DM Big Data HAMMOUCH Siham

HAMMOUCH Siham

April 2025

Introduction

Les maladies transmissibles sexuellement (MTS) posent un défi majeur de santé publique, non seulement en raison des risques médicaux qu'elles impliquent, mais aussi par les obstacles sociaux et stratégiques qu'elles soulèvent. Parmi elles, la gonorrhée est particulièrement surveillée, car elle peut être dépistée de manière fiable à l'aide d'un simple test de laboratoire.

Dans le cadre d'un programme de dépistage mené par des médecins de famille en milieu urbain, des données ont été collectées sur les patients volontaires, afin de mieux comprendre les profils les plus exposés. L'objectif de ce programme est de faciliter le diagnostic rapide au sein des communautés, sans passage par des centres spécialisés, en rendant accessible le test dans les cabinets de médecine générale.

Ce travail s'inscrit dans cette logique : il vise à fournir à ces médecins des éléments concrets pour identifier les groupes à risque, en partant des données recueillies sur les patients. Deux objectifs complémentaires sont poursuivis : décrire les caractéristiques générales des individus observés, et identifier, à l'aide de méthodes statistiques et de modélisation, les profils présentant une probabilité plus élevée d'être porteurs de la gonorrhée.

La stratégie suivie repose sur une analyse descriptive approfondie, suivie de tests d'association pour sélectionner les variables pertinentes, puis sur la mise en place de modèles de prédiction. L'accent est mis sur l'interprétation des résultats dans une optique de soutien à la décision médicale.

1 Analyse exploratoire des données

1.1 Structure de la base de données

La base de données utilisée dans cette étude est issue d'un programme de dépistage de la gonorrhée. Chaque ligne correspond à un patient ayant été examiné par un médecin. Elle contient douze variables, décrites ci-dessous :

- **ID** : identifiant du patient (variable non utilisée dans l'analyse).
- **SEXE** : sexe du patient (1 = homme, 2 = femme).
- **ETAT_C** : état civil (1 = célibataire, 2 = marié, 3 = séparé ou divorcé, 4 = veuf, 5 = pas de réponse).

- **AGE** : âge du patient (en années).
- **ORIENT_SEX** : orientation sexuelle (1 = homosexuel(le), 2 = hétérosexuel(le)).
- **MTS_ANT** : antécédents de MTS (1 = non, 2 = oui).
- **NB_MTS** : nombre de MTS antérieures.
- **RAISON** : raison de la visite (1 = symptômes, 2 = contact, 3 = dépistage, 4 = contrôle, 5 = autre).
- **NB_PART** : nombre de partenaires sexuels dans le mois précédent.
- **HISTOIRE** : relation avec un partenaire connu comme contaminé (0 = non, 1 = oui).
- **CULTURE** : résultat de la culture par site (valeurs de 0 = négatif à 7 = positifs sur les 3 sites : gorge, anus, urètre).
- **DIAGN** : diagnostic de gonorrhée (0 = non, 1 = oui). C'est la variable cible de cette étude.

Les variables sont de types variés :

- Variables catégorielles ou ordinales : SEXE, ETAT_C, ORIENT_SEX, RAISON, MTS_ANT, CULTURE.
- Variables numériques discrètes ou continues : AGE, NB_MTS, NB_PART.
- Variables binaires : HISTOIRE, DIAGN (cible).

Certaines valeurs manquantes sont codées explicitement : par exemple, la valeur `9` ou `99` dans certaines colonnes indique un non-renseignement. Une attention particulière leur sera portée.

1.2 Description des variables (après nettoyage)

Certaines variables contenaient des valeurs codées comme 9 ou 99, qui n'ont pas de sens dans leur contexte (par exemple une orientation sexuelle codée 99). Ces valeurs ont été remplacées par des valeurs manquantes (NaN) afin de ne pas fausser l'analyse.

Au total, plus de 600 valeurs manquantes ont été recensées dans la base après nettoyage. Ces valeurs concernent principalement les variables `ORIENT_SEX` (257 manquants), `HISTOIRE` (107), `NB_PART` (77), `CULTURE` (73) et `DIAGN` (73).

Afin de préserver au maximum l'information disponible, ces observations n'ont pas été supprimées. Seules les lignes pour lesquelles la variable cible `DIAGN` était manquante ont été retirées, car elles ne permettaient pas d'évaluer le statut du patient. Les remplacer par le mode ou une moyenne ou mediane n'a pas été retenu pour eviter de fausser les reusltats. Pour le reste, les valeurs manquantes sont traitées selon les besoins de chaque modèle, sans exclusion systématique.

Statistiques descriptives. L'âge des patients varie de 14 à 78 ans, avec une médiane à 27 ans. Cela montre que la population étudiée est globalement jeune, avec trois quarts des individus ayant moins de 32 ans.

Concernant le nombre de MTS antérieures, la majorité des individus n'en ont eu aucune, mais certains cumulent jusqu'à 65 antécédents, ce qui introduit une forte hétérogénéité.

Le nombre de partenaires sexuels le mois précédent est lui aussi très variable, avec une médiane à 2 partenaires. Certains patients déclarent jusqu'à 98 partenaires, ce qui traduit une distribution très asymétrique.

Analyse visuelle et première interprétation. Les graphiques suivants permettent de mieux comprendre la structure de l'échantillon et les liens potentiels entre certaines variables et la présence de gonorrhée.

