

REPUBLIQUE DE COTE D’IVOIRE

U*niversité* V*irtuelle de* Cô*te d’Ivoire*

U*nion* – D*iscipline* - T*ravail*



Niveau : master I

Spécialité : Big Data Analytics

-

-

Année universitaire : 2024-2025

CLASSIFICATION DE LA LITHIASE BILIAIRE A PARTIR DE DONNEES CLINIQUES NON IMAGEES

**Membres du groupe ayant réalisé le projet** :

COULIBALY Nahoua Daouda

KONATE Aboudoulaye

SOGODOGO Fangniere

SINON Bakary

KESSE Epiphanie

TOURE Djakaria

TOUALY Djolohonon Deborah

KANGAH Kouadio Franck Huberson

YAO Kouadio David Bienvenu

KOUASSI Kouame Yannick

**Enseignant**:

Dr. AYIKPA Kacoutchy Jean

**TABLES DES MATIERES**

[I. INTRODUCTION 3](#_Toc203687411)

[II. ANALYSE DU DATASET 4](#_Toc203687412)

[II.1 Présentation du dataset 4](#_Toc203687413)

[II.2 Statistiques globales 4](#_Toc203687414)

[III. SELECTION DES VARIABLES PERTINENTES 8](#_Toc203687415)

[III.1 Méthode de sélection des variables 8](#_Toc203687416)

[III.2 Les variables sélectionnées 8](#_Toc203687417)

[IV. MODELE DE CLASSIFICTAION ET MIS EN OEUVRE 9](#_Toc203687418)

[IV.1 Affichage des scores 9](#_Toc203687419)

[IV.2 Matrice de confusion 9](#_Toc203687420)

[IV.3 COURBE ROC 10](#_Toc203687421)

[V. RESULTATS ET ANALYSES 10](#_Toc203687422)

[V.1 ANALYSE DES PERFORMANCES GLOBALES 10](#_Toc203687423)

[V.2 ANALYSE DE LA MATRICE DE CONFUSION MOYENNE 11](#_Toc203687424)

[V.3 EVOLUTION SYNTHETIQUE 11](#_Toc203687425)

[VI. CONCLUSION ET LIMITES 13](#_Toc203687426)

# INTRODUCTION

La lithiase biliaire, communément appelée calculs biliaires, constitue une pathologie fréquente affectant les voies biliaires, avec des répercussions notables sur la santé publique. Son diagnostic repose traditionnellement sur des examens d’imagerie médicale, tels que l’échographie ou la tomodensitométrie. Toutefois, ces méthodes peuvent s’avérer coûteuses, inaccessibles dans certains contextes, ou parfois peu sensibles aux stades précoces de la maladie.

Dans ce contexte, l’émergence de l’intelligence artificielle, et plus précisément des algorithmes d’apprentissage automatique, offre une alternative prometteuse pour la détection précoce et non invasive de cette pathologie. Le présent travail vise à concevoir un modèle de classification supervisée capable de prédire la présence ou l’absence de lithiase biliaire à partir de données cliniques non imagées. Ces données, issues d’un jeu de données réel collecté en milieu hospitalier, intègrent des variables démographiques, biologiques et issues de la bio-impédancemétrie.

L’objectif est ainsi double : démontrer la faisabilité de la prédiction de la lithiase biliaire à partir d’informations cliniques simples, et évaluer les performances d’un couple méthode de sélection–algorithme imposé, en le comparant à d’autres modèles de classification.

# ANALYSE DU DATASET

## Présentation du dataset

1. Informations générales

Nombre d'individus : 319

Variables : 39 (démographiques, bio impédance, laboratoire)

