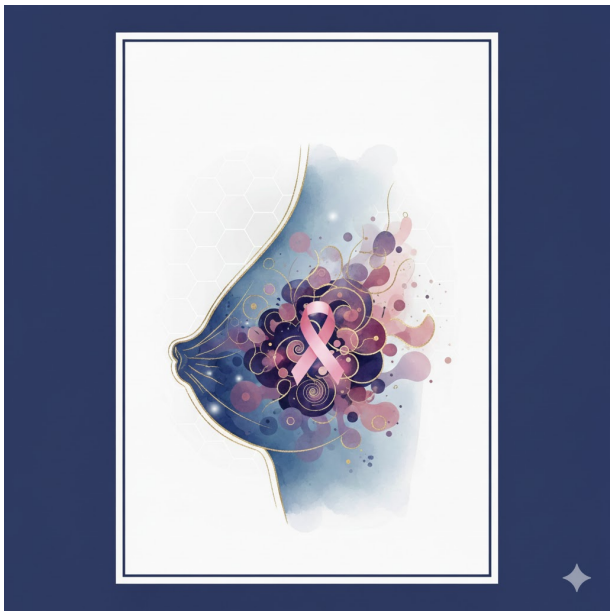

RAPPORT DE PROJET

Intelligence Artificielle et Santé :

Analyse Médicale et Prédiction du Diagnostic par Classification

Présenté par :

EL-WALI IKRAM

Sous la direction de :

M LAGHLIMI

Année Universitaire 2024-2025

Table des matières

Chapitre 1

Introduction et Contexte Métier

1.1 Le Problème

Dans le domaine médical, la fatigue des radiologues, la variabilité entre praticiens et la complexité des images peuvent entraîner des erreurs de diagnostic dans la détection du cancer du sein. Ce dernier représente la forme de cancer la plus fréquente chez les femmes, et un diagnostic précoce constitue le facteur clé pour améliorer les chances de survie. Pour remédier à ce problème, il est donc proposé de développer un assistant basé sur l'intelligence artificielle capable de fournir un second avis médical. Cet assistant analyserait les caractéristiques extraites des images cellulaires afin de prédire si une tumeur est maligne (cancéreuse) ou bénigne (non cancéreuse).

1.2 Objectif du Projet

Dans ce projet, nous avons développé un modèle de machine learning basé sur l'algorithme Random Forest pour classifier des tumeurs à partir d'images. L'objectif principal est d'évaluer la performance du modèle à l'aide de matrices de confusion et de graphiques illustrant ses résultats. Bien que ce travail ne constitue pas un diagnostic médical direct, il fournit une évaluation quantitative de la capacité de l'IA à distinguer les tumeurs malignes des bénignes, ce qui constitue une étape importante vers le développement d'outils d'aide au diagnostic.

1.3 Enjeu Critique : L'Asymétrie des Coûts d'Erreur

Comme souligné dans notre analyse, les erreurs n'ont pas le même impact :

— **Définition du concept :** L'**asymétrie des coûts d'erreur** signifie que toutes les erreurs de classification n'ont pas le même impact. Dans le contexte de la détection du cancer du sein, les conséquences d'une fausse négative et d'une fausse positive

sont très différentes.

- **Exemples d'erreurs et de classifications correctes :**
  - **Vrai positif (TP) :** la tumeur est maligne et le modèle la prédit correctement. C'est le scénario idéal pour un traitement rapide.
  - **Vrai négatif (TN) :** la tumeur est bénigne et le modèle la prédit correctement, évitant des examens inutiles.
  - **Fausse négative (FN) :** la tumeur est maligne mais le modèle la prédit bénigne. Cette erreur est **critique**, car elle peut retarder le traitement et mettre la vie du patient en danger.
  - **Fausse positive (FP) :** la tumeur est bénigne mais le modèle la prédit maligne. Cette erreur entraîne surtout une anxiété accrue et des examens supplémentaires, mais sans danger immédiat pour la santé.
- **Importance dans le diagnostic :** Le coût d'une **fausse négative** est donc beaucoup plus élevé que celui d'une fausse positive. Les matrices de confusion générées permettent de visualiser et de quantifier ces erreurs et succès (TP et TN), et d'évaluer le compromis entre précision et rappel.
- **Lien avec le projet ML :** Le modèle Random Forest développé doit être **optimisé pour réduire les fausses négatives**, même si cela entraîne un nombre légèrement plus élevé de fausses positives. Cette asymétrie guide la **priorité des décisions du modèle** pour maximiser la sécurité des patients.

