Classification de tumeur du sein

BELKALEM Rayane

AMROUNI Moh-Ouali

SOUPAYA Raphael

3A IAMD

15 décembre 2025

Table des matières

1 Contexte du projet

Ce projet porte sur la classification automatique des tumeurs du sein à partir du jeu de données *Breast Cancer Wisconsin (Diagnostic)*. Ce jeu de données, largement utilisé dans la littérature en apprentissage automatique, regroupe 569 observations de tumeurs du sein, chacune décrite par 30 caractéristiques morphologiques issues d'analyses microscopiques. Les variables mesurent différentes propriétés des cellules, telles que la taille, la texture, la compacité ou encore l'irrégularité des contours.

Chaque échantillon est associé à un diagnostic clinique : **bénin** ou **malin**. L'objectif du projet est de prédire ce diagnostic à partir des caractéristiques morphologiques des tumeurs, en s'appuyant sur plusieurs modèles de classification supervisée.

Le travail demandé couvre l'ensemble de la démarche de modélisation : préparation et exploration statistique des données, mise en place de plusieurs modèles de classification (perceptron, régression logistique, SVM linéaire, réseau de neurones), comparaison de stratégies d'évaluation (séparation entraînement/test et validation croisée), et analyse des résultats obtenus.

2 Preparation des données

Le jeu de données a été importé à partir de la plateforme Kaggle au format CSV. Une première étape de préparation a consisté à nettoyer et structurer les données avant toute analyse. Deux colonnes ont été supprimées : l'identifiant **id**, qui ne contient aucune information utile pour la classification, et la colonne **Unnamed : 32**, entièrement vide.

La variable cible **diagnosis** a été encodée en valeurs numériques afin d'être compatible avec les algorithmes de classification : la classe **B** (bénin) a été codée 0, et la classe **M** (malin) a été codée 1. Aucune valeur manquante n'a été détectée dans les 569 observations, ce qui permet de conserver l'intégralité des données.

Les 30 variables explicatives restantes sont toutes numériques. Elles sont organisées en trois groupes cohérents :

- les mesures **mean** représentent la valeur moyenne observée sur l'ensemble des cellules d'une tumeur ;
- les mesures **se** quantifient la variabilité entre les tumeurs ;
- les mesures **worst** correspondent aux valeurs extrêmes observées, c'est-à-dire aux cellules les plus atypiques.

Cette structure est particulièrement adaptée à l'analyse de la morphologie cellulaire, car

elle permet de capturer à la fois les caractéristiques globales de la tumeur, l'hétérogénéité entre les cellules, et la présence éventuelle de cellules présentant des anomalies extrêmes, souvent révélatrices d'une tumeur maligne. Les données préparées sont ensuite utilisées dans l'analyse exploratoire.

3 Analyse exploratoire des données

3.1 Description globale des données

Les mesures moyennes révèlent une forte variabilité dans la taille des cellules. Des variables telles que **radius_mean**, **perimeter_mean** et **area_mean** présentent des amplitudes très élevées. Par exemple, **area_mean** s'étend de 143 à 2501, indiquant des différences importantes entre tumeurs de petite taille et tumeurs beaucoup plus volumineuses. Ce type de dispersion est cohérent avec le cas clinique, les tumeurs malignes étant souvent plus étendues et présentant une croissance plus agressive.

Les mesures relatives à la forme des cellules, comme **compactness_mean**, **concavity_mean** et **concave points_mean**, présentent des distributions asymétriques. Dans ces cas, la médiane est nettement inférieure à la moyenne, et les valeurs maximales sont largement supérieures aux quartiles. Cela signifie que la plupart des tumeurs possèdent des cellules relativement régulières, mais qu'un sous-ensemble présente des déformations importantes. Ces valeurs extrêmes ne constituent pas un bruit statistique mais peuvent traduire la présence de cellules très irrégulières, un signe fréquent des tumeurs malignes.

Les mesures d'erreur standard (**se**) apportent une information complémentaire en quantifiant l'hétérogénéité des cellules au sein d'une même tumeur. Certaines variables, telles que **radius_se** ou **concavity_se**, illustrent particulièrement bien ce phénomène. Par exemple, la valeur médiane de **radius_se** est de 0.324 mais son maximum atteint 2.873, c'est presque dix fois plus. De même, **concavity_se** présente une médiane de 0.0259 mais un maximum de 0.396. Ces écarts témoignent de tumeurs dont les cellules présentent des tailles ou des formes très hétérogènes, ce qui est typique des tissus malins.

Enfin, les mesures **worst** correspondent aux valeurs maximales observées pour chacune des caractéristiques et représentent donc les cellules les plus atypiques au sein de chaque tumeur. Leur amplitude est particulièrement élevée : par exemple, **area_worst** varie de 185 à 4254, et **concavity_worst** peut atteindre 1.252. Ces valeurs extrêmes reflètent directement la présence de cellules présentant des anomalies marquées, fréquemment associées à un caractère malin.

De manière générale, l'analyse descriptive montre que les tumeurs malignes tendent à se distinguer à la fois par leur taille, par l'irrégularité de leurs contours et par une forte hétérogénéité cellulaire. Les variables issues des trois groupes (**mean**, **se** et **worst**) fournissent donc des dimensions essentielles pour discriminer efficacement les tumeurs bénignes et malignes.

3.2 Comparaison entre tumeurs bénignes et malignes

Afin de confirmer les tendances observées dans la description globale du jeu de données, nous avons comparé la distribution de plusieurs variables clés (issues de la sous-section précédente) entre les tumeurs bénignes (classe 0) et les tumeurs malignes (classe 1). Les boxplots permettent de visualiser les différences entre les deux groupes, tandis que le calcul des moyennes par classe fournit une interprétation quantitative de ces écarts.

Variables liées à la taille : Les figures obtenues pour **radius_mean** et **area_mean** mettent en évidence un écart net entre les deux types de tumeurs. Les tumeurs bénignes présentent en moyenne des cellules plus petites (radius_mean $\approx$ 12.15, area_mean $\approx$ 462.79), tandis que les tumeurs malignes se distinguent par des dimensions nettement plus élevées (radius_mean $\approx$ 17.46, area_mean $\approx$ 978.37). Ces résultats confirment que les tumeurs malignes tendent à être plus volumineuses, ce qui est cohérent avec le comportement biologique d'une croissance agressive.

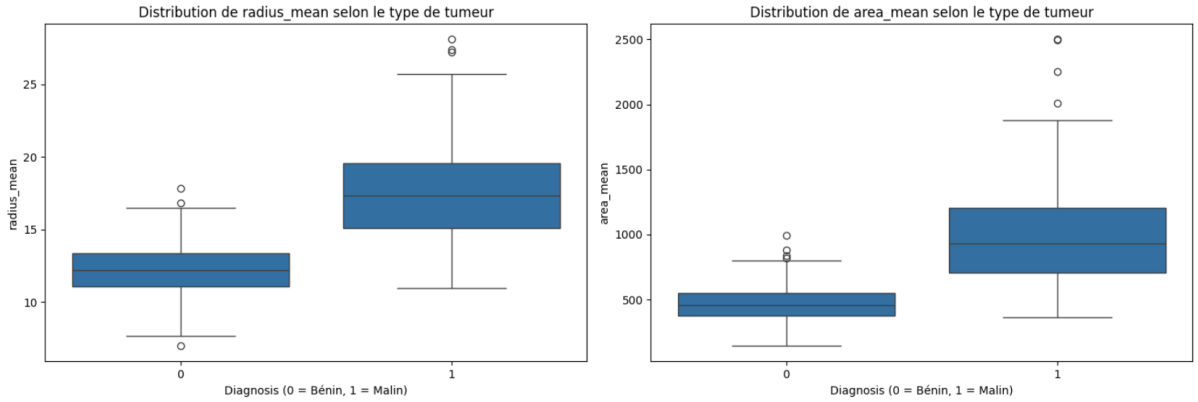

FIGURE 1 – Box plot de radius et area mean

Variables décrivant la forme : Les différences sont également marquées pour des variables telles que **concavity_mean** et **compactness_mean**. Pour les tumeurs bénignes, la concavité moyenne reste faible (0.046), signe de contours cellulaires relativement réguliers. À l'inverse, les tumeurs malignes présentent une concavité trois fois plus élevée (0.161), ce qui reflète la déformation et l'irrégularité de leurs cellules. Le même

phénomène est observé pour compactness_mean, où les valeurs des tumeurs malignes sont presque doublées. Ces résultats renforcent l'idée que l'asymétrie et l'irrégularité des contours constituent des indicateurs morphologiques pertinents de malignité.