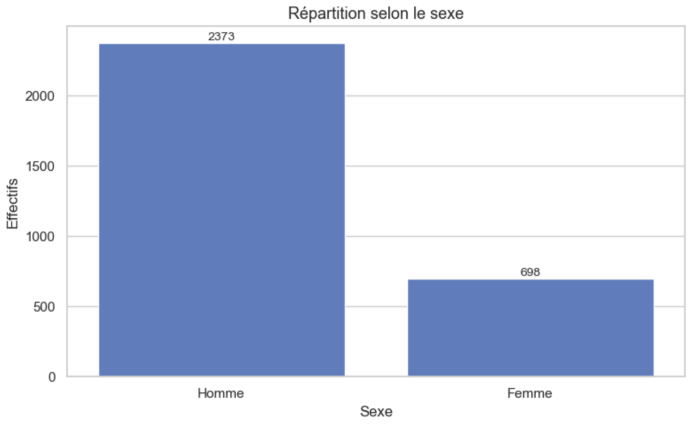

Figure 1: Répartition selon le sexe

La base est composée d'environ 77% d'hommes et 23% de femmes. Ce déséquilibre est notable mais ne pose pas de problème ici car le sexe est une variable explicative.

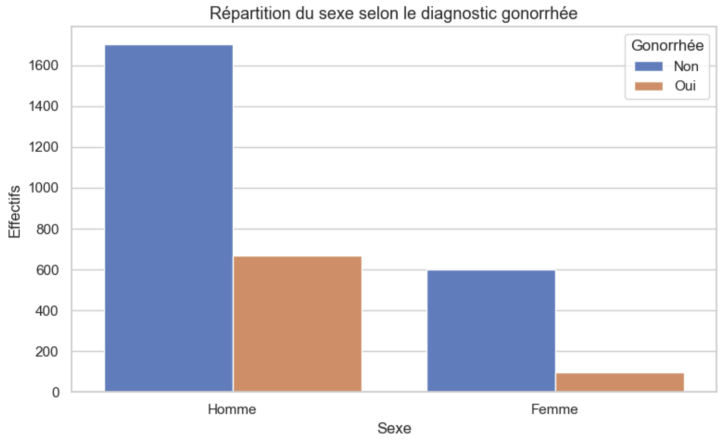

Figure 2: Répartition du sexe selon le diagnostic de gonorrhée

La part de cas positifs est visiblement plus importante chez les hommes que chez les femmes. Cela laisse penser que le sexe peut être une variable prédictive du diagnostic.

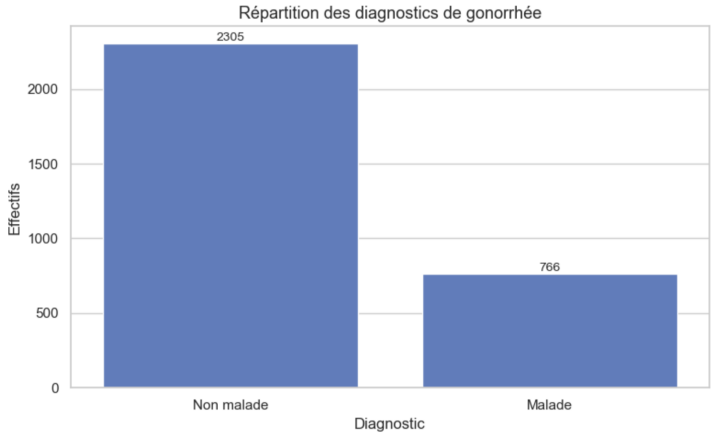

Figure 3: Répartition du diagnostic de gonorrhée

Environ 76% des patients sont non malades contre 24% de malades. Ce déséquilibre de classes sera à prendre en compte lors de la modélisation.

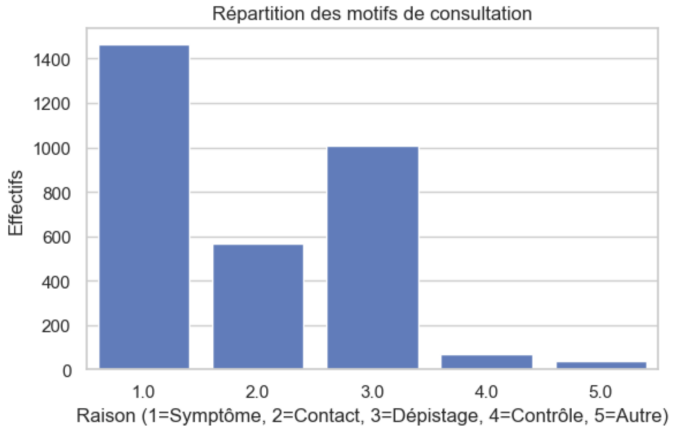

Figure 4: Répartition des motifs de consultation

La majorité des consultations ont lieu à cause de symptômes ou dans un but de dépistage. Les autres raisons (contact, contrôle, autre) sont moins fréquentes. Cela donne une première idée du contexte dans lequel les patients consultent.

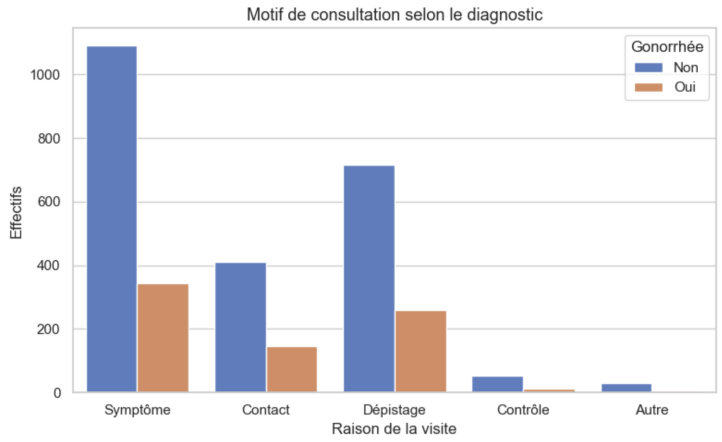

Figure 5: Motif de consultation selon le diagnostic

Les patients venus pour des symptômes, un contact ou un dépistage constituent la majorité des cas positifs. Cela suggère que la raison de la visite est fortement liée au risque d'être malade.