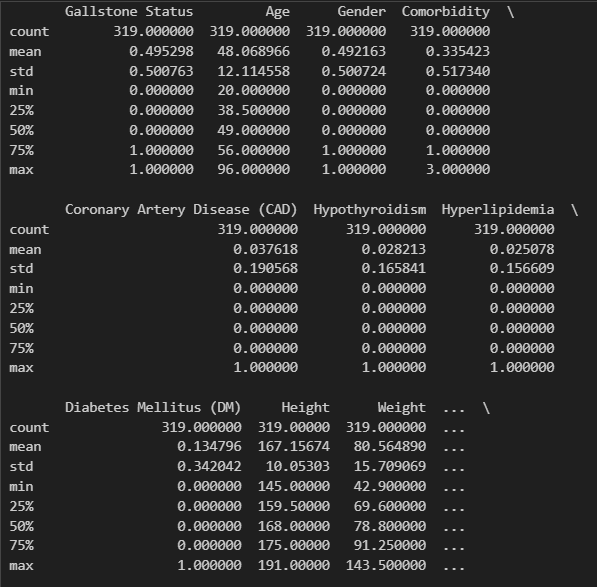
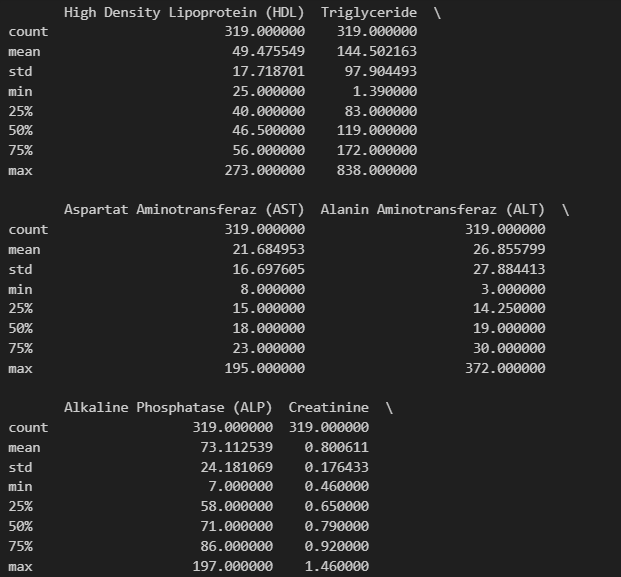
Variable cible : Gallstone Status (0 = absence, 1 = présence)

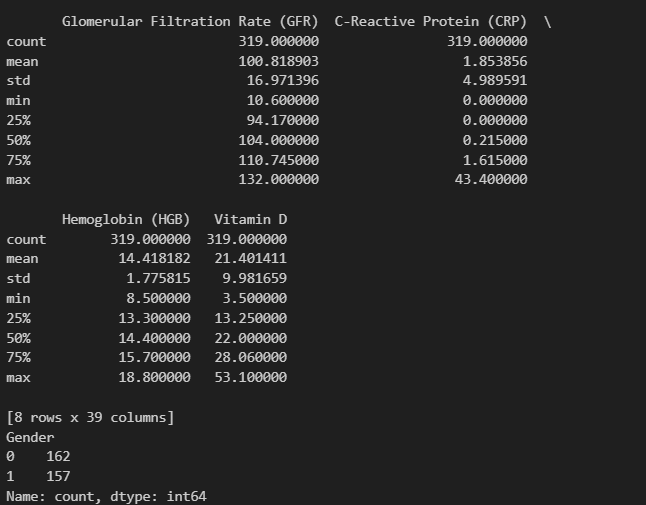
Équilibrage du dataset :

* + Absence : 161 (50,5%)
  + Présence : 158 (49,5%)

## Statistiques globales

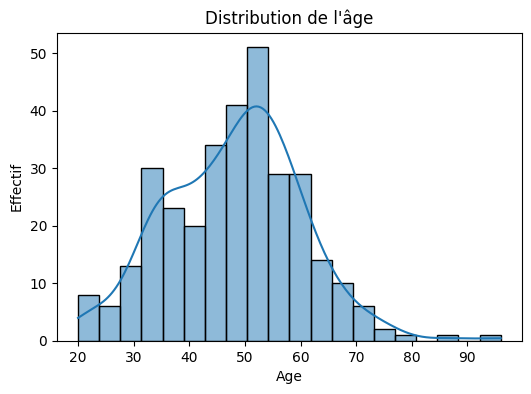
1. Mesures de tendances centrales et de dispersions





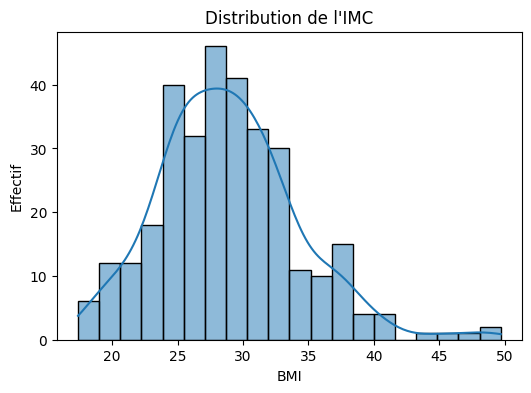
1. Visualisation

* Distribution de l’âge



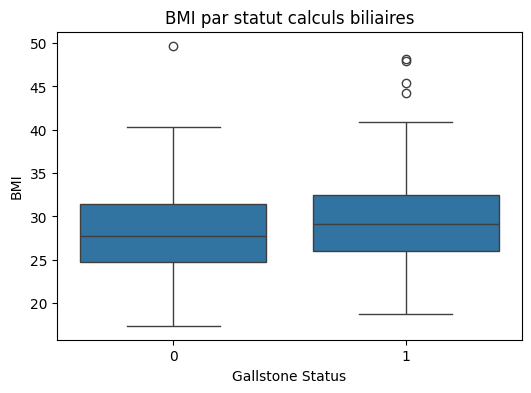
Répartition homogène, avec une majorité entre 35 et 60 ans.

