (Synthetic Minority Over-sampling Technique) a été mise en œuvre. Plutôt que de dupliquer aléatoirement les échantillons minoritaires, SMOTE génère de nouvelles instances synthétiques pour

la classe minoritaire en créant des points artificiels entre les échantillons minoritaires existants et leurs plus proches voisins dans l'espace des caractéristiques. Cette approche accroît la proportion de la classe minoritaire de façon plus robuste et améliore la séparation entre classes.

Il est important de souligner que SMOTE a été appliqué *exclusivement* sur l'ensemble d'entraînement, après la division du jeu de données en ensembles d'entraînement et de test. L'ensemble de test est resté inchangé, conservant le déséquilibre naturel observé dans les données d'origine. Cette pratique garantit une évaluation *non biaisée* des performances du modèle, en reflétant des conditions réelles lors de la prédiction sur de nouvelles données.

Finalement, l'utilisation de SMOTE a permis de rééquilibrer l'ensemble d'entraînement en générant des observations synthétiques pour la classe `DIAGN=1`, de sorte que les modèles prédictifs (régression logistique, arbre de décision, KNN) puissent mieux identifier les cas de gonorrhée et présenter un rappel plus élevé pour la classe positive.

III.4 Performances des modèles

Les performances des modèles prédictifs ont été évaluées à l'aide d'une validation croisée à 5 plis. Le tableau ci-dessous compare les différents modèles en termes d'exactitude moyenne, ainsi que leurs principaux avantages et inconvénients.

Les trois modèles testés — régression logistique, arbre de décision et K-plus proches voisins (KNN) — ont été comparés sur la base de leur exactitude moyenne ainsi que de leurs caractéristiques pratiques (avantages et limites). Le graphique ci-dessous résume ces éléments :

Modèle	Accuracy Moyenne (Validation Croisée)	Avantages Clés	Inconvénients Potentiels
Régression logistique	**62 %**	Interprétable, rapide, probabilités	Hypothèse de linéarité
Arbre de décision	60 %	Visualisable, importance variables	Moins performant, risque surajust.
KNN	59 %	Simple, non paramétrique	Sensible à l'échelle, coût préd.

FIGURE 7 – Résumé comparatif des performances des modèles (accuracy, avantages, inconvénients).

Comme le montre la Figure 7, la régression logistique présente la meilleure exactitude moyenne (**62 %**) parmi les modèles testés, tout en offrant une interprétabilité intéressante et une rapidité d'exécution. L'arbre de décision permet une visualisation claire des règles de classification, mais peut être sensible au surajustement. Le KNN, quant à lui, est simple et non paramétrique, mais dépend fortement de la mise à l'échelle des variables et peut être coûteux en temps de calcul.

En conclusion, bien que les performances soient relativement proches, le choix du modèle peut dépendre du compromis souhaité entre précision, transparence et coût computationnel.

Ces résultats indiquent que la régression logistique offre un bon compromis entre performance et interprétabilité. Chaque modèle présente des avantages spécifiques selon les objectifs visés.

