Université de Lille

Master 2 Mathématiques, Finance et Actuariat

Compte rendu du projet de Data Science

Auteurs : Timothée Kuntzmann & Valentin Martiaux

Professeur : Olivier Bouaziz

Date : 16 novembre 2025

Université de Lille

Table des matières

Table des figures

Liste des tableaux

Statistiques Descriptives sur la Base de données

1.1 Description et présentation des données

La base de données a pour thématique le cancer du sein et contient au total 569 observations de cellules. Chaque observation est caractérisée par 30 variables numériques représenant différentes caractéristiques des cellules (rayon, périmètre, aire, texture, etc...), calculées en moyenne (_mean), en écart-type (_se) et en valeur maximale (_worst).

Les principales variables du dataset sont les suivantes (selon le fichier wdbc.names) :

a) **radius** : moyenne des distances entre le centre et les points du périmètre
b) **texture** : écart-type des valeurs en niveaux de gris
c) **perimeter** : périmètre
d) **area** : aire
e) **smoothness** : variation locale des longueurs de rayon
f) **compactness** : $\frac{perimeter^2}{area} - 1.0$
g) **concavity** : sévérité des portions concaves du contour
h) **concave points** : nombre de portions concaves du contour
i) **symmetry** : symétrie
j) **fractal dimension** : « approximation de la côte » -1

La variable cible est le diagnostique (Diagnostic), à valeurs dans B,M que l'on recode de manière binaire afin de mieux l'étudier, avec la transformation suivante :

— 0 : Benign (Bénin)
— 1 : Malignant (Malin)

1.2 Matrice de Corrélation

La matrice de corrélation permet d'observer les relations entre les variables. Les variables fortement corrélées peuvent être redondantes et prises en compte lors de la réduction de dimension. Comme le montre l'image 1.1, certaines variables sont fortement corrélées entre elles, ce qui peut justifier l'utilisation de techniques de réduction de dimensionnalité comme l'ACP par la suite.

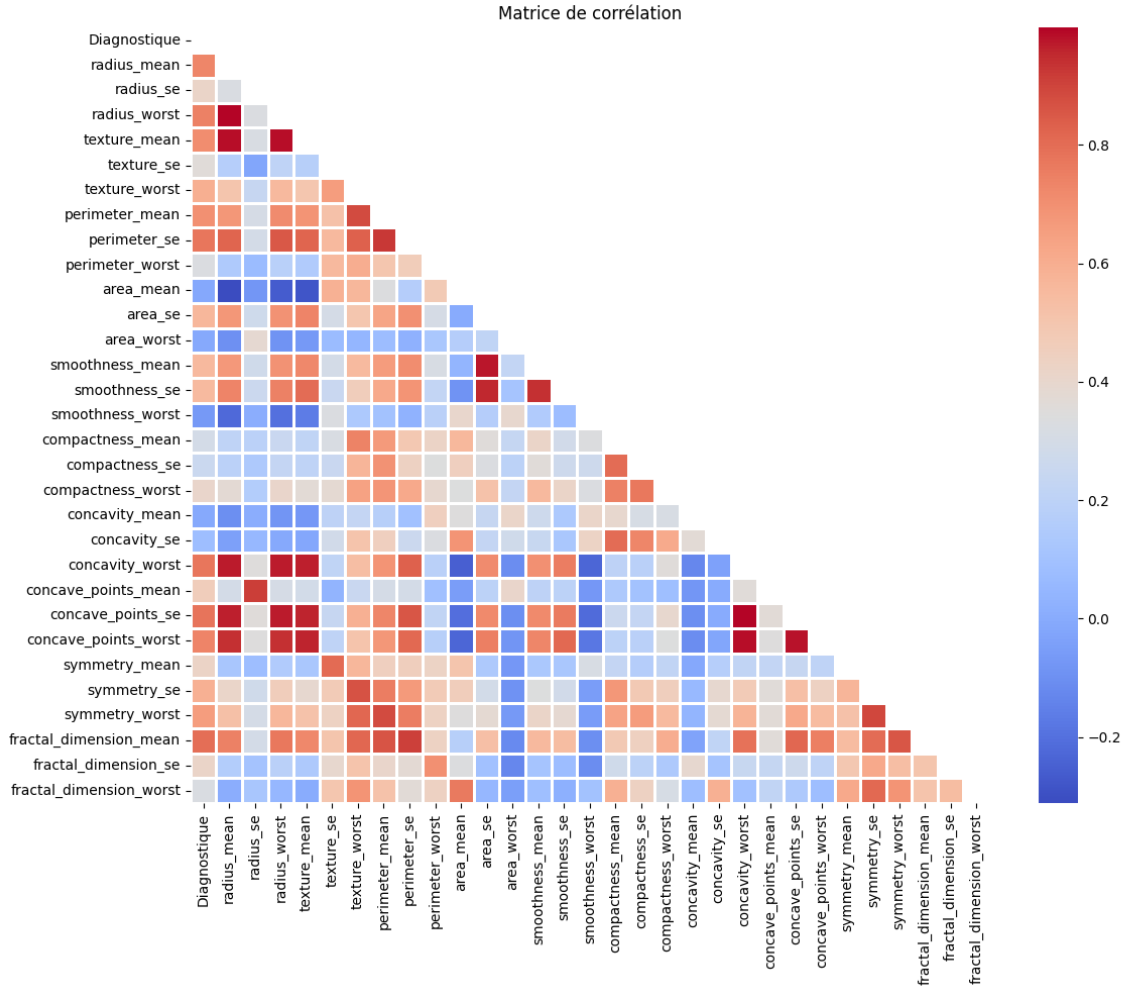


FIGURE 1.1 – Matrice de corrélation entre les variables du dataset

Les variables présentant une forte corrélation sont les variables qui concernent les rayons et la concavité, ainsi que le périmètre et la dimension fractale notamment. On retrouve également assez souvent des corrélations entre les même variables dans leurs catégories se et worst. Les variable d'aire moyenne et concavité moyenne sont peu corrélées avec les autres variables .

