

ANALYSE DE :WISCONSIN DIAGNOSTIC BREAST CANCER (WDBC)

TARAK DARZA

Table des matières

1. Compréhension des Données

Exploration initiale : Examiner la structure des données, y compris le nombre de cas, les types de variables et les valeurs manquantes.

Statistiques descriptives : Calculer des statistiques résumées pour chaque variable, telles que la moyenne, la médiane, l'écart-type, etc.

2. Nettoyage des Données

Gestion des valeurs manquantes : Décider de la manière de traiter les valeurs manquantes (suppression, imputation, etc.).

Correction des erreurs : Identifier et corriger les erreurs ou incohérences dans les données (valeurs aberrantes, fautes de frappe, etc.).

3. Analyse Exploratoire des Données (AED)

Visualisation : Utiliser des graphiques (histogrammes, boîtes à moustaches, scatter plots) pour observer la distribution et les relations entre les variables.

Corrélation : Évaluer la corrélation entre les variables pour identifier les relations potentielles.

Analyse Bivariée

4. Préparation des Données

Sélection des caractéristiques : Choisir les variables les plus pertinentes pour la prédiction du diagnostic.

Normalisation/Standardisation : Appliquer des transformations pour que les données soient sur une échelle comparable.

Division du jeu de données : Séparer les données en ensembles d'entraînement et de test pour valider les performances du modèle.

5. Modélisation

Choix du modèle : Sélectionner un ou plusieurs algorithmes de machine learning (forêt aléatoire, SVM, régression logistique, etc.).

Entraînement du modèle : Appliquer l'algorithme sur l'ensemble d'entraînement pour construire le modèle.

Optimisation : Ajuster les hyperparamètres du modèle pour améliorer ses performances.

6. Évaluation du Modèle

Validation croisée : Utiliser cette technique pour estimer la performance du modèle sur des données non vues.

Métriques de performance : Calculer des métriques telles que la précision, le rappel, le score F1 et l'AUC pour évaluer la qualité des prédictions.

7. Interprétation des Résultats

Analyse des erreurs : Examiner les cas mal classifiés pour comprendre les limites du modèle.

Importance des caractéristiques : Identifier les variables qui contribuent le plus à la prédiction.

8. Déploiement du Modèle

Intégration du modèle : Incorporer le modèle dans une application ou un système de workflow existant.

Surveillance et maintenance : Suivre les performances du modèle au fil du temps et le mettre à jour au besoin.

Introduction :

Le Diagnostic du Cancer du Sein du Wisconsin (WDBC) est une base de données bien établie et largement utilisée dans le domaine de la recherche médicale pour étudier les caractéristiques des cellules cancéreuses du sein.

Cette base de données contient des mesures cliniques de cellules issues de biopsies de tissus mammaires, accompagnées de leur étiquetage comme étant bénignes ou malignes.

L'analyse approfondie de ces données peut fournir des informations cruciales pour la compréhension des mécanismes au cancer du sein et pour le développement de méthodes de diagnostic plus efficaces.

Dans le cadre de ce projet, l'objectif est d'explorer en profondeur les données du WDBC afin d'identifier les caractéristiques les plus discriminantes entre les tumeurs bénignes et malignes.

Pour ce faire, j'utilise des techniques d'analyse de données et de machine learning, en mettant l'accent sur l'utilisation des langages de programmation R et Python, deux langages puissants et largement adoptés dans le domaine de la science des données et de la bioinformatique.

L'intégration de R et Python dans ce projet permettra non seulement une exploration exhaustive des données, mais aussi la mise en œuvre de modèles prédictifs sophistiqués pour la classification des tumeurs.

En combinant les forces de ces deux langages, je pourrais bénéficier d'une gamme étendue d'outils et de bibliothèques spécialisées, ainsi que d'une flexibilité accrue dans le processus d'analyse.

Cette approche multidisciplinaire me permettra d'approfondir ma compréhension du cancer du sein tout en développant des compétences essentielles en programmation et en analyse de données.

Cette introduction présente brièvement le contexte du projet, ses objectifs et l'approche méthodologique qui sera adoptée, en mettant en avant l'utilisation des langages R et Python pour mener à bien l'analyse des données.

1-Compréhension des Données

1.1 Description du jeu de données :

Le jeu de données du Diagnostic du Cancer du Sein du Wisconsin (WDBC) comprend des mesures cliniques détaillées extraites de biopsies de tissus mammaires, ainsi que des étiquettes de classification indiquant si les cellules sont bénignes ou malignes.

Ces mesures comprennent des informations telles que le rayon, la texture, la compacité, la symétrie et d'autres caractéristiques des noyaux cellulaires, chacune étant quantifiée à l'aide de techniques d'imagerie médicale.

Au total, le jeu de données contient des informations sur plusieurs centaines de cas. Il contient un total de 569 cas (ou lignes) et 32 variables

fournissant ainsi une base solide pour une analyse statistique approfondie visant à identifier les caractéristiques distinctives des tumeurs malignes par rapport aux tumeurs bénignes.

La majeure partie de la donnée étudiée a été extraite du site suivant :

<https://lipn.univ-paris13.fr/~grozavu/PredA/dataProject/2%20-%20breast-cancer-wisconsin/>

J'ai choisi de récupérer tous les variables décrivant la population.

1.2-Description des variables :

Nom de Variable	Type	Description de variable
id	int64	identifiant unique pour chaque patient
diagnosis	object	Catégorie du diagnostic - (M) ou (B).
radius_mean	float64	moyenne des distances du centre à des points sur le périmètre
texture_mean	float64	standard deviation des échelles de gris
perimeter_mean	float64	Périmètre moyen des cellules tumorales.
area_mean	float64	Surface moyenne des cellules tumorales
smoothness_mean	float64	variation locale des longueurs des rayons
compactness_mean	float64	Compacité moyenne
concavity_mean	float64	Concavité moyenne
concave points_mean	float64	nombre de portions concaves du contour
symmetry_mean	float64	Symétrie moyenne
fractal_dimension_mean	float64	Dimension fractale moyenne
radius_se	float64	Erreur standard du rayon
texture_se	float64	Erreur standard de la texture
perimeter_se	float64	Erreur standard du périmètre
area_se	float64	Erreur standard de la surface
smoothness_se	float64	Erreur standard du lissage
compactness_se	float64	Erreur standard de la compacité
concavity_se	float64	Erreur standard de la concavité

concave points_se	float64	Erreur standard des points concaves.
symmetry_se	float64	Erreur standard de la symétrie
fractal_dimension_se	float64	Erreur standard de la dimension fractale
radius_worst	float64	moyenne des trois plus grands rayons
texture_worst	float64	moyenne des trois plus grandes textures
perimeter_worst	float64	Plus grand périmètre
area_worst	float64	Plus grande surface
smoothness_worst	float64	Plus grand lissage.
compactness_worst	float64	Plus grande compacité
concavity_worst	float64	Plus grande concavité
concave points_worst	float64	Plus grand nombre de points concaves
symmetry_worst	float64	Plus grande symétrie
fractal_dimension_worst	float64	Plus grande dimension fractale
Unnamed: 32	float64	

1.3-Visualisation et Exploration initiale :

La première étape de l'analyse des données (WDBC) sous Python consiste à obtenir une vue globale des données.

*Après l'importation des bibliothèques nécessaires telles que Pandas, NumPy et Matplotlib, le jeu de données est chargé dans une structure de données, généralement un DataFrame Pandas. Ensuite, une inspection initiale des données est effectuée avec la fonction **def lire_csv***

*pour afficher les premières lignes du notre fichier data je fais appel à ma fonction **def afficher_lignes et def info_data** pour obtenir des informations sur les types de données et les valeurs manquantes.*

Les Trois Groupes de Variables :

Means: Les valeurs moyennes de ces caractéristiques pour chaque noyau de cellule, calculées à partir de l'image.

SE (Standard Error): L'erreur standard de ces mesures, qui donne une idée de la variabilité de chaque mesure entre les noyaux de cellules dans l'échantillon.

Worst (ou Largest Mean): La moyenne des trois plus grandes valeurs (les pires cas) pour chacune de ces caractéristiques parmi tous les noyaux de cellules de l'image.

La différence entre ces groupes est liée à la manière dont les mesures sont résumées :

Means donne la valeur moyenne pour toutes les cellules de l'échantillon, reflétant les propriétés typiques.

SE fournit une mesure de la variation ou de la dispersion des valeurs moyennes, indiquant la fiabilité de la moyenne.

Worst résume le pire cas trouvé dans les mesures, donnant une indication des valeurs extrêmes qui pourraient être particulièrement pertinentes pour l'identification des tumeurs malignes.

str(data) :

data.frame': 569 obs. of 33 variables: Cela indique que data est un data frame avec 569 observations et 33 variables. Ensuite, chaque ligne suivante représente une variable dans le data frame, avec les informations suivantes :Le nom de la variable.

Le type de données de la variable (int, num, chr, logi).

2. Nettoyage des Données

Identification des données manquantes : Identifiez les valeurs manquantes dans l'ensemble de données.

Sur ce jeu de données, il n'y a pas de valeur manquante.

Détection des valeurs aberrantes : Recherchez et traitez les valeurs aberrantes qui pourraient fausser vos analyses.

techniques statistiques telles que les écarts à la moyenne ou en utilisant des méthodes graphiques comme les diagrammes en boîte ou les diagrammes de dispersion que je veux les faire sur l'étape suivante

Validation des données : Vérification de la validité des données en vous assurant qu'elles respectent les contraintes et les règles métier. Par exemple, vérifiez que les dates sont dans le bon format, que les valeurs numériques sont dans les plages attendues..

Normalisation et standardisation : Si nécessaire, normalisez ou standardisez les données pour les mettre sur une même échelle. Cela peut être utile lorsqu'on travaille avec des algorithmes sensibles à l'échelle des variables, comme les modèles basés sur la distance.

Traitement des doublons : Identifiez et supprimez les doublons dans le jeu de données, en s'assurant de conserver uniquement les entrées uniques.

Conversion des types de données : les types de données de chaque variable sont corrects. Par exemple, les dates devraient être traitées comme des objets de date, les variables catégorielles devraient être converties en variables indicatrices si nécessaire..

Segmentation des données : Si mon ensemble de données est volumineux, je peux envisager de le segmenter en sous-ensembles plus gérables pour faciliter le nettoyage et l'analyse.

Documentation : Documentation de toutes les étapes que j'ai suivies pour nettoyer les données, y compris les décisions prises et les techniques utilisées.

