Est-il possible de prédire si un patient est atteint d'une tumeur au cerveau ?

Tuân Le Minh, Justine Pouget, Ines Abbache

7 juin 2024 - Groupe B

Résumé

Dans le cadre de l'UV SY09 - Data Mining nous avons mené un travail d'analyse du jeu de données "Brain Tumor". Ce dernier provient du site Kaggle. Les tumeurs cérébrales causent environ 189 000 décès par an dans le monde [2]. Ainsi, nous avons choisi de travailler sur la thématique de détection d'une présence de tumeur au cerveau, cherchant à répondre à la question suivante : *est-il possible de détecter une tumeur cérébrale grâce aux caractéristiques obtenues par imagerie médicale du patient ?*

Notre jeu de données

Le jeu de données est composé de caractéristiques d'image d'IRM. Il comprend *5 caractéristiques de premier ordre* (annexe 1), représentant les mesures statistiques décrivant la distribution des intensités en nuances de gris des pixels d'une image sans tenir compte des relations entre les pixels adjacents, ainsi que *8 caractéristiques de second ordre* représentant les relations entre les pixels adjacents. Ces variables sont toutes quantitatives à l'exception de la variable de prédiction "Class" qui est qualitative : 0 pour absence de tumeur, 1 pour présence de tumeur. Ce *dataset* est constitué de *3762 entrées* et de *15 colonnes*.

Analyse des variables

Dans cette sous-partie, nous allons nous intéresser plus précisément à l'analyse de nos variables afin de mieux les appréhender.

Concernant les statistiques de premier ordre, il s'agit d'un histogramme mesurant la distribution des niveaux de gris dans une image.

Pour commencer, la *moyenne* représente la valeur centrale des intensités en nuances de gris des pixels, similaire à la moyenne en statistique. La *variance* mesure l'étendue de la distribution des niveaux de gris par rapport à la moyenne, tandis que l'*écart-type*, en étant sa racine carrée, indique la dispersion des valeurs.

L'*asymétrie* quantifie la dissymétrie de l'histogramme autour de la moyenne, tandis que *l'aplatissement* (*Kurtosis*) évalue la forme de l'histogramme en indiquant la concentration des valeurs autour de cette moyenne.

Les statistiques de second ordre, obtenues via une matrice de cooccurrence [1], révèlent la répartition des niveaux de gris dans les images. Ces attributs, détaillés en Annexe 1, incluent *Energy*, *ASM* (Angular Second Moment) et *Homogeneity*. Nous savons également que les trois ont des formules mathématiques très proches, car elles dépendent de l'intensité des pixels [2]. Ces métriques évaluent l'uniformité des images et peuvent fournir des informations similaires.

Analyse des données

Afin de mieux comprendre notre jeu de données, nous avons réalisé plusieurs visualisations et analyses préliminaires. Cependant, nous avons remarqué que la colonne "Coarseness" affiche des valeurs identiques pour tous les individus. Après analyse, nous avons constaté que cette variable prend des valeurs qui sont toutes très proches de zéro (*7.46e-155*), ce qui indique que la variable ne fournit aucune information pertinente. Par conséquent, nous avons décidé de supprimer cette colonne pour la suite de notre étude.

Nous avons examiné les distributions en fonction de la variable prédictive en réalisant les boxplots de la figure 1. La plupart des boxplots sont "déaxés", indiquant que chaque variable explique une quantité significative d'informations. Nous remarquons également que, dans notre dataset, les valeurs sont normalisées, car la corrélation n'est pas comprise entre -1 et 1.