Chapitre 2

Exploration et Préparation des Données

2.1 Source des Données

Le dataset utilisé est le "Breast Cancer Wisconsin Diagnostic". Il contient 569 instances avec 30 caractéristiques numériques calculées à partir d'images numérisées d'aspiration à l'aiguille fine (FNA).

2.2 Étude des caractéristiques des cellules

Chaque cellule est caractérisée par un ensemble de mesures physiques décrivant sa morphologie. Parmi ces caractéristiques figurent notamment le **rayon moyen**, qui correspond à la moyenne des distances entre le centre de la cellule et les points situés sur son périmètre, ainsi que la **concavité**, qui mesure le degré d'irrégularité et la gravité des portions concaves du contour. D'autres paramètres, tels que la **symétrie** et la **dimension fractale**, permettent de décrire plus finement la forme et la complexité de la cellule, fournissant ainsi des informations essentielles pour la distinction entre tumeurs bénignes et malignes.

2.3 Visualisation et Interprétation

Les histogrammes permettent d'analyser la distribution de chacune des variables numériques du jeu de données et d'obtenir une première compréhension de leur comportement statistique. L'examen de ces graphiques montre que certaines variables, telles que *area_mean* ou *perimeter_mean*, présentent une distribution fortement asymétrique, avec une concentration importante des observations sur certaines plages de valeurs et une longue traîne vers les valeurs élevées. Cette asymétrie peut indiquer une hétérogénéité marquée entre les observations, notamment entre les tumeurs bénignes et malignes. À l'inverse, d'autres variables, comme