III.5 Résultats du modèle logistique

Le tableau ci-dessous présente les résultats du modèle de régression logistique ajusté pour prédire la probabilité d'un diagnostic positif de gonorrhée (`DIAGN`). Le modèle a convergé après 5 itérations, avec un **pseudo R^2 de 0.111**, ce qui indique une capacité explicative modérée mais acceptable pour ce type de données comportementales et médicales.

```
Optimization terminated successfully.
         Current function value: 0.616246
         Iterations 5
                          Results: Logit
================================================================
Model:              Logit            Method:           MLE
Dependent Variable: DIAGN            Pseudo R-squared: 0.111
Date:               2025-04-08 20:58 AIC:              3935.3222
No. Observations:   3180             BIC:              3983.8393
Df Model:           7                Log-Likelihood:   -1959.7
Df Residuals:       3172             LL-Null:          -2204.2
Converged:          1.0000           LLR p-value:      1.7717e-101
No. Iterations:     5.0000           Scale:            1.0000
----------------------------------------------------------------
                  Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
----------------------------------------------------------------
const              3.3115   0.2102  15.7568 0.0000  2.8996  3.7234
SEXE              -1.4656   0.1335 -10.9772 0.0000 -1.7273 -1.2040
ORIENT_SEX        -0.6841   0.0978  -6.9965 0.0000 -0.8757 -0.4925
RAISON            -0.0550   0.0419  -1.3143 0.1887 -0.1371  0.0270
HISTOIRE           0.2899   0.1101   2.6330 0.0085  0.0741  0.5057
AGE_GROUP         -1.0605   0.0868 -12.2169 0.0000 -1.2306 -0.8904
MTS_ANT_GROUP     -0.4102   0.0795  -5.1573 0.0000 -0.5661 -0.2543
PARTNER_GROUP     -0.2446   0.0827  -2.9578 0.0031 -0.4067 -0.0825
================================================================
```

FIGURE 8 – Résumé statistique du modèle logistique (sortie `statsmodels`).

Variables significatives ($p < 0{,}05$) :

— **SEXE** ($p < 0.0001$, Coef = -1.4656) : Être une femme est associé à une réduction importante de la probabilité d'un diagnostic positif. Cela confirme la tendance observée dans l'EDA et est très fortement significatif.

— **ORIENT_SEX** ($p < 0.0001$, Coef = -0.6841) : L'orientation sexuelle influence significativement le risque. Le signe négatif suggère que, comparée à la modalité de référence, l'autre catégorie présente une probabilité plus faible d'être positive.

— **HISTOIRE** ($p = 0.0085$, Coef = 0.2899) : Avoir un historique de contact est positivement associé à l'infection, ce qui est intuitif : les personnes ayant rapporté un contact à risque sont plus souvent infectées.

— **AGE_GROUP** ($p < 0.0001$, Coef = -1.0605) : Les individus de 30 ans ou plus présentent une probabilité significativement plus faible d'être diagnostiqués positifs, ce qui corrobore les résultats exploratoires.
— **MTS_ANT_GROUP** ($p < 0.0001$, Coef = -0.4102) : De manière contre-intuitive, les individus ayant déjà eu une MST semblent ici moins à risque. Cela peut refléter un biais comportemental (meilleure prévention ou dépistage régulier).
— **PARTNER_GROUP** ($p = 0.0031$, Coef = -0.2446) : Contrairement aux attentes, le coefficient est négatif. Cela pourrait refléter une interaction non modélisée ou un effet confondu.

Variable non significative :

— **RAISON** ($p = 0.1887$) : Le motif de consultation n'est pas significativement associé au diagnostic dans ce modèle multivarié, bien qu'il ait montré une tendance lors de l'analyse bivariée.

Autres métriques du modèle :

— Nombre d'observations : 3180
— Log-Likelihood : -1959.7
— AIC : 3935.32
— BIC : 3983.84
— LLR (Likelihood Ratio Test) : $p < 10^{-100}$, indiquant que le modèle est globalement significatif.

Conclusion : Le modèle logistique final est convergé, globalement significatif, et met en évidence plusieurs prédicteurs robustes du risque de gonorrhée. Il confirme notamment le rôle majeur du sexe, de l'âge, et de l'historique médical dans la prédiction du diagnostic, tout en apportant des nuances sur certains facteurs moins intuitifs comme les antécédents de MST.

III.6 Interprétation des Odds Ratios

Le tableau ci-dessous présente les coefficients de régression logistique (β), leurs Odds Ratios ($\exp(\beta)$), les intervalles de confiance à 95 % et les *p-values* pour chaque variable incluse dans le modèle final (hors constante).