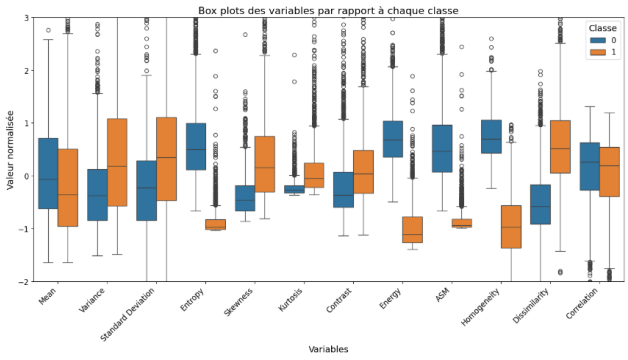


FIGURE 1 – Box plots des variables par rapport à chaque classe

Pour confirmer notre analyse, nous avons effectué un *test de Wilcoxon-Mann-Whitney*, adapté à des variables continues et à une variable prédictive qualitative correspondant à deux populations indépendantes. Les résultats montrent que les variables ont tous des p-values très faibles *** ($p<0.001$).

Les variables *Entropy*, *Energy*, *ASM*, et *Homogeneity* ont des statistiques de Kruskal-Wallis élevées et des p-valeurs très faibles, indiquant des différences marquées entre les classes 0 et 1. *Dissimilarity*, *Kurtosis* et *Skewness* présentent également des différences significatives avec des p-valeurs très faibles. *Contrast*, *Variance* et *Standard Deviation* montrent des différences significatives mais moins prononcées. Enfin, *Mean* et *Correlation* affichent des différences significatives, bien que moins marquées.

Finalement, cette première approche sur nos données nous a permis d'identifier, quelles variables influencent le fait qu'une image indique la présence d'une tumeur ou non.

1 Traitement préalable des données

Puisque nous souhaitons étudier la catégorisation d'une image présentant une tumeur ou non grâce aux autres données, nous avons donc décidé de supprimer du dataframe la colonne *"Class"*, qui prédisait déjà cette information. Comme dit précédemment, nous avons décidé de supprimer la colonne correspondant à Coarseness qui ne nous apportait pas beaucoup d'informations. Nous avons également supprimé toutes les valeurs `NaN`. Finalement, avant de débuter nos analyses, nous avons standardisé nos données en centrant chaque colonne par rapport à sa moyenne.

2 Analyse descriptive

2.1 Distribution des classes

Dans la colonne 'Class' le 0 représente la non-présence de tumeur et le 1 la présence de tumeur. Nous pouvons observer que nous avons une proportion de cas de tumeur de *55.26%* contre *44.74%*. Ainsi, les classes semblent relativement bien équilibrées.

2.2 Matrice de corrélation

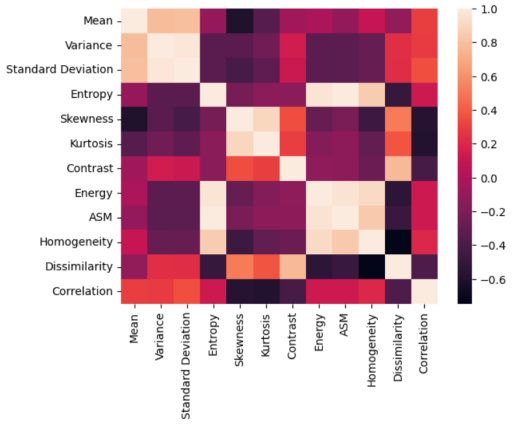


FIGURE 2 – Matrice de corrélation

Pour *Standard Deviation* et *Variance*, la corrélation entre les deux était évidemment attendue, comme le montre la figure 2. Il était également prévu de trouver des corrélations entre *Energy*, *ASM* (Angular Second Moment)

et *Homogeneity*, conformément à l'analyse des données réalisée en première partie. Globalement, nous observons que le nombre de variables corrélées dépasse celui des variables non corrélées, ce qui est avantageux pour l'*Analyse en Composantes Principales (ACP)*, puisque nous cherchons à maximiser la compression d'informations.