1.3 Analyse de l'équilibre des classes

Le dataset présente le déséquilibre suivant entre les classes B et M :

— **Benign (Bénin)** : 357 observations (environ 63%)
— **Malignant (Malin)** : 212 observations (environ 37%)
— **Ratio** : 1.684 (Benign/Malignant)

Ce déséquilibre, bien que modéré, nécessite une attention particulière lors de l'entraînement des modèles. Un split stratifié est utilisé pour maintenir la distribution des classes dans les ensembles d'entraînement et de test. Certains modèles ont également utilisé des techniques d'équilibrage comme `class_weight='balanced'` ou des méthodes d'oversampling comme SMOTE ou NearMiss.

Méthodologie et évaluation des Modèles

2.1 Métriques utilisées

Plusieurs métriques ont été utilisées pour évaluer les performances des modèles :

TABLE 2.1 – Définitions des principales métriques de performance

Métrique	Définition	Formule
Accuracy (Précision globale)	Proportion d'observations correctement classées	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	Proportion de prédictions positives qui sont réellement positives	$\frac{TP}{TP+FP}$
Recall (Sensibilité)	Proportion de cas positifs réellement détectés	$\frac{TP}{TP+FN}$
Specificity	Proportion de cas négatifs réellement détectés	$\frac{TN}{TN+FP}$
F1-score	Moyenne harmonique de la précision et du rappel	$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$
AUC (Area Under Curve)	Aire sous la courbe ROC, mesure la capacité du modèle à distinguer les classes	–

Ces métriques permettent d'évaluer les modèles sous différents angles, ce qui est particulièrement important dans un contexte médical où les faux négatifs (manquer un cancer présent) sont à éviter.

2.2 Validation croisée

Le challenge en Machine Learning est d'entraîner des modèles capables de se généraliser à des données inconnues. La validation croisée est une méthode statistique stable et robuste permettant d'évaluer la capacité de généralisation d'un modèle au travers d'un découpage répété du jeu d'entraînement.

Par exemple, dans une k-fold cross-validation :

1. On subdivise les données en k parties de tailles approximativement égales, appelées *folds* ;
2. On entraîne le modèle en utilisant $k-1$ folds et on l'évalue sur le $k^{\text{ème}}$ fold ;
3. On répète l'opération k fois et on fait la moyenne des scores obtenus.

Cette méthode présente deux avantages principaux :

1. Elle permet à la phase d'entraînement d'utiliser l'ensemble des données, ce qui est particulièrement utile lorsque le jeu de données est de petite taille ;
2. La partition multiple des données fournit des informations sur la sensibilité du modèle à la sélection du jeu d'entraînement. Une variance élevée entre les folds indique une instabilité du modèle, souvent due à un *overfitting* (le modèle mémorise le bruit), à un manque de données ou à une forte hétérogénéité de celles-ci.

2.3 GridSearchCV

La validation croisée est ensuite implicitement utilisée dans la recherche des paramètres optimaux de nos modèles à l'aide de GridSearchCV. Il s'agit d'un algorithme visant à essayer de multiples combinaisons de paramètres et retenir ceux qui produisent la sensibilité la plus élevée sur les sets de validation.

2.4 Seuil optimal

Le seuil de décision optimal a été déterminé à partir de la courbe ROC, qui met en relation le taux de vrais positifs (TPR) et le taux de faux positifs (FPR) pour différents seuils de probabilité. Afin d'équilibrer les performances en sensibilité et en spécificité, le critère retenu repose sur l'indice de Youden optimisé, défini par :

$$J_\omega = \omega \times TPR - (1 - \omega) \times FPR$$

Cette version pondérée permet d'ajuster la contribution (ω fixé à 0,8) relative de la sensibilité selon les besoins du modèle (ici minimiser les faux négatifs). Le seuil correspondant au maximum de cet indice est utilisé pour générer les prédictions finales. Les performances associées (AUC, accuracy, précision, sensibilié, F1-score), ainsi que la matrice de confusion et la courbe ROC, sont ensuite évaluées et exportées afin de comparer les modèles de manière cohérente.

LDA

3.1 Description du modèle

L'Analyse Discriminante Linéaire (LDA) suppose que les données suivent une distribution gaussienne multivariée avec une matrice de covariance commune $\mathbf{\Sigma}$. Cette méthode est adaptée à notre problème car les caractéristiques biologiques suivent généralement des distributions approximativement normales et l'hypothèse de covariance commune simplifie l'estimation avec 30 variables corrélées.

La fonction de décision pour une classe k est :

$$\delta_k(\mathbf{x}) = \mathbf{x}^T \mathbf{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^T \mathbf{\Sigma}^{-1} \boldsymbol{\mu}_k + \log(\pi_k) \tag{3.1}$$

où $\boldsymbol{\mu}_k$ est le vecteur moyen de la classe k et π_k la probabilité a priori. La classification attribue l'observation à la classe qui maximise $\delta_k(\mathbf{x})$.