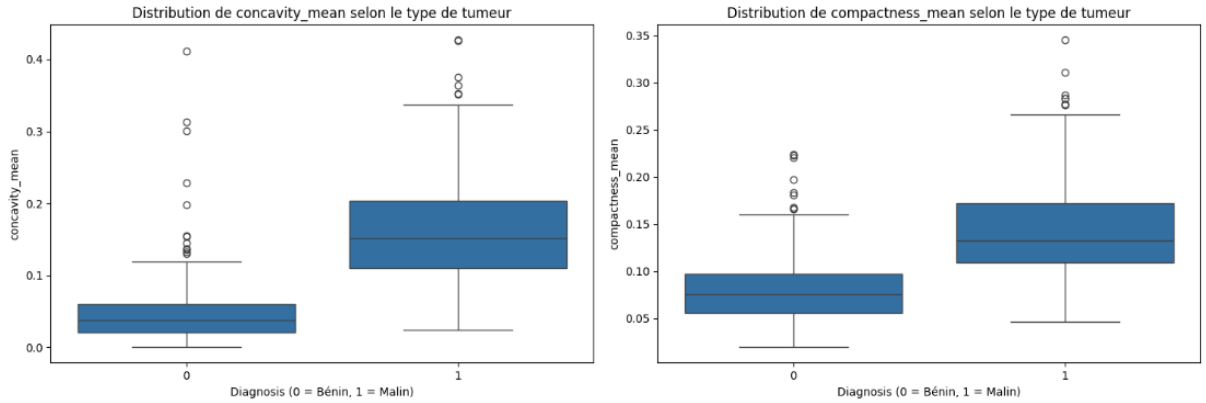

FIGURE 2 – Box plot de concavity et compactness mean

Variabilité intra-tumeur : Les variables d'erreur standard confirment également une séparation nette entre les deux classes. Par exemple, **radius_se** passe d'une valeur moyenne de 0.28 pour les tumeurs bénignes à 0.60 pour les tumeurs malignes. De même, **concavity_se** augmente quasiment du simple au double. Ces différences montrent que les tumeurs malignes présentent une plus grande hétérogénéité morphologique, autrement dit des cellules beaucoup moins uniformes les unes par rapport aux autres. Cette hétérogénéité est caractéristique des cellules anormales qui coexistent avec d'autres plus proches d'un fonctionnement normal.

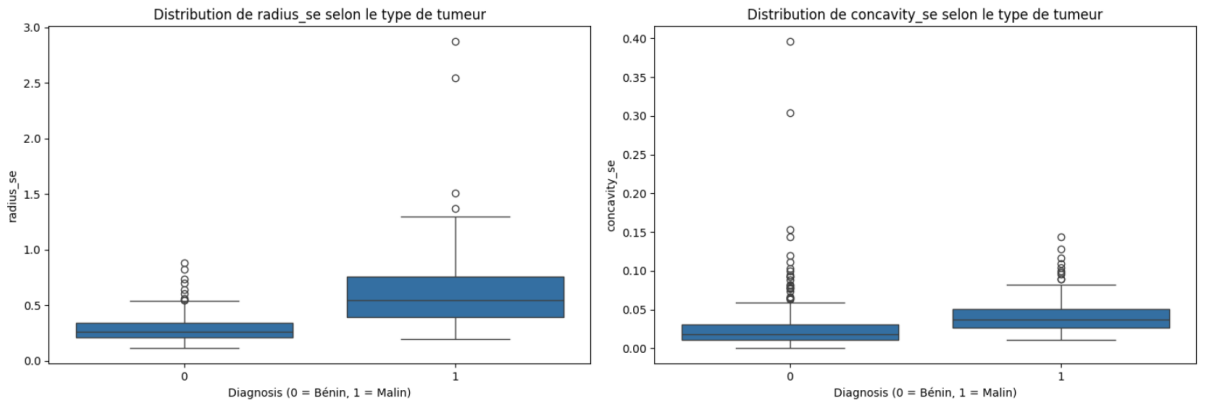

FIGURE 3 – Box plot de radius et concavity se

Valeurs extrêmes (*worst*) : Les différences les plus marquées apparaissent dans les

mesures **worst**, qui reflètent les anomalies maximales observées pour chaque tumeur. Par exemple, **radius_worst** passe en moyenne de 13.38 pour les tumeurs bénignes à plus de 21.13 pour les tumeurs malignes, et **area_worst** augmente de 558.90 à 1422.29. La variable **concavity_worst** illustre particulièrement bien cette tendance, avec des valeurs moyennes multipliées par presque trois entre les deux classes. Ces mesures extrêmes traduisent directement la présence de cellules fortement atypiques, signature classique des tumeurs agressives.

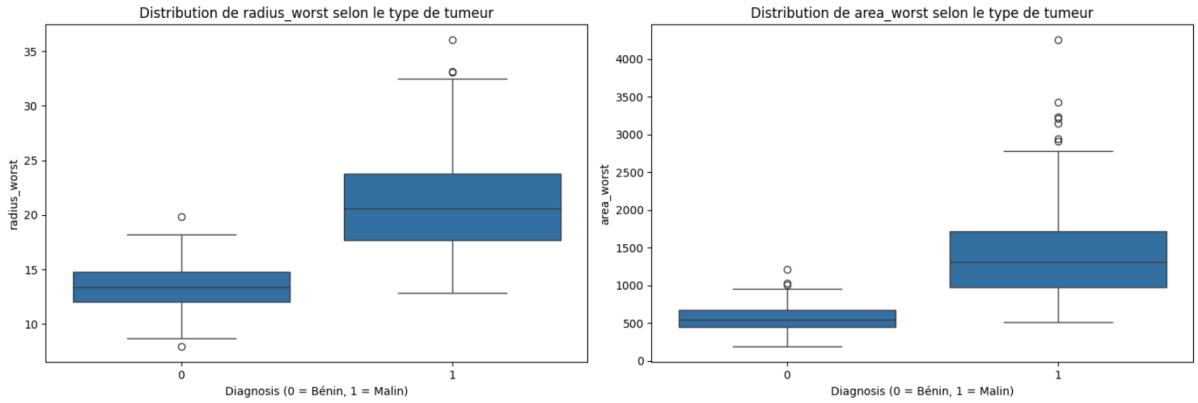

FIGURE 4 – Box plot de area et radius worst

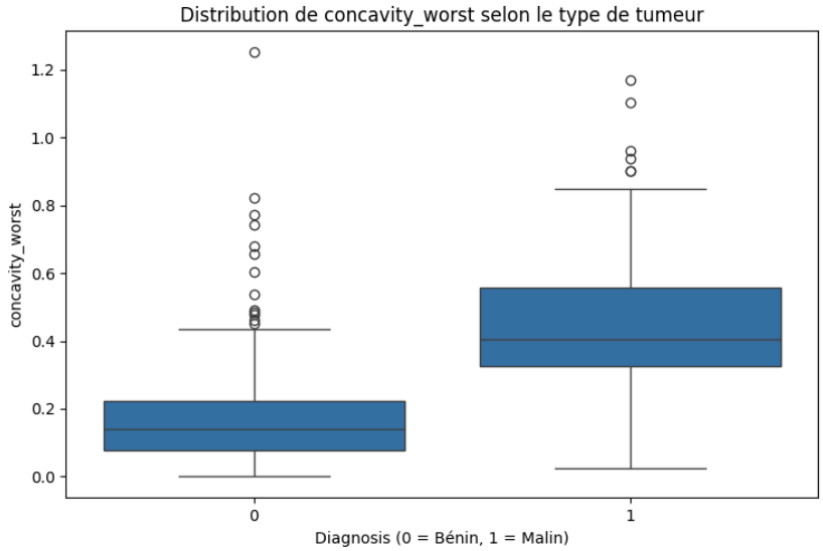

FIGURE 5 – Box plot de concavity worst

Synthèse : L'ensemble de ces observations confirme les tendances décrites dans l'analyse descriptive : les tumeurs malignes sont généralement plus grandes, présentent des contours plus irréguliers et manifestent une hétérogénéité cellulaire beaucoup plus importante que les tumeurs bénignes. Ces différences nettes expliquent pourquoi les variables de taille, de forme et les mesures extrêmes constituent de bon prédicteurs dans les modèles de classification supervisée.