2.3 Analyse en Composantes Principales (ACP)

Nous avons réalisé une analyse en composantes principales (*ACP*) (figure 4) pour explorer la structure de notre jeu de données sur les caractéristiques des images pour la détection de tumeurs cérébrales. Voici un résumé des principales étapes et résultats de notre analyse :

2.3.1 Méthodologie

Nous avons extrait un total de *12 composantes principales* à partir de nos données standardisées. Ces composantes ont été ordonnées de manière décroissante, en fonction de la quantité de variance qu'elles expliquent dans les données.

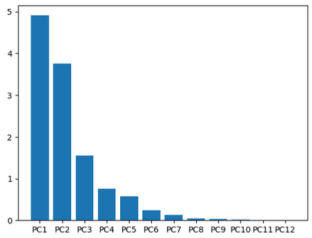


FIGURE 3 – Inerties expliquées des différentes composantes principales

2.3.2 Résultats

Comme nous pouvons le voir sur la figure 3, la première composante principale explique *40.88%* de la variance totale des données. De même, la deuxième composante principale explique *31.30%* de la variance totale.

2.3.3 Interprétation

Le cercle des corrélations en Annexe 13 révèle que, dans le premier plan factoriel *Kurtosis* et *Skewness* contribuent le plus au deuxième axe. *Standard Deviation* est plutôt anticorrélée à *Entropy*. De même, *Correlation* et *Dissimilarity* sont également assez anticorrélées. Finalement, *Homogeneity* et *Mean* sont elles corrélées.

Ces observations semblent cohérentes avec les grandeurs réelles de notre jeu de données, puisqu'une grande dispersion des valeurs de pixels conduit à une entropie élevée. Cette observation est cohérente avec les propriétés des images et des textures[2].

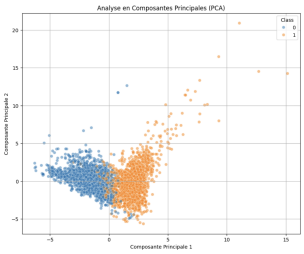


FIGURE 4 – ACP

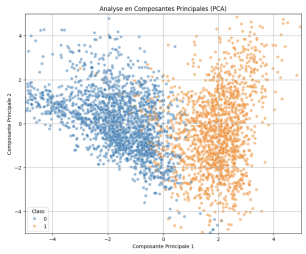


FIGURE 5 – Zoom sur la frontière

Finalement, l'*Analyse en Composantes Principales (ACP)* a été utilisée pour déterminer la méthode du coude (voir Figure 3), à partir de laquelle nous avons conclu qu'il serait efficace de réduire la dimensionnalité à deux composantes. Dans notre cas, l'ACP a réussi à réduire efficacement la dimensionnalité de nos données tout en préservant l'essentiel de leur structure. Les deux premières composantes principales capturent une part significative de la variance (*plus de 70%*), ce qui souligne leur importance pour comprendre les interactions entre les variables. Cette réduction dimensionnelle a également permis une visualisation plus claire, facilitant ainsi notre transition vers le *clustering* et la future création de *modèles d'apprentissage*. De plus, cette réduction de dimensionnalité aide à éviter le fléau de la dimension, améliorant ainsi la performance et la stabilité de nos futurs modèles.

3 Visualisation en cluster

La suite du projet a impliqué la création de clusters sur la variable indiquant la *présence ou l'absence* de tumeur, utilisant la méthode des *K-means*. Bien que les classes des données

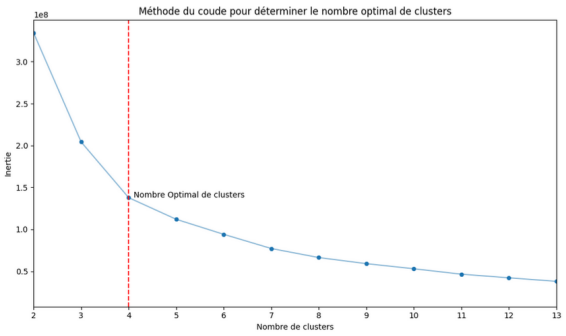


FIGURE 6 – Inertie en fonction du nombre de cluster - Méthode du coude.

soient déjà connues et que la clusterisation soit une *méthode non supervisée*, cette approche a été jugée intéressante pour la visualisation des données. Malgré l'intention initiale de former deux groupes, il a été enrichissant d'observer les résultats de la méthode du coude, qui indiquait le nombre idéal de clusters (voir Figure 6).