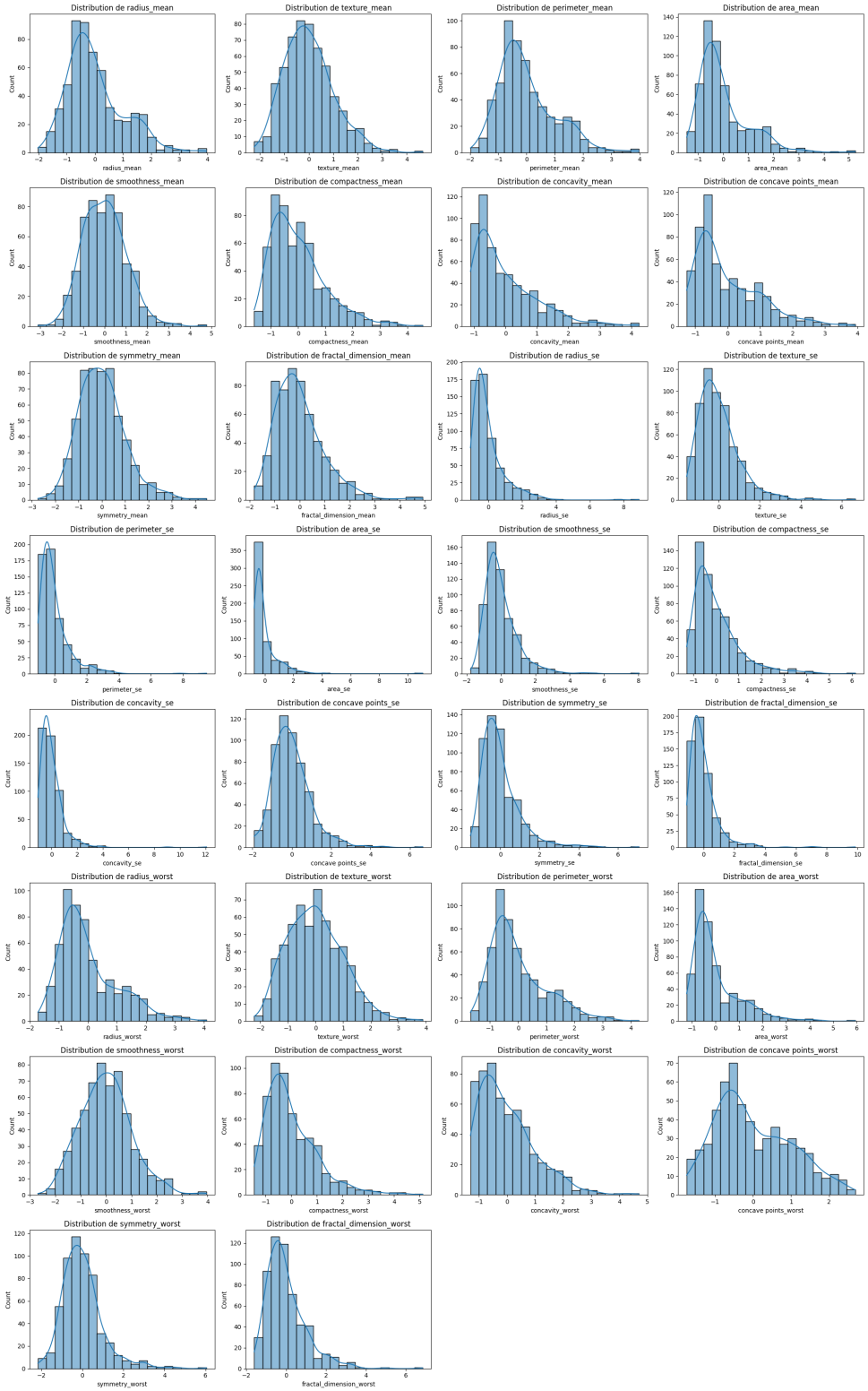

FIGURE 2.1 – Distribution des variables numériques

smoothness_mean, affichent une distribution plus régulière et relativement proche d'une distribution normale, suggérant une variabilité plus homogène au sein du jeu de données. L'analyse de ces distributions est essentielle pour identifier la présence éventuelle de valeurs extrêmes (*outliers*) susceptibles d'influencer les performances des modèles de machine learning. Ces observations constituent une étape clé de l'analyse exploratoire des données, car elles permettent de guider les choix méthodologiques ultérieurs, notamment en ce qui concerne l'application de transformations, la normalisation ou la standardisation des variables avant la phase de modélisation.

Les boxplots constituent un outil graphique essentiel pour analyser la dispersion des données, la position de la médiane ainsi que la présence de valeurs extrêmes pour chaque variable numérique du jeu de données. Leur analyse met en évidence que la majorité des variables présentent des *outliers*, ce qui est une caractéristique courante dans les données médicales, en particulier celles liées aux mesures morphologiques des tumeurs.

Les variables associées à la taille des tumeurs, telles que le rayon (*radius*), le périmètre (*perimeter*) et l'aire (*area*), se distinguent par une étendue importante et une forte dispersion. Cette observation confirme l'existence d'une grande variabilité dans les dimensions des tumeurs présentes dans le dataset, ce qui peut refléter des différences significatives entre les tumeurs bénignes et malignes.

En revanche, les variables décrivant des caractéristiques plus fines, comme la texture, la compacité ou la concavité, présentent généralement des distributions plus resserrées autour de la médiane. Cette concentration indique une variabilité plus limitée de ces paramètres au sein des observations.

Ces visualisations jouent un rôle clé dans l'analyse exploratoire des données, car elles permettent d'identifier les valeurs extrêmes susceptibles d'influencer les performances des modèles de machine learning. Elles fournissent ainsi des indications précieuses pour décider de l'application éventuelle de transformations des variables ou de méthodes spécifiques de traitement des outliers avant la phase de modélisation.