Il est important de noter que d'autres caractéristiques, telles que **perimeter**, **smoothness**, **symmetry**, **fractal dimension** ou encore **texture**, existent également sous leurs déclinaisons *mean*, *se* et *worst*. Ces variables n'ont pas été étudiées individuellement dans le détail parce qu'une première inspection statistique suggère qu'elles sont fortement liées aux caractéristiques déjà analysées. Par exemple, les variables de taille (*radius*, *perimeter*, *area*) évoluent généralement de manière cohérente, tandis que les variables de forme (*compactness*, *concavity*, *concave points*) décrivent souvent des aspects similaires.

3.3 Corrélation entre les features

Afin de vérifier cette intuition, nous allons désormais examiner la matrice de corrélation des différentes caractéristiques. Cette analyse permettra de confirmer la présence de groupes de variables redondantes, d'identifier celles qui sont les plus informatives pour la classification, et de justifier les choix effectués dans cette première exploration.

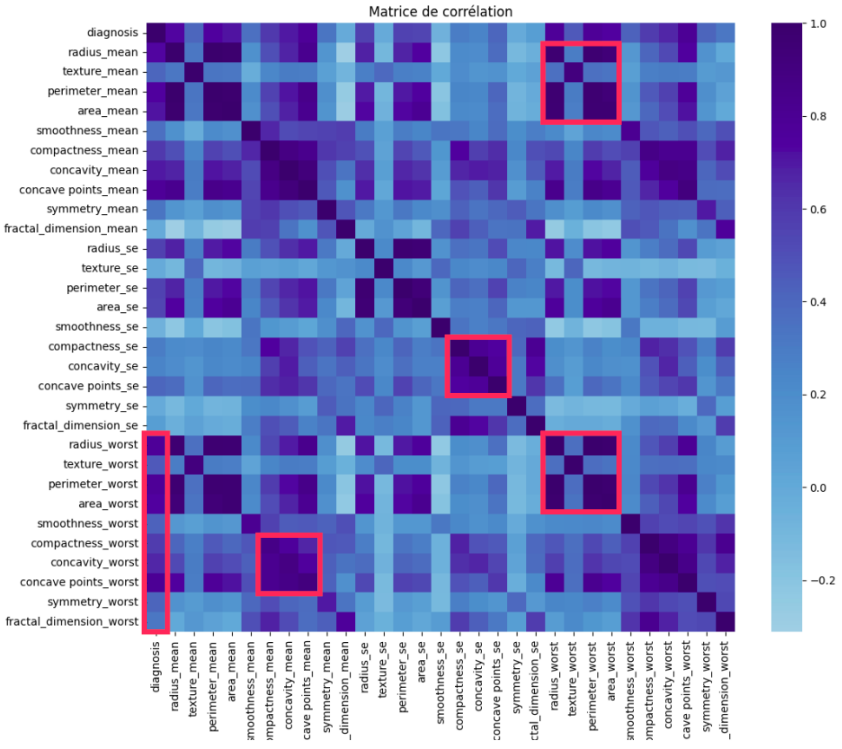

FIGURE 6 – Matrice de corrélation des caractéristiques

La matrice de corrélation met en évidence plusieurs structures remarquables, soulignées par les zones entourées. La première région, située dans la partie supérieure, correspond aux corrélations entre les mesures moyennes de taille et de texture (**radius_mean**, **texture_mean**, **perimeter_mean**, **area_mean**) et leurs homologues *worst* (**radius_worst**, **texture_worst**, **perimeter_worst**, **area_worst**). Cette zone montre que les tumeurs qui présentent des valeurs moyennes élevées de taille ou de texture sont également celles qui possèdent les cellules les plus extrêmes. Autrement dit, lorsqu'une tumeur est globalement volumineuse, elle tend aussi à contenir des cellules particulièrement grandes, ce qui est cohérent avec le comportement attendu des tumeurs malignes.

Une deuxième zone, centrée sur les variables **compactness_se**, **concavity_se**, **concave points_se** et **symmetry_se**, met en évidence un bloc de corrélations fortes au sein des mesures d'erreur standard liées à la forme. Ce cluster indique que la variabilité intra-tumeur des contours cellulaires agit de manière conjointe : lorsqu'une tumeur présente une forte variabilité de concavité, elle présente également une variabilité importante de compacité, de concave points et de symétrie. Cette cohérence interne confirme que l'hétérogénéité morphologique est global.

Une troisième région, située en bas à droite de la matrice, correspond aux corrélations entre les mesures *worst* de taille et de texture (**radius_worst**, **texture_worst**, **perimeter_worst**, **area_worst**). Ces variables sont fortement corrélées entre elles, ce qui signifie que les cellules les plus extrêmes d'une tumeur sont simultanément anormales sur plusieurs dimensions (rayon, périmètre, aire, texture). Les anomalies sévères ne se limitent donc pas à une seule mesure mais affectent plusieurs caractéristiques géométriques en même temps.

La colonne correspondante à la variable **diagnosis**, entourée dans la partie inférieure gauche, montre que les corrélations les plus élevées avec la classe *maligne* sont obtenues pour les variables *worst*. En particulier, **area_worst**, **radius_worst**, **perimeter_worst**, ainsi que **concavity_worst** et **concave points_worst** apparaissent comme les caractéristiques les plus informatives pour discriminer tumeurs bénignes et malignes. Cela confirme l'hypothèse formulée lors de l'analyse descriptive, selon laquelle les valeurs extrêmes des mesures morphologiques jouent un rôle central dans la classification.

Enfin, la zone située à droite de cette colonne met en évidence des corrélations élevées entre les mesures de forme *worst* (**compactness_worst**, **concavity_worst**, **concave points_worst**) et leurs versions moyennes (**compactness_mean**, **concavity_mean**, **concave points_mean**). Cela signifie que les tumeurs dont les contours sont déjà irréguliers en moyenne ont également tendance à présenter des cellules encore plus déformées. Les mesures *worst* amplifient ainsi les tendances observées sur les mesures *mean*, ce qui

renforce leur pouvoir discriminant.

Dans l'ensemble, ces différentes zones confirment les conclusions précédentes : les tumeurs malignes se caractérisent par une augmentation conjointe de la taille, de l'irrégularité des contours et de l'hétérogénéité cellulaire, et ce sont principalement les valeurs extrêmes des caractéristiques qui portent l'information la plus discriminante pour la classification.

3.4 Visualisation des distributions par classe

Pour compléter l'analyse précédente, nous avons représenté la distribution de quelques variables clés en distinguant explicitement les deux classes (0 : tumeur bénigne, 1 : tumeur maligne). Les histogrammes sont accompagnés d'une estimation de densité, ce qui permet de visualiser finement la forme des distributions et le degré de recouvrement entre les deux groupes.

La Figure ci-dessous illustre la distribution de **radius mean**. On observe que les tumeurs bénignes sont majoritairement concentrées autour de valeurs comprises entre 10 et 14, alors que les tumeurs malignes se situent plutôt entre 15 et 20. Le recouvrement entre les deux distributions reste limité, essentiellement dans une zone centrale autour de 14–15, ce qui confirme que le rayon moyen des cellules est une caractéristique très discriminante entre les deux types de tumeurs.

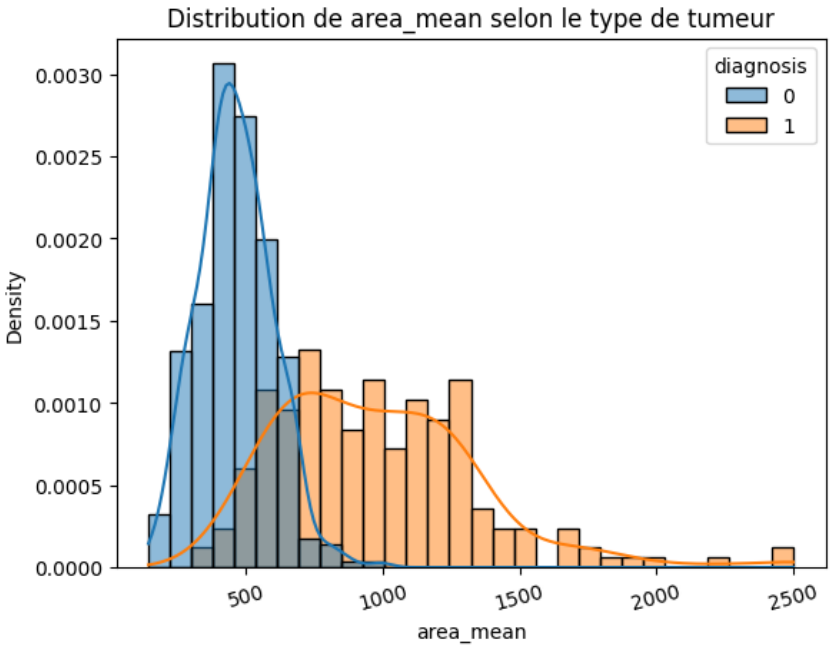

FIGURE 7 – Distribution de *radius mean* selon le type de tumeur

La distribution de **are mean** montre un comportement similaire. Les tumeurs bénignes

présentent des aires moyennes majoritairement inférieures à 700, avec un pic autour de 500, tandis que les tumeurs malignes sont décalées vers des valeurs nettement plus élevées, souvent supérieures à 800 voire 1000. La zone de recouvrement est là encore relativement réduite, et la queue de distribution des tumeurs malignes s'étend vers des valeurs très élevées. Cela confirme que la taille globale de la tumeur constitue un indicateur important de malignité.