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_0, \quad \mathbf{w} = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0), \quad w_0 = -\tfrac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0)^T \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) + \log \frac{\pi_1}{\pi_0}$$

Décision : classe 1 si $f(\mathbf{x}) > 0$.

3.2 Résultats

TABLE 3.1 – Matrice de confusion - LDA

	Prédit B	**Prédit M**
Vrai Bénin	72	0
Vrai Malignant	2	40

TABLE 3.2 – Métriques de performance - LDA

Métrique	**B**	**M**
Precision	0.97	1.00
Recall	1.00	0.95
F1-score	0.99	0.98
Accuracy	0.98	
Macro avg	0.98	
Weighted avg	0.98	

Le modèle LDA obtient une excellente performance avec une précision de 98%. Il prédit parfaitement les cas bénins (72 sur 72), mais commet 2 faux négatifs sur les cas malins (2 sur 42). L'AUC est de 0.997, ce qui indique une excellente capacité de discrimination.

QDA

4.1 Description du modèle

L'Analyse Discriminante Quadratique (QDA) étend le LDA en permettant à chaque classe d'avoir sa propre matrice de covariance $\boldsymbol{\Sigma}_k$. Cette flexibilité est utile car les cellules bénignes et malignes peuvent avoir des dispersions différentes dans l'espace des caractéristiques biologiques.

La fonction de décision quadratique est :

$$\delta_k(\mathbf{x}) = -\frac{1}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) + \log(\pi_k) \tag{4.1}$$

Le terme quadratique représente la distance de Mahalanobis, qui prend en compte la forme et l'orientation de chaque distribution. La frontière de décision devient ainsi quadratique au lieu de linéaire.

$$\mathbf{x}^T(\boldsymbol{\Sigma}_i^{-1} - \boldsymbol{\Sigma}_j^{-1})\mathbf{x} + 2(\boldsymbol{\mu}_j^T\boldsymbol{\Sigma}_j^{-1} - \boldsymbol{\mu}_i^T\boldsymbol{\Sigma}_i^{-1})\mathbf{x} + c = 0,$$

où c regroupe les constantes (termes quadratiques en $\boldsymbol{\mu}_k$, déterminants et priors).

4.2 Résultats

TABLE 4.1 – Matrice de confusion - QDA

	Prédit B	**Prédit M**
Vrai Bénin	66	6
Vrai Malignant	1	41

Le modèle QDA présente une performance légèrement inférieure au LDA avec une précision de 94%. Il commet 6 faux positifs et 1 faux négatif. L'AUC est de 0.989, restant très élevée. La flexibilité supplémentaire du QDA ne semble pas améliorer les performances dans ce cas, probablement en raison d'un nombre limité d'observations.

Régression Logistique

5.1 Description du modèle

La régression logistique modélise la probabilité d'appartenance à une classe via la fonction sigmoïde.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

Et la variable dépendante ayant seulement des valeurs binaires

$$Y = \begin{cases} 0 & \text{si Classe 1} \\ 1 & \text{si Classe 2} \end{cases}$$

Ensuite on applique la fonction multi-linéaire aux variables d'entrée

$$z = \left(\sum_{i=1}^{n} w_i x_i \right) + b$$

Ici x_i est la $i_{\text{ème}}$ observation de X, $w_i = [w_1, w_2, w_3, \cdots, w_m]$ sont les poids ou coefficients et b est le terme de biais. Finalement on peut représenter l'expression comme le produit scalaire des poids et du biais. Finalement on peut représenter l'expression comme le produit scalaire des poids et du biais :

$$z = WX + b$$

La régression logistique applique ensuite la fonction sigmoïde à pour la convertir en probabilité :

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Ce modèle est adapté à notre problème car il fournit des probabilités interprétables cliniquement, les coefficients reflètent l'impact de chaque caractéristique sur le risque de cancer, et il fonctionne bien avec des variables corrélées.

On construit 3 modèles variants de la régression logistique, dont les paramètres sont optimisés par la fonction GridsearchCV, qui subdivise le set d'entraînement en entraînement et en validation pour la recherche de paramètres optimaux en validation croisée. Le premier modèle constitue un modèle simple, sans algorithme de sampling. Le second emploie une fonction de sous-échantillonnage visant à réduire la classe majoritaire pour équilibrer la proportion de classes. Le troisième repose sur la technique inverse, c'est-à-dire synthétiser de nouvelles données pour la classe minoritaire via k-plus proches voisins.