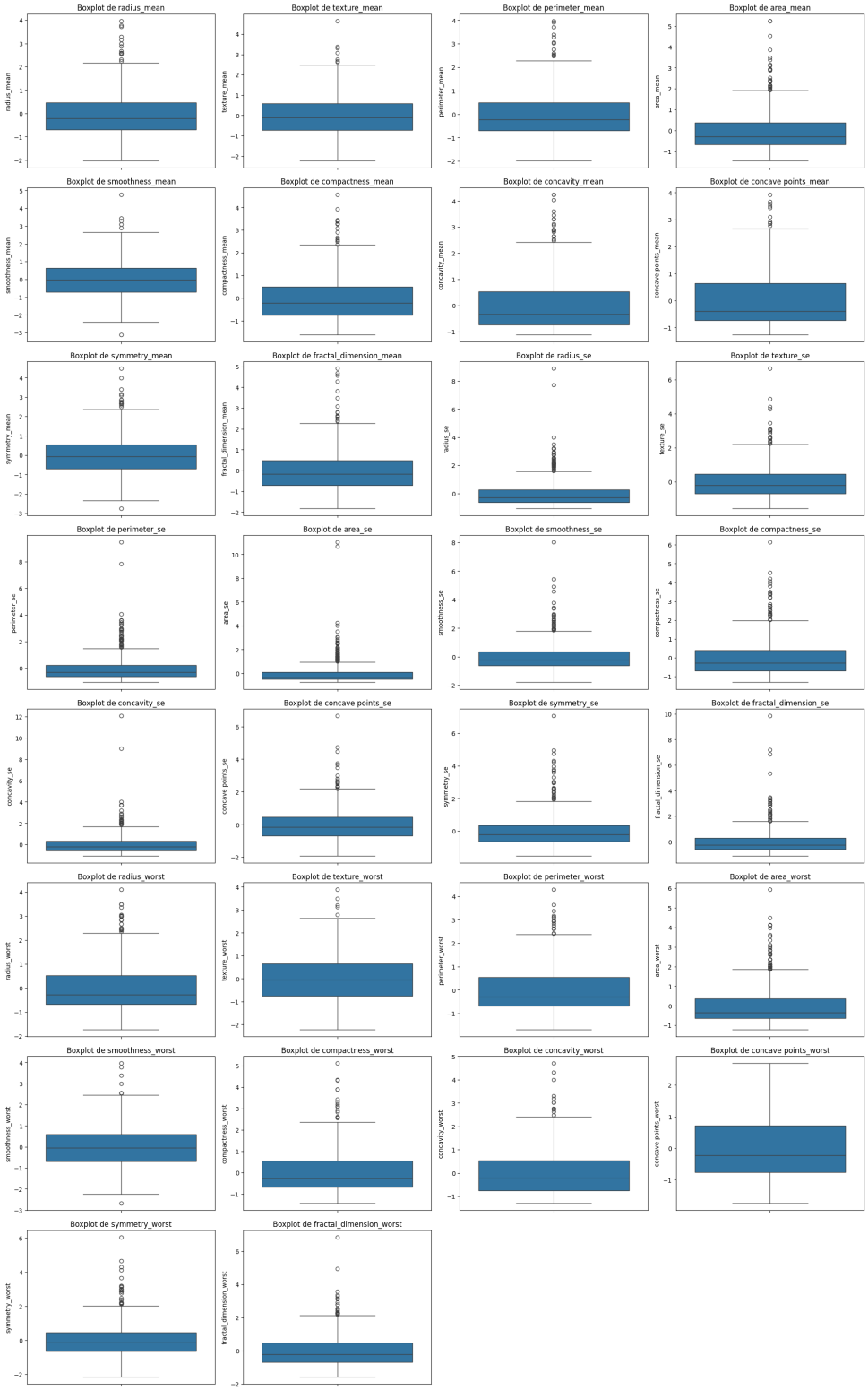

FIGURE 2.2 – Boxplots pour détecter les outliers

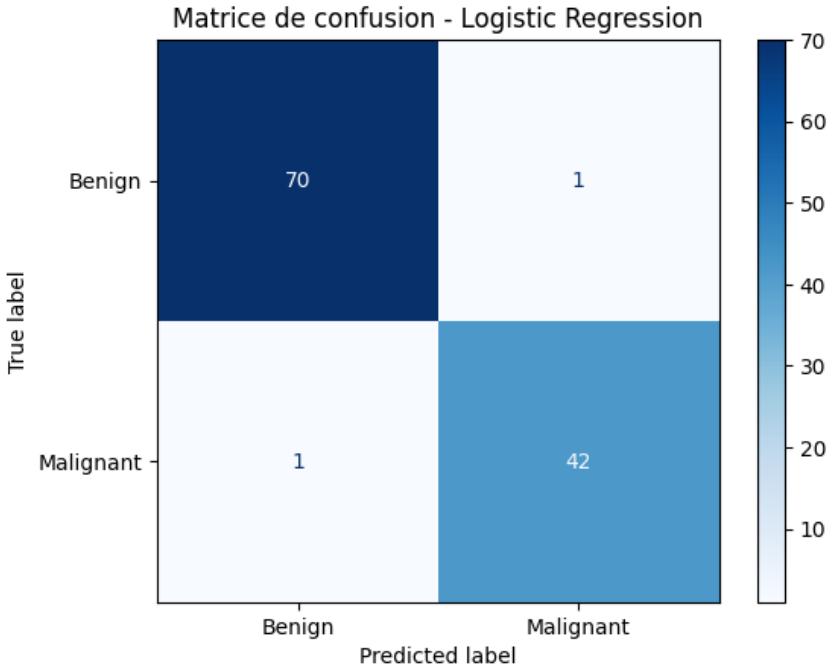

FIGURE 2.3 – matrice de confusion : logistic regression

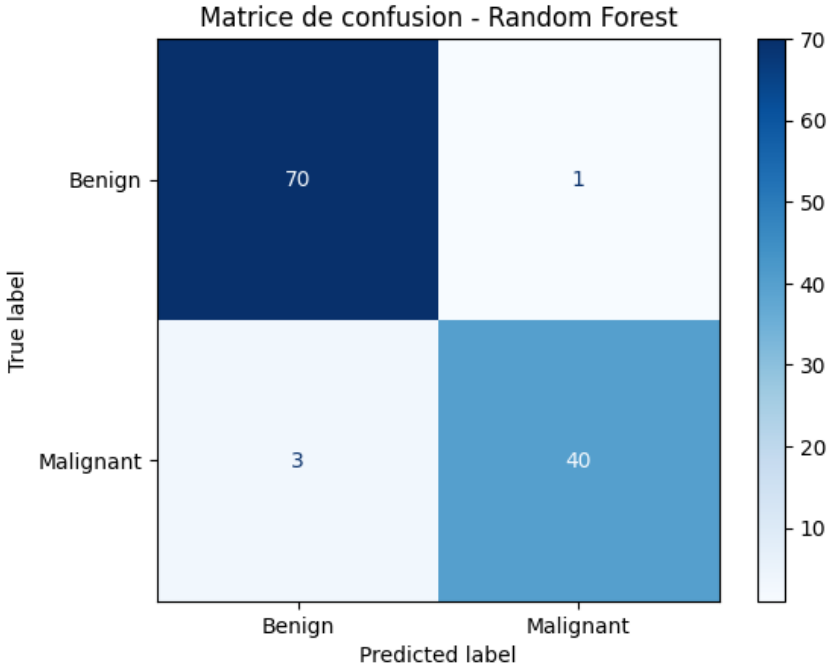

FIGURE 2.4 – matrice de confusion : rendom forest

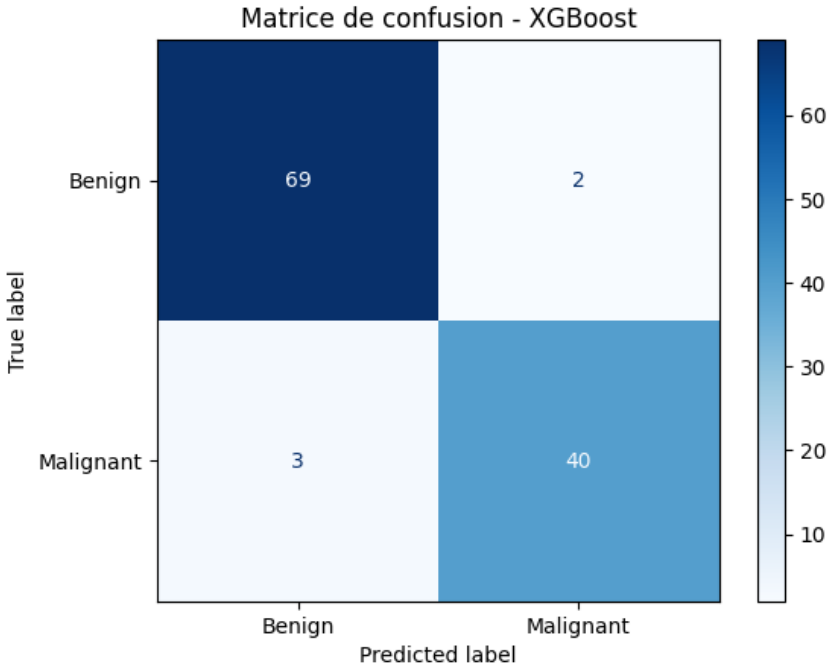

FIGURE 2.5 – matrice de confusion : XG BOOST

Chapitre 3

Préparation et Transformation des Données

3.1 Nettoyage des données

Le nettoyage des données consiste à :

— Supprimer les doublons.

— Corriger les erreurs de saisie ou valeurs aberrantes.

— Standardiser les formats de données (dates, chaînes de caractères, etc.).

3.2 Gestion des valeurs manquantes

La gestion des valeurs manquantes est essentielle pour éviter que le modèle ne soit biaisé ou inefficace. Les méthodes courantes sont :

— Supprimer les lignes ou colonnes avec un trop grand nombre de valeurs manquantes.

— Remplacer les valeurs manquantes par la moyenne, la médiane ou le mode.