Les résultats semblent révéler une cassure à *4 clusters*. L'existence d'un troisième et quatrième cluster peut être interprétée comme correspondant à une zone ambiguë pour le troisième entre *présence et absence* de tumeur (voir annexe 12). Malheureusement, la représentation du quatrième cluster est très difficilement interprétable. Cette observation souligne les limites de cette méthode lorsque la distribution des données n'est pas nettement définie.

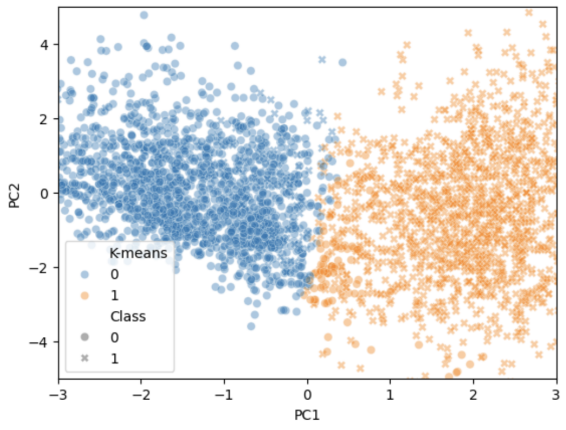


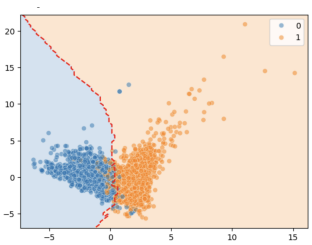
FIGURE 7 – Zoom sur les 2 clusters

Finalement, le choix de deux clusters s'est avéré être le plus approprié pour notre étude, bien qu'il ait été enrichissant d'explorer d'autres possibilités. Nous avons validé ce choix à l'aide de *l'indice de Rand*, qui a donné un taux de satisfaction de *81.674%*, un résultat très satisfaisant. Il serait donc intéressant de comparer cette méthode non supervisée avec des méthodes supervisées pour évaluer leurs performances respectives dans des contextes similaires.

4 Création de modèles

Dans cette partie, nous allons mettre en œuvre divers modèles pour aborder notre question initiale : *est-il possible de détecter une tumeur cérébrale grâce aux caractéristiques obtenues par imagerie médicale du patient ?* Nous explorerons plusieurs modèles pour identifier celui qui convient le mieux à nos données et qui offre les meilleures performances. Nous avons pris soin de séparer les données en ensembles d'entraînement et de test, et avons également utilisé la technique de *validation croisée* pour renforcer la fiabilité de nos résultats.

4.1 K-plus proches voisins



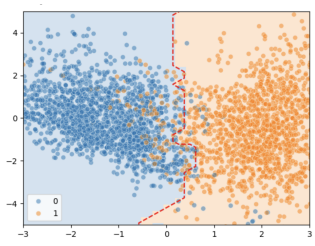


FIGURE 8 – KNN FIGURE 9 – KNN-Zoom

Nous avons développé un modèle en utilisant l'algorithme des k-plus proches voisins (KNN). Pour optimiser ce modèle, nous avons employé l'optimisation par grille via `GridSearchCV` de *scikit-learn*, qui a suggéré que le meilleur choix pour le nombre de voisins était $k = 25$. La performance du modèle a été évaluée par une validation croisée à *5 plis*, résultant en un taux de réussite moyen de *95,53%*, ce qui est très satisfaisant. Les détails de la performance du modèle sont illustrés dans la Figure 8 et une vue détaillée est présentée dans la Figure 9.

4.2 Régression Logistique

Nous avons choisi d'utiliser une régression logistique pour notre problématique, car notre

variable cible est binaire (présence de tumeur ou non).