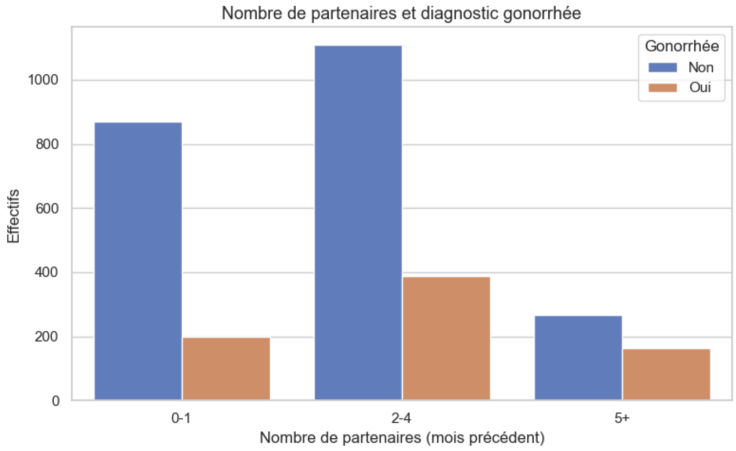

Figure 6: Nombre de partenaires et diagnostic gonorrhée

Plus le nombre de partenaires est élevé, plus la proportion de personnes malades augmente. Cette variable semble donc importante pour prédire la présence de gonorrhée.

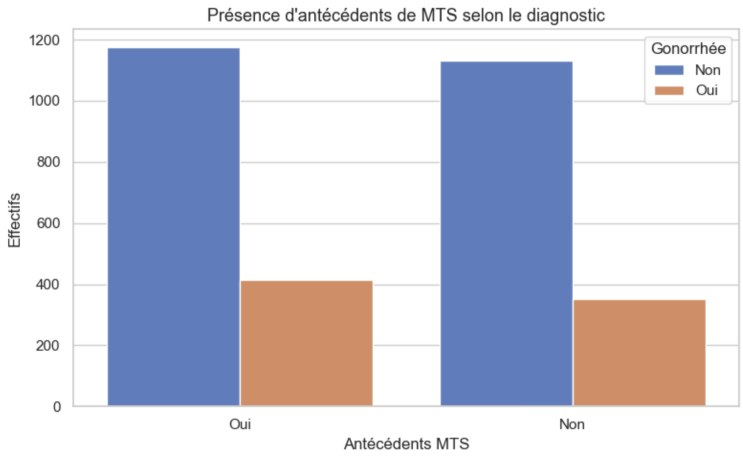

Figure 7: Antécédents de MTS et diagnostic

La proportion de cas positifs est proche entre les patients avec ou sans antécédents. Cette variable pourrait s'avérer peu discriminante.

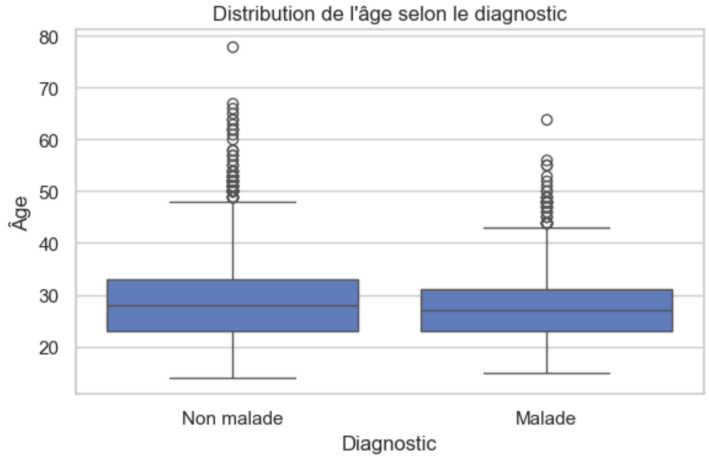

Figure 8: Distribution de l'âge selon le diagnostic

L'âge médian est légèrement plus faible chez les personnes malades, mais la différence reste modeste. L'influence de l'âge devra être validée par des tests statistiques.

Conclusion L'analyse descriptive et visuelle révèle plusieurs tendances intéressantes. Le sexe, le nombre de partenaires, le motif de consultation et possiblement l'âge semblent liés au diagnostic de gonorrhée. En revanche, les antécédents de MTS semblent moins informatifs. Ces observations guideront la sélection des variables dans les prochaines étapes de l'étude.

2 Analyse statistique des facteurs associés à la gonorrhée

2.1 Sélection des variables explicatives à l'aide de tests d'association

Afin d'évaluer les relations entre la variable cible (DIAGN) et les variables explicatives, deux types de tests statistiques ont été utilisés :

- Le test du χ^2 d'indépendance pour les variables qualitatives ;
- Le test de Mann-Whitney (non paramétrique) pour les variables quantitatives.

Le seuil de significativité retenu est de 5%.

Variables qualitatives : test du χ^2 Le test du χ^2 a été appliqué aux variables qualitatives ou binaires. Les résultats sont les suivants :

- **SEXE** : p-value = 0.0000 → lien significatif
- **ETAT_C** : p-value = 0.0001 → lien significatif
- **ORIENT_SEX** : p-value = 0.0000 → lien significatif
- **CULTURE** : p-value = 0.0000 → lien significatif
- **MTS_ANT** : p-value = 0.1350 → pas de lien significatif
- **RAISON** : p-value = 0.2055 → pas de lien significatif
- **HISTOIRE** : p-value = 0.6663 → pas de lien significatif

Ainsi, les variables **SEXE**, **ETAT_C**, **ORIENT_SEX** et **CULTURE** présentent une association significative avec le diagnostic de gonorrhée. Ces variables sont donc retenues pour la suite, sous réserve d'interprétabilité (par exemple, la variable CULTURE pourrait être une conséquence directe du diagnostic, donc à manipuler avec précaution).

