Classification de tumeur du sein

BELKALEM Rayane

AMROUNI Moh-Ouali

SOUPAYA Raphael

3A IAMD

6 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.

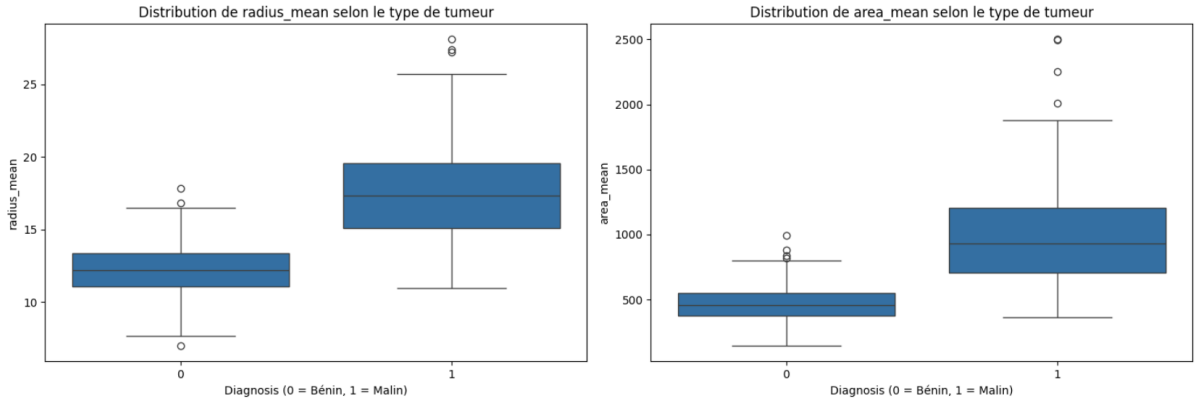

FIGURE 1 – Box plot de radius et area mean

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.

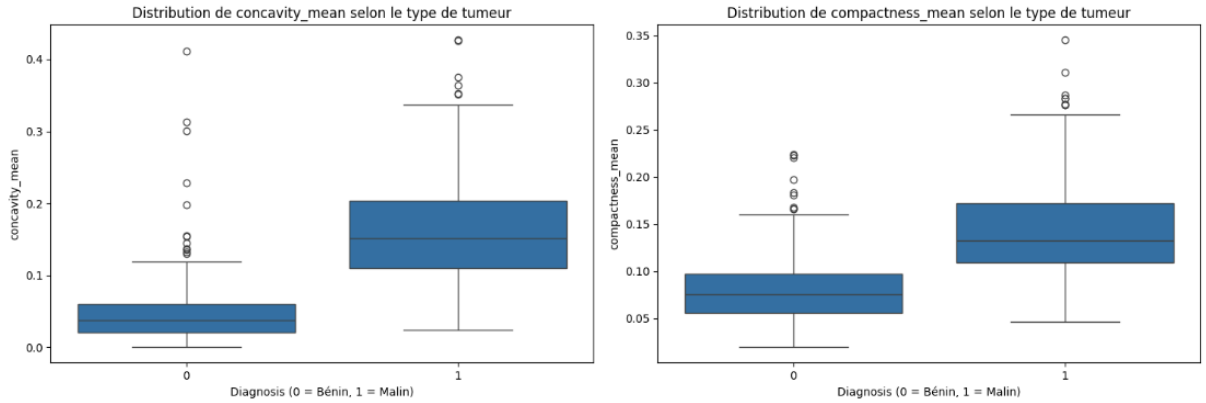

FIGURE 2 – Box plot de concavity et compactness mean

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.

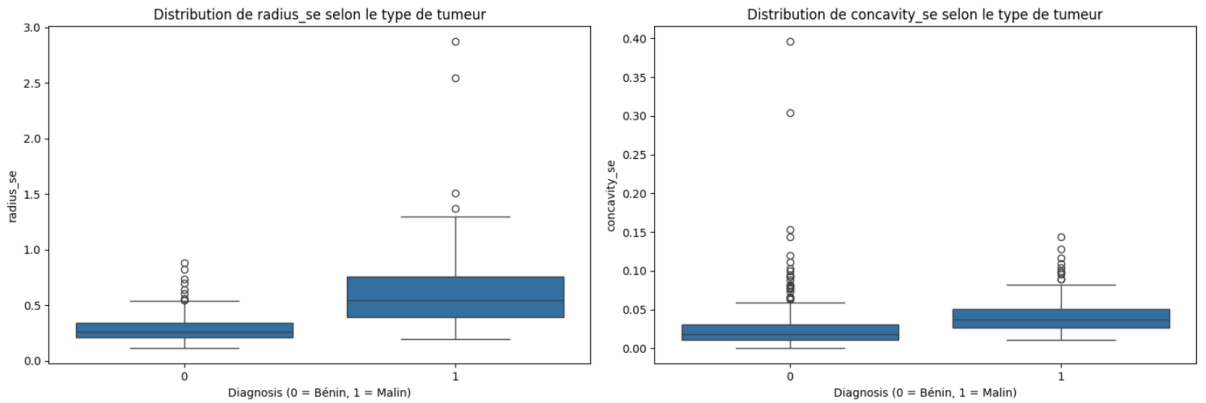

FIGURE 3 – Box plot de radius et concavity se

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.