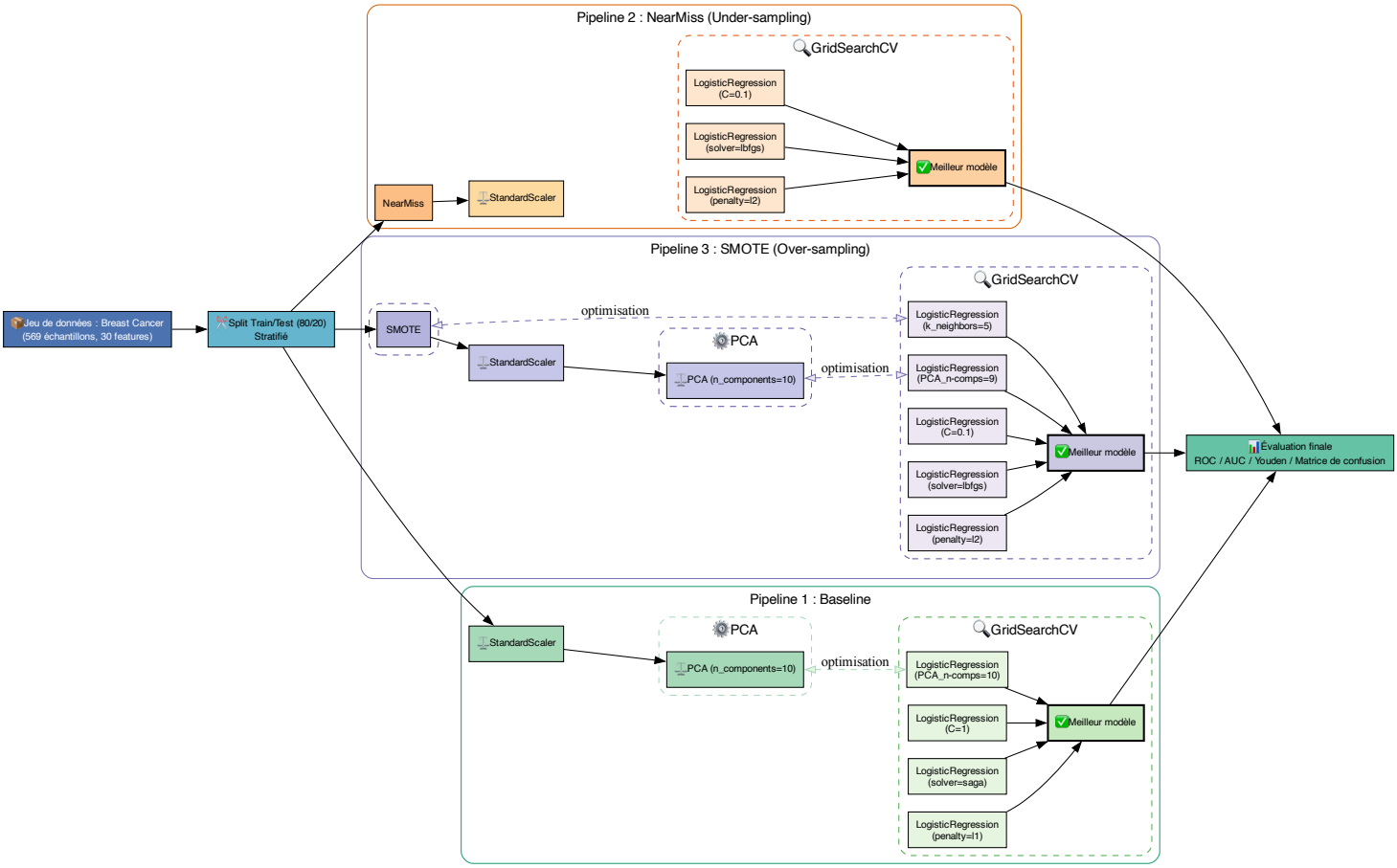


FIGURE 5.1 – Schéma des modèles de régressions logistiques

Ensuite il nous a paru utile de vérifier si la réduction de dimension (via PCA) pouvait améliorer les prédictions de nos modèles ou du moins conserver des performances similaires avec un nombre d'axes grandement réduit :

1. Modèle baseline (sans ré-échantillonage) : Réduire la dimensionalité permet de se débarasser d'une partie du bruit qui affecte les prédictions du modèle et nous permet ici de retirer un faux négatif.
2. Near Miss : la réduction de dimension peut projeter les données dans un espace où la frontière entre classes est moins bien préservée, car PCA maximise la variance globale et non la séparation des classes. On observe que la performance de la régression logistique à légèrement diminué car certaines informations discriminantes sont perdues. En effet on a d'abord réduit la classe majoritaire (bénin) puis réduit la dimensionalité, cela est conforme avec l'appariton d'un faux positif supplémentaire.
3. SMOTE : La réduction de dimension conserve les directions de plus grande variance, qui sont préservés par les nouveaux point synthétisés par SMOTE. On observe une amélioration de la performance car les axes se basent sur des données plus abondantes et mieux équlibrées.

5.2 Résultats

TABLE 5.1 – Comparaison des métriques de performance pour les 6 modèles

Métrique	Logit Baseline	Logit NM	Logit SMOTE	Logit-PCA Baseline	Logit-PCA NM	Logit-PCA SMOTE
Sensibilité B	0.986	**1.0**	0.903	0.986	0.903	0.917
Sensibilité M	0.952	0.976	**1.0**	0.976	**1.0**	**1.0**
Précision B	0.973	0.986	**1.0**	0.986	**1.0**	**1.0**
Précision M	0.976	**1.0**	0.857	0.976	0.857	0.875
F1-score B	0.979	**0.993**	0.949	0.986	0.949	0.957
F1-score M	0.964	**0.988**	0.923	0.976	0.923	0.933
Accuracy	0.974	**0.991**	0.939	0.982	0.939	0.947
Sensibilité moy. pond.	0.974	**0.991**	0.939	0.982	0.939	0.947
AUC	0.980	**0.998**	0.995	0.997	0.997	0.997

La régression logistique obtient d'excellentes performances en retenant les meilleurs modèles (ceux qui ne produisent aucun faux négatif) : Logit SMOTE, Logit PCA NM (PCA Near Miss), Logit PCA SMOTE. Parmi eux, ceux avec PCA produisent un AUC supérieur. Le Logit PCA SMOTE semble être ainsi le meilleur modèle au prix de 6 cas bénins classés comme malins. Son AUC de 0.997 est très élevé et confirme la qualité de prédiction du modèle.