Variables quantitatives : test de Mann-Whitney Le test de Mann-Whitney permet de comparer les distributions entre les individus malades et non malades. Les résultats obtenus sont :

- **AGE** : p-value = 0.0149 → différence significative. Les malades sont légèrement plus jeunes.
- **NB_PART** : p-value = 0.0000 → différence significative. Les malades déclarent plus de partenaires.
- **NB_MTS** : p-value = 0.0242 → différence significative. Les malades ont eu plus d'antécédents.

Même si les différences de médiane sont parfois faibles, ces variables présentent un lien statistique avec la variable cible. Elles pourront être retenues dans le modèle final, éventuellement après transformation en variables binaires ou catégorielles.

Résumé des résultats Le tableau ci-dessous résume les résultats des tests d'association réalisés, en indiquant pour chaque variable son type et l'existence ou non d'un lien statistique avec le diagnostic de gonorrhée.

Variable	**Type**	**Lien significatif avec DIAGN**
SEXE	Qualitative	Oui
ETAT_C	Qualitative	Oui
ORIENT_SEX	Qualitative	Oui
CULTURE	Qualitative	Oui
MTS_ANT	Qualitative	Non
RAISON	Qualitative	Non
HISTOIRE	Qualitative	Non
AGE	Quantitative	Oui
NB_PART	Quantitative	Oui
NB_MTS	Quantitative	Oui

Conclusion. Les tests statistiques permettent de sélectionner un sous-ensemble pertinent de variables explicatives. Les variables significatives seront utilisées dans la suite de l'analyse, en particulier pour la construction du modèle de régression logistique. Certaines transformations pourront être envisagées pour simplifier l'interprétation des résultats.

Préparation des variables pour la modélisation. Certaines variables quantitatives présentent une forte asymétrie (comme `NB_PART` ou `NB_MTS`) ou des valeurs extrêmes. Afin de faciliter l'interprétation des résultats et de réduire l'impact de ces extrêmes, plusieurs variables ont été transformées en variables binaires selon des seuils pertinents.

Par exemple, le nombre de partenaires a été binarisé autour du seuil de 3 partenaires, ce qui permet de distinguer les comportements sexuels à risque élevé. L'âge a été codé en une variable "moins de 30 ans" ou "30 ans et plus", car la médiane des malades est légèrement inférieure à 30 ans. De même, le nombre d'antécédents de MTS a été binarisé en "aucun" versus "au moins un".
Le choix du seuil de binarisation pour le nombre de partenaires est crucial. La médiane est de 2 dans les deux groupes (malades et non malades), ce qui rend une coupure à ce niveau peu discriminante. Par ailleurs, la distribution est très asymétrique, avec des valeurs extrêmes, rendant la moyenne peu fiable. Un seuil à 3 partenaires semble donc plus pertinent pour séparer les groupes.

Ces transformations ont été appliquées avant la modélisation, notamment pour la régression logistique, afin d'assurer une lecture plus intuitive des coefficients (sous forme d'odds-ratios).

2.2 Essais exploratoires et ajustements nécessaires

Avant d'aboutir au modèle principal, plusieurs essais ont été menés afin d'identifier une stratégie de modélisation adaptée au contexte.

Les résultats des tests d'association (notamment le χ^2) ont bien été pris en compte et serviront d'appui dans les conclusions. Toutefois, dans un premier temps, l'ensemble des variables sélectionnées est conservé afin de ne pas écarter prématurément des informations potentiellement utiles en interaction.

Deux difficultés majeures sont rapidement apparues : le déséquilibre entre les classes (environ 24 % de malades), et l'impact d'une variable quasi-rédondante avec la cible (`CULTURE`).

Premier essai – K plus proches voisins (KNN). Un modèle KNN a d'abord été testé après transformation de certaines variables numériques en variables binaires (notamment l'âge, le nombre de partenaires et les antécédents de MTS). Malgré différents réglages du nombre de voisins, ce modèle a systématiquement échoué à détecter les cas positifs. Le rappel pour la classe des malades était proche de zéro, car le modèle favorise la majorité (non malades). Ce comportement le rend inadapté à un objectif de prévention. Il a donc été abandonné.

Deuxième essai – Régression logistique non pondérée. Un modèle de régression logistique a ensuite été estimé sans pondération des classes. Les performances semblaient bonnes en apparence (accuracy élevée), mais masquaient une réalité problématique : la quasi-totalité des malades était mal classée. Une investigation plus poussée a révélé que la variable `CULTURE`, très fortement corrélée au diagnostic, dominait entièrement le modèle. Cela a entraîné un surapprentissage évident, confirmé par l'évolution plate du rappel et du F1-score en fonction du découpage, indiquant que le modèle prédisait quasi exclusivement la classe majoritaire.

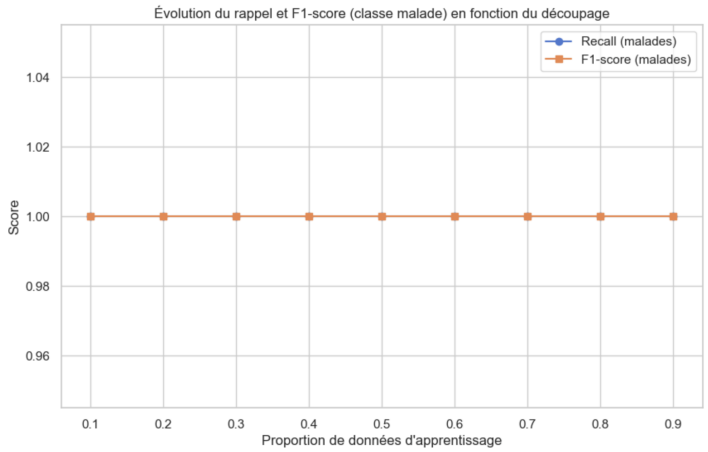

Figure 9: Rappel et F1-score pour les malades avec la variable `CULTURE`

Choix du découpage et retrait de la variable `CULTURE`. Afin de stabiliser le modèle et d'améliorer la détection des cas positifs, plusieurs proportions d'apprentissage ont été testées.