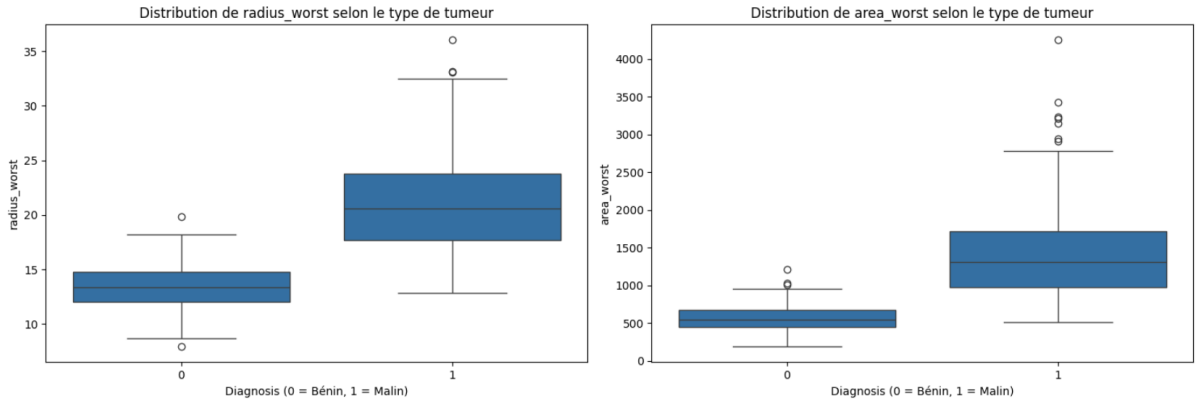

FIGURE 4 – Box plot de area et radius worst

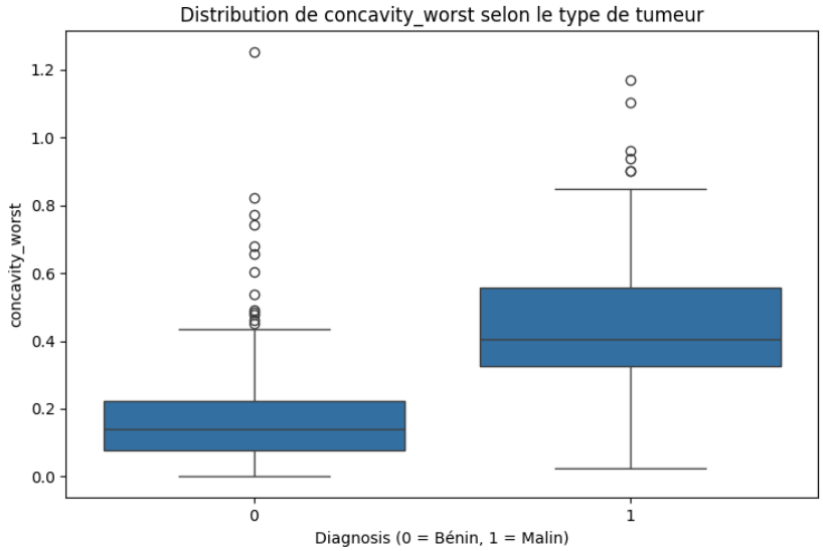

FIGURE 5 – Box plot de concavity worst

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.

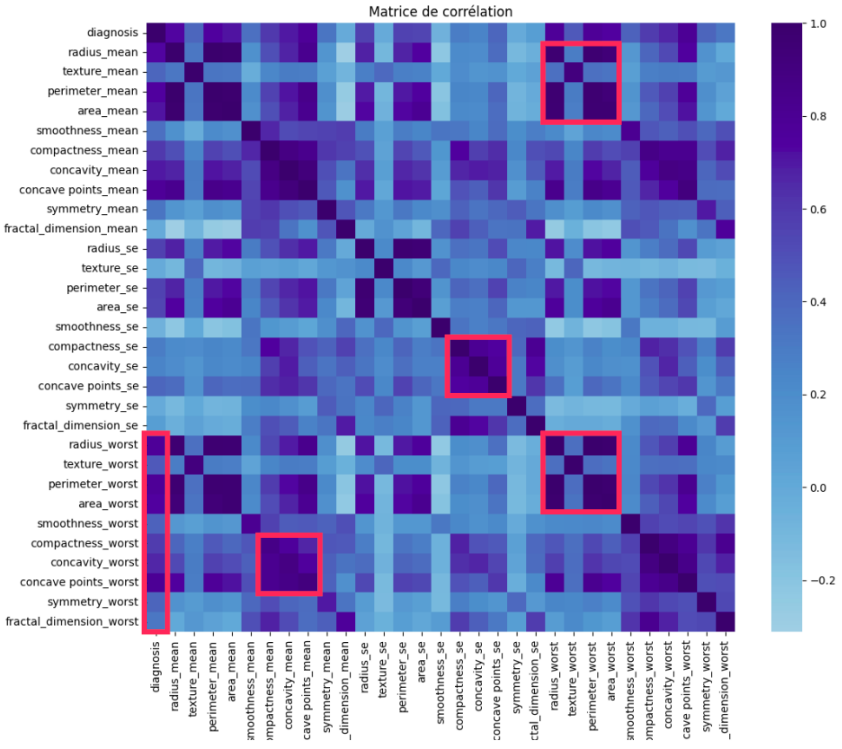

FIGURE 6 – Matrice de corrélation des caractéristiques

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.