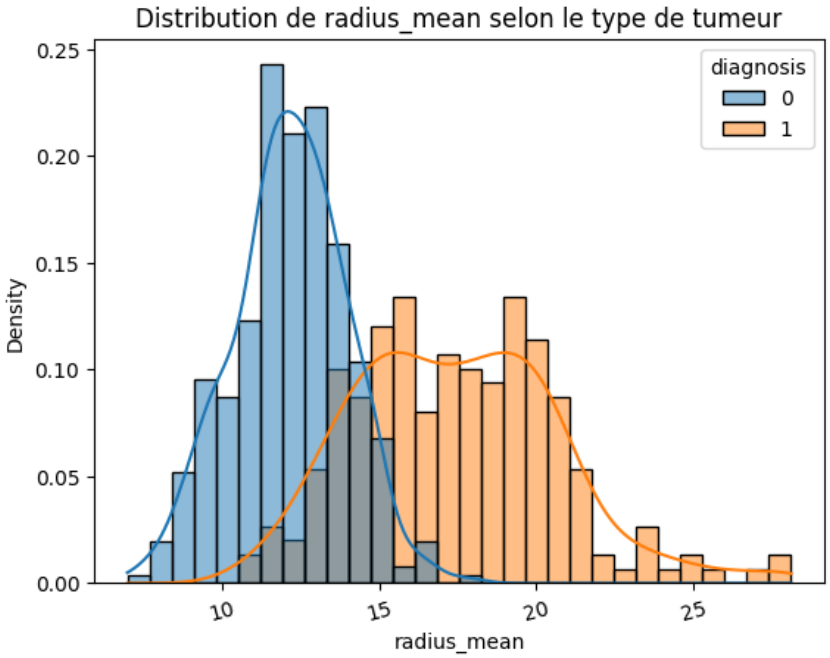

FIGURE 8 – Distribution de *area mean* selon le type de tumeur

La variable **concavity mean**, qui mesure l'irrégularité des contours cellulaires, présente une séparation encore plus nette. Les tumeurs bénignes sont fortement concentrées près de 0, avec la plupart des valeurs inférieures à 0.1, ce qui correspond à des contours globalement réguliers. À l'inverse, les tumeurs malignes sont décalées vers des concavités moyennes plus élevées, souvent comprises entre 0.1 et 0.3, avec très peu de recouvrement entre les deux distributions. Cette figure illustre bien le rôle central des caractéristiques de forme dans la détection des tumeurs malignes : des contours très irréguliers sont rarement observés dans les tumeurs bénignes.

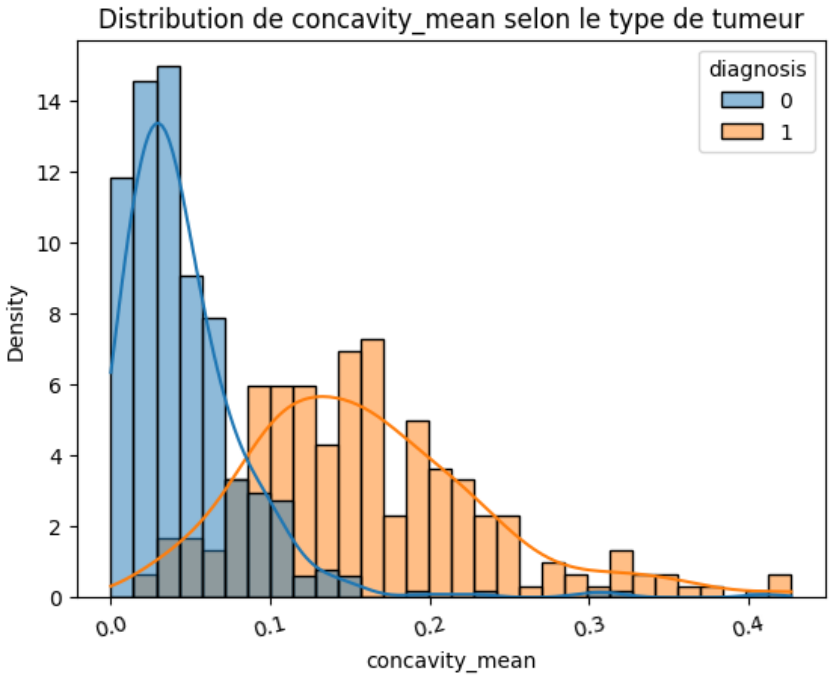

FIGURE 9 – Distribution de *concavity mean* selon le type de tumeur

Ces trois distributions confirment les conclusions tirées à partir des boxplots : les tumeurs malignes se distinguent nettement par des valeurs plus élevées de taille et d'irrégularité des contours, et le recouvrement relativement faible entre les deux classes suggère que ces variables devraient être très informatives pour les modèles de classification supervisée.

3.5 Projection des données dans un plan 2D

Afin de visualiser globalement la structure du jeu de données dans un espace de dimension réduite, nous avons appliqué une **Analyse en Composantes Principales (PCA)** sur l'ensemble des 30 variables, après standardisation. Les deux premières composantes principales capturent une grande partie de la variance totale du jeu de données. La Figure suivante représente chaque tumeur projetée dans le plan défini par ces deux composantes, en colorant les points selon le diagnostic (0 : bénigne, 1 : maligne).

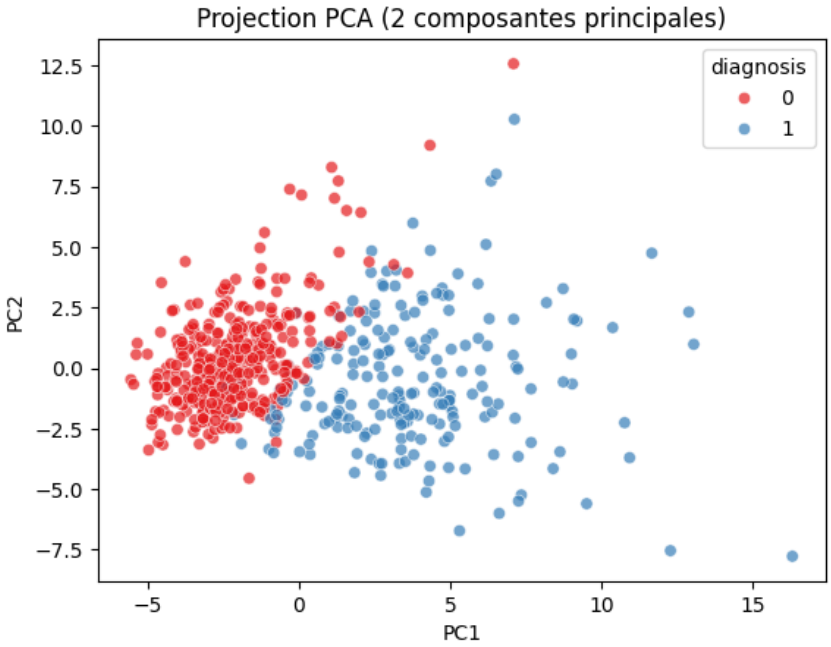

FIGURE 10 – Projection PCA (deux premières composantes principales) avec coloration par type de tumeur

On observe que les tumeurs bénignes forment un nuage de points relativement compact, tandis que les tumeurs malignes occupent une région plus étendue et globalement séparée dans l'espace des deux premières composantes. Même si une zone de recouvrement subsiste, la frontière entre les deux groupes apparaît déjà assez marquée dès cette projection bidimensionnelle, ce qui indique que les caractéristiques morphologiques contiennent une information suffisante pour séparer les classes de manière quasi linéaire. Cette visualisation vient donc renforcer les résultats obtenus précédemment sur les distributions univariées et la matrice de corrélation, et confirme que le jeu de données se prête bien à l'utilisation de modèles de classification supervisée, en particulier de classifieurs linéaires.