Régression Logistique : Réseau Neuronal

6.1 Description du modèle

On a souhaité implémenter dans notre modèle une technique de régularisation dite early stopping. Il s'agit pour des modèles notamment de réseaux neuronnaux de stopper l'entraînement dès que l'erreur sur le jeu de validation atteint son minimum (pour un nombre de batches donnés). Par ailleurs, il est nécessaire de construire le modèle de régression logistique sous la forme d'un réseau neuronal avec une seule couche (l'output layer) afin d'y greffer un algorithme d'early stopping. Dans les faits, employer un réseau neuronal unicouche avec fonction d'activation sigmoïde revient exactement à construire un modèle de régression logistique : $h_{W,b}(X) = \phi(XW + b)$ avec $\phi = \sigma$

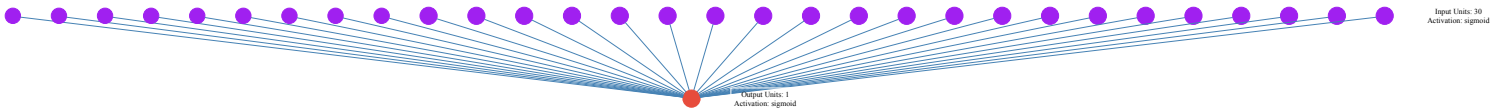


FIGURE 6.1 – Schéma du réseau Neuronal pour la régression logisitque

Avec la binary cross-entropy loss function, c'est-à-dire la fonction de coût à minimiser. On détermine les poids optimaux de chaque noeud par back-propagation, c'est-à-dire calculer les contribution de chaque noeud à la perte finale par discrimination. On calcule ainsi les gradients grâce à la règle de la chaîne.

6.2 Résultats

TABLE 6.1 – Comparaison des métriques de performance pour les 2 modèles

Métrique	Early-Stopping	Early-Stopping NearMiss
Sensibilité B	0.958	**0.972**
Sensibilité M	**1.0**	0.976
Précision B	**1.0**	0.986
Précision M	0.933	**0.953**
F1-score B	**0.979**	**0.979**
F1-score M	**0.966**	0.965
Accuracy	**0.974**	**0.974**
Sensibilité moy. pond.	**0.974**	**0.974**
AUC	**0.999**	0.996

Le modèle avec Early-Stopping présente des performances remarquables : il ne commet aucun faux négatif et seulement trois faux positifs, tout en détectant l'ensemble des cas malins. Sa sensibilité moyenne pondérée atteint 97%. Toutefois, comme attendu, l'utilisation de la méthode de ré-échantillonnage NearMiss dégrade les performances du modèle. En effet, bien que le réseau neuronal utilisé, sans couche cachée, donc équivalent à une régression logistique, puisse apprendre

des relations linéaires entre les variables, il nécessite néanmoins un volume d'échantillons suffisant pour estimer correctement ces relations. La réduction du nombre d'exemples par NearMiss supprime une partie de l'information utile et pénalise ainsi la stabilité et la performance du modèle.

TABLE 6.2 – Top 10 des plus grands coefficients du modèle de régression logistique

Variable	**Odds Ratio**
worst radius	2.1978
worst area	2.0431
worst compactness	1.6524
worst smoothness	1.5926
mean texture	1.5652
radius standard error	1.5537
mean radius	1.5468
worst symmetry	1.5285
worst concave points	1.4793
mean concave points	1.4565

KNN

7.1 Description du modèle

Le K-Nearest Neighbors (KNN) classe une observation selon la classe majoritaire parmi ses k plus proches voisins. Ce modèle non paramétrique est adapté car il ne fait aucune hypothèse sur la distribution des données biologiques et peut capturer des frontières de décision non linéaires complexes. Une validation croisée a sélectionné $k = 3$ avec un score de 0.929.

7.2 Résultats

TABLE 7.1 – Matrice de confusion - KNN ($k = 3$)

	Prédit B	**Prédit M**
Vrai Bénin	68	4
Vrai Malignant	1	41

TABLE 7.2 – Métriques de performance - KNN

Métrique	**B**	**M**
Precision	0.99	0.91
Recall	0.94	0.98
F1-score	0.96	0.94
Accuracy	0.96	
Macro avg	0.95	
Weighted avg	0.96	

Le modèle KNN avec $k = 3$ obtient une précision de 96%. Il prédit très bien les cas bénins (68 sur 72) mais commet 1 faux négatif sur les cas malins. L'AUC de 0.982 reste élevée mais légèrement inférieure aux autres modèles. Le KNN montre une bonne capacité à identifier les cas bénins et une bonne sensibilité pour les cas malins.

Random Forest (RF)

8.1 Description du modèle

Les modèles de forêts aléatoire font partie des modèles de Machine Learning dits „Ensemble Methods“, il s'agit d'une collection d'arbres décisionnels, chaque arbre étant légèrement différents de ses voisins, la prédiction finale s'obtient par vote majoritaire des composantes (arbres) de l'ensemble. Au lieu de rechercher le feature optimal lors du fractionnement d'un noeud (comme le ferait un arbre décisionnel classique), il recherche le meilleur feature parmi un sous-ensemblre aléatoire de features. On obtient au final un modèle non paramétrique, et qui répond au problème classique de surapprentissage. Derrière cette idée, chaque arbre fait individuellement de bonnes prédictions, mais surapprend une partie des données ; en construisant plusieurs arbres, on compense leur excès individuels.