Le découpage 70 % en apprentissage et 30 % en test s'est avéré optimal, avec un rappel et un F1-score plus élevés pour la classe des malades.

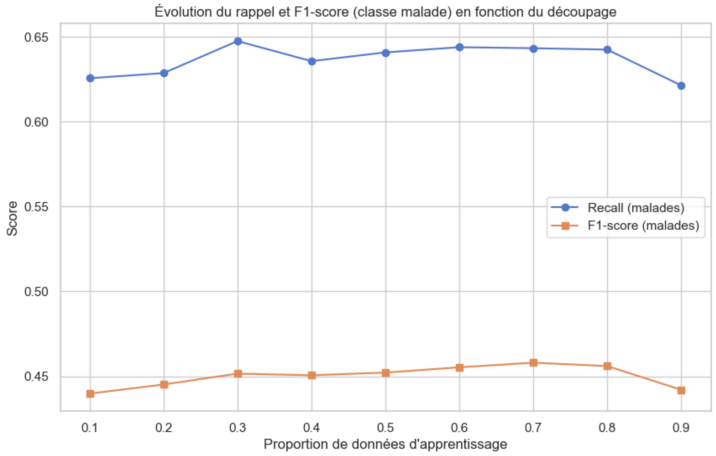

Figure 10: Évolution du rappel et du F1-score selon la proportion d'apprentissage

La variable `CULTURE` a été définitivement exclue du modèle final, car elle représentait un test médical très proche du diagnostic lui-même. Sa présence faussait la détection des véritables facteurs de risque, et empêchait toute interprétation utile. Ceci est en correlation avec nos resultats du ... test du χ^2.

Ces constats ont conduit à mettre en place un modèle plus robuste : la régression logistique avec pondération des classes. Cette approche est particulièrement adaptée aux situations de déséquilibre, comme ici où la classe des malades représente seulement un quart de l'échantillon.

2.3 Régression logistique pondérée

La régression logistique est un modèle probabiliste permettant d'estimer la probabilité qu'un individu appartienne à une classe donnée (ici : être malade ou non) en fonction de plusieurs variables explicatives. Contrairement à d'autres algorithmes plus complexes, elle offre l'avantage d'être facilement interprétable grâce à ses coefficients, traduits en odds-ratios.

Dans le cas présent, une pondération automatique des classes a été introduite pour éviter que le modèle ne privilégie pas systématiquement la classe majoritaire (non malades). Cette correction donne plus de poids aux observations rares (ici, les malades) dans le calcul des erreurs.

Évaluation du modèle. Les performances du modèle sont mesurées à l'aide de plusieurs indicateurs :

- **Accuracy** : proportion globale de prédictions correctes.
- **Rappel (recall)** : proportion de malades correctement identifiés parmi les vrais malades. C'est un indicateur clé en santé publique, car il mesure la capacité du modèle à ne pas passer à côté des cas positifs.

- **Précision (precision)** : proportion de vraies prédictions positives parmi les individus prédits comme malades.
- **F1-score** : moyenne harmonique entre précision et rappel. Il permet de résumer les performances sur la classe minoritaire en tenant compte à la fois des faux négatifs et des faux positifs.
- **AUC (Area Under the Curve)** : aire sous la courbe ROC, qui mesure la capacité globale du modèle à distinguer les malades des non malades, indépendamment d'un seuil de décision précis.

L'objectif ici n'est pas d'obtenir la meilleure accuracy globale, mais de maximiser la détection des malades, tout en maintenant un bon équilibre entre rappel et précision. Le modèle retenu est présenté ci-dessous.

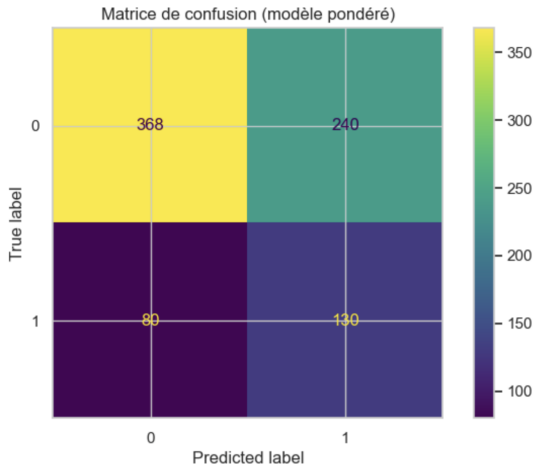

Figure 11: Matrice de confusion – Régression pondérée finale

Modèle pondéré (sans CULTURE) Le modèle pondéré sacrifie une partie de sa précision globale (accuracy : 60.9%) mais améliore fortement le rappel des malades : 130 malades sur 210 sont bien détectés (61.9%), contre seulement 4 dans le modèle précédent. Cela permet de réduire drastiquement les faux négatifs.