Cependant, la forte corrélation entre nos variables pose problème pour l'application de la régression logistique, ce qui se traduit par des coefficients très petits ou très grands. Pour pallier cela, nous avons utilisé les résultats de l'ACP, qui a pour but de supprimer les corrélations entre variables. Nous acceptons ainsi une certaine perte d'information en faveur de la stabilité du modèle.

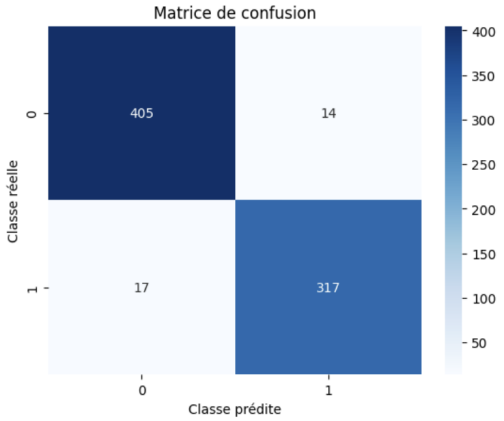


FIGURE 10 – Matrice de confusion - Régression Logistique

La validation croisée et la cohérence des coefficients ont confirmé la pertinence de notre modèle. Notre validation croisée à 5 plis nous donne une précision sur nos données de tests d'environ *98,83%*, ce qui signifie qu'il prédit très bien la présence ou l'absence de tumeur avec une faible imprécision. La matrice de confusion (figure 10) montre également qu'il y a très peu de faux positifs ou de faux négatifs.

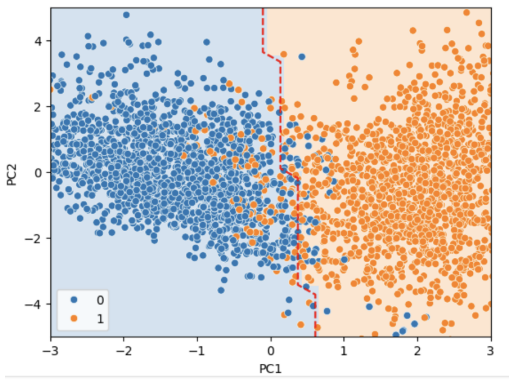


FIGURE 11 – Zoom sur la frontière de décision

5 Conclusion

Finalement, notre travail d'analyse et de création nous a permis d'explorer différents champs pour répondre à notre problématique initiale de détection de tumeurs à l'aide des caractéristiques présentes dans des images médicales. Nous avons démontré l'efficacité de plusieurs méthodes, notamment les KNN, les arbres binaires et la régression logistique, en utilisant des techniques comme l'ACP pour améliorer la stabilité et la performance de nos modèles. La régression logistique reste le meilleur modèle en termes de performances. Ces résultats encourageants suggèrent des perspectives intéressantes pour les futures recherches et développements dans le domaine de la détection de tumeurs.

Le fait que notre jeu de données soit essentiellement composé de variables quantitatives rend l'interprétation des variables plus difficile.

Il est cependant important de noter qu'aucune fonction de perte asymétrique n'est définie. Or, il serait possible de considérer que prédire une non-tumeur alors que l'individu est atteint d'une tumeur est plus grave que de prédire l'opposé. Formuler une fonction de perte appropriée au contexte d'utilisation serait pertinent pour un modèle plus juste.