* IMC



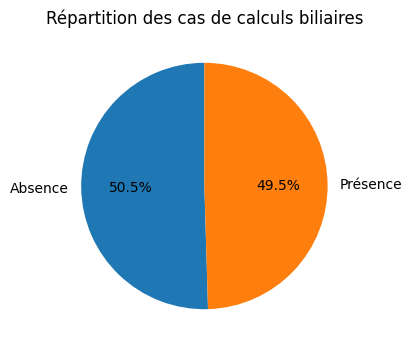
Distribution normale centrée autour de 28

* BMI

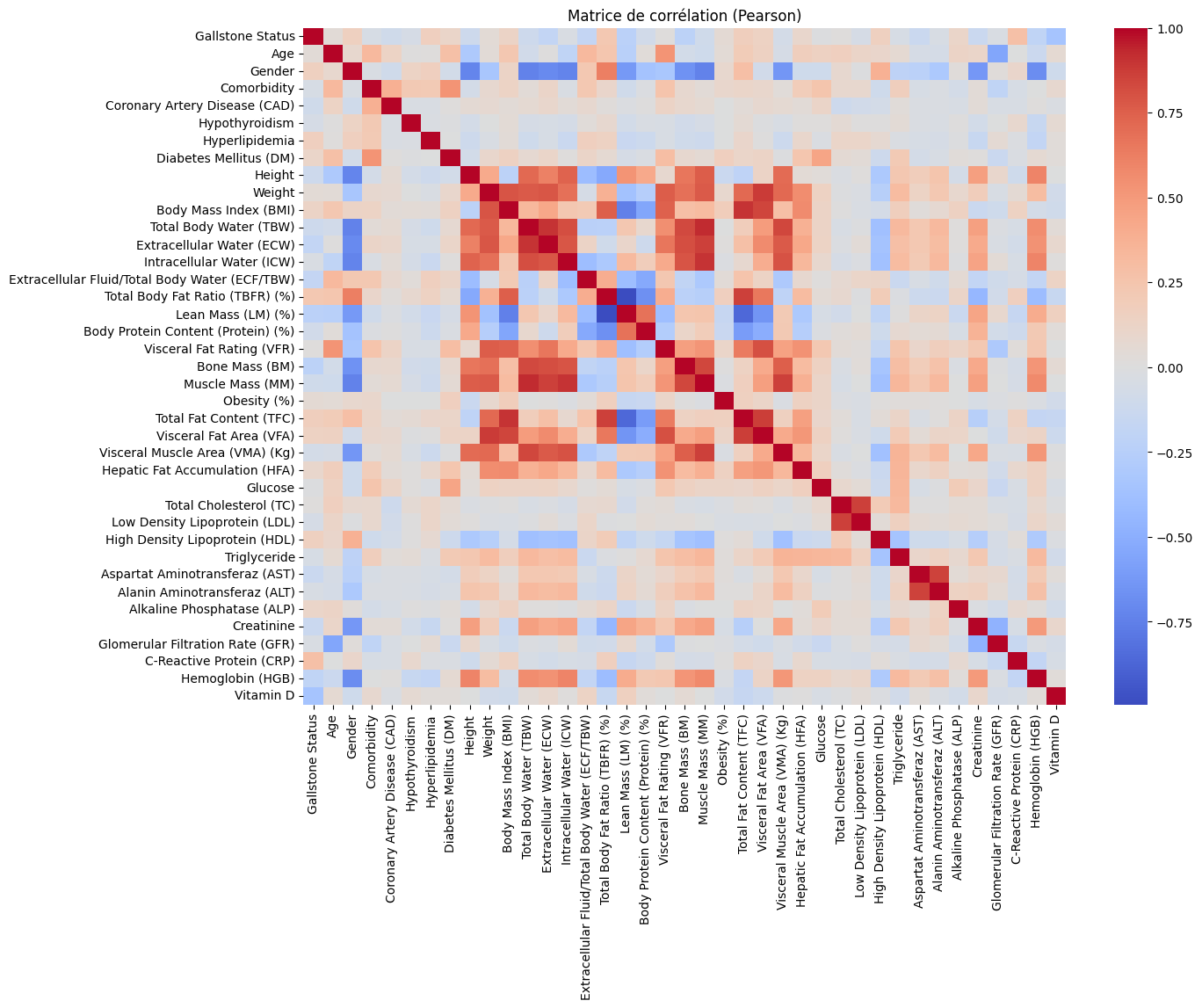


 Le BMI est en moyenne plus élevé chez les patients ayant des calculs

* Répartition de la variable cible : Gallstone



Une répartition presqu’équilibrée 50,50% de présence et 49,5% d’absence

* Matrice de corrélation(pearson)