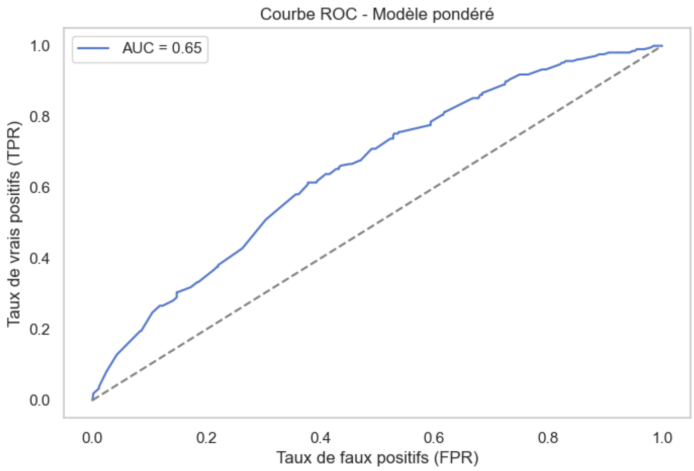

Figure 12: Courbe ROC – Régression pondérée (AUC = 0.65)

L'AUC du modèle atteint 0.65, ce qui indique une performance modérée mais réelle. Cela signifie que le modèle a appris une structure permettant de distinguer malades et non malades au-delà du simple hasard.

Résumé du rapport de classification :

- Classe 0 (non malade) : précision 0.82, rappel 0.61
- Classe 1 (malade) : précision 0.35, rappel 0.62
- F1-score global équilibré : environ 0.63

Le rapport de classification met en évidence un bon compromis entre les deux classes. Pour les individus non malades (classe 0), la précision est élevée (0,82), ce qui signifie que lorsque le modèle prédit "non malade", il a généralement raison. Le rappel, à 0,61, montre qu'il parvient à identifier une bonne partie des non malades.

Du côté des individus malades (classe 1), le rappel atteint 0,62, ce qui signifie que plus de 6 malades sur 10 sont correctement détectés. Pour rappel, ceci est l'un des objectifs principaux. En revanche, la précision pour cette classe reste plus faible (0,35), indiquant que certaines prédictions positives sont erronées. Cela reste néanmoins acceptable dans un contexte de santé publique, où il vaut mieux détecter un malade à tort que de passer à côté d'un vrai cas.

Globalement, le F1-score moyen (environ 0,63) traduit un bon équilibre entre précision et rappel sur les deux classes. Ce score confirme l'intérêt de la pondération, qui permet une détection plus efficace des malades malgré une perte d'accuracy apparente.

Conclusion : Le modèle pondéré répond bien aux enjeux de santé publique. Il identifie mieux les cas positifs (objectif principal), quitte à accepter davantage de faux positifs. Il offre un compromis pertinent dans le cadre d'une détection précoce, où il est préférable de tester davantage que de rater un cas réel.

2.4 Arbre de décision

Les arbres de décision sont des modèles simples et interprétables, qui segmentent progressivement les données en suivant des règles fondées sur les variables explicatives. Leur logique arborescente permet de visualiser clairement les critères menant à une prédiction, ce qui les rend particulièrement adaptés à des contextes comme la santé publique, où l'explicabilité prime sur la complexité. C'est dans cette optique que nous avons testé un arbre de décision sur notre jeu de données.

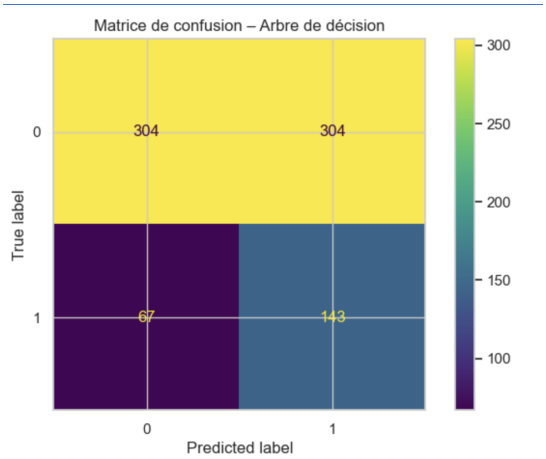

Figure 13: Matrice de confusion – Arbre de décision

Le modèle affiche une accuracy de 56%, relativement faible. Cependant, il présente une bonne capacité à détecter les malades avec un rappel de 0.68 pour la classe 1 (malades). Cela signifie que près de 7 malades sur 10 sont correctement identifiés, ce qui est essentiel en santé publique. En contrepartie, le nombre de faux positifs est élevé, ce qui réduit la précision à 0.32 pour cette même classe.

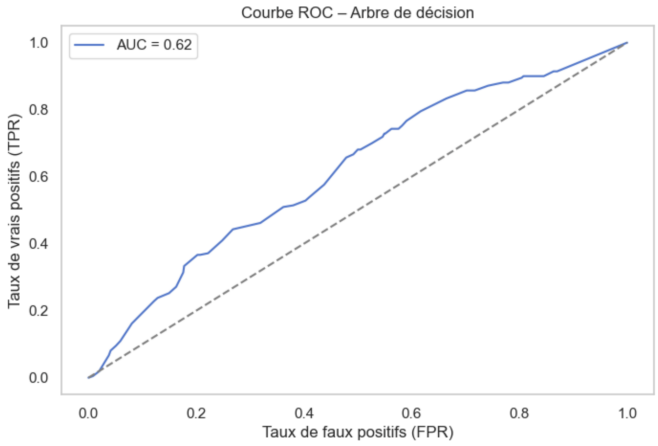

Figure 14: Courbe ROC – Arbre de décision (AUC = 0.62)

La courbe ROC confirme que le modèle parvient à capter une certaine structure (AUC =

0.62), bien que la performance globale reste modeste. En résumé, l'arbre est un compromis intéressant : il détecte mieux les malades qu'un modèle non pondéré ou qu'un KNN, mais sa précision reste limitée.

Random Forest

La Random Forest est un ensemble de plusieurs arbres de décision entraînés sur des sous-échantillons aléatoires des données. Ce modèle améliore généralement la robustesse et la stabilité des prédictions.