3. Analyse Exploratoire des Données (AED)

1-Statistiques descriptives :

Indicateur Clé :diagnosis

Ensemble, cela me donne les pourcentages des niveaux 'B' et 'M' de la variable diagnosis avec une précision de 4 chiffres. Voici le résultat que vous avez obtenu :

'B': 62.74%

'M': 37.26%

Répartition des diagnostics : Environ 62.74% des observations sont étiquetées comme "B" (bénignes) et environ 37.26% sont étiquetées comme "M" (malignes). Cela suggère qu'il y a une répartition inégale entre les diagnostics bénins et malins dans l'échantillon analysé.

si un modèle de prédiction ou une méthode de diagnostic est développé à partir de cet ensemble de données, il devra être évalué en tenant compte de cette répartition inégale des diagnostics pour éviter tout biais potentiel.

Calcul de moyenne et de médiane des variables

	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fractal_dimension_mean
count	569	569	569	569	569	569	569	569	569	569
mean	14,12729174	19,28964851	91,96903339	654,8891037	0,096360281	0,104340984	0,088799316	0,048919146	0,181161863	0,06279761
std	3,524048826	4,301035768	24,29898104	351,9141292	0,014064128	0,052812758	0,079719809	0,038802845	0,027414281	0,007060363
min	6,981	9,71	43,79	143,5	0,05263	0,01938	0	0	0,106	0,04996
min%	11,7	16,17	75,17	420,3	0,08637	0,06492	0,02956	0,02031	0,1619	0,0577
50%	13,37	18,84	86,24	551,1	0,09587	0,09263	0,06154	0,0335	0,1792	0,06154
75%	15,78	21,8	104,1	782,7	0,1053	0,1304	0,1307	0,074	0,1957	0,06612
max	28,11	39,28	188,5	2501	0,1634	0,3454	0,4268	0,2012	0,304	0,09744
radius_se	569	569	569	569	569	569	569	569	569	569
count	569	569	569	569	569	569	569	569	569	569
mean	0,405172056	1,216852427	2,866059227	40,33707909	0,007040979	0,025478139	0,031893716	0,011796137	0,020542299	0,003794904
std	0,277312733	0,551648393	2,021854554	45,49100552	0,003002518	0,017908179	0,03018606	0,006170285	0,008266372	0,002646071
min	0,1115	0,3602	0,757	6,802	0,001713	0,002252	0	0	0,007882	0,0008948
min%	0,2324	0,8339	1,606	17,85	0,005169	0,01308	0,01509	0,007638	0,01516	0,002248
50%	0,3242	1,108	2,287	24,53	0,00638	0,02045	0,02589	0,01093	0,01873	0,003187
75%	0,4789	1,474	3,357	45,19	0,008146	0,03245	0,04205	0,01471	0,02348	0,004558
max	2,873	4,885	21,98	542,2	0,03113	0,1354	0,396	0,05279	0,07895	0,02984
radius_worst	569	569	569	569	569	569	569	569	569	569
count	569	569	569	569	569	569	569	569	569	569
mean	16,26918981	25,6772232	107,2612127	880,5831283	0,132368594	0,254265044	0,272188483	0,114606223	0,290075571	0,083945817
std	4,83324158	6,146257623	33,60254227	569,3569927	0,022832429	0,157336489	0,208624281	0,065732341	0,061867468	0,018061267
min	7,93	12,02	50,41	185,2	0,07117	0,02729	0	0	0,1565	0,05504
min%	13,01	21,08	84,11	515,3	0,1166	0,1472	0,1145	0,06493	0,2504	0,07146
50%	14,97	25,41	97,66	686,5	0,1313	0,2119	0,2267	0,09993	0,2822	0,08004
75%	18,79	29,72	125,4	1084	0,146	0,3391	0,3829	0,1614	0,3179	0,09208
max	36,04	49,54	251,2	4254	0,2226	1,058	1,252	0,291	0,6638	0,2075

La plage des valeurs pour chaque variable numérique varie considérablement :

radius_mean, les valeurs vont de 6.981 à 28.11, ce qui indique une grande variation dans les tailles des tumeurs.

radius_mean, *texture_mean*, *perimeter_mean*, *area_mean* :

Elles ont des moyennes comprises entre environ 14 et 91, ce qui suggère une variabilité dans les tailles et les formes des tumeurs.

Les valeurs maximales montrent qu'il existe des tumeurs très grandes dans l'ensemble de données, avec des rayons pouvant atteindre jusqu'à 28.11.

Distribution des Données :

Pour plusieurs variables, la moyenne et la médiane sont proches, suggérant une distribution relativement symétrique. Cependant, des écarts entre ces deux mesures pour certaines variables peuvent indiquer des distributions asymétriques ou la présence de valeurs aberrantes.

les variables montrent une large gamme de moyennes et de médianes, indiquant des échelles différentes. Cela souligne l'importance de la normalisation ou de la standardisation des données avant de les utiliser dans des analyses comparatives ou des modèles prédictifs.

Taille et Forme des Tumeurs :

Les variables liées à la taille des tumeurs (*radius_mean*, *perimeter_mean*, *area_mean* et leurs équivalents *_worst*) montrent une augmentation significative de la moyenne par rapport à la médiane dans leurs mesures "worst", ce qui peut indiquer une tendance vers des valeurs extrêmes plus élevées dans les cas les plus sévères de tumeurs.

La présence de concavités (*concavity_mean*, *concave points_mean*, et leurs équivalents *_worst*) avec des valeurs moyennes et médianes indique que de nombreuses tumeurs présentent des irrégularités significatives, ce qui est souvent associé à des tumeurs malignes.

Texture et Lissage :

La texture (*texture_mean* et *texture_worst*) montre une variabilité modérée, avec des valeurs médianes légèrement inférieures aux moyennes, suggérant une distribution avec une queue plus lourde vers les valeurs plus élevées.

Le lissage (*smoothness_mean* et *smoothness_worst*) reste relativement constant, mais augmente dans les mesures "worst", ce qui peut refléter des changements dans la texture de la surface de la tumeur à mesure qu'elle devient plus agressive.

Symétrie et Dimension Fractale :

La symétrie et la dimension fractale (*symmetry_mean*, *fractal_dimension_mean*, et leurs équivalents *_worst*) montrent que ces caractéristiques augmentent également dans les mesures "worst", ce qui peut indiquer une complexité et une asymétrie croissantes dans les cas de tumeurs plus avancées.

L'analyse des distributions, la visualisation des données, pourraient être les prochaines étapes pour approfondir mes compréhension de ces données.

Calcul de variance et Ecart_type

Variance des variables		Écart-type des variables	
radius_mean	12.418920	radius_mean	3.524049
texture_mean	18.498909	texture_mean	4.301036
perimeter_mean	590.440480	perimeter_mean	24.298981
area_mean	123843.554318	area_mean	351.914129
smoothness_mean	0.000198	smoothness_mean	0.014064
compactness_mean	0.002789	compactness_mean	0.052813
concavity_mean	0.006355	concavity_mean	0.079720
concavepoints_mean	0.001506	concavepoints_mean	0.038803
symmetry_mean	0.000752	symmetry_mean	0.027414
fractal_dimension_mean	0.000050	fractal_dimension_mean	0.007060
radius_se	0.076902	radius_se	0.277313
texture_se	0.304316	texture_se	0.551648
perimeter_se	4.087896	perimeter_se	2.021855
area_se	2069.431583	area_se	45.491006
smoothness_se	0.000009	smoothness_se	0.003003
compactness_se	0.000321	compactness_se	0.017908
concavity_se	0.000911	concavity_se	0.030186
concavepoints_se	0.000038	concavepoints_se	0.006170
symmetry_se	0.000068	symmetry_se	0.008266
fractal_dimension_se	0.000007	fractal_dimension_se	0.002646
radius_worst	23.360224	radius_worst	4.833242
texture_worst	37.776483	texture_worst	6.146258
perimeter_worst	1129.130847	perimeter_worst	33.602542
area_worst	324167.385102	area_worst	569.356993
smoothness_worst	0.000521	smoothness_worst	0.022832
compactness_worst	0.024755	compactness_worst	0.157336
concavity_worst	0.043524	concavity_worst	0.208624
concavepoints_worst	0.004321	concavepoints_worst	0.065732
symmetry_worst	0.003828	symmetry_worst	0.061867
fractal_dimension_worst	0.000326	fractal_dimension_worst	0.018061

Ces deux mesures sont essentielles pour comprendre la dispersion des données autour de la moyenne.

Variance :

Mesure le degré de dispersion des données. Une variance élevée indique que les données sont plus éparpillées autour de la moyenne. Une variance plus faible suggère que les données sont plus proches de la moyenne.

Écart-type :

La racine carrée de la variance, fournissant une mesure de dispersion dans les mêmes unités que les données.

Variables de Taille (radius_mean, perimeter_mean, area_mean, et leurs équivalents_worst) :

Ces variables montrent à la fois des variances et des écarts-types élevés, ce qui indique une large gamme dans les tailles des tumeurs. Les valeurs "worst" ont tendance à avoir une dispersion encore plus grande, ce qui pourrait refléter la variabilité accrue dans les cas de tumeurs plus avancées ou agressives.

Caractéristiques de Surface (smoothness_mean, compactness_mean, concavity_mean, concavepoints_mean, et leurs équivalents_worst) :

La dispersion des caractéristiques de surface varie, avec généralement des variances et des écarts-types plus faibles pour les mesures de "mean" par rapport aux mesures de "worst". Cela suggère que les caractéristiques de surface deviennent plus hétérogènes dans les tumeurs classées comme pires cas. Texture (texture_mean, texture_worst):

Les mesures de texture présentent une dispersion modérée à élevée, indiquant une variabilité

dans la sensation ou l'apparence de la surface des tumeurs à travers les échantillons.

Symétrie et Dimension Fractale (symmetry_mean, fractal_dimension_mean, et leurs équivalents_worst) :
Ces mesures montrent une dispersion relativement faible à modérée, suggérant moins de variabilité parmi les tumeurs pour ces caractéristiques, bien que certaines augmentations dans la dispersion soient observées dans les mesures "worst".

2-Analyse de Distribution :Histogramme et Boite de moustache

L'analyse des histogrammes pour l'ensemble de données simulées, inspirée par le Wisconsin Diagnostic Breast Cancer (WDBC), révèle une distribution principalement normale des variables, suggérant une homogénéité dans la variabilité des caractéristiques mesurées.

Cette observation initiale souligne l'importance de procédures de normalisation ou de standardisation dans la préparation des données, afin d'assurer une comparabilité et une efficacité maximales dans les techniques de modélisation ultérieures.

Cette exploration fournit une base solide pour comprendre la distribution des caractéristiques au sein du WDBC, tout en mettant en évidence les étapes critiques de prétraitement des données, indispensables à l'élaboration de modèles analytiques robustes et précis. Cette démarche s'avère essentielle pour avancer vers une compréhension plus nuancée des nuances inhérentes au diagnostic du cancer du sein, facilitant ainsi le développement de solutions diagnostiques plus efficaces.