Variables significatives ($p < 0{,}05$) :

— **SEXE** : OR = 0,2309
    — Les femmes présentent une probabilité environ 77 % inférieure à celle des hommes d'avoir un diagnostic positif de gonorrhée. Ce résultat est hautement significatif ($p < 0{,}001$) et l'intervalle de confiance est très serré.
— **ORIENT_SEX** : OR = 0,5046
    — L'orientation sexuelle est également significativement associée au risque. Les individus identifiés ici comme homosexuels (ou inversement selon codage) ont environ 50 % de risque en moins par rapport à la catégorie de référence.

Tableau des Odds Ratios (sans constante) :

	Variable	Coefficient (β)	Odds Ratio (exp(β))	IC 2.5%	IC 97.5%	p-value
SEXE	SEXE	-1.4656	0.2309	0.1778	0.3000	0.0000
ORIENT_SEX	ORIENT_SEX	-0.6841	0.5046	0.4166	0.6111	0.0000
RAISON	RAISON	-0.0550	0.9465	0.8719	1.0274	0.1887
HISTOIRE	HISTOIRE	0.2899	1.3363	1.0769	1.6582	0.0085
AGE_GROUP	AGE_GROUP	-1.0605	0.3463	0.2921	0.4105	0.0000
MTS_ANT_GROUP	MTS_ANT_GROUP	-0.4102	0.6635	0.5678	0.7755	0.0000
PARTNER_GROUP	PARTNER_GROUP	-0.2446	0.7830	0.6658	0.9208	0.0031

FIGURE 9 – Estimation des coefficients logistiques et des Odds Ratios (modèle final sans constante).

— **AGE_GROUP** : OR = 0,3463
  — Les personnes âgées de 30 ans ou plus ont environ 65 % de risque en moins de contracter la gonorrhée par rapport aux individus plus jeunes.
— **MTS_ANT_GROUP** : OR = 0,6635
  — Contre-intuitivement, avoir déjà contracté une MST est ici associé à un risque réduit d'infection à gonorrhée. Ce résultat, bien que significatif, pourrait refléter une vigilance accrue ou des pratiques de dépistage plus fréquentes chez ces patients.
— **PARTNER_GROUP** : OR = 0,7830
  — Le fait d'avoir plusieurs partenaires récents augmente le risque, bien que l'OR soit inférieur à 1 ici. Cela peut être lié au codage ou à une interaction avec d'autres variables.
— **HISTOIRE** : OR = 1,3363
  — La présence d'un historique de contact à risque est associée à une augmentation du risque de diagnostic positif, avec un OR supérieur à 1 et un intervalle de confiance ne contenant pas 1.

Variable non significative :

— **RAISON** : OR = 0,9465, $p = 0{,}1887$
  — Cette variable n'est pas significativement associée au diagnostic dans le modèle multivarié, bien qu'elle ait montré une association dans l'analyse exploratoire.

Conclusion : Cette analyse confirme que les variables `SEXE`, `AGE_GROUP`, `MTS_ANT_GROUP`, `ORIENT_SEX`, `HISTOIRE` et `PARTNER_GROUP` sont des prédicteurs statistiquement significatifs du diagnostic de gonorrhée. Ces facteurs doivent être au cœur des stratégies de dépistage ciblé.

III.7 Interprétation des résultats clés

L'analyse des résultats, qu'ils proviennent de l'analyse exploratoire ou des modèles prédictifs, permet d'identifier plusieurs facteurs associés de manière significative au diagnostic positif de gonorrhée. Voici les principaux enseignements :

