2023-2024



Analyse de Tumeur Cérébrale par Apprentissage Automatique

- Projet Machine Learning -

Professeur :

M. Haja Zakaria

Participant du Groupe :

Ouerdi Ilyas

Es-sahli Soufiane

1

Structure de rapport

1. Introduction ………………………………………………………………………. 3

* + Contexte Médical
* Les Tumeurs Cérébrales
* Diagnostic des Tumeurs

2. Apprentissage Automatique ………………………………………………. 4

* + Définition de l'Apprentissage Automatique

3. Collecte des Données …………………………………………………………. 4

* + Source des Données

4. Prétraitement des Données………………………………………………… 5

* + Le Traitement des Données avant leur Utilisation

5. Entraînement………………………………………………………………….….. 6

* + Méthodologie d'Entraînement

6. Évaluation des Modèles………………………………………….…………… 8

* + Critères d'Évaluation
  + Résultats et Analyses

7. Déploiement et Usage Clinique…………………………………………. 12

* Intégration dans la Pratique Clinique
* Considérations Éthiques

8. Conclusion…………………………………..……………………………………. 12

* Résumer

2

1.Introduction

Les Tumeurs Cérébrales :

Les tumeurs cérébrales sont des masses anormales de cellules qui se développent dans le cerveau. Elles peuvent être bénignes (non cancéreuses) ou malignes (cancéreuses) et peuvent se former à partir de divers types de cellules cérébrales, notamment les cellules gliales et les neurones. Les tumeurs cérébrales peuvent entraîner divers symptômes en fonction de leur taille, de leur emplacement et de leur agressivité, tels que des maux de tête, des convulsions, des changements de comportement et des problèmes de coordination. Le traitement des tumeurs cérébrales peut impliquer une combinaison de chirurgie, de radiothérapie et de chimiothérapie, en fonction du type et de la localisation de la tumeur, ainsi que de l'état de santé général du patient.

Diagnostic des Tumeurs :

Le diagnostic des tumeurs cérébrales peut être réalisé en utilisant diverses caractéristiques extraites des images médicales, telles que la moyenne 'Mean', la variance 'Variance', l'écart type 'Standard Deviation', l'entropie 'Entropy', l’asymétrie 'Skeness', la kurtosis 'Kurtosis', le contraste 'Contrast', l'énergie 'Energy', l'ASM 'Angular Second Moment', l'homogénéité 'Homogeneity', la dissimilarité 'Dissimilarity', la corrélation 'Correlation' et la grossièreté 'Coarseness'.

3

2.Apprentissage Automatique

Définition de l'Apprentissage Automatique :

L’apprentissage automatique, encore appelé Machine Learning en anglais, est un sous-domaine de l’intelligence artificielle qui consiste en un ensemble de méthodes permettant aux ordinateurs d’apprendre à partir de données sans être explicitement programmée. Contrairement aux ordinateurs traditionnels qui suivent les instructions spécifiques qui leur sont programmées, les systèmes créés à partir de l’apprentissage automatique utilisent des algorithmes pour analyser des données, détecter des modèles et en tirer des conclusions ; ces conclusions sont utilisées pour prendre des décisions ou faire des prévisions.

3.Collecte des Données

Source des Données :

La source de données “Brain Tumor” réalisée par ”Jakesh Bohaju” sur Kaggle est composée d’une base de données vaste des images IRM de cerveaux où chaque image confirme la présence d’une tumeur, avec des attributs associés. Plusieurs types d’attributs sont extraits des images IRM de cerveaux existants. Même si ces attributs varient légèrement, ils traitent les informations de manière similaire. Tous les hymnes de plusieurs variétés signifient la moyenne 'Mean', la variance 'Variance', l'écart type 'Standard Deviation', l'entropie 'Entropy', l’asymétrie 'Skeness', la kurtosis 'Kurtosis', le contraste 'Contrast', l'énergie 'Energy', l'ASM 'Angular Second Moment', l'homogénéité 'Homogeneity', la dissimilarité 'Dissimilarity', la corrélation 'Correlation' et la grossièreté 'Coarseness'.

Lien : <https://www.kaggle.com/datasets/jakeshbohaju/brain-tumor>

4

4.Prétraitement des Données

Le Traitement des Données avant leur Utilisation :

La fonction 'df.isnull().sum()' est un outil précieux en analyse de données et en prétraitement des données dans le langage de programmation Python, en particulier lorsqu'on travaille avec la bibliothèque pandas. Cette expression permet de vérifier rapidement et efficacement la présence de valeurs nulles dans un DataFrame, L'output de cette fonction fournit un résumé concis du nombre de valeurs nulles dans chaque colonne du DataFrame.