Je choisis de traiter les données par lot de dix, où chaque groupe de dix correspond à un critère spécifique.

Cette approche me permet de structurer et d'analyser efficacement les informations en les regroupant selon des caractéristiques définies. En divisant les données en ensembles distincts de taille fixe, je facilite la gestion et l'interprétation des données, ce qui me permet d'extraire des insights pertinents et précis pour mes analyses.

Ce processus méthodique me permet également d'optimiser les ressources et de minimiser les erreurs potentielles, assurant ainsi la qualité et la fiabilité de mes résultats.

En prenant en considération les moyennes dans un premier groupe, et en me basant sur les résultats fournis par le box plot, je remarque une similitude entre les **variables perimeter_mean et area_mean**.

Cette similitude attire mon attention quant à une potentielle corrélation entre ces deux variables. Bien qu'il soit important de noter que cette observation ne soit pas toujours exacte, généralement, lorsque les fonctionnalités sont corrélées entre elles, il est envisageable d'en éliminer une sans perdre d'informations significatives. Ainsi, cette constatation initiale suscite une réflexion quant à l'analyse approfondie de la relation entre perimeter_mean et area_mean afin de déterminer leur corrélation et, éventuellement, de rationaliser le modèle en éliminant une des deux variables si nécessaire.

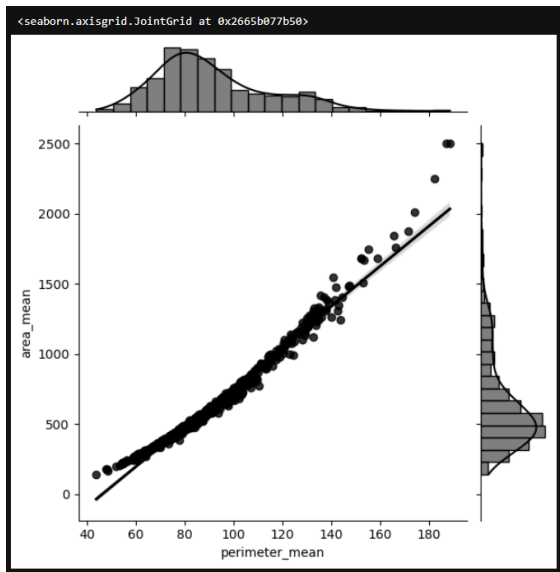
J'ai entrepris d'analyser la corrélation entre les variables perimeter_mean et area_mean en utilisant la fonction suivante :

`sns.jointplot(x='perimeter_mean', y='area_mean', data=features, kind="reg", color="black")`.

L'application de la régression linéaire à ces données révèle une corrélation positive entre ces deux variables.

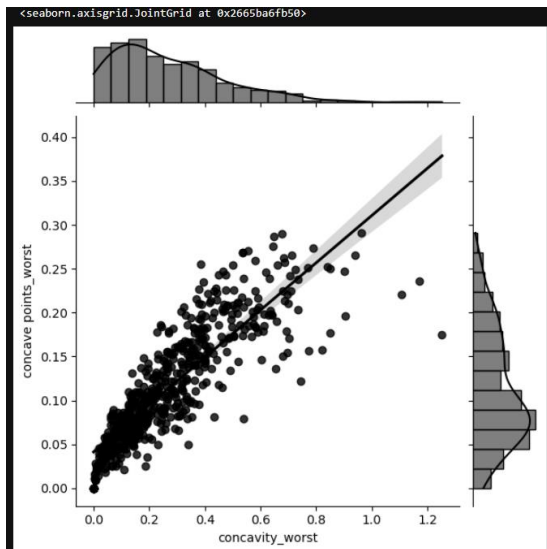
La distribution des points sur le graphique indique un alignement satisfaisant autour de la ligne de régression, suggérant ainsi une forte corrélation linéaire entre les variables. Cette observation suggère que ces deux caractéristiques sont vraisemblablement liées de manière significative.

Cette cohérence renforce mon confiance dans la corrélation identifiée entre `perimeter_mean` et `area_mean`.



J'ai également entrepris une analyse similaire sur les **variables `concavity_worst` et `concave point_worst`**, qui présentent des caractéristiques apparentées et semblent également être fortement corrélées. Cette investigation a été menée à l'aide de la même méthodologie, utilisant la fonction `sns.jointplot` avec les variables d'intérêt et en appliquant la régression linéaire pour évaluer leur corrélation.

Les résultats de cette analyse révèlent une corrélation positive entre `concavity_worst` et `concave point_worst`, comme indiqué par la distribution des points alignés autour de la ligne de régression. Cette cohérence dans la répartition des données suggère une relation significative entre ces deux variables, renforçant l'hypothèse de leur similarité et de leur corrélation élevée.



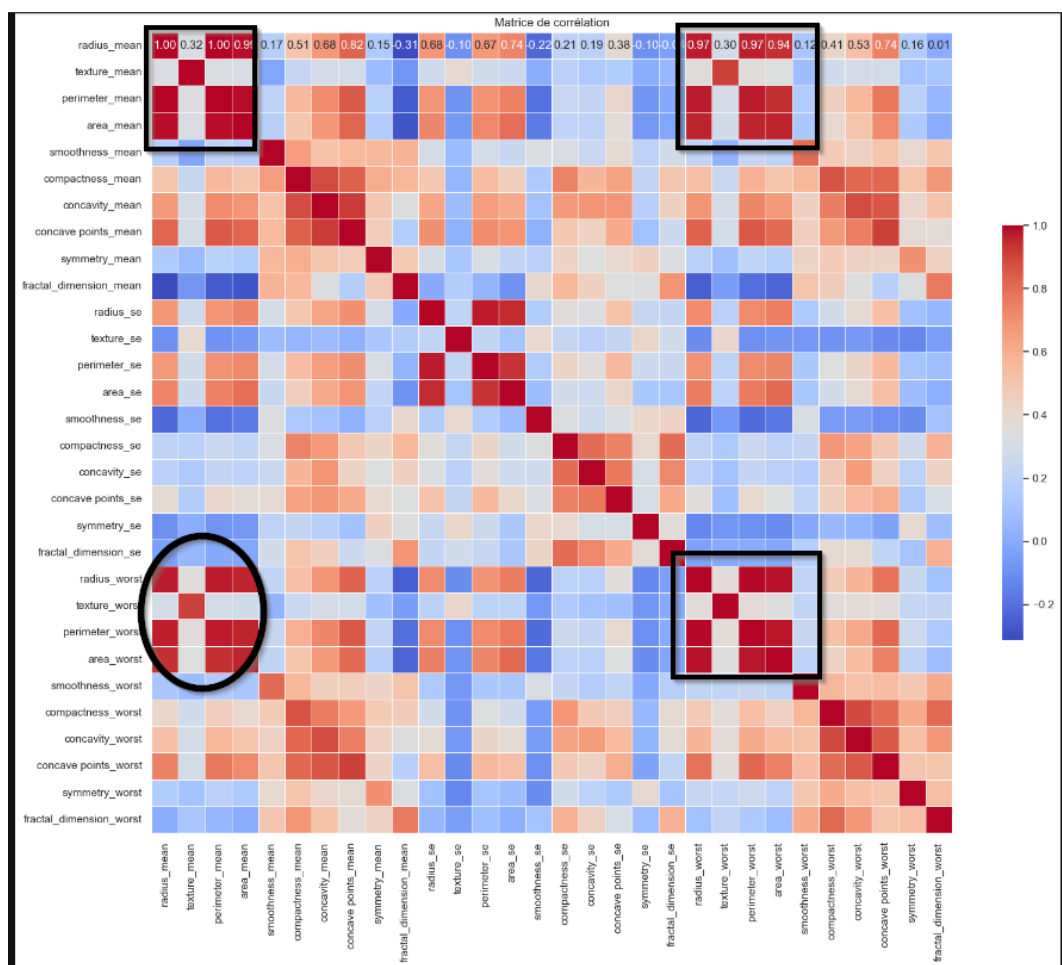
4- Analyse Bivariée

Après avoir mené une analyse exploratoire approfondie des données et tiré des conclusions significatives, je suis désormais prêt à entreprendre une analyse bivariable des données. Cette étape implique d'examiner les relations entre les différentes paires de variables pour mieux comprendre leur interdépendance et leur impact sur le phénomène étudié.

Dans le but de réduire le nombre de variables tout en conservant le maximum d'informations possible, je prévois d'appliquer la méthode de l'Analyse en Composantes Principales (ACP) à mon ensemble de données. L'ACP est une technique puissante qui permet de transformer un ensemble de variables corrélées en un ensemble de variables non corrélées, appelées composantes principales. Ces composantes principales capturent l'essentiel de la variabilité des données tout en réduisant la dimensionnalité de l'ensemble de données.

L'objectif ultime de cette démarche est de parvenir à un ensemble de variables plus restreint tout en conservant la diversité et la richesse des informations contenues dans le jeu de données initial. En réduisant le nombre de variables, je cherche à simplifier l'analyse tout en préservant les relations essentielles entre les variables, ce qui devrait aboutir à un modèle plus robuste et plus interprétable. En somme, l'application de l'ACP vise à optimiser la représentation de mon ensemble de données en sélectionnant les variables les plus informatives tout en réduisant la redondance et le bruit, afin de créer un modèle plus efficace et plus complet.

1-matrice de corrélation



Corrélations Fortes:

Il y a des corrélations très élevées (proches de 1) entre radius_mean, perimeter_mean, et area_mean, ainsi qu'entre leurs équivalents '_worst'.

==> Cela est attendu car ces mesures sont géométriquement liées (le périmètre et l'aire d'un cercle sont directement liés à son rayon).

Ces corrélations élevées suggèrent une redondance d'information, ce qui pourrait justifier une réduction de dimensionnalité ou la sélection de caractéristiques pour éviter la multicollinéarité dans les modèles de machine learning.

Corrélations entre Caractéristiques de Texture et Taille : Il existe des corrélations modérées entre les mesures de texture (texture_mean, texture_worst) et les mesures géométriques. Cela peut indiquer une relation entre la texture de la tumeur et sa taille ou sa croissance.

Indicateurs de Concavité et Points Concaves : Les caractéristiques liées à la concavité (concavity_mean, concave points_mean et leurs équivalents '_worst') montrent également de fortes corrélations entre elles et avec les mesures géométriques.

==> Cela suggère que les tumeurs avec des régions plus concaves ont tendance à être plus grandes et peuvent être plus susceptibles d'être malignes.

Relation entre la Lissité et les Dimensions Fractales : Les mesures de lissité (smoothness_mean, smoothness_worst) et les dimensions fractales (fractal_dimension_mean, fractal_dimension_worst) présentent des corrélations positives modérées avec d'autres caractéristiques, ce qui indique que la complexité de la surface de la tumeur pourrait être liée à d'autres propriétés de la tumeur.