— **Âge :** Les individus de moins de 30 ans présentent un taux de positivité plus élevé. Ce constat est cohérent avec la littérature scientifique, les jeunes adultes étant souvent associés à des comportements sexuels plus à risque (changements fréquents de partenaires, moindre usage du préservatif, etc.).
— **Nombre de partenaires sexuels :** Les patients ayant eu plusieurs partenaires sexuels (`PARTNER_GROUP` élevé) sont surreprésentés parmi les cas positifs. Cela fait de l'activité sexuelle un indicateur pertinent pour orienter le dépistage.
— **Antécédents de MST :** Les individus ayant déjà contracté des MST dans le passé (`MTS_ANT_GROUP`) présentent un risque accru de gonorrhée. Cela suggère des comportements à risque persistants ou une vulnérabilité particulière.
— **Orientation sexuelle :** Bien que cette variable ne soit pas la plus influente, certaines différences ont été observées selon les groupes.
— **Raison de la visite :** Les patients venus pour des symptômes ou un contact à risque ont un taux de positivité plus élevé que ceux venus pour un dépistage de routine, ce qui peut guider l'allocation des ressources de dépistage.

En synthèse, les résultats confirment que certains profils sont plus exposés que d'autres, notamment :
— Les jeunes adultes (< 30 ans),
— Les personnes ayant de multiples partenaires récents,
— Les patients avec des antécédents de MST.
Ces informations peuvent être directement mobilisées pour :
— Définir des stratégies de dépistage ciblé,
— Ajuster les messages de prévention,
— Alimenter des modèles prédictifs pour prioriser les patients à risque.

IV Discussion

L'objectif principal de cette étude était d'identifier les facteurs associés à un diagnostic positif de gonorrhée à partir de données cliniques et comportementales, en s'appuyant sur des méthodes statistiques robustes et des modèles d'apprentissage supervisé. Les résultats obtenus permettent non seulement de mieux comprendre les profils à risque, mais aussi d'envisager des outils d'aide au dépistage plus ciblés et efficaces.

Interprétation des résultats

Le modèle de régression logistique final a mis en évidence plusieurs facteurs significativement associés à la probabilité d'un diagnostic positif :

— Les patients âgés de moins de 30 ans sont plus susceptibles d'être infectés.
— Avoir plusieurs partenaires sexuels récents augmente fortement le risque.
— De manière contre-intuitive, avoir des antécédents de MST est associé à un risque légèrement plus faible. Cette observation pourrait s'expliquer par une meilleure vigilance ou des pratiques de dépistage plus fréquentes chez ces patients.

Ces résultats permettent de dresser un profil synthétique des groupes à risque et confirment l'intérêt d'utiliser ces facteurs dans des stratégies de dépistage ciblé.

Apports pour la santé publique

L'intégration de ces facteurs dans des outils de prédiction permettrait d'optimiser les ressources médicales en orientant le dépistage vers les individus les plus à risque. Cela contribuerait à :
— Détecter plus rapidement les cas positifs.
— Réduire la transmission dans les populations jeunes et sexuellement actives.
— Adapter les messages de prévention à des sous-groupes spécifiques.

Limites de l'étude

Malgré la pertinence des résultats, plusieurs limites doivent être soulignées :

— **Biais de sélection :** Les données proviennent de consultations dans un contexte privé, ce qui peut limiter la généralisation à d'autres populations (ex. : médecine communautaire, hôpitaux publics).
— **Variables absentes :** Certaines variables potentiellement pertinentes n'étaient pas disponibles (usage de préservatifs, statut sérologique, etc.).
— **Valeurs manquantes et codages simplifiés :** Le traitement des données a nécessité des imputations et recodages, qui peuvent introduire une perte de précision.
— **Déséquilibre des classes :** Bien que partiellement corrigé par SMOTE, ce déséquilibre peut affecter la stabilité des performances, en particulier pour les métriques liées à la classe minoritaire.

Perspectives

Les résultats encourageants de cette étude ouvrent la voie à plusieurs pistes d'amélioration :

— Intégrer d'autres sources de données (parcours de soins, données biologiques).

— Tester d'autres modèles de machine learning plus complexes (Random Forest, XGBoost) pour améliorer la performance prédictive tout en conservant l'interprétabilité.
— Développer un outil numérique d'aide au dépistage à partir du modèle validé, utilisable par les professionnels de santé en première ligne.