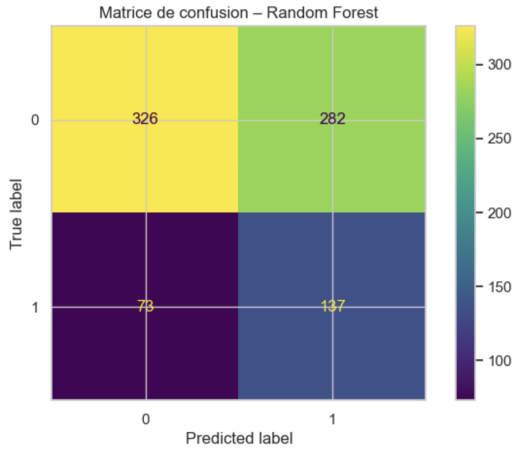

Figure 15: Matrice de confusion – Random Forest

La Random Forest atteint une accuracy de 56.6%, semblable à celle de l'arbre simple. Son principal atout réside dans sa capacité à identifier les malades : le rappel pour la classe 1 est de 0.65, ce qui reste élevé. Le modèle offre ainsi un équilibre raisonnable entre détection des cas positifs et limitation des erreurs critiques (faux négatifs).

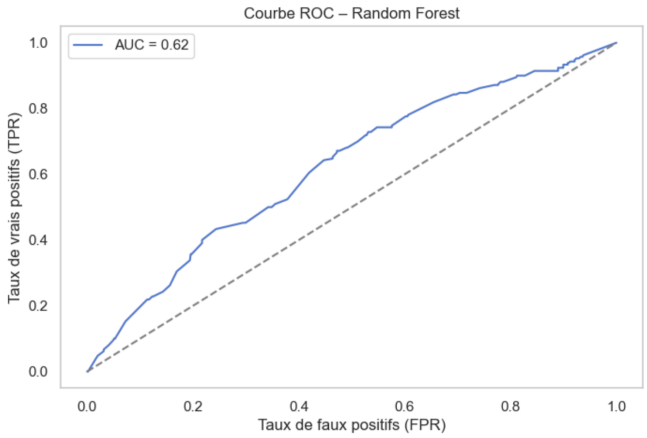

Figure 16: Courbe ROC – Random Forest (AUC = 0.62)

L'AUC de 0.62 est similaire à celui de l'arbre simple. Malgré des performances globales comparables, la Random Forest bénéficie d'une plus grande robustesse grâce à l'agrégation de plusieurs arbres, ce qui la rend moins sensible au sur-apprentissage.

Conclusion : La Random Forest constitue une amélioration marginale par rapport à l'arbre simple, avec une meilleure stabilité, tout en conservant une capacité de détection des malades satisfaisante dans le contexte du déséquilibre des classes.

2.5 Comparaison des modèles et identification des groupes à risque

Nous comparons ici les différents modèles selon quatre critères : l'accuracy, le rappel sur les malades (classe 1), le F1-score (malades) et l'AUC.

	Modèle	Accuracy	Recall (malades)	F1-score (malades)	AUC
1	Régression logistique (pondérée)	0.609	0.619	0.448	0.653
4	Random Forest	0.566	0.652	0.436	0.622
3	Arbre de décision	0.546	0.681	0.435	0.616
0	Régression logistique non pondérée	0.747	0.019	0.037	0.656
2	KNN	0.743	0.000	0.000	0.553

Table 1: Comparaison des performances des modèles

La régression logistique simple et le KNN affichent des accuracies élevées mais ne détectent presque aucun malade. Ces modèles sont donc à exclure dans un contexte de santé publique.

La régression logistique pondérée offre un bon équilibre et une excellente interprétabilité.

L'arbre de décision donne le meilleur rappel (68%) mais présente une précision plus faible.

La Random Forest est un compromis robuste, mais moins lisible.

Nous retenons la combinaison de deux modèles :

- la régression logistique pondérée pour interpréter les effets des variables
- l'arbre de décision pour visualiser les profils à risque

	Modèle	Accuracy	Recall (malades)	F1-score (malades)	AUC	Précision (malades)	Malades détectés
0	Régression logistique (pondérée)	0.610	0.620	0.450	0.650	0.350	130
1	Arbre de décision	0.550	0.680	0.440	0.620	0.320	143

Table 2: Comparaison des modèles pondérés

2.5.1 Profils à risque selon la régression logistique pondérée

	Variable	Coefficient	Odds-ratio
7	NB_PART	0.458187	1.581204
8	HISTOIRE	0.283060	1.327184
4	MTS_ANT	0.159239	1.172619
1	ETAT_C	0.021751	1.021990
6	RAISON	-0.014334	0.985768
5	NB_MTS	-0.241886	0.785146
0	SEXE	-0.418310	0.658158
2	AGE	-0.472838	0.623231
3	ORIENT_SEX	-0.792710	0.452617

Table 3: Coefficients et odds-ratios (régression logistique pondérée)

Les facteurs les plus associés à un risque accru sont : NB_PART (1.58), HISTOIRE (1.33) et MTS_ANT (1.17). Inversement, ORIENT_SEX (0.45), AGE (0.62) et SEXE (0.66) réduisent le risque.