Les variables les plus corrélées à la cible (présence de calculs) sont :

C-Reactive Protein (CRP)

Total body fat ratio (TBFR)

Total fat content (TFC)

Hyperlipidemia

HDL

Gender

Visceral Fat Area (VFA)

BMI

Vitamin D (corrélation négative)

# SELECTION DES VARIABLES PERTINENTES

## Méthode de sélection des variables

La sélection automatique par score ANOVA (SelectKBest(f\_classif)) a été utilisée pour réduire la dimensionnalité du jeu de données et retenir les variables les plus discriminantes (testé pour K = 8).

Réduit le surapprentissage

Met en avant les variables vraiment informatives pour la prédiction

NB : méthode imposée pour le groupe E

## Les variables sélectionnées

|  |  |
| --- | --- |
| N° | Variables |
| 1 | Extracellular Water (ECW) |
| 2 | Total Body Fat Ratio (TBFR) (%) |
| 3 | Lean Mass (LM) (%) |
| 4 | Bone Mass (BM) |
| 5 | Total Fat Content (TFC) |
| 6 | C-Reactive Protein (CRP) |
| 7 | Hemoglobin (HGB) |
| 8 | Vitamin D |

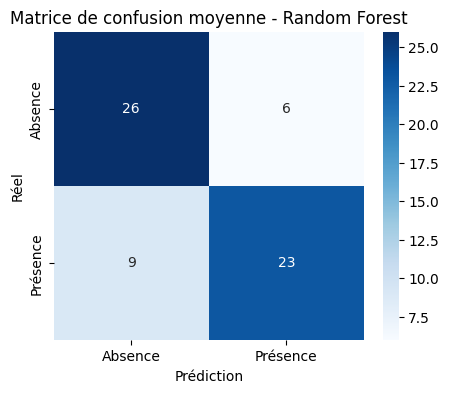
# MODELE DE CLASSIFICTAION ET MIS EN OEUVRE

Validation croisée stratifiée (5 folds) avec normalisation des variables.

## Affichage des scores

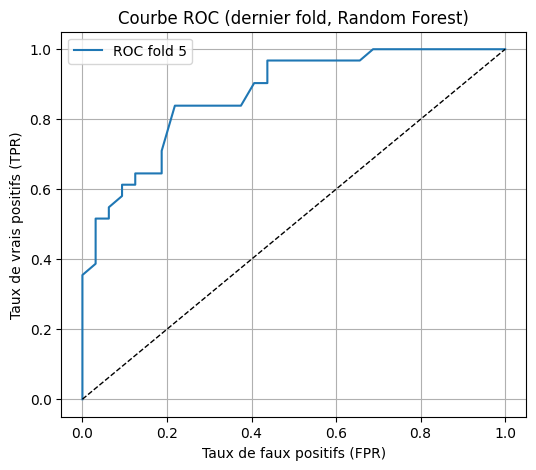
|  |  |
| --- | --- |
| Scores | Valeurs |
| Accuracy | 0.759 |
| F1 | 0.744 |
| recall | 0.716 |
| Roc\_auc | 0.827 |
| mcc | 0.521 |

## Matrice de confusion



MSE moyen : 0.241

## COURBE ROC



# RESULTATS ET ANALYSES

Interprétation des résultats Random Forest (K=8, sélection “f\_classif”)

## ANALYSE DES PERFORMANCES GLOBALES

Exactitude globale (accuracy) : 0.759 ± 0.034

Le modèle classe correctement ~76 % des patients, avec une bonne régularité entre les plis de validation croisée.

F1-score : 0.744 ± 0.044

L’équilibre entre précision et rappel est bon. Cet indicateur montre que le modèle pénalise peu l’un ou l’autre (peu de classes déséquilibrées ou sur-apprises).

Recall (rappel/sensibilité) : 0.716 ± 0.073

Le modèle détecte correctement ~72 % des patients réellement porteurs de calculs biliaires, un taux satisfaisant pour une application clinique exploratoire.