Potentiel pour la Prédiction de Diagnostic : Les caractéristiques qui montrent des corrélations fortes avec de nombreuses autres (par exemple, concave points_worst) pourraient être particulièrement utiles pour prédire le diagnostic (maligne vs bénigne), car elles capturent des informations importantes sur la tumeur.

Implications pour la Modélisation:

Sélection de Caractéristiques : La présence de corrélations élevées suggère que certaines caractéristiques peuvent être redondantes. La sélection de caractéristiques ou l'extraction de caractéristiques (par exemple, via une analyse en composantes principales, PCA) pourrait être bénéfique pour simplifier le modèle sans perdre d'information significative.

2-Test d'extraction des variables significative sans ACP

Après avoir mené une analyse approfondie des données et obtenu des informations précieuses grâce à des techniques telles que l'analyse exploratoire et la matrice de corrélation, je suis désireux de procéder à une extraction de variables significatives sans recourir à l'Analyse en Composantes Principales (ACP). En me basant sur les conclusions tirées de mes analyses précédentes, ainsi que sur les informations fournies par le modèle obtenu à partir de la matrice de corrélation, je souhaite identifier les variables les plus pertinentes pour mon étude.

En utilisant une approche sélective, je vais évaluer chaque variable en fonction de son importance et de sa contribution potentielle à la compréhension du phénomène étudié. Ceci inclura une analyse attentive des relations entre les variables, en mettant l'accent sur celles qui présentent des corrélations significatives avec la variable cible ou qui offrent des insights pertinents pour la question de recherche.

Cette approche me permettra de sélectionner un sous-ensemble de variables qui capturent efficacement l'essentiel de la variabilité des données et qui sont les plus susceptibles d'influencer les résultats de manière significative.

En éliminant les variables redondantes ou moins importantes, je pourrai simplifier l'analyse tout en préservant la qualité et la pertinence des résultats obtenus. En fin de compte, cette méthode d'extraction de variables significatives sans recourir à l'ACP me permettra de construire un modèle précis et plus interprétable, tout en optimisant l'utilisation des ressources disponibles.

Dans un premier temps, ma stratégie consiste à filtrer la matrice de corrélation afin de ne conserver que les variables présentant une corrélation supérieure à **0,85**. Ce seuil de corrélation élevé est choisi délibérément pour identifier les relations les plus fortes entre les différentes variables de mon ensemble de données. En appliquant ce filtre, je cible spécifiquement les associations les plus robustes et les plus significatives, dans le but de réduire la complexité de l'analyse tout en préservant les informations les plus pertinentes.

Ce processus de filtration permettra de mettre en évidence les paires de variables qui sont fortement interdépendantes, ce qui peut être particulièrement utile pour identifier des motifs ou des relations importantes dans les données. En se concentrant sur les corrélations les plus élevées, je peux orienter mon analyse vers les aspects les plus cruciaux de mon ensemble de données, facilitant ainsi l'interprétation et la prise de décision subséquente.

	Variable1	Variable2	Correlation
2	radius_mean	perimeter_mean	True
3	radius_mean	area_mean	True
20	radius_mean	radius_worst	True
22	radius_mean	perimeter_worst	True
23	radius_mean	area_worst	True
51	texture_mean	texture_worst	True
63	perimeter_mean	area_mean	True
67	perimeter_mean	concave points_mean	True
80	perimeter_mean	radius_worst	True
82	perimeter_mean	perimeter_worst	True
83	perimeter_mean	area_worst	True
110	area_mean	radius_worst	True
112	area_mean	perimeter_worst	True
113	area_mean	area_worst	True
156	compactness_mean	concavity_mean	True
175	compactness_mean	compactness_worst	True
187	concavity_mean	concave points_mean	True
206	concavity_mean	concavity_worst	True
207	concavity_mean	concave points_worst	True
232	concave points_mean	perimeter_worst	True
237	concave points_mean	concave points_worst	True
312	radius_se	perimeter_se	True
313	radius_se	area_se	True
373	perimeter_se	area_se	True
622	radius_worst	perimeter_worst	True
623	radius_worst	area_worst	True
683	perimeter_worst	area_worst	True
776	compactness_worst	concavity_worst	True
807	concavity_worst	concave points_worst	True

Après j'effectue un clustering hiérarchique basé sur la matrice de corrélation pour identifier les groupes de variables qui sont fortement corrélées entre elles.:

Cluster 1:
['smoothness_se']

Cluster 2:
['symmetry_se']
Cluster 3:
['texture_se']
Cluster 4:
['texture_mean', 'texture_worst']
Cluster 5:
['symmetry_mean']
Cluster 6:
['symmetry_worst']
Cluster 7:
['smoothness_worst', 'smoothness_mean']
Cluster 8:
['fractal_dimension_mean']
Cluster 9:
['fractal_dimension_worst']
Cluster 10:
['compactness_se', 'fractal_dimension_se']
Cluster 11:
['concavity_se']
Cluster 12:
['concave points_se']
Cluster 13:
['concavity_worst', 'compactness_worst']
Cluster 14:
['concave points_mean', 'concave points_worst', 'concavity_mean', 'compactness_mean']
Cluster 15:
['perimeter_mean', 'area_worst', 'radius_worst', 'area_mean', 'perimeter_worst', 'radius_mean']
Cluster 16:
['radius_se', 'area_se', 'perimeter_se']

Cette opération de filtrage initial de la matrice de corrélation a abouti à l'identification de 16 clusters de variables fortement corrélées. Cependant, dans un souci de précision et de cohérence avec les hypothèses mises en évidence lors de l'analyse exploratoire des données (AED), je prévois de réduire davantage ce nombre de clusters. Cette étape supplémentaire de réduction sera guidée par les insights issus de l'AED, ainsi que par les hypothèses préliminaires qui ont émergé lors de l'examen initial des données.

En examinant de plus près les clusters identifiés et en tenant compte des motifs et des relations entre les variables qui les composent, je chercherai à regrouper les clusters qui semblent être redondants ou qui partagent des caractéristiques similaires. Cette consolidation des clusters permettra de simplifier la structure de données tout en préservant les informations les plus essentielles et pertinentes pour mon analyse.

*** radius_mean, perimeter_mean and area_mean : JE SELECTE area_mean EN SE BASANT SUR LA FIGURE de swarm plots area_mean ==> les données sont plus au moins mieux séparé**

*** compactness_mean, concavity_mean and concave points_mean sont bien cooreler entre eux : ==> choix : concavity_mean**

*** texture_mean and texture_worst sont bien cooreler entre eux : ==> choix texture_mean**

*** radius_se, perimeter_se et area_se sont bien cooreler entre eux : ==> choix : area_se.**

*** compactness_se, concavity_se et concave points_se sont bien cooreler entre eux : ==> choix concavity_se.**

*** radius_worst, perimeter_worst et area_worst sont bien cooreler entre eux : ==> choix : area_worst**

*** compactness_worst, concavity_worst et concave points_worst sont bien cooreler entre eux : ==> choix : concavity_worst**

*** area_worst and area_mean sont bien cooreler entre eux : ==> choix : area_mean.**

Liste à supprimer

perimeter_mean, radius_mean, compactness_mean, concave points_mean, radius_se, perimeter_se, radius_worst, perimeter_worst, compactness_worst, concave points_worst, compactness_se, concave points_se, texture_worst, area_worst

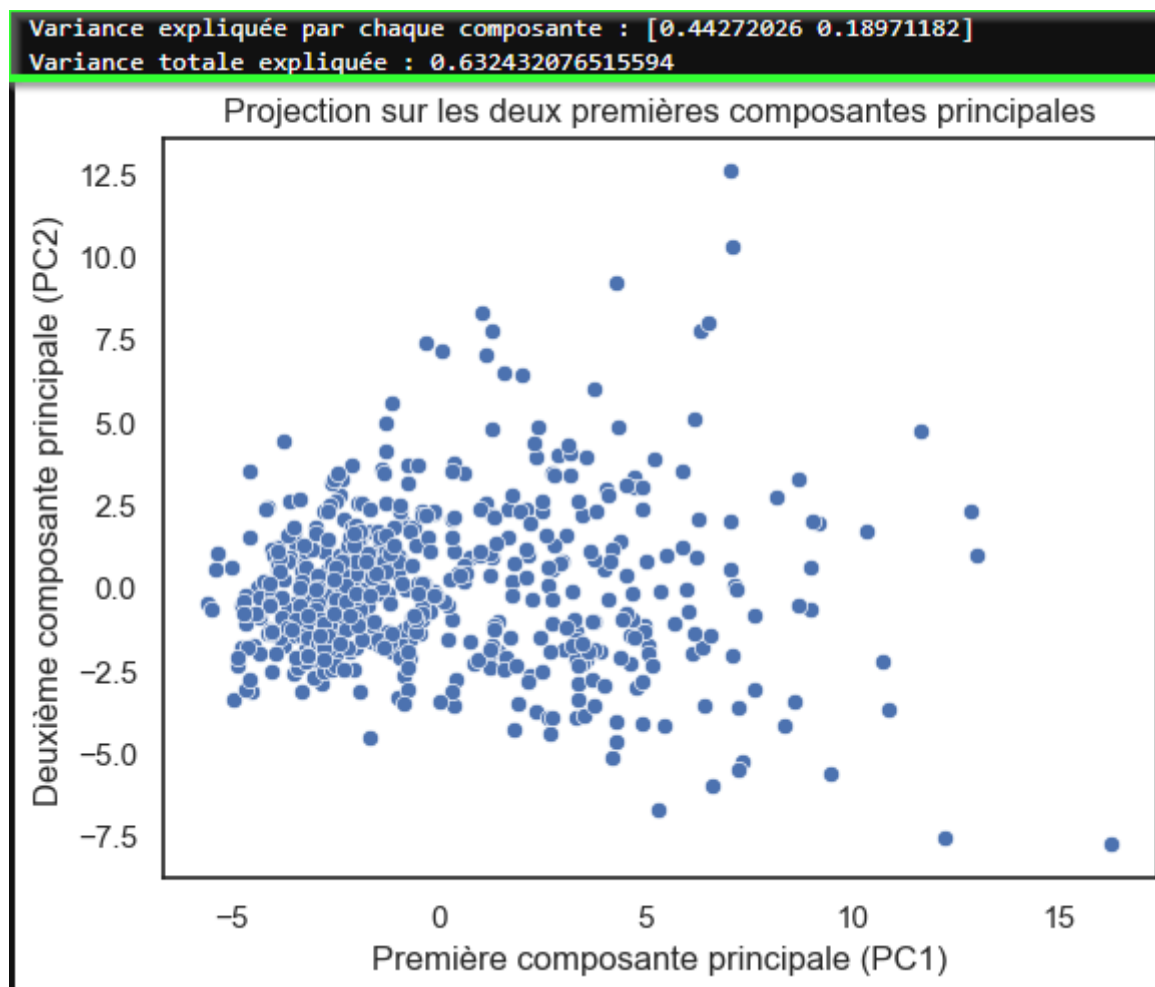
Désormais, je projette de mettre de côté le processus de réduction de variables entrepris jusqu'à présent, afin d'explorer une autre approche, à savoir l'Analyse en Composantes Principales (ACP). Cette méthode

offre une stratégie alternative pour réduire la dimensionnalité des données en transformant les variables initiales en un ensemble de composantes principales non corrélées.