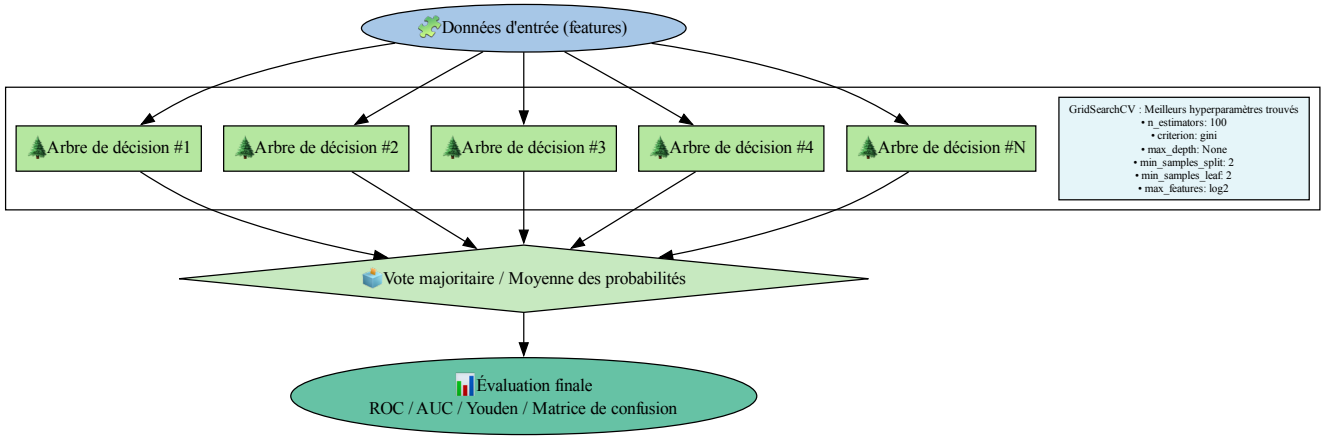


FIGURE 8.1 – Schéma du modèle de forêts aléatoires

8.2 Résultats

TABLE 8.1 – Comparaison des métriques de performance Random Forest Classifier

Métrique	Random Forest Classifier
Sensibilité B	0.972
Sensibilité M	1.0
Précision B	1.0
Précision M	0.955
F1-score B	0.99
F1-score M	0.98
Accuracy	0.98
Sensibilité moy. pond.	0.98
AUC	0.998

Le Random Forest obtient les meilleures performances parmi tous les modèles testés. Il ne commet aucun faux négatif et seulement 2 faux positifs, respectivement une sensibilité maligne de 1.0 et une sensibilité bénine de 0.972. Cela signifie qu'il a été capable de prédire l'intégralité des cas malins et 70/72 des cas bénins.

Discussions et Conclusions

9.1 Comparaison des modèles

TABLE 9.1 – Comparaison des performances des modèles

Modèle	Accuracy	AUC	FP	FN
LDA	**0.98**	0.997	**0**	2
QDA	0.94	0.989	6	1
Régression Logistique	0.97	**0.999**	3	**0**
KNN ($k = 3$)	0.96	0.982	4	1
Random Forest	**0.98**	0.998	2	**0**

9.1.1 Analyse des résultats

Les résultats montrent que tous les modèles obtiennent de très bonnes performances, avec des AUC supérieures à 0.98. La figure 9.2 présente les courbes ROC pour tous les modèles, permettant une comparaison visuelle de leurs capacités discriminantes. Les meilleurs modèles sont la régression logistique avec Early-Stopping (Réseau Neuronal) et le Random Forest avec une AUC de 0.999 et 0,998. Cette mesure garantit une excellente performance globale, quel que soi le seuil sélectionné.

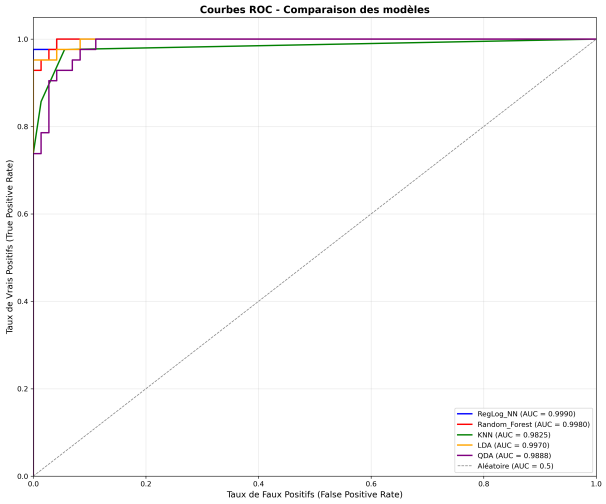


FIGURE 9.1 – Courbes ROC pour tous les modèles de classification. Plus la courbe est proche du coin supérieur gauche, meilleure est la performance du modèle.

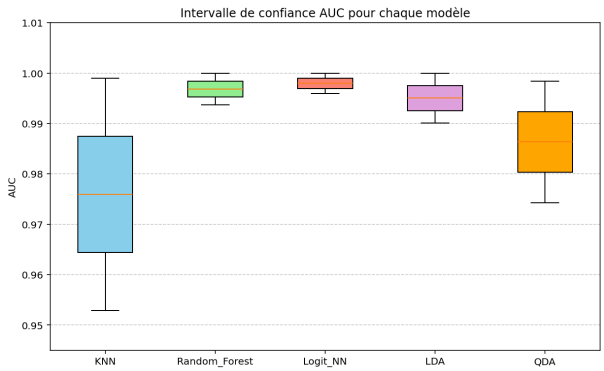


FIGURE 9.2 – Boxplot des intervalles de confiances pour les AUC de la quasi majorité des modèles

Le LDA prédit parfaitement les cas bénins (0 faux positifs) avec seulement 2 faux négatifs. La régression logistique prédit tous les cas malins, tout comme Random Forest (0 faux négatif). L'avantage semble être (sur notre échantillon) au modèle Random Forest puisqu'il commet un faux positif de moins que la régression logistique.