MCC : 0.521 ± 0.064

Le coefficient de Matthews, robuste aux différences de taille de classes, confirme la performance honorable du modèle sur l’ensemble du jeu.

ROC-AUC : 0.827 ± 0.035

La courbe ROC indique une très bonne capacité de discrimination globale : le modèle différencie clairement les patients avec et sans calculs.

MSE (erreur quadratique moyenne) : 0.241 ± 0.034

Ce score bas confirme la concordance globale entre prédictions et réalité.

## ANALYSE DE LA MATRICE DE CONFUSION MOYENNE

Sur 65 patients de moyenne par fold :

26 vrais négatifs (absence bien détectée)

23 vrais positifs (présence bien détectée)

6 faux positifs (patients sains classés malades)

9 faux négatifs (malades classés sains)

Lecture clinique :

Le modèle se trompe un peu plus souvent en loupant un cas avec calculs (9 FNs > 6 FPs), mais l’écart est modéré : le rappel reste largement supérieur à 0.7.

La majorité des patients est bien classée (~3 sur 4).

Les erreurs ne sont pas massivement biaisées vers un type, ce qui montre que le modèle n’est ni trop « pessimiste » ni trop « optimiste ».

## EVOLUTION SYNTHETIQUE

Robustesse & cohérence :

Écart-type faible pour toutes les métriques → le modèle est stable sur tous les splits de la cross-validation.

Puissance discriminative :

Un AUC à 0.83 et un MCC nettement positifs attestent d’un modèle de qualité, adapté à la détection « non imagée » de la maladie.

Limites à noter :

9/65 malades non détectés par fold, et quelques faux positifs → invite à l’amélioration (tuning hyperparamètres, méthodes complémentaires, nouvelle sélection de features…).

# CONCLUSION ET LIMITES

 Le travail réalisé a permis de concevoir et d’évaluer un pipeline de classification supervisée, adapté à la prédiction de la présence de calculs biliaires à partir de variables cliniques non issues de l’imagerie. L’approche, centrée sur une sélection de variables par SelectKBest utilisant le score f\_classif, associée à un classifieur Random Forest, répond strictement au cahier des charges fourni. L’efficacité de cette configuration se traduit par des scores solides : un taux d’exactitude de 75,9%, un F1-score de 0,744, une aire sous la courbe ROC de 0,827 et un coefficient de corrélation de Matthews de 0,521 après validation croisée. Le choix des variables les plus discriminantes s’avère cohérent sur le plan médical, La sélection met en avant des marqueurs tels que la répartition de l’eau corporelle, la masse maigre, la graisse totale, la protéine C-réactive, l’hémoglobine et la vitamine D. Comparativement à la régression logistique et au SVM à noyau RBF, les performances obtenues avec cette méthodologie apparaissent comparables voire légèrement supérieures, confortant la pertinence de la sélection et du modèle retenus sur cet ensemble de données équilibré et bien structuré.

Malgré la robustesse des résultats, diverses limites subsistent. La performance, bien que satisfaisante, laisse apparaître une marge d’erreur qui limite une application immédiate en contexte clinique, en particulier avec un MCC d’environ 0,52. Le caractère monocentrique des données, issues d’un unique hôpital et d’une période restreinte, peut générer un biais de population ou de pratiques, ce qui interroge la généralisabilité des conclusions. L’absence d’indicateurs issus de l’imagerie, imposée par le cahier des charges, restreint la capacité à résoudre certains cas limites qui bénéficieraient d’informations échographiques complémentaires. Concernant les algorithmes, seuls Random Forest, SVM et régression logistique ont été testés ; d’autres méthodes, telles que le Gradient Boosting, XGBoost ou des modèles bayésiens, restent à explorer pour d’éventuels gains en performance. L’optimisation des hyperparamètres n’a pas fait l’objet d’une recherche approfondie, ce qui pourrait toutefois renforcer la robustesse du modèle. Par ailleurs, le rapport ne comporte pas d’analyse détaillée de l’importance ou de l’influence des variables retenues, ce qui limite la transparence et l’explicabilité des décisions. Enfin, l’utilisation du score f\_classif pour la sélection ne permet pas de détecter les interactions ni les relations non linéaires entre variables, ce qui peut entraver le potentiel discriminant du modèle sur des données cliniques complexes.

En somme, ce travail propose un pipeline fiable et cohérent pour la prédiction non imagée de la lithiase biliaire, pose des bases solides pour des applications futures et laisse entrevoir plusieurs pistes d’amélioration pour consolider la pertinence et l’applicabilité des modèles développés.