Références

[1] JuliaImage. (2023). Matrice de cooccurrence des niveaux de gris. `https://juliaimages.org/ImageFeatures.jl/stable/tutorials/glcm/#Mean`

[2] Haralick, R. M., Shanmugam, K., & Dinstein, I. (1973). Textural Features for Image Classification. IEEE Transactions on Systems, Man, and Cybernetics, SMC-3(6), 610–621. `https://doi.org/10.1109/TSMC.1973.4309314`

[3] Archives of Public Health (2024). Burden and trends of brain and central nervous system cancer from 1990 to 2019 at the global, regional, and country levels. `https://archpublichealth.biomedcentral.com/articles/10.1186/s13690-022-00965-5`

Annexes 1

Nom	*Description*	*Type de feature*
Moyenne (Mean)	Moyenne des intensités des pixels.	1er ordre
Variance (Standard Deviation)	Dispersion autour de la moyenne.	1er ordre
Écart-Type (Standard Deviation)	Racine carrée de la variance.	1er ordre
Asymétrie (Skewness)	Asymétrie autour de la moyenne.	1er ordre
Aplatissement (Kurtosis)	"Pointicité" de la distribution des pixels	1er ordre
Contraste (Contrast)	Variations texture et couleur.	2ème ordre
Énergie (Energy)	Mesure l'uniformité de l'image	2ème ordre
ASM (Angular Second Moment)	Uniformité des pixels.	2ème ordre
Entropie (Entropy)	Complexité, désordre.	2ème ordre
Homogénéité (Homogeneity)	Proximité des valeurs.	2ème ordre
Dissimilarité (Dissimilarity)	Différences pixels voisins.	2ème ordre
Corrélation (Correlation)	Corrélation des pixels.	2ème ordre
Rugosité (Coarseness)	Rugosité de la texture.	2ème ordre
Class	1 = Tumor, 0 = Non-Tumor	Variable prédictive

TABLE 1 – Description des colonnes du jeu de donnée

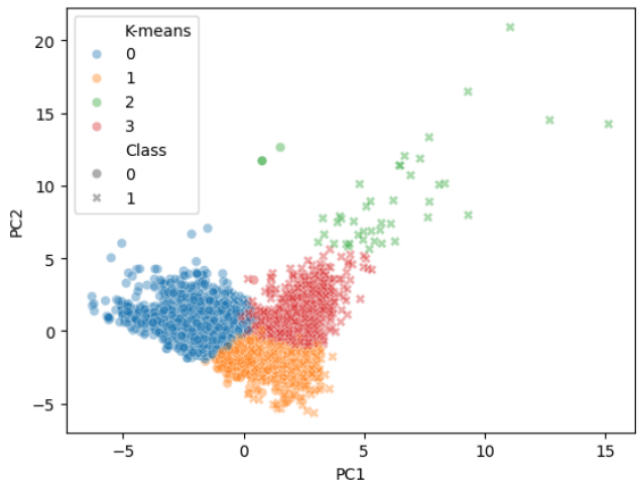


FIGURE 12 – K-means avec 4 clusters

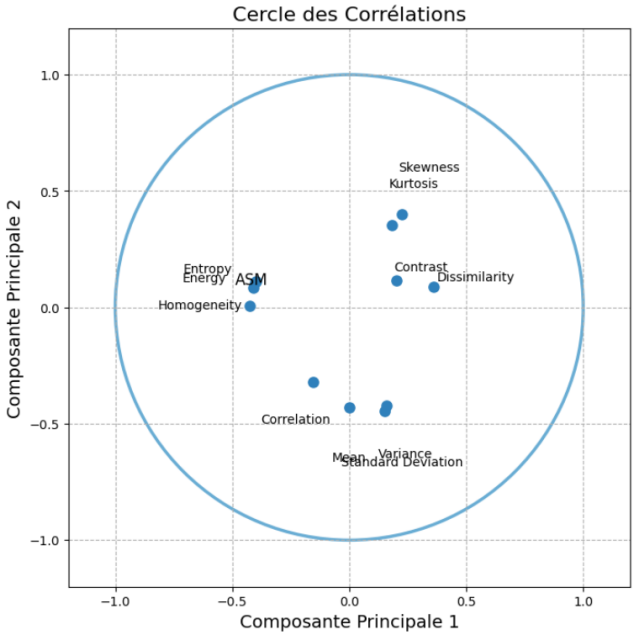


FIGURE 13 – Cercle de corrélation - 1er plan factoriel