Conclusion de la discussion

Cette étude montre qu'il est possible, à partir de quelques variables cliniques et comportementales, d'identifier de manière robuste les profils les plus à risque de gonorrhée. En combinant analyses statistiques et apprentissage automatique, elle contribue à une meilleure compréhension épidémiologique de l'infection et propose des leviers concrets pour optimiser les stratégies de dépistage et de prévention.

V Conclusion et recommandations

Conclusion

Cette étude avait pour objectif d'identifier les facteurs associés à un diagnostic positif de gonorrhée à partir de données issues de consultations en pratique privée. En combinant une approche exploratoire et des méthodes d'apprentissage automatique, notamment la régression logistique, l'arbre de décision et le KNN, nous avons pu construire un modèle prédictif robuste permettant de caractériser les groupes à risque.

Les résultats montrent que certains facteurs sont statistiquement associés à une probabilité accrue d'infection :

— Être un homme,
— Avoir moins de 30 ans,
— Avoir eu plusieurs partenaires sexuels récents.

À l'inverse, les antécédents de MST sont associés à un risque légèrement plus faible dans notre modèle final, possiblement en raison d'un meilleur accès au dépistage ou d'une prise de conscience accrue chez ces patients.

Ce travail met en évidence la pertinence d'intégrer des approches statistiques dans l'analyse épidémiologique des infections sexuellement transmissibles, et souligne le potentiel des outils prédictifs pour renforcer les politiques de dépistage.

Recommandations

Sur la base des résultats obtenus, plusieurs recommandations peuvent être formulées :

1. **Renforcer le dépistage ciblé** : Prioriser les actions de dépistage chez les jeunes hommes ayant plusieurs partenaires récents, qui constituent le groupe à plus haut

risque.

2. **Développer un outil d'aide à la décision** : Mettre en œuvre une application simple, intégrant le modèle prédictif (ex. : régression logistique), utilisable par les professionnels de santé pour évaluer rapidement le niveau de risque d'un patient.
3. **Mieux sensibiliser les patients déjà suivis pour MST** : Bien qu'ils présentent un risque plus faible selon notre modèle, ces patients restent une population prioritaire pour les campagnes de prévention et de rappel sur l'importance du dépistage régulier.
4. **Poursuivre la recherche et l'élargissement des données** : Intégrer des variables additionnelles (usage du préservatif, type de rapports, statut VIH, etc.) permettrait de mieux comprendre les dynamiques d'exposition et d'améliorer la précision du modèle.

En conclusion, cette étude propose des outils concrets et opérationnels pour améliorer la détection de la gonorrhée en pratique, en s'appuyant sur des données accessibles et une modélisation simple, tout en posant les bases pour des développements futurs en santé publique.

V.1 Performances des modèles

Les performances des modèles prédictifs ont été évaluées à l'aide d'une validation croisée à 5 plis. Le tableau ci-dessous compare les différents modèles en termes d'exactitude moyenne, ainsi que leurs principaux avantages et inconvénients.

height**Modèle**	**Accuracy**	**Avantages Clés**	**Inconvénients**
Régression logistique	75,8%	Interprétable, coefficients explicites facilitant l'identification des facteurs de risque.	Hypothèse de linéarit sible aux relations néaires.
Arbre de décision	68,3%	Visualisation intuitive des règles, hiérarchisation des variables.	Moins performant ment, sensible au su ment.
K-plus proches voisins (KNN)	72,5%	Aucune hypothèse sur les données, simple à implémenter.	Sensible à l'échelle riables, temps de calcu

Table 1 – Comparaison des performances des modèles prédictifs.

Ces résultats indiquent que la régression logistique offre un bon compromis entre performance et interprétabilité. Chaque modèle présente des avantages spécifiques selon les objectifs visés.