4 Modélisation et évaluation des modèles

4.1 Protocole expérimental

À partir des données préparées précédemment, nous avons mis en place plusieurs modèles de classification supervisée pour prédire le diagnostic (**bénin** ou **malin**) à partir des 30 caractéristiques morphologiques. Dans tout ce qui suit, on note :

— $X \in R^{n \times d}$ la matrice des variables explicatives, où n est le nombre de tumeurs et $d = 30$ le nombre de caractéristiques ;

— $y \in \{0, 1\}^n$ le vecteur des étiquettes, avec 0 pour les tumeurs bénignes et 1 pour les tumeurs malignes.

Afin de rendre les différentes caractéristiques comparables et d'éviter qu'une variable à grande échelle domine les autres lors de l'optimisation, nous appliquons systématiquement une **standardisation** des variables :

$$x_{ij}^{\text{std}} = \frac{x_{ij} - \mu_j}{\sigma_j},$$

où μ_j et σ_j sont respectivement la moyenne et l'écart-type de la j-ème caractéristique sur l'ensemble d'apprentissage. Cette normalisation est intégrée dans un *pipeline* scikit-learn, ce qui garantit que les paramètres de normalisation sont appris uniquement sur les données d'entraînement, puis réutilisés tels quels pour transformer les données de test.

Pour éviter la répétition de code et assurer un traitement homogène de tous les modèles, nous avons défini plusieurs **fonctions utilitaires** :

— `decrire_matrice_confusion` : à partir d'une matrice de confusion binaire

$$\text{CM} = \begin{pmatrix} \text{TN} & \text{FP} \\ \text{FN} & \text{TP} \end{pmatrix},$$

cette fonction identifie les quatre types d'issues possibles (vrai négatif, faux positif, faux négatif, vrai positif) et les interprète dans le contexte médical (par exemple, un faux négatif correspond à une tumeur maligne prédite comme bénigne, ce qui est particulièrement critique).

— `eval_holdout` : réalise l'entraînement d'un modèle sur un ensemble d'apprentissage, puis l'évalue sur un ensemble de test distinct. Elle calcule notamment l'**accuracy** (taux de bonnes classifications)

$$\widehat{\text{acc}} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \mathbf{1}\big(\hat{y}_i = y_i\big),$$

le **taux d'erreur** $(1 - \widehat{\mathrm{acc}})$, le rapport de classification (précision, rappel, F1-score par classe) et affiche la matrice de confusion sous forme graphique.

— `eval_cv` : met en œuvre une **validation croisée** à $k = 5$ plis stratifiés. Le jeu de données est partitionné en cinq sous-ensembles de taille similaire et de proportions de classes comparables. Pour chaque pli, un modèle est entraîné sur quatre sous-ensembles et évalué sur le cinquième ; les prédictions issues de tous les plis sont ensuite agrégées afin de construire une matrice de confusion globale et d'estimer une accuracy moyenne plus robuste vis-à-vis du choix du découpage.

Ces fonctions sont utilisées en conjonction avec des **pipelines** de la forme :

$$\texttt{Pipeline} = \big[\texttt{StandardScaler} \to \texttt{Classifieur}\big],$$

où le `Classifieur` est tour à tour un perceptron linéaire, une régression logistique, un SVM linéaire et un réseau de neurones MLP. Cette architecture garantit que le même protocole de normalisation et d'évaluation est appliqué à chaque famille de modèles, ce qui rend les comparaisons possibles.

Deux **stratégies d'évaluation** sont systématiquement considérées :

— **Stratégie hold-out** : le jeu de données est séparé une seule fois de manière aléatoire en un ensemble d'entraînement (80 % des observations) et un ensemble de test (20 %). Le modèle est ajusté sur l'ensemble d'apprentissage et évalué sur les données de test, qui fournissent une estimation ponctuelle de sa capacité de généralisation.

— **Validation croisée 5-fold** : l'ensemble des données est utilisé à la fois pour l'apprentissage et l'évaluation selon le principe de la validation croisée. Chaque observation est prédite une fois à partir d'un modèle entraîné sans l'avoir vue, ce qui permet de réduire la variance de l'estimateur de performance par rapport à un simple découpage unique.

4.2 Perceptron linéaire

Le **perceptron linéaire** est l'un des modèles les plus simples pour la classification binaire. Il cherche un hyperplan linéaire $\{x \in R^d : w^\top x + b = 0\}$ tel que les exemples soient séparés au mieux dans l'espace des caractéristiques. L'apprentissage se fait par une mise à jour itérative des poids w et du biais b :

$$w \leftarrow w + \eta\, y_i x_i, \qquad b \leftarrow b + \eta\, y_i,$$

lorsqu'un exemple (x_i, y_i) est mal classé, où $\eta > 0$ est le taux d'apprentissage. Dans notre implémentation, le perceptron est encapsulé dans un pipeline de la forme :

$$\texttt{StandardScaler} \rightarrow \texttt{Perceptron(max\_iter=1000, tol=1e-3)},$$

ce qui garantit une normalisation préalable des entrées et un nombre maximal d'itérations suffisamment élevé pour la convergence.

Évaluation hold-out

Dans le cadre de la stratégie hold-out, le perceptron linéaire est entraîné sur les 80 % de données d'apprentissage puis évalué sur les 20 % de données de test. La matrice de confusion associée est représentée à la Figure 11. Elle montre 67 tumeurs bénignes correctement classées, 5 fausses alertes (bénignes prédites malignes), 40 tumeurs malignes correctement détectées et 2 tumeurs malignes manquées. Le modèle génère donc un peu plus de faux positifs que de faux négatifs : il tend à "sur-diagnostiquer" légèrement la malignité, ce qui est moins dangereux cliniquement qu'un grand nombre de faux négatifs, mais reste moins précis que les modèles plus avancés.

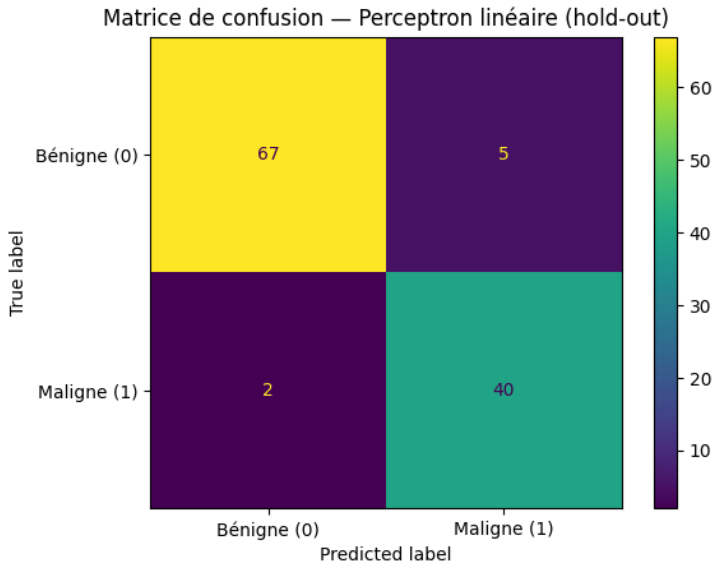

FIGURE 11 – Matrice de confusion du perceptron linéaire (stratégie hold-out)

Validation croisée 5-fold

La même architecture est ensuite évaluée en validation croisée à 5 plis. À chaque itération, un nouveau perceptron est appris sur quatre plis et évalué sur le pli restant. L'utilisation de `cross_val_predict` permet d'obtenir une prédiction pour chaque observation à partir d'un modèle entraîné sans cette observation. En agrégeant ces prédictions, on obtient la

matrice de confusion globale présentée Figure 12, avec 348 vrais négatifs, 9 faux positifs, 202 vrais positifs et 10 faux négatifs. Globalement, le perceptron garde un bon comportement mais reste le modèle qui commet le plus d'erreurs parmi ceux étudiés.