Le KNN et le QDA présentent des performances légèrement inférieures, avec respectivement 96% et 94% de précision, mais tous les modèles restent très performants avec des AUC supérieures à 0.98.

9.2 Discussion des résultats & Interprétation médicale

L'analyse des variables dominantes révèle des caractéristiques cohérentes avec la biologie du cancer du sein (cf. figure 10.6) . Les variables les plus importantes identifiées par le Random Forest et la régression logistique incluent principalement les mesures de concavité (`concave_points_worst`, `concavity_worst`), de taille (`radius_worst`, `perimeter_worst`, `area_worst`) et de texture (`texture_mean`).

D'un point de vue médical, ces résultats semblent pertinents : les cellules malignes présentent généralement des contours plus irréguliers et concaves, reflétant une croissance invasive. Les mesures *worst* (maximales) sont particulièrement discriminantes car elles capturent les régions anormales de la tumeur, souvent caractéristiques d'une malignité. La texture, mesurée par l'écart-type des valeurs en niveaux de gris, reflète la densité cellulaire plus désorganisée dans les tumeurs malignes.

Ces caractéristiques morphologiques sont utilisées en pratique par les médecins, parmi d'autres bien-sûr, qui sont évidemment au-delà du spectre de nos connaissances en tant qu'étudiants en mathématiques. Notre modèle tente cependant d'automatiser cette analyse médicale en une analyse quantitative, permettant une aide à la décision rapide et efficace. La bonne performance des modèles (selon les taux de faux négatifs et l'AUC) suggère que les variables du dataset contiennent suffisamment d'information pour distinguer assez efficacement les tumeurs bénignes des malignes, pouvant aboutir à des applications d'aide au diagnostic dans le domaine de l'imagerie médicale.

Source : American Cancer Society, "What Are Neoplasms and Tumors?", 2025.

9.3 Conclusion

À la lumière des modèles étudiés, il semble raisonnable de sélectionner ceux qui ne produisent aucun faux négatif. D'un point de vue clinique le médecin privilégie un modèle qui considère des bénins comme malins plutôt que l'inverse. Deux modèles se démarquent, la régression logistique via réseau-neuronal ainsi que le modèle de forêts aléatoires.

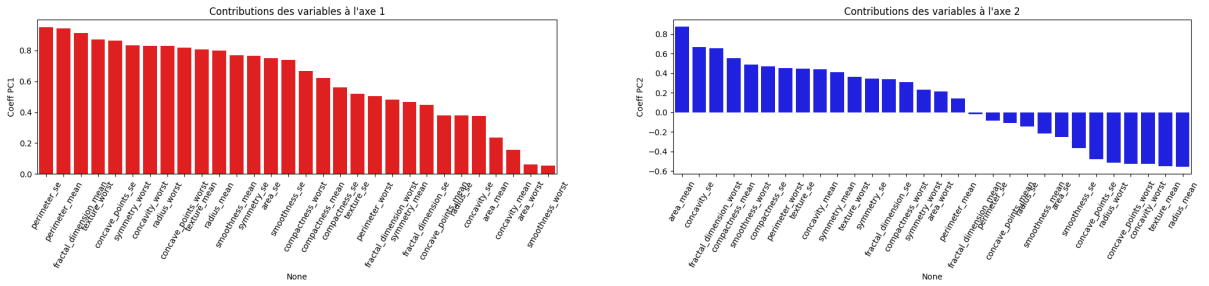
La régression logistique via réseau de neurones conserve l'interprétabilité du modèle linéaire classique tout en bénéficiant d'une optimisation par descente de gradient et d'une régularisation implicite grâce à l'early stopping. Le Random Forest, quant à lui, excelle grâce à sa nature non linéaire et ensembliste : il agrège les décisions de nombreux arbres indépendants, réduisant ainsi la variance et les erreurs liées à la sélection d'un seul modèle. Il capture efficacement les interactions complexes entre les variables (par exemple entre la texture, la concavité et la taille des noyaux cellulaires), souvent présentes dans les données biologiques réelles. De plus, il reste peu sensible aux valeurs aberrantes et à la colinéarité entre variables. En pratique, ces deux

approches se complètent : la régression logistique fournit une interprétation directe et quantifiable du risque (utile en contexte médical), tandis que le Random Forest assure une sensibilité prédictive maximale. Leur performance quasi parfaite (AUC 0.999) indique que les signaux présents dans les données sont à la fois linéairement séparables et renforcés par des interactions non linéaires, expliquant pourquoi ces deux modèles atteignent les meilleurs résultats.