Profils à risque selon l'arbre de décision

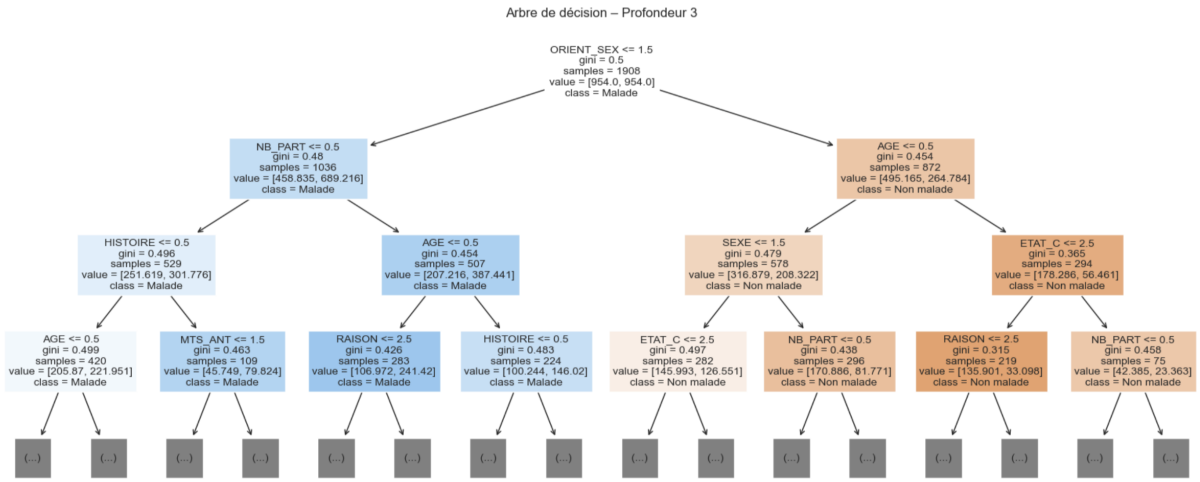

Figure 17: Extrait d'arbre de décision - Profondeur 3

Branche 1 : Homosexuel jeune, peu de partenaires, pas d'exposition connue.
Branche 2 : Homosexuel âgé, peu de partenaires, motivation incertaine.
Branche 3 : Jeune homme homosexuel avec plusieurs partenaires.

Ces profils sont en accord avec les résultats de la régression logistique pondérée.

Comparaison des variables clés entre régression logistique et arbre

	Variable	Odds-ratio	Importance
8	ORIENT_SEX	0.453	0.350
3	ETAT_C	1.022	0.138
4	RAISON	0.986	0.118
6	SEXE	0.658	0.076
7	AGE	0.623	0.075
0	NB_PART	1.581	0.068
5	NB_MTS	0.785	0.066
2	MTS_ANT	1.173	0.063
1	HISTOIRE	1.327	0.046

Table 4: Comparaison croisée des variables selon les modèles

ORIENT_SEX ressort comme très discriminant dans les deux approches. **NB_PART** est très influent en régression, modérément mobilisé par l'arbre. **RAISON** et **ETAT_C** : peu significatifs en régression mais structurants dans l'arbre, ce qui souligne l'intérêt de combiner les deux approches.

Les deux modèles retenus (régression logistique pondérée et arbre de décision) offrent un bon compromis entre performance et interprétabilité. Ils permettent non seulement de détecter une part significative de malades, mais aussi d'identifier clairement les profils les plus à risque. Ces profils seront décrits ci-dessous, en tenant compte à la fois de leur fréquence et de leur implication potentielle dans la transmission.

Conclusion

L'objectif de ce travail était double :

- décrire les individus observés dans le cadre du programme de dépistage de la gonorrhée,
- et identifier les groupes à risque pour mieux cibler les efforts de prévention.

L'analyse menée, fondée sur des tests statistiques et des modèles de classification, a permis d'atteindre ces deux objectifs.

Les résultats montrent que les jeunes hommes homosexuels ayant plusieurs partenaires sexuels récents constituent le groupe le plus à risque. Cette caractérisation s'appuie à la fois sur la régression logistique pondérée et sur l'arbre de décision. Dans le modèle logistique, la variable `NB_PART` présente un odds-ratio de 1.58, ce qui signifie, toutes choses égales par ailleurs, que le fait d'avoir trois partenaires ou plus multiplie par 1.58 les chances d'être diagnostiqué malade. De même, la variable `ORIENT_SEX` a un odds-ratio de 0.45, ce qui indique que les personnes hétérosexuelles ont une probabilité deux fois plus faible d'être malades que les homosexuelles. L'âge inférieur à 30 ans, enfin, est associé à un odds-ratio de 0.62, renforçant l'idée d'un sur-risque chez les jeunes adultes.

Ces tendances sont confirmées par l'arbre de décision, dont les branches principales isolent très tôt les individus jeunes, homosexuels, et ayant plusieurs partenaires comme les cas les plus à surveiller. Le modèle segmente notamment des profils comme "jeune homme homosexuel avec plusieurs partenaires" ou "homosexuel peu exposé mais âgé", permettant de visualiser clairement les chemins menant à un diagnostic positif.

Ces groupes sont donc identifiés comme à risque non pas sur la base d'intuitions, mais sur la base de deux approches complémentaires qui convergent sur les mêmes profils. D'autres facteurs, comme le fait d'avoir été en contact avec un partenaire contaminé (`HISTOIRE`) ou d'avoir eu des antécédents de MTS, ressortent également comme associés à une probabilité plus élevée d'être malade (odds-ratios de 1.32 et 1.17 respectivement).

À l'inverse, les femmes, les hétérosexuels et les individus âgés de 30 ans ou plus apparaissent comme relativement protégés dans les deux approches. Leurs odds-ratios sont inférieurs à 1, et ils sont rarement sélectionnés dans les branches menant à un diagnostic positif dans l'arbre.

Ces conclusions ont une portée concrète : elles peuvent guider les médecins de famille dans leur approche du dépistage, en les aidant à repérer les profils à surveiller en priorité. Elles peuvent également orienter les politiques de prévention, en ciblant les campagnes vers les groupes les plus exposés, tout en maintenant une équité dans l'accès au test.

En combinant la robustesse de la régression logistique pondérée et l'interprétabilité de l'arbre de décision, ce travail propose une démarche fondée sur les données, claire et directement mobilisable dans un cadre de santé publique.