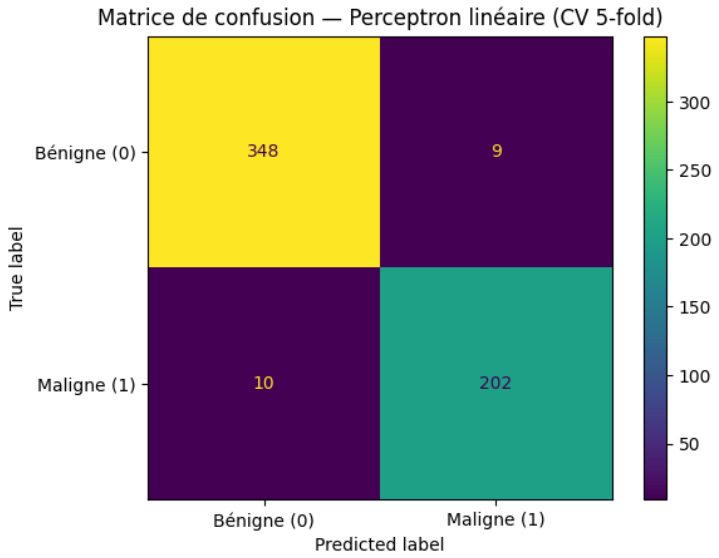

FIGURE 12 – Matrice de confusion du perceptron linéaire (validation croisée 5-fold)

4.3 Régression logistique

La **régression logistique** est un modèle probabiliste linéaire qui associe à chaque observation une probabilité d'appartenir à la classe maligne. Le modèle suppose que

$$P(Y = 1 \mid X = x) = \sigma(w^\top x + b) \quad \text{avec} \quad \sigma(z) = \frac{1}{1 + e^{-z}},$$

et les paramètres (w, b) sont estimés par maximisation de la vraisemblance, équivalente à la minimisation de la *log-loss* (perte logistique). Dans scikit-learn, nous utilisons une régularisation L2 de norme $\lambda\|w\|_2^2$, contrôlée par le paramètre $C = 1/\lambda$, afin de limiter le sur-apprentissage.
Le pipeline utilisé est de la forme :

`StandardScaler` → `LogisticRegression(penalty=L2, max_iter=1000)`.

Évaluation hold-out

En stratégie hold-out, le modèle est appris sur l'échantillon d'entraînement et évalué sur l'échantillon de test. La matrice de confusion qui en résulte est illustrée sur la Figure 13. On y observe 71 tumeurs bénignes correctement classées, 1 fausse alerte, 40 tumeurs malignes correctement détectées et 2 faux négatifs. Par rapport au perceptron, la régression

logistique réduit nettement le nombre total d'erreurs et reste particulièrement fiable pour la classe bénigne tout en gardant un nombre très limité de tumeurs malignes manquées.

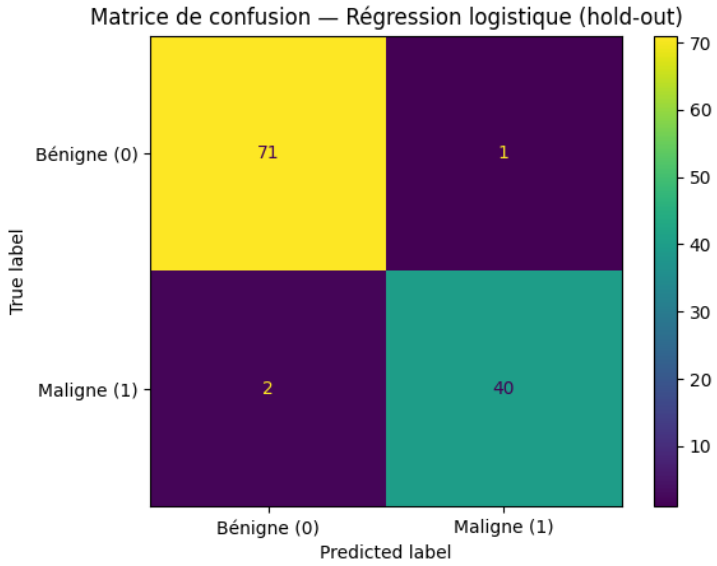

FIGURE 13 – Matrice de confusion de la régression logistique (stratégie hold-out)

Validation croisée 5-fold

La régression logistique est ensuite évaluée en validation croisée 5-fold sur l'ensemble des observations. La matrice de confusion globale, Figure 14, regroupe 354 vrais négatifs, 3 faux positifs, 203 vrais positifs et 9 faux négatifs. Les tumeurs bénignes sont donc presque toutes correctement classées, et la plupart des tumeurs malignes sont correctement détectées. La régression logistique apparaît comme un modèle très stable, avec un compromis intéressant entre faux positifs et faux négatifs.

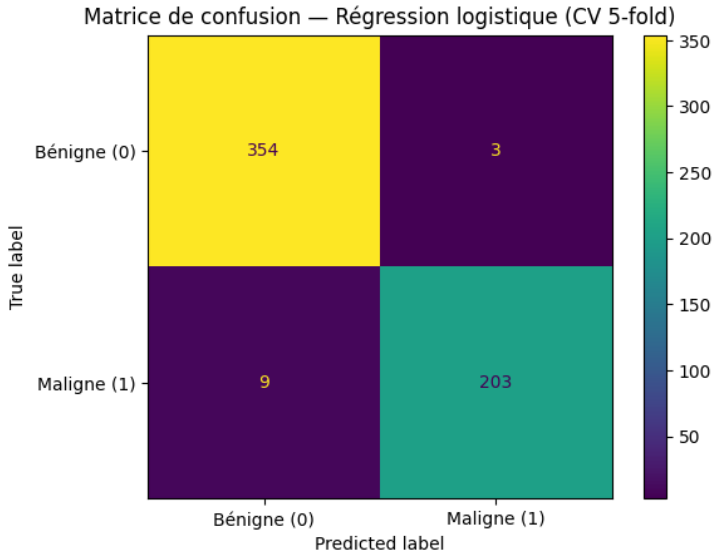

FIGURE 14 – Matrice de confusion de la régression logistique (validation croisée 5-fold)

4.4 SVM linéaire

Le **SVM linéaire** (*Support Vector Machine*) vise à trouver un hyperplan de séparation

$$\{x \in R^d : w^\top x + b = 0\}$$

qui maximise la marge entre les exemples des deux classes. En présence de données non parfaitement séparables, on introduit des variables d'écart ξ_i et on résout le problème de marge souple :

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \quad \text{sous} \quad y_i(w^\top x_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0,$$

où le paramètre $C > 0$ contrôle le compromis entre taille de la marge et pénalisation des erreurs. Dans scikit-learn, cette optimisation est effectuée via l'implémentation `LinearSVC` avec une perte de type hinge.
Le pipeline étudié est :

$$\texttt{StandardScaler} \rightarrow \texttt{LinearSVC(C=1.0)}.$$

Évaluation hold-out

Dans la stratégie hold-out, le SVM est ajusté sur l'échantillon d'apprentissage et évalué sur l'échantillon de test. La matrice de confusion obtenue est présentée Figure 15. On obtient 71 vrais négatifs, 1 faux positif, 37 vrais positifs et 5 faux négatifs. Le SVM est

donc très conservateur pour la classe bénigne (quasi pas de faux positifs), mais au prix d'un nombre de faux négatifs plus élevé : plusieurs tumeurs malignes sont classées à tort comme bénignes, ce qui est problématique dans un contexte de dépistage.

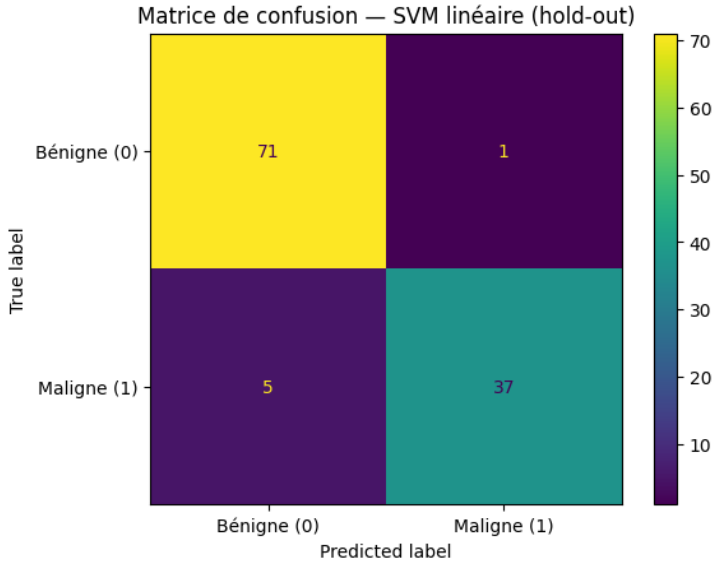

FIGURE 15 – Matrice de confusion du SVM linéaire (stratégie hold-out)

Validation croisée 5-fold

En validation croisée 5-fold, on répète l'apprentissage du SVM sur différents sous-ensembles et on agrège les prédictions de tous les plis pour construire la matrice de confusion globale (Figure 16). Celle-ci contient 352 vrais négatifs, 5 faux positifs, 203 vrais positifs et 9 faux négatifs. Le SVM reste donc un modèle très performant, mais légèrement moins équilibré que la régression logistique ou le MLP en termes de faux négatifs.