Annexe

10.1 Application de la PCA

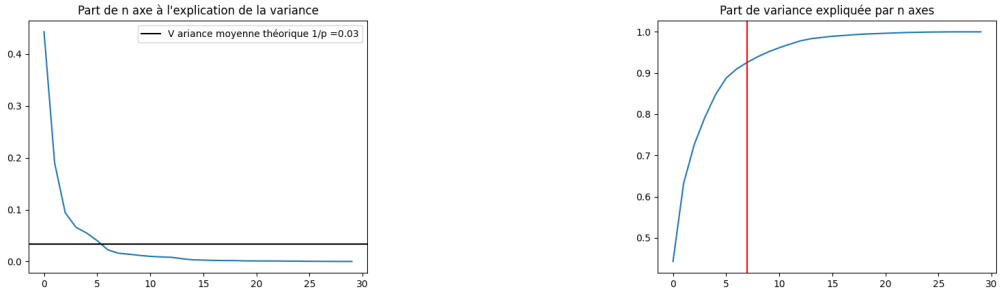
L'Analyse en Composantes Principales (PCA) transforme les variables corrélées en composantes principales non corrélées, ordonnées par variance expliquée. Cette technique est utile pour notre problème car elle réduit la redondance parmi les 30 variables corrélées, permet de visualiser la séparation des classes dans un espace de dimension réduite, et peut améliorer les performances en réduisant le bruit.



(a) Contributions axe 1 (b) Contributions axe 2

FIGURE 10.1 – Contributions des variables aux deux premières composantes principales

Les figures 10.2a et 10.2b montrent respectivement la variance expliquée par chaque axe et la variance cumulée. Les figures 10.3a et 10.3b permettent de sélectionner le nombre optimal d'axes, fixé à 9 dans ce projet.



(a) Variance par axe (b) Variance cumulée

FIGURE 10.2 – Analyse de la variance expliquée par la PCA

(a) AUC vs nombre d'axes (b) Critères AIC et BIC

FIGURE 10.3 – Sélection du nombre optimal d'axes PCA

La régression logistique appliquée aux données transformées par PCA (10 axes) obtient une AUC de 0.997, proche de celle sans PCA (0.980). La projection sur les deux premiers axes

(figure 10.4) montre une bonne séparation visuelle entre les classes. Les cercles des corrélations (figures **??** et **??**) et les contributions des variables (figures 10.1a, 10.1b et 10.5) identifient les caractéristiques les plus importantes.

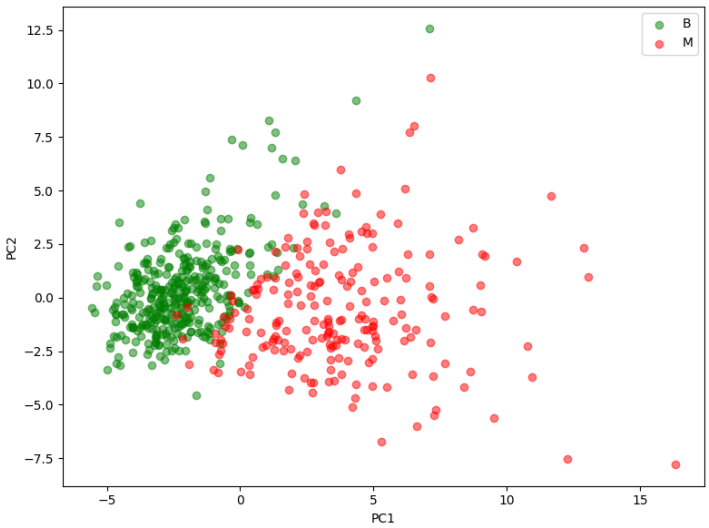


FIGURE 10.4 – Projection des patients sur le plan formé par les deux premiers axes PCA, colorée selon le diagnostic

On observe une bonne séparation entre les classes.

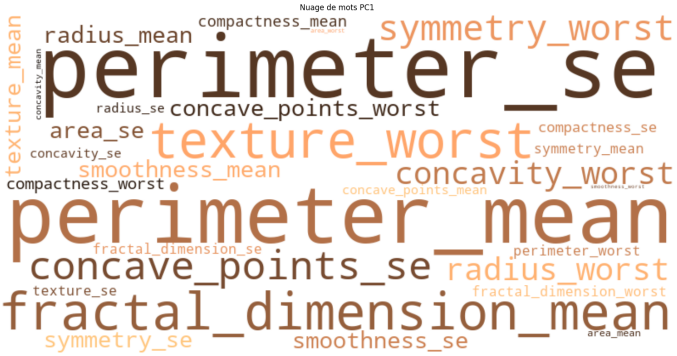


FIGURE 10.5 – Nuage de mots représentant les contributions des variables à la première composante principale (PC1)

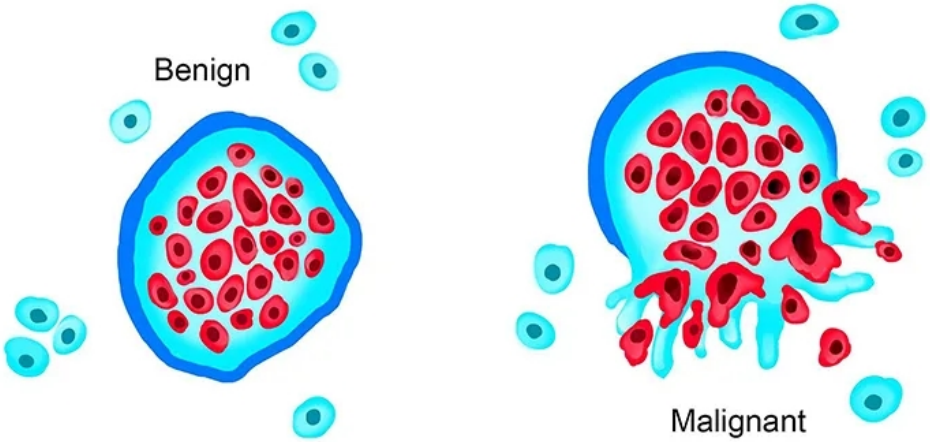


FIGURE 10.6 – Tumeur Bénigne vs Maligne