Output :

Image 0

Class 0

Mean 0

Variance 0

Standard Deviation 0

Entropy 0

Skewness 0

Kurtosis 0

Contrast 0

Energy 0

ASM 0

Homogeneity 0

Dissimilarity 0

Correlation 0

Coarseness 0

dtype: int64

L'instruction 'dataset = df.drop('Image', axis=1) ' est une opération courante en prétraitement des données lorsque l'on travaille avec des DataFrames dans Python. Cette ligne de code vise à supprimer une colonne spécifique, en l'occurrence la colonne nommée "Image", du DataFrame df. Ce processus de suppression des colonnes non pertinentes ou redondantes est une étape importante du prétraitement des données, car il simplifie le DataFrame en se

5

concentrant sur les variables les plus pertinentes pour l'analyse ou là

modélisation ultérieure. Cela peut également réduire la complexité des données, améliorer les performances des modèles, et rendre l'analyse plus facile à interpréter.

Output :

Index(['Class', 'Mean', 'Variance', 'Standard Deviation', 'Entropy',

'Skewness', 'Kurtosis', 'Contrast', 'Energy', 'ASM', 'Homogeneity',

'Dissimilarity', 'Correlation', 'Coarseness'],

dtype='object')

5.Entraînement

Régression logistique :

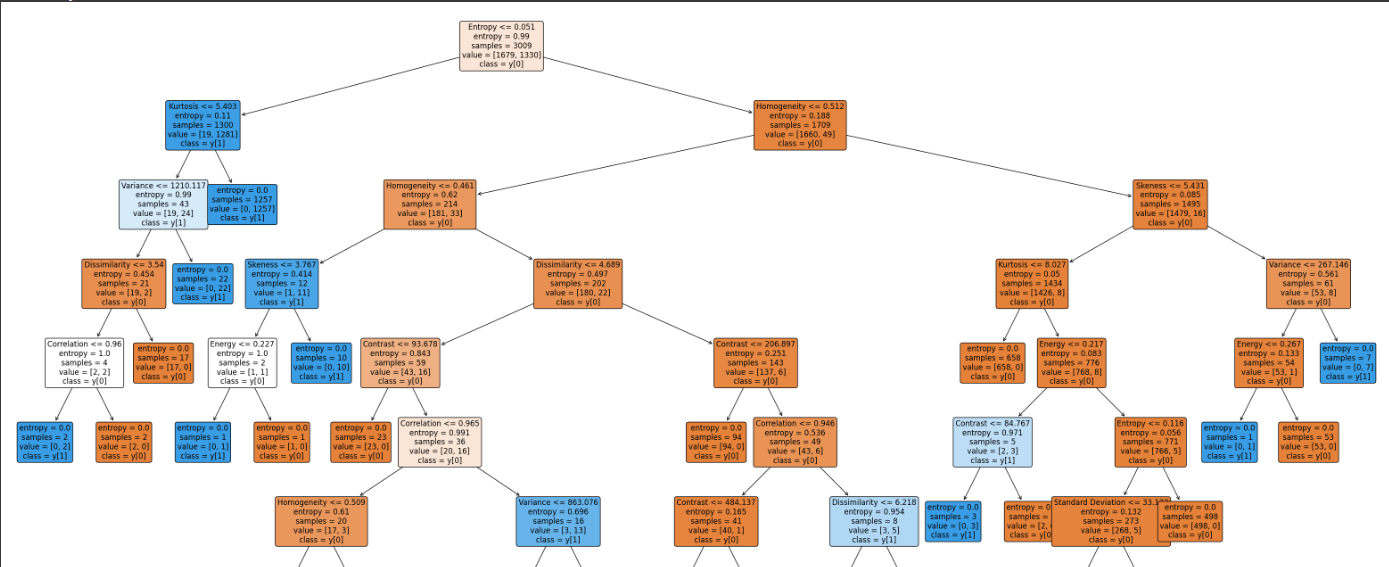
Notre modèle de régression logistique a été construit en utilisant un ensemble d'entraînement constitué de caractéristiques extraites à partir de scans cérébraux, représentées dans un DataFrame dataset. Les caractéristiques, à l'exception de la variable cible "Class", ont été utilisées pour entraîner le modèle à prédire si une observation donnée correspond à une tumeur cérébrale ou non. L'ensemble d'entraînement a été divisé en sous-ensembles d'apprentissage (X\_train, y\_train) et de test (X\_test, y\_test) à l'aide de la fonction train\_test\_split, avec 80% des données attribuées à l'entraînement et 20% aux tests, garantissant ainsi une évaluation impartiale des performances du modèle. Le modèle a été construit en utilisant la classe LogisticRegression de la bibliothèque scikit-learn, une méthode efficace pour la classification binaire. En ajustant le modèle aux données d'entraînement à l'aide de la méthode fit(), le modèle a appris à identifier les schémas complexes dans les caractéristiques qui sont indicatives de la présence d'une tumeur cérébrale. Les prédictions du modèle sur l'ensemble de test (y\_pred2) ont été