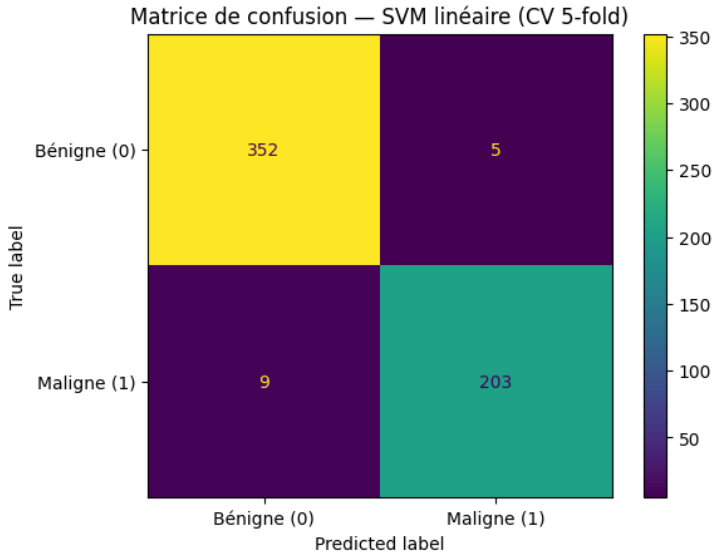

FIGURE 16 – Matrice de confusion du SVM linéaire (validation croisée 5-fold)

4.5 Réseau de neurones MLP

Le **réseau de neurones** utilisé est un perceptron multicouche (*Multi-Layer Perceptron*, MLP) de type *feed-forward*. L'architecture choisie comporte une couche cachée avec un nombre modéré de neurones (50 dans nos expériences) et une fonction d'activation non linéaire de type ReLU :

$$\mathrm{ReLU}(z) = \max(0, z).$$

L'approximation de la fonction de décision est réalisée par composition linéaire–non linéaire, ce qui permet de capturer des séparations plus complexes que les modèles purement linéaires. Les paramètres (poids et biais) sont estimés par descente de gradient stochastique, à l'aide de l'optimiseur Adam, en minimisant une fonction de perte adaptée à la classification binaire.
Le pipeline adopté est :

`StandardScaler` → `MLPClassifier(hidden_layer_sizes=(50,), activation='relu')`.

Évaluation hold-out

Dans le cadre du hold-out, le MLP est entraîné sur l'échantillon d'apprentissage standardisé puis évalué sur l'échantillon de test. La matrice de confusion correspondante, montrée Figure 17, présente 69 vrais négatifs, 3 faux positifs, 41 vrais positifs et seulement 1 faux négatif. Le réseau de neurones se distingue donc par un très faible nombre de tumeurs malignes manquées, au prix d'un nombre de fausses alertes légèrement supérieur à celui de la régression logistique.

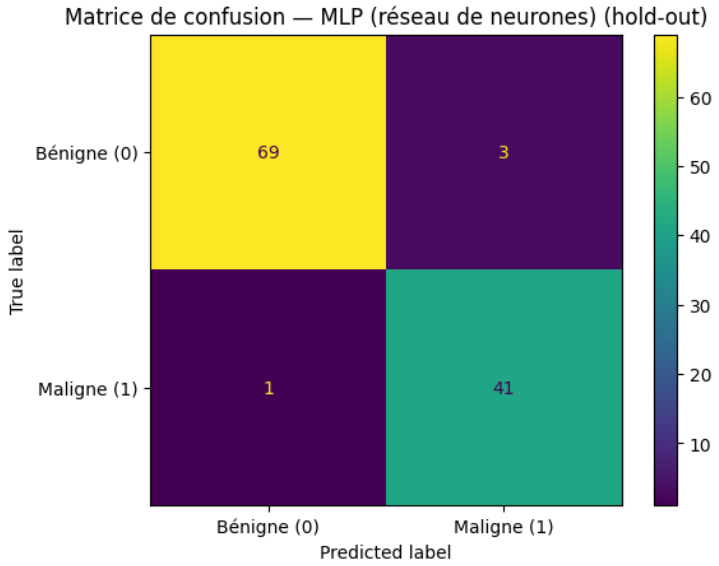

FIGURE 17 – Matrice de confusion du MLP (stratégie hold-out)

Validation croisée 5-fold

En validation croisée 5-fold, le réseau de neurones est réentraîné sur des sous-échantillons différents et ses prédictions sont agrégées pour produire la matrice de confusion globale illustrée Figure 18. Celle-ci contient 355 vrais négatifs, 2 faux positifs, 204 vrais positifs et 8 faux négatifs. Parmi tous les modèles testés, le MLP est celui qui obtient la meilleure combinaison : très peu de tumeurs bénignes faussement alarmées et un nombre de tumeurs malignes manquées relativement faible.

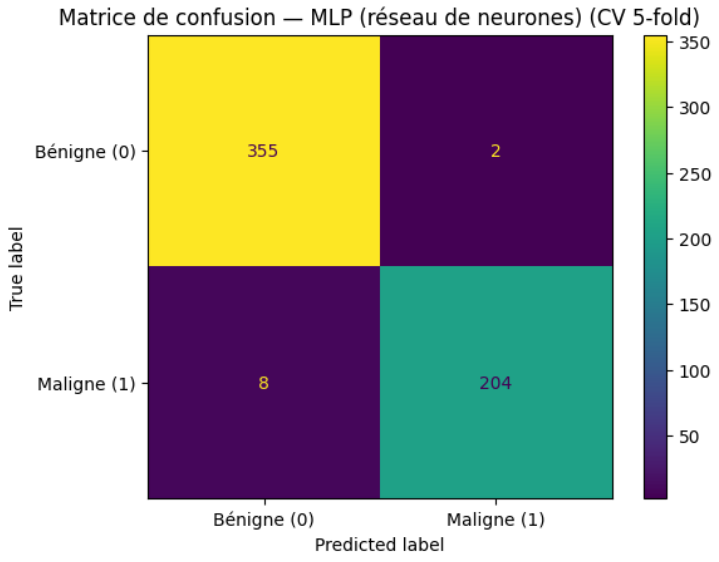

FIGURE 18 – Matrice de confusion du MLP (validation croisée 5-fold)

4.6 Comparaison des stratégies d'évaluation et des modèles

Pour chaque modèle, les fonctions d'évaluation fournissent les principales métriques de performance (accuracy, taux d'erreur, précision, rappel, F1-score) en stratégie hold-out et en validation croisée 5-fold. Afin de synthétiser ces résultats, nous regroupons les valeurs d'accuracy et de taux d'erreur dans deux tableaux distincts :

- un tableau pour la stratégie hold-out, où chaque ligne correspond à un modèle évalué sur l'unique ensemble de test ;
- un tableau pour la validation croisée 5-fold, où chaque ligne correspond à un modèle et les valeurs représentent les performances globales agrégées sur les cinq plis.

Les résultats en stratégie hold-out (ensemble de test unique) sont récapitulés dans le Tableau 1.

Modèle	**Accuracy (%)**	**Taux d'erreur (%)**
Perceptron linéaire	93.86	6.14
Régression logistique	97.37	2.63
SVM linéaire	94.74	5.26
MLP (réseau de neurones)	96.49	3.51

TABLE 1 – Performances des modèles en stratégie hold-out (évaluation sur un seul ensemble de test)

Dans cette configuration, la régression logistique est le modèle linéaire le plus performant, devant le SVM et le perceptron. Le MLP obtient déjà des résultats très compétitifs, légèrement inférieurs à la régression logistique en termes d'accuracy globale, mais avec un nombre de faux négatifs particulièrement faible, ce qui est un avantage important pour le dépistage.

Les performances moyennes obtenues par validation croisée 5-fold sur l'ensemble du jeu de données sont présentées dans le Tableau 2.