— Utiliser des techniques avancées d'imputation basées sur des modèles prédictifs.

3.3 Transformation des données

La transformation des données permet de préparer les données pour le modèle :

— Normalisation ou standardisation des variables numériques.

— Encodage des variables catégorielles (one-hot encoding, label encoding).

— Création de nouvelles fonctionnalités (feature engineering) si nécessaire.

3.4 Séparation des données

Pour évaluer correctement les modèles, les données sont divisées en :

— Ensemble d'apprentissage : pour entraîner le modèle.
— Ensemble de test : pour évaluer la performance du modèle.

3.5 Résumé de la préparation des données

La préparation des données assure que les modèles de *Machine Learning* reçoivent des données propres, complètes et adaptées, améliorant ainsi leur performance et leur robustesse.

Chapitre 4

Modélisation et Choix de l'Algorithme

4.1 Le Random Forest (Forêt Aléatoire)

Nous avons privilégié l'algorithme Random Forest en raison de sa capacité à modéliser des relations complexes et non linéaires entre les variables explicatives et la variable cible. Basé sur le principe du *Bagging* (*Bootstrap Aggregating*), le Random Forest consiste à construire une multitude d'arbres de décision à partir de sous-échantillons aléatoires du jeu de données. Chaque arbre est entraîné indépendamment, et la prédiction finale du modèle est obtenue par un mécanisme de vote majoritaire, ce qui permet de réduire le risque de surapprentissage et d'améliorer la robustesse globale du modèle. Cette approche est particulièrement adaptée aux données médicales, souvent caractérisées par une forte variabilité et des interactions complexes entre les variables, rendant le Random Forest pertinent pour la classification des tumeurs bénignes et malignes.

4.1.1 Avantages pour le Diagnostic Médical

Contrairement aux modèles linéaires classiques, tels que la régression logistique, le Random Forest offre la possibilité d'estimer l'importance relative des variables dans le processus de décision. Cette caractéristique permet d'identifier les critères morphologiques qui contribuent le plus à la classification des tumeurs en bénignes ou malignes. L'analyse de l'importance des variables apporte ainsi une dimension interprétable au modèle, facilitant la compréhension des mécanismes sous-jacents à ses prédictions. Dans un contexte médical, cette transparence est particulièrement précieuse, car elle permet aux professionnels de santé de mieux appréhender les facteurs influençant la décision du modèle et d'envisager son utilisation comme un outil d'aide au diagnostic plutôt que comme un système de décision autonome.

4.2 Importance des Variables

L'analyse de l'importance des variables met en évidence que les caractéristiques *mean concave points* et *mean area* figurent parmi les facteurs les plus discriminants dans l'identification des tumeurs malignes. Ces variables jouent un rôle déterminant dans le processus de classification, car elles traduisent respectivement le degré d'irrégularité des contours de la tumeur et sa taille moyenne. Leur contribution élevée au modèle souligne leur pertinence clinique et confirme que certaines caractéristiques morphologiques sont particulièrement informatives pour distinguer les tumeurs bénignes des tumeurs malignes.

Chapitre 5

Évaluation et Conclusion

Ce projet a permis d'explorer l'utilisation des techniques de machine learning pour la classification des tumeurs du sein à partir de caractéristiques extraites d'images médicales. À travers une analyse exploratoire approfondie des données, incluant des histogrammes et des boxplots, nous avons mis en évidence la diversité des distributions des variables ainsi que la présence de valeurs extrêmes, éléments essentiels à prendre en compte avant toute phase de modélisation.

Plusieurs modèles de classification, notamment la régression logistique, le Random Forest et XGBoost, ont été mis en œuvre et évalués. Les résultats obtenus montrent que ces modèles sont globalement capables de distinguer efficacement les tumeurs bénignes des tumeurs malignes. L'analyse des matrices de confusion a permis d'aller au-delà des métriques globales telles que l'accuracy, en identifiant la nature des erreurs de classification et en soulignant l'importance de l'asymétrie des coûts d'erreur dans un contexte médical.

Bien que ce travail ne constitue pas un diagnostic médical, il met en évidence le potentiel des modèles de machine learning comme outils d'aide à la décision pour les professionnels de santé. Des perspectives d'amélioration incluent l'optimisation des hyperparamètres, l'intégration de nouvelles caractéristiques issues des images et l'évaluation des modèles sur des jeux de données plus larges et plus diversifiés, afin de renforcer la robustesse et la fiabilité des résultats obtenus.