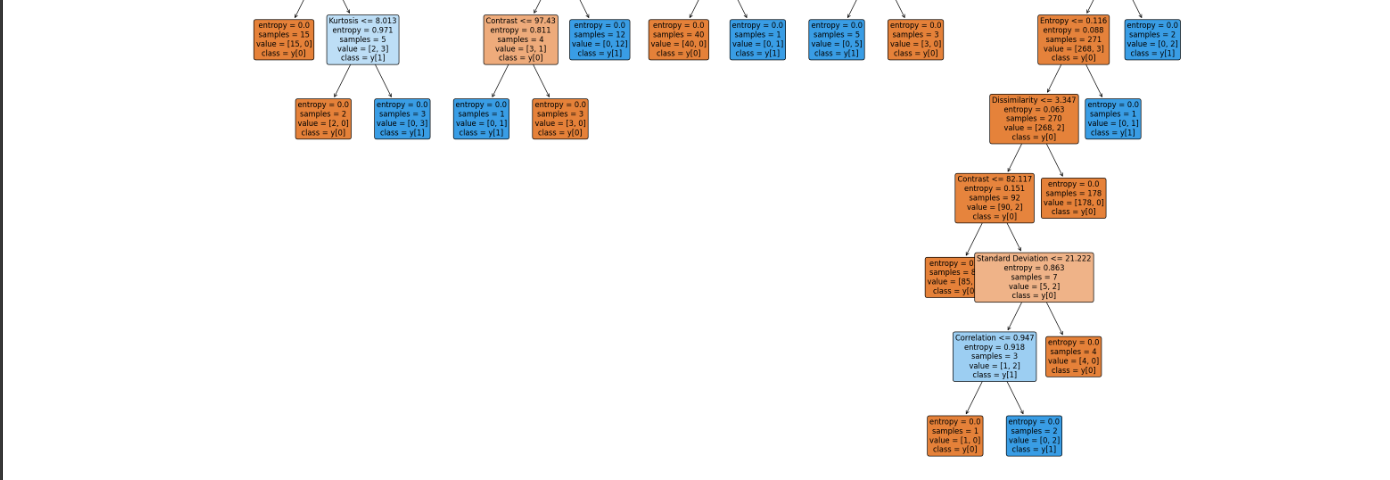
6

prédire avec précision la probabilité qu'une observation appartienne à l'une des deux classes possibles, facilitant ainsi le diagnostic des tumeurs cérébrales et contribuant à des décisions médicales éclairées.

Classificateur d'arbres :

Un classificateur d'arbres de décision est initialisé, configuré pour diviser les nœuds en fonction du critère d'entropie. Ce classificateur est ensuite entraîné à l'aide des données d'entraînement, apprenant à faire des prédictions en fonction des caractéristiques fournies et de leurs étiquettes cibles associées.





La taille imposante de l'arbre de décision résulte de la complexité des données utilisées par le classificateur. En effet, la largeur des colonnes ou des caractéristiques utilisées pour former l'arbre influence directement sa structure. 7

Lorsque les données présentent une grande diversité ou une multitude de caractéristiques, l'arbre de décision tend à devenir plus étendu pour capturer les nuances subtiles de ces données.

6. Évaluation des Modèles

Critères d'Évaluation :

Pour évaluation de modèle de régression logistique, on a utilisé plusieurs critères :

* Accuracy : on a calculé à l'aide de accuracy\_score(y\_test, y\_pred2). Cela mesure la proportion de prédictions correctes parmi toutes les prédictions.
* Classification Report : on a utilisé classification\_report(y\_test, y\_pred2) pour obtenir un rapport détaillé sur la précision
* Matrice de Confusion : Matrice de Confusion : Vous avez déjà calculé la matrice de confusion avec metrics.confusion\_matrix(y\_test, y\_pred2). Cela montre le nombre de prédictions correctes et incorrectes pour chaque classe, offrant ainsi une vue détaillée des performances du modèle.

Pour évaluation de modèle Classificateur d’arbres, on a utilisé :

* Scores de Validation Croisée : on a calculé