Modèle	**Accuracy (%)**	**Taux d'erreur (%)**
Perceptron linéaire	96.66	3.34
Régression logistique	97.89	2.11
SVM linéaire	97.54	2.46
MLP (réseau de neurones)	98.24	1.76

TABLE 2 – Performances moyennes des modèles en validation croisée 5-fold (scores globaux agrégés sur les cinq plis)

La validation croisée confirme les tendances observées avec le hold-out : le perceptron est

le moins performant mais reste déjà très correct, alors que la régression logistique et le SVM linéaire forment un premier groupe de modèles très efficaces. Le réseau de neurones MLP offre les meilleurs résultats globaux, avec le taux d'erreur moyen le plus faible. D'un point de vue applicatif, la combinaison *faible taux d'erreur global + nombre raisonnable de faux négatifs* fait de la régression logistique, du SVM linéaire et surtout du MLP des candidats pertinents pour l'aide au diagnostic automatique des tumeurs du sein.

4.7 Analyse des courbes ROC et de l'aire sous la courbe (AUC)

En complément des mesures globales (accuracy, taux d'erreur, matrices de confusion), nous avons tracé les **courbes ROC** (*Receiver Operating Characteristic*) des quatre modèles sur le **jeu de test hold-out**. Pour chaque modèle, on calcule un score continu de "malignité" pour chaque tumeur (probabilité prédite de la classe 1 lorsqu'elle est disponible, ou bien fonction de décision), puis on fait varier le seuil de décision de 0 à 1. Pour chaque valeur de seuil, on obtient une paire :

$$(\mathrm{FPR}, \mathrm{TPR}) = \big(\text{taux de faux positifs},\ \text{taux de vrais positifs}\big),$$

où la FPR (*False Positive Rate*) correspond à la proportion de tumeurs bénignes prédites malignes, et la TPR (*True Positive Rate* ou sensibilité) à la proportion de tumeurs malignes correctement détectées. La courbe ROC est la trajectoire de ces points lorsque le seuil varie.

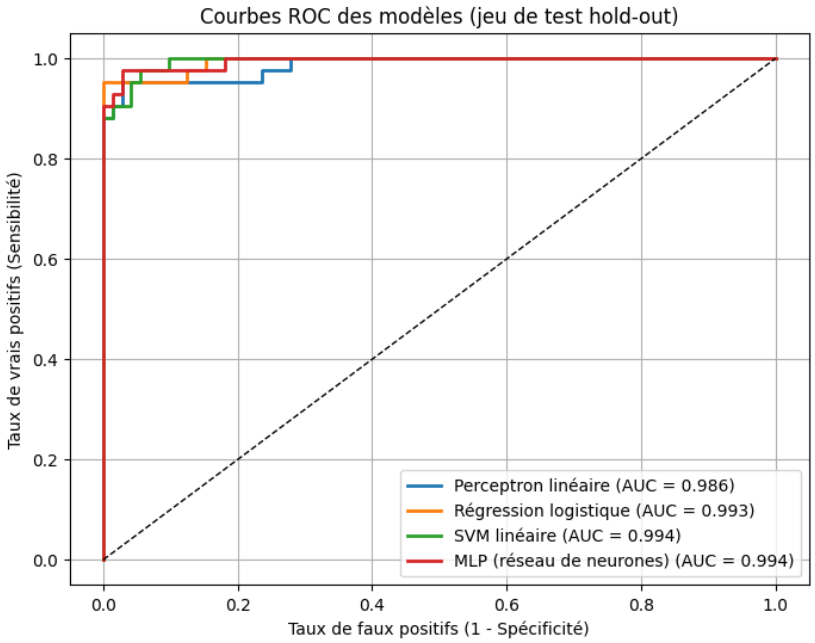

FIGURE 19 – Courbes ROC des différents modèles sur le jeu de test (stratégie hold-out)

L'aire sous la courbe ROC (**AUC**, *Area Under the Curve*) fournit un indicateur synthétique du pouvoir discriminant d'un modèle, indépendant du seuil de décision choisi. Elle peut être interprétée comme la probabilité qu'un modèle attribue un score plus élevé à une tumeur maligne qu'à une tumeur bénigne tirées au hasard. Dans nos expériences, les AUC obtenues sur le jeu de test sont toutes très élevées (AUC $>$ 0,98) : le perceptron linéaire atteint environ 0.986, la régression logistique 0.993, le SVM linéaire 0.994 et le MLP 0.994. Ces valeurs confirment que le jeu de données est bien séparé dans l'espace des caractéristiques et que tous les modèles étudiés disposent d'un excellent pouvoir de discrimination entre tumeurs bénignes et malignes.

Sur la Figure 19, les quatre courbes sont très proches du coin supérieur gauche, ce qui correspond à une sensibilité élevée pour un taux de faux positifs très faible. On observe un léger avantage pour la régression logistique, le SVM linéaire et le MLP, dont les courbes dominent celle du perceptron sur la majeure partie de la plage de FPR. En pratique, cela signifie que, pour un même niveau de faux positifs accepté (par exemple 1–2 % de tumeurs bénignes injustement signalées comme suspectes), ces trois modèles offrent généralement un taux de détection des tumeurs malignes (*rappel*) supérieur à celui du perceptron.

Du point de vue clinique, ces courbes ROC sont particulièrement utiles car elles permettent de choisir, pour un modèle donné, un **seuil de décision adapté aux contraintes médicales**. Dans un contexte de dépistage, on cherchera à travailler dans une zone de la courbe où la sensibilité est très élevée, quitte à tolérer quelques faux positifs supplémentaires. Les résultats obtenus montrent que la régression logistique, le SVM linéaire et le MLP offrent tous trois un excellent compromis : ils permettent de maintenir un taux de faux positifs faible tout en conservant une probabilité de détection des tumeurs malignes très élevée. Le MLP et le SVM, légèrement au-dessus des autres en termes d'AUC, apparaissent ainsi comme des candidats particulièrement intéressants lorsqu'on souhaite privilégier la sensibilité sans trop dégrader la spécificité.

5 Conclusion

5.1 La nature des données

Dans ce projet, nous avons étudié le problème de la classification automatique des tumeurs du sein à partir du jeu de données *Breast Cancer Wisconsin (Diagnostic)*. L'objectif était de prédire le caractère bénin ou malin d'une tumeur à partir de mesures morphologiques extraites d'images cellulaires, en mettant en œuvre différentes méthodes de classification supervisée et en comparant leurs performances selon plusieurs protocoles d'évaluation.

L'analyse exploratoire des données a montré que les tumeurs malignes se distinguent nettement des tumeurs bénignes par des cellules plus grandes, des contours plus irréguliers et une hétérogénéité morphologique accrue. Les variables de type *worst* se sont révélées particulièrement discriminantes, ce qui a été confirmé par l'analyse de corrélation, où ces variables présentent les liens les plus forts avec la variable cible **diagnosis**. La présence de fortes corrélations entre de nombreuses caractéristiques a également mis en évidence une redondance importante de l'information, suggérant que des modèles capables de gérer la multicolinéarité sont bien adaptés à ce problème.

5.2 La classification

Les expériences de classification ont montré que l'ensemble des modèles testés atteint de bonnes performances globales. Toutefois, des différences notables apparaissent lorsqu'on compare les méthodes. Le perceptron, bien que simple, reste limité par son incapacité à capturer des séparations plus complexes. La régression logistique et le SVM linéaire obtiennent des résultats très solides, avec des performances élevées et stables, confirmant que le problème est en grande partie linéairement séparable, comme suggéré par la projection PCA.

Parmi les modèles étudiés, le SVM linéaire et la régression logistique apparaissent comme les plus appropriés pour cette tâche. Ils offrent un excellent compromis entre performance, robustesse et interprétabilité. Le réseau de neurones (MLP) peut atteindre des performances comparables, voire légèrement supérieures dans certains cas, mais au prix d'une complexité accrue et d'un risque plus élevé de surapprentissage, ce qui limite son intérêt dans un contexte où les données sont relativement bien structurées et de taille modérée.

5.3 L'évaluation

En ce qui concerne l'évaluation, la validation croisée s'est révélée plus fiable que le simple découpage entraînement/test, en fournissant des estimations de performance plus stables et moins dépendantes d'une partition particulière des données. L'analyse des matrices de confusion a également mis en évidence l'importance de minimiser les faux négatifs, c'est-à-dire les tumeurs malignes mal classées, un enjeu crucial dans un contexte médical.