L'objectif premier de cette démarche est de déterminer le nombre optimal de variables à inclure dans l'analyse, en se basant sur les composantes principales qui capturent la variance maximale des données. Une fois cette étape achevée, je prévois d'incorporer ces variables sélectionnées dans les modèles d'Intelligence Artificielle (IA) que j'entends construire, afin d'évaluer leur performance et leur efficacité par rapport au processus de réduction de variables précédemment entrepris.

Cette transition vers l'ACP me permettra de comparer les résultats obtenus par cette méthode avec ceux du test de réduction de variables initial. En examinant les différences et les similitudes entre ces deux approches, je pourrai évaluer leur pertinence et leur applicabilité dans le contexte spécifique de mon analyse de données. Cette comparaison sera essentielle pour déterminer la méthode la plus appropriée et la plus efficace pour réduire la dimensionnalité des données et améliorer les performances des modèles d'IA ultérieurs.

3-Application de L' ACP



Les deux premières composantes principales expliquent environ **63.24%** de la variance totale dans cette jeu de données.

Plus spécifiquement, la première composante principale (**PC1**) explique **44.27%** de la variance, tandis que la deuxième composante principale (**PC2**) en explique **18.97%**.

La première composante principale est la plus significative, capturant plus de la variance dans les données que la deuxième. Cela suggère que PC1 représente les caractéristiques ou les variations les plus dominantes dans vos données

Bien que 63.24% de la variance expliquée soit significative, cela signifie aussi qu'environ 36.76% de la variance n'est pas capturée par ces deux composantes.

je veux envisager d'inclure davantage de composantes pour capturer plus de variance (par exemple, jusqu'à ce que la variance totale expliquée atteigne un seuil comme 80%, 90%, ou 95%).

Pour augmenter la proportion de variance expliquée dans une ACP, je peux envisager d'ajouter plus de composantes principales.

Calcule la variance cumulée expliquée par un nombre plus élevé de composantes principales.

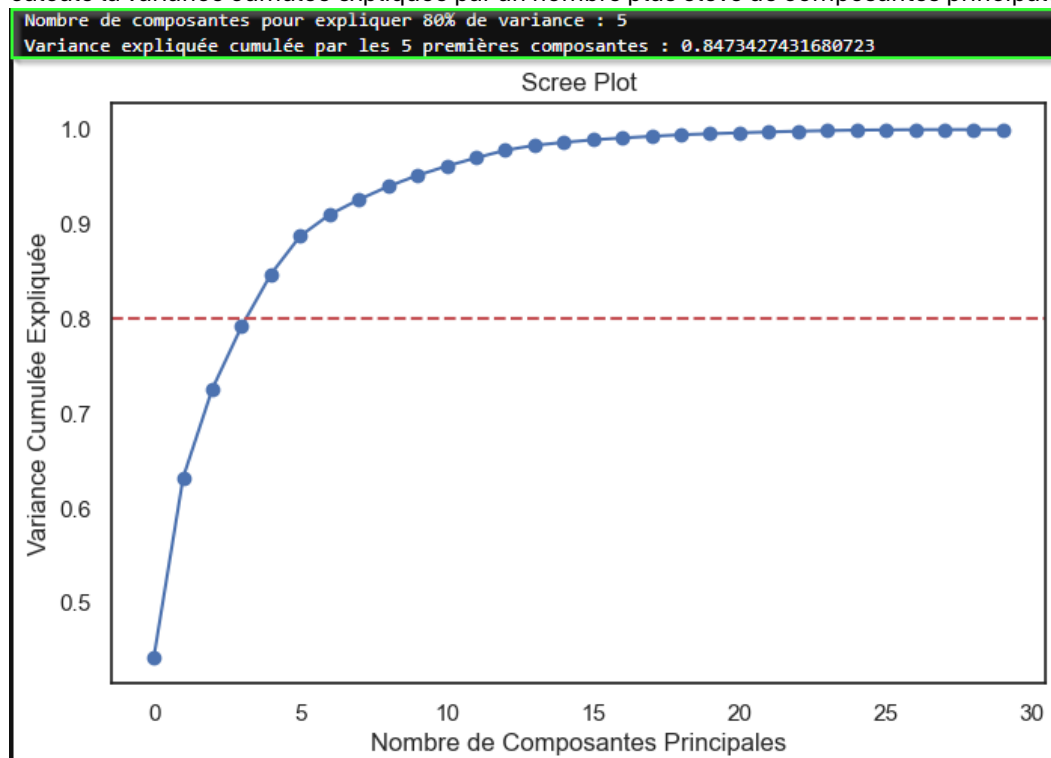
Choisir plus de composantes : Au lieu de limiter l'ACP à deux composantes principales, je peux choisir un nombre plus élevé. Généralement, on examine le "screen plot" pour déterminer le nombre de composantes à retenir. Ce graphique montre la variance expliquée en fonction du nombre de composantes et aide à identifier un "coude" qui suggère un nombre raisonnable de composantes principales à retenir.

Examiner le Scree Plot : Générer un screen plot à partir de votre ACP pour voir comment la variance expliquée s'accumule avec chaque composante principale supplémentaire.

Calcul de la Variance Cumulée : Calculer la variance cumulée expliquée par un nombre progressivement plus grand de composantes principales jusqu'à ce que vous atteigniez un niveau satisfaisant (80-90%).

Interpréter les Composantes : Assurez-vous de pouvoir interpréter les composantes supplémentaires, car elles peuvent devenir de plus en plus difficiles à interpréter avec l'augmentation de leur nombre.

calcule la variance cumulée expliquée par un nombre plus élevé de composantes principales



Selon ce graphique, 5 composantes principales sont nécessaires pour expliquer 80% de la variance des données. Cela signifie qu'avec cinq dimensions, une grande partie de l'information contenue dans les données d'origine est préservée.

Après environ 5 composantes, la courbe atteint un plateau où chaque composante supplémentaire n'ajoute qu'une petite augmentation à la variance cumulée expliquée. Cela suggère que la majorité des informations utiles est déjà capturée dans les cinq premières composantes.

Décision sur le nombre de composantes : En pratique, ce scree plot pourrait être utilisé pour décider combien de composantes principales retenir pour une analyse plus poussée. Dans ce cas, choisir 5 composantes semble être un compromis entre la simplicité du modèle (un plus petit nombre de composantes) et la préservation de l'information (un pourcentage plus élevé de variance expliquée).

Test des différentes valeurs de seuil

Nombre de composantes principales nécessaires pour atteindre :

- 80.0% de la variance expliquée: 5 composantes principales
- 90.0% de la variance expliquée: 7 composantes principales
- 95.0% de la variance expliquée: 10 composantes principales
- 99.0% de la variance expliquée: 17 composantes principales

Après avoir mené des analyses rigoureuses sur les résultats obtenus à partir des composantes principales dérivées de l'Analyse en Composantes Principales (ACP), j'ai décidé de sélectionner un nombre de composantes principales égal à **7**. Cette décision découle d'une évaluation minutieuse des résultats, où j'ai constaté que ce nombre de composantes principales était suffisant pour représenter **90%** de la variance totale des données. En optant pour ce nombre spécifique de composantes principales, je vise à atteindre un équilibre optimal entre la réduction de la dimensionnalité des données et la préservation de l'information essentielle contenue dans le jeu de données. Cette sélection de 7 composantes principales a démontré une réduction significative du nombre de variables tout en conservant une proportion considérable de l'information originale, ce qui constitue un compromis idéal pour les objectifs de mon analyse.

5. Modélisation

À cette étape, je m'apprête à initier la modélisation de mon jeu de données, une fois que j'aurai achevé les étapes préliminaires de nettoyage et de visualisation des données. Mon intention est d'appliquer ce jeu de données prétraité à plusieurs modèles d'apprentissage automatique afin d'explorer leurs performances respectives et de comparer leurs capacités prédictives.

Cette démarche implique l'utilisation de diverses techniques de modélisation, telles que les modèles linéaires, les arbres de décision, les méthodes ensemblistes.

Chaque modèle sera entraîné sur une partie des données et évalué sur une autre partie, selon des métriques de performance appropriées telles que l'exactitude, la précision, le rappel, ou encore l'aire sous la courbe ROC.

L'objectif principal de cette comparaison entre modèles est de déterminer celui qui offre les meilleures performances en termes de capacité prédictive et de généralisation des données. En utilisant une approche rigoureuse et méthodique.

Cette étape de modélisation revêt une importance cruciale dans le processus d'analyse des données, car elle permet de transformer les informations préalablement extraites en connaissances exploitables et prédictives. En réalisant une comparaison approfondie entre différents modèles, je serai en mesure de prendre des décisions éclairées quant au choix du modèle le plus approprié pour mes besoins d'analyse et de prédiction.

1-Séparation des données en ensembles d'entraînement et de test

La première étape essentielle dans le processus de modélisation des données est la séparation de l'ensemble de données en deux ensembles distincts : l'ensemble d'entraînement et l'ensemble de test. Cette division est cruciale pour évaluer objectivement les performances du modèle. L'ensemble d'entraînement est utilisé pour entraîner le modèle, c'est-à-dire pour ajuster ses paramètres aux données disponibles. En revanche, l'ensemble de test est réservé pour évaluer la capacité du modèle à généraliser sur des données qu'il n'a pas encore vues, permettant ainsi de mesurer sa performance en situation réelle. Il est impératif de veiller à ce que la séparation soit effectuée de manière aléatoire et à ce que les deux ensembles soient représentatifs de la distribution des données initiales afin d'éviter tout biais dans l'évaluation du modèle.

2-Application de modèle Random Forest

Précision du modèle RandomForest : **0.956140350877193**

Rapport de classification du modèle RandomForest :

	precision	recall	f1-score	support
B	0.97	0.96	0.96	71
M	0.93	0.95	0.94	43
accuracy	0.96	114		
macro avg	0.95	0.96	0.95	114
weighted avg	0.96	0.96	0.96	114

Précision Globale (Accuracy) :

La précision globale du modèle est de 95.61% (0.956140350877193 arrondi à 95.61). Cela signifie que le modèle a correctement prédit 95.61% des cas dans l'ensemble de test.

Mesures par Classe :

Pour la classe B (bénins) :

Précision : **97%** - Parmi toutes les prédictions pour la classe B, 97% étaient correctes.

Rappel : **96%** - Parmi tous les cas réels de la classe B, le modèle a correctement identifié 96% d'entre eux.

F1-score : **96%** - Une mesure qui combine la précision et le rappel pour donner une idée de l'efficacité générale du modèle pour cette classe.

Support : **71** - Nombre total de cas réels dans l'ensemble de test pour la classe B.

Pour la classe M (malins) :

Précision : **93%** - Parmi toutes les prédictions pour la classe M, 93% étaient correctes.

Rappel : **95%** - Parmi tous les cas réels de la classe M, le modèle a correctement identifié 95% d'entre eux.

F1-score : **94%** - Une mesure combinée de la précision et du rappel pour cette classe.

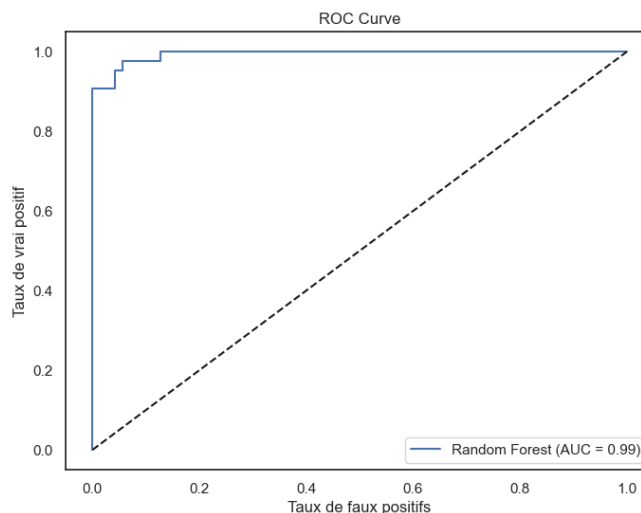
Support : **43** - Nombre total de cas réels dans l'ensemble de test pour la classe M.

3. Évaluation du Modèle

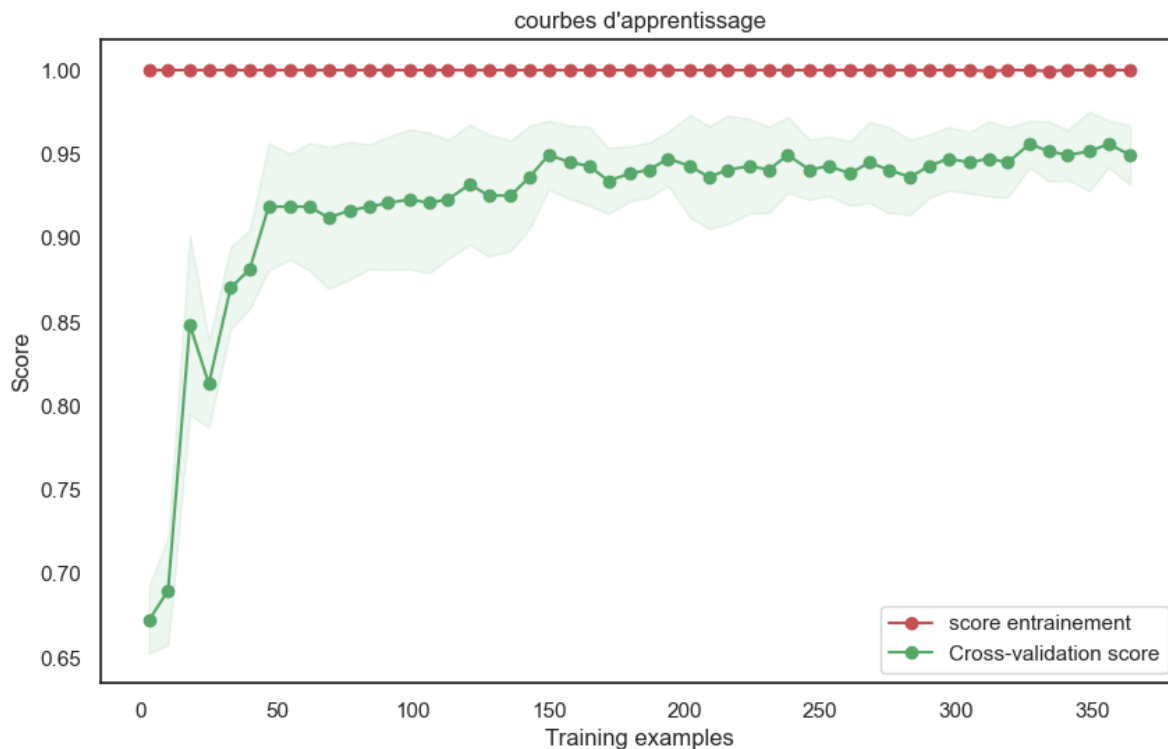
La courbe se rapproche fortement du coin supérieur gauche, ce qui indique une performance élevée du modèle.

L'AUC (Area Under the Curve) est de 0.99, ce qui est très proche de 1.

ce modèle a une excellente capacité à distinguer entre les classes positives (malin) et négatives (bénin). Une AUC de 1 représente un modèle parfait, tandis qu'une AUC de 0.5 indiquerait une performance équivalente à une sélection aléatoire. La ligne en pointillé représente la performance d'un modèle aléatoire. Une courbe ROC qui se trouve bien au-dessus de cette ligne indique que le modèle est significativement meilleur que le hasard.



4- Interprétation des Résultats



Le score (ligne rouge) commence très haut et reste constant et proche de 1 à mesure que le nombre d'exemples d'entraînement augmente. Cela suggère que le modèle est capable de bien s'adapter aux données d'entraînement à travers toute la gamme des tailles d'entraînement.

Le score de validation croisée (ligne verte) commence plus bas pour les petits ensembles d'entraînement mais augmente rapidement et se stabilise autour d'un score légèrement inférieur à celui du score d'entraînement, bien qu'il y ait toujours un écart entre les deux. Cela indique que le modèle généralise également bien, mais pas aussi parfaitement qu'il s'adapte aux données d'entraînement.

L'écart entre les scores d'entraînement et de validation suggère qu'il pourrait y avoir un léger surajustement, car le modèle performe mieux sur les données d'entraînement que sur les données non vues. Cependant, l'écart n'est pas très grand, ce qui signifie que le surajustement n'est pas sévère.

Il est également important de noter que la performance du modèle sur les données de validation est très élevée (au-dessus de 0.9), ce qui est généralement un très bon signe.

Pour adresser l'écart entre l'entraînement et la validation, on peut :

Collecter plus de données, si possible.

Réduire la complexité du modèle, en ajustant les hyperparamètres pour contrôler la profondeur de l'arbre ou en augmentant le nombre minimum d'échantillons requis pour effectuer une division dans les arbres de la forêt.

La courbe d'apprentissage semble indiquer que l'ajout de plus de données d'entraînement ne serait pas très bénéfique, car les performances sur l'ensemble de validation se sont stabilisées. Cela pourrait être un signe que le modèle a déjà appris autant qu'il le peut sur la structure du problème à partir des données fournies.

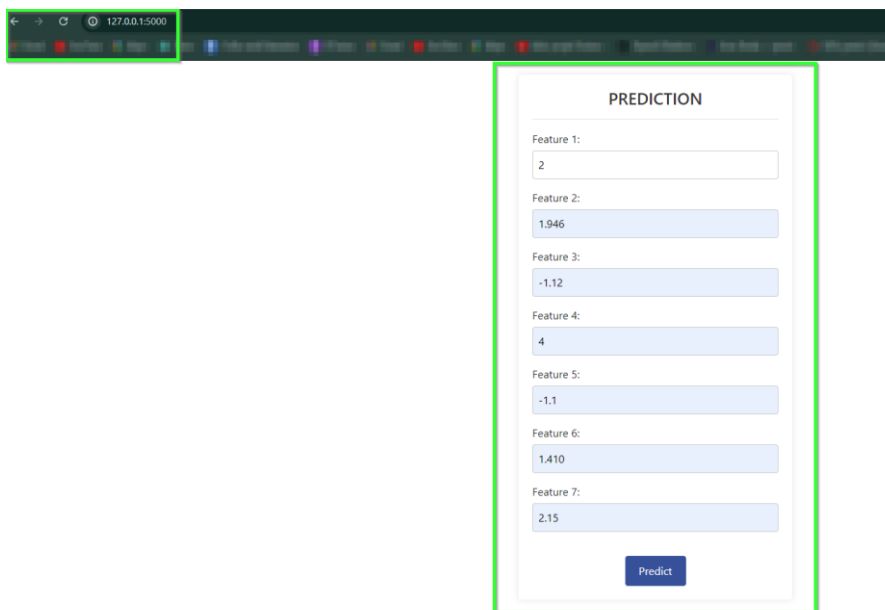
5- Déploiement du Modèle

Après l'élaboration et l'entraînement d'un modèle de machine learning utilisant les données du Wisconsin Breast Cancer Database (WDBC), j'ai procédé à l'exportation de ce modèle pour permettre la prédiction de nouvelles valeurs.

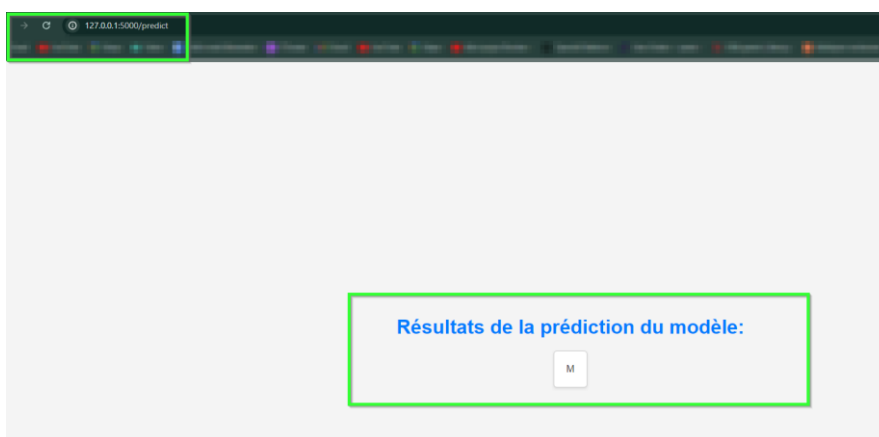
Cette étape cruciale vise à rendre le modèle aisément accessible et opérationnel pour des applications concrètes. Dans le cadre de cette démarche, un script de machine learning a été élaboré et exporté, facilitant ainsi l'interaction avec le modèle prédictif.

Pour évaluer l'efficacité et la réactivité du modèle dans un environnement opérationnel, un système composé de deux pages web a été intégré. La première page constitue une interface utilisateur intuitive, conçue pour la saisie des sept variables explicatives (ou caractéristiques) pertinentes au modèle. Cet environnement interactif est doté d'un bouton 'Prédire', qui, une fois activé, déclenche le processus de prédiction.

L'activation de ce bouton initie la transmission des données saisies vers le modèle de machine learning exporté, lequel effectue la prédiction des valeurs cibles. La seconde page web joue un rôle crucial dans ce dispositif : elle réceptionne les résultats de la prédiction et les affiche de manière claire et concise à l'utilisateur. Cette architecture permet non seulement de valider la performance du modèle dans un contexte réel mais aussi de fournir une expérience utilisateur fluide et informative, facilitant ainsi la compréhension et l'interprétation des résultats de prédiction.



The screenshot shows a web browser window with the address bar displaying '127.0.0.1:5000'. The main content area is titled 'PREDICTION' and contains seven input fields labeled 'Feature 1:' through 'Feature 7:'. The values entered in the fields are: Feature 1: 2, Feature 2: 1.946, Feature 3: -1.12, Feature 4: 4, Feature 5: -1.1, Feature 6: 1.410, and Feature 7: 2.15. A blue 'Predict' button is located at the bottom of the form.



The screenshot shows a web browser window with the address bar displaying '127.0.0.1:5000/predict'. The main content area is titled 'Résultats de la prédiction du modèle:' in blue text. Below the title is a single input field containing the letter 'M'.

6. Test des Autres Modèle

En phase finale de ce projet, j'envisage d'explorer et de tester d'autres modèles de machine learning pour élargir ma compréhension des différentes techniques de prédiction disponibles et de leurs forces respectives.

L'objectif de cette démarche est de déterminer quelle méthode serait la plus efficace et la plus pertinente pour traiter l'ensemble de données spécifique en question.

En évaluant les performances de différents modèles, tels que les réseaux de neurones, les machines à vecteurs de support, ou les modèles de régression, je compte identifier les caractéristiques qui optimisent les prédictions dans ce contexte particulier. La comparaison systématique de ces modèles, basée sur des critères tels que la précision, le rappel, l'aire sous la courbe ROC (Receiver Operating Characteristic) et d'autres métriques pertinentes, permettra de dégager une vue d'ensemble claire de leurs capacités prédictives.

1- Support Vector Machine (SVM)

Les SVM sont basées sur le concept de trouver l'hyperplan qui sépare de manière optimale les différentes classes dans l'espace des caractéristiques. L'objectif est de maximiser la marge entre les points de données les plus proches de chaque classe, connus sous le nom de vecteurs de support.

Kernel trick : Les SVM utilisent une astuce mathématique appelée le "kernel trick" qui permet de traiter des données linéairement non séparables en les projetant dans un espace de dimension supérieure où elles peuvent être séparées linéairement. Les fonctions noyau (kernel) couramment utilisées incluent le noyau linéaire, polynomial, RBF (Radial Basis Function) et sigmoid.

Le Support Vector Machine (SVM) une fois le modèle entraîné, j'ai le testé sur l'ensemble de test X_{test} et calculé la précision, obtenue à 0.9824561403508771 ou environ 98.25%.

Cette haute précision indique que le modèle SVM a excellé dans la classification des données de test, ce qui suggère une bonne adaptation du modèle aux données et une capacité élevée à généraliser à partir des données d'entraînement pour faire des prédictions précises.

Nombre total de vecteurs de support : 42

Nombre de vecteurs de support par classe : [23 19]

Validation croisée k-fold

La validation croisée k-fold divise l'ensemble des données en k sous-ensembles. Le modèle est ensuite entraîné sur k-1 sous-ensembles et testé sur le sous-ensemble restant. Ce processus est répété k fois, chaque sous-ensemble servant exactement une fois comme ensemble de test.

Cela permet d'utiliser toutes les données disponibles pour l'entraînement et le test, offrant ainsi une évaluation complète et fiable de la performance du modèle.

La validation croisée k-fold sur votre modèle SVM avec un noyau linéaire, utilisant 5 folds. La précision obtenue en validation croisée est de 0.97 (ou 97%) avec un écart-type de 0.02 (ou 2%).

Interprétation de la précision en validation croisée

Une précision moyenne de 97% indique que le modèle est très performant sur l'ensemble d'apprentissage. Cela suggère que le modèle SVM, avec la configuration linéaire spécifiée, est bien adapté aux données et capable de généraliser efficacement à partir des différentes parties de l'ensemble de données.

Faible écart-type : Un écart-type de 2% montre que la performance du modèle est stable à travers les différents folds.

Une faible variation signifie que le modèle est fiable et a une performance consistante indépendamment de la portion spécifique de données utilisée pour l'entraînement ou le test dans le processus de validation croisée.

La stabilité du modèle à travers les différents folds de la validation croisée est un bon indicateur de sa robustesse, réduisant les risques de surajustement (overfitting).

les résultats de la validation croisée indiquent que votre modèle SVM actuel est à la fois performant et stable, ce qui est idéal pour une application pratique.

La performance des différents modèles de machine learning : le Support Vector Machine (SVM) et le Random Forest (RF) peut varier en fonction des caractéristiques spécifiques des données. Dans le cas de l'ensemble des données du Wisconsin Breast Cancer Database (WDBC), je pense que les raisons pour lesquelles le modèle SVM pourrait surpasser le Random Forest :

1. Linéarité des données

SVM est particulièrement efficace pour les jeux de données où les classes peuvent être séparées linéairement ou avec une marge claire, même dans un espace de dimension supérieure (après transformation par un noyau). Si les données de WDBC présentent une séparabilité linéaire, cela pourrait expliquer pourquoi SVM fonctionne très bien.

Random Forest, en tant que modèle basé sur des arbres de décision, est excellent pour capturer des relations non linéaires et des interactions complexes entre les caractéristiques, mais il pourrait être moins efficace si la meilleure séparation des données est essentiellement linéaire.

2. Sensibilité au bruit et aux valeurs aberrantes

SVM est conçu pour maximiser la marge entre les classes, ce qui le rend moins sensible aux valeurs aberrantes ou au bruit près de la frontière de décision.

Random Forest peut parfois être affecté par le bruit et les valeurs aberrantes, surtout si elles conduisent à des divisions moins optimales dans les arbres de décision.

3. Complexité du modèle et surajustement

SVM peut être plus facile à réguler pour éviter le surajustement grâce à des paramètres comme le paramètre de régularisation C.

Random Forest a tendance à bien fonctionner avec des ensembles de données de grande taille et peut gérer le surajustement grâce à des mécanismes comme le bagging, mais dans certains cas, si le modèle est trop complexe, il peut mal généraliser sur des données inédites.

4. Taille de l'ensemble des données

Les performances de SVM peuvent être très bonnes sur des ensembles de données de taille moyenne, ce qui est souvent le cas avec WDBC.

Random Forest performe bien sur de grandes tailles d'ensemble de données mais peut être moins performant sur des ensembles plus petits ou lorsque les données sont linéairement séparables.

la performance du SVM sur les données de WDBC par rapport au Random Forest peut être attribuée à des facteurs comme la linéarité des données, la sensibilité au bruit, la gestion de la complexité du modèle et la nature de l'ensemble des données. Il est toujours bénéfique d'évaluer plusieurs modèles pour un problème donné afin de déterminer lequel est le plus adapté aux caractéristiques spécifiques des données.

2- Régression logistique

modèle de régression logistique sur vos données, avec un nombre maximal de 1000 itérations pour l'entraînement. Le modèle a ensuite été évalué sur l'ensemble de test X_test, et j'ai obtenu une précision de 0.9824561403508771, soit environ 98.25%.

Matrice de confusion :

```
[[70  1]
```

```
 [ 1 42]]
```

Vrai Positif (TP) : 42

Vrai Négatif (TN) : 70

Faux Positif (FP) : 1

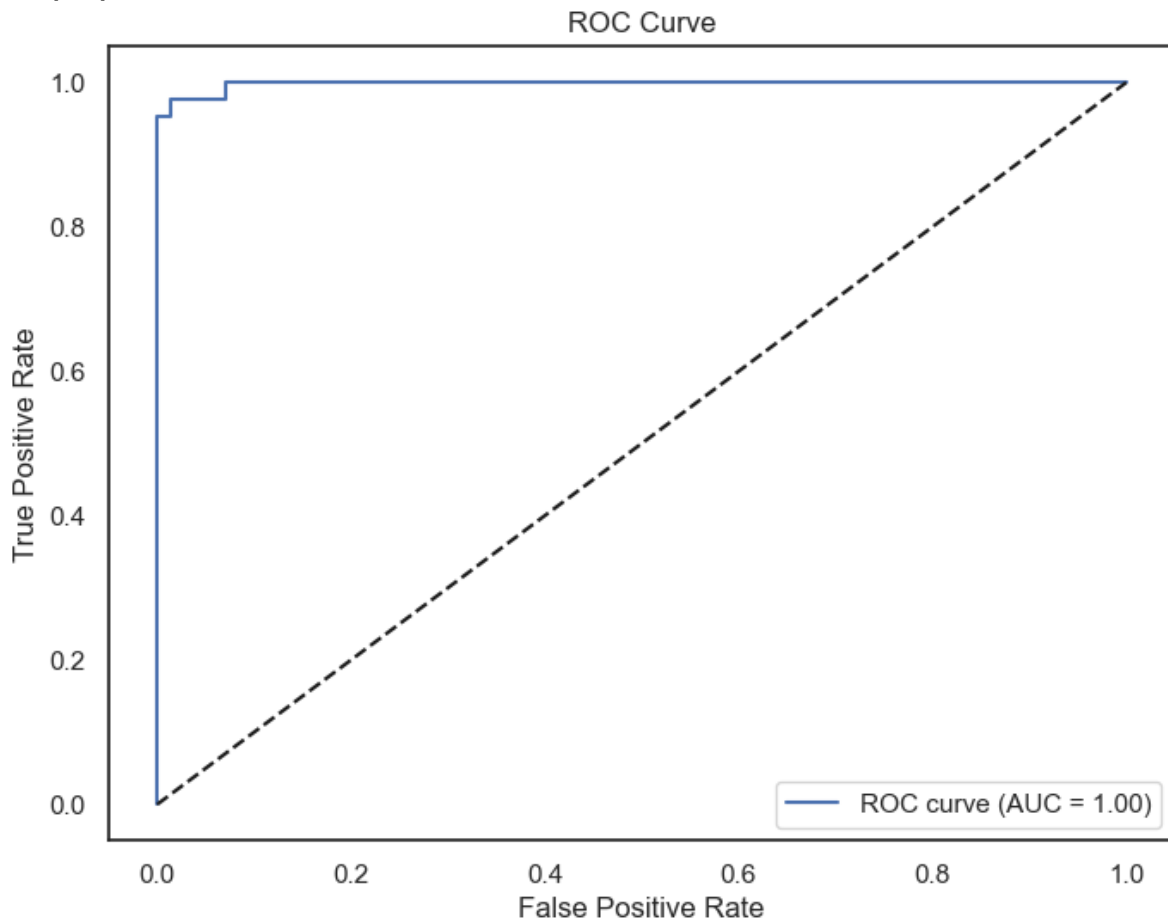
Faux Négatif (FN) : 1

Le modèle a correctement prédit la majorité des cas, avec 70 vrais négatifs et 42 vrais positifs, ce qui reflète la haute précision du modèle.

Il y a très peu d'erreurs, avec seulement 1 faux positif et 1 faux négatif, indiquant une bonne capacité à distinguer les deux classes.

La matrice de confusion confirme l'excellente performance de votre modèle de régression logistique sur l'ensemble de test. Il montre une grande efficacité à prédire correctement les cas des deux classes avec un taux très faible d'erreurs

Graphique ROC



indique une AUC (Area Under the Curve) parfaite de 1.00, ce qui signifie que votre modèle de régression logistique a une performance exceptionnelle sur les données testées, avec une capacité parfaite de distinguer entre les classes bénignes ('B') et malignes ('M').

Données très bien séparables : le modèle pourrait être sur un ensemble de données où les classes sont linéairement séparables d'une manière qui est presque parfaite.

Surajustement : Le modèle pourrait être trop bien ajusté aux données d'entraînement, au point qu'il a capturé du bruit qui n'est pas généralisable.

C'est moins probable avec la régression logistique et les données de test, mais cela reste une considération.

Taille de l'ensemble de test : si l'ensemble de test est très petit, il pourrait ne pas bien représenter la difficulté de la tâche de classification.

Examinant ces résultats avec une validation croisée sur l'ensemble des données pour s'assurer que la performance est consistante et fiable.

Validation croisée

Précision pour chaque fold:

[0.97368421 0.95614035 0.97368421 0.97368421 0.98230088]

Précision moyenne: 0.97

Écart-type de la précision: 0.01

Les résultats de la validation croisée pour le modèle de régression logistique sont très bons. J'ai obtenu des précisions élevées dans tous les 5 folds de la validation croisée, avec une précision moyenne de 0.97 (97%) et un écart-type très faible de 0.01 (1%), ce qui indique que la performance du modèle est à la fois élevée et consistante à travers les différentes sous-parties de vos données.

Que signifient ces résultats ?

Haute précision moyenne : le modèle est très précis en moyenne, ce qui est excellent pour la plupart des applications de classification.

Faible écart-type : La faible variation de la précision entre les folds signifie que la performance de ce modèle est stable et qu'il n'est pas trop dépendant d'un sous-ensemble particulier des données.

Implications :

Généralisabilité : Un modèle avec une telle précision moyenne élevée et un faible écart-type est susceptible de bien se généraliser à de nouvelles données non vues.

Robustesse : La robustesse du modèle est indiquée par la cohérence des scores de précision à travers les folds, suggérant qu'il n'y a pas de surajustement significatif aux données spécifiques de l'un des folds.

Analyse d'erreurs : Bien que les scores de précision soient élevés, il est toujours utile d'examiner les erreurs commises par le modèle pour voir s'il y a des modèles ou des types d'instances systématiquement mal classifiés.

Analyse d'erreurs :

Vrais Positifs (TP): 205

Vrais Négatifs (TN): 352

Faux Positifs (FP): 5

Faux Négatifs (FN): 7

3- Synthèse des modèles

Les résultats obtenus avec ce modèle de régression logistique sur les données du Wisconsin Breast Cancer Database (WDBC) indiquent une performance très élevée, avec une précision moyenne de 97% et une matrice de confusion indiquant un faible nombre d'erreurs (faux positifs et faux négatifs).

Pour analyser en détail les raisons pour lesquelles ce modèle s'adapte bien et comment il se compare au modèle SVM, je dois examiner plusieurs aspects :

Pourquoi la régression logistique s'adapte bien aux données WDBC :

Linéarité : La régression logistique fonctionne bien avec des données où la relation entre les caractéristiques et la probabilité logistique est linéaire, ce qui semble être le cas avec les données WDBC.

Taille de l'échantillon : Avec un grand nombre de caractéristiques, la régression logistique a tendance à bien performer, surtout si le nombre d'échantillons est suffisamment grand pour estimer les poids de manière fiable.

Équilibre de classe : Si les classes sont relativement équilibrées, comme dans le cas des données WDBC, la régression logistique n'est pas confrontée à des problèmes liés à l'imbalance des classes, ce qui est souvent un avantage par rapport à d'autres modèles.

Probabilités de sortie : La régression logistique fournit non seulement des prédictions de classe, mais aussi des probabilités, ce qui peut être utile pour des décisions cliniques où la probabilité de maladie est importante.

Avantages par rapport au modèle SVM :

Un avantage majeur de la régression logistique par rapport au SVM est son interprétabilité. Les coefficients du modèle peuvent être directement liés aux chances log (ratio), ce qui donne des aperçus explicatifs des résultats.

Efficacité : La régression logistique peut être plus efficace à entraîner et à exécuter sur de grands jeux de données, car elle ne nécessite pas de calculer les distances ou produits scalaires entre les points de données, contrairement au SVM avec des noyaux non linéaires.

Contrairement au SVM standard, la régression logistique fournit des probabilités de classe intrinsèques, qui peuvent être très utiles pour les décisions de seuillage ou dans les contextes où la probabilité est requise.

Simplicité : La régression logistique a moins d'hyperparamètres à régler (comme le terme de régularisation 'C') comparée au SVM, qui peut avoir besoin d'un réglage fin pour le paramètre de régularisation, le choix du noyau, et les paramètres spécifiques du noyau.

SVM est souvent préféré pour sa capacité à gérer les espaces de grande dimension et à trouver des marges maximales, ce qui peut être bénéfique dans les cas où les données ne sont pas linéairement séparables. Cependant, pour les données WDBC, qui peuvent être linéairement séparables ou proches de l'être, la régression logistique semble être un choix tout aussi approprié.

la régression logistique s'adapte bien aux données WDBC et offre des avantages en termes d'interprétabilité et de simplicité par rapport au SVM, surtout lorsque les données sont linéairement séparables. Cela dit, le choix entre régression logistique et SVM peut aussi dépendre de considérations pratiques telles que la taille des données, la nécessité de probabilités de classe, et les exigences en matière d'interprétabilité du modèle.

Random Forest : Ce modèle est un ensemble d'arbres de décision et a tendance à bien fonctionner sur des ensembles de données avec des interactions complexes et non linéaires entre les caractéristiques. Si Random Forest ne surpasse pas la régression logistique, cela pourrait indiquer que les relations linéaires sont suffisantes pour capturer la dynamique des données, ou que les paramètres de Random Forest n'ont pas été optimisés de manière adéquate.

Avantages de la régression logistique par rapport à Random Forest :

Performance : Si la régression logistique a une meilleure performance, cela peut être dû à une meilleure séparation linéaire des données, ce que le modèle peut exploiter efficacement.

Vitesse et simplicité : La régression logistique est généralement plus rapide à entraîner et à prédire que Random Forest, surtout lorsque le nombre de caractéristiques est grand.

Interprétabilité : Les coefficients de la régression logistique peuvent être directement interprétés en termes d'impact sur la probabilité logistique, ce qui est utile pour comprendre l'importance des différentes caractéristiques.

Prédiction de probabilités : La régression logistique fournit naturellement des probabilités de classe, tandis que Random Forest nécessite une calibration supplémentaire pour obtenir des probabilités fiables.

7- Conclusion du Projet d'Analyse des Données WDBC

Ce projet, consacré à l'analyse des données du Wisconsin Breast Cancer Database (WDBC), a traversé toutes les étapes essentielles d'une étude approfondie en science des données, depuis l'acquisition des données jusqu'au déploiement de modèles de machine learning en production. Ce parcours méthodique et structuré m'a permis d'affiner mes compétences en programmation et en machine learning tout en approchant un problème réel de manière rigoureuse et professionnelle.

Téléchargement et Nettoyage des Données

Le projet a commencé par le téléchargement des données WDBC, suivi d'un nettoyage méticuleux pour assurer la qualité et la fiabilité de l'analyse. Cette phase a inclus la gestion des valeurs manquantes, la correction des anomalies et l'uniformisation des formats, établissant ainsi une base solide pour les investigations ultérieures.

Visualisation et Exploration des Données

Des techniques de visualisation des données ont été employées pour explorer les relations entre les caractéristiques et pour comprendre la distribution des classes diagnostiques. Ces analyses exploratoires ont aidé à identifier les tendances significatives et les potentiels facteurs prédictifs, facilitant ainsi la modélisation statistique.

Réduction de Dimensionnalité et Analyse des Composantes Principales (ACP)

L'application de techniques de réduction de la dimensionalité, y compris l'ACP, a été un pivot dans l'analyse, permettant de condenser l'information contenue dans de nombreuses variables en un ensemble réduit de composantes principales. Cette transformation a non seulement clarifié les structures sous-jacentes des données mais a également optimisé les performances computationnelles des modèles de prédiction.

Application et Comparaison de Modèles de Machine Learning

Plusieurs modèles de machine learning, y compris la régression logistique, les machines à vecteurs de support (SVM) et le Random Forest, ont été appliqués et évalués. Cette diversité d'approches a permis une compréhension plus nuancée des forces et des limitations de chaque modèle face aux données WDBC, avec une analyse comparative qui a souligné la supériorité de la régression logistique dans ce contexte spécifique.

Déploiement et Accessibilité en Ligne

La réussite du déploiement des modèles via des interfaces web interactives a marqué la phase finale du projet. Ces interfaces, hébergées dans des conteneurs pour garantir leur accessibilité, ont concrétisé le projet en outils pratiques et utilisables, permettant aux utilisateurs de simuler des prédictions en temps réel et de visualiser les résultats de manière intuitive.

Impact sur le Développement Professionnel

Ce projet a été une opportunité exceptionnelle pour approfondir mes connaissances en programmation et en machine learning. En naviguant à travers les diverses phases de l'analyse des données, de la préparation et de la modélisation, jusqu'au déploiement de solutions informatiques, j'ai acquis une compréhension holistique des dynamiques du domaine et des exigences techniques et théoriques associées.

Le projet d'analyse des données WDBC s'est avéré être une entreprise exhaustive et enrichissante, renforçant non seulement ma compréhension théorique et pratique du machine learning mais aussi ma capacité à traduire les analyses complexes en solutions concrètes et accessibles. Ce voyage, de la collecte de données à la mise en production de modèles analytiques, illustre l'impact transformationnel que la science des données peut avoir dans la résolution de problèmes concrets, marquant une étape significative dans mon parcours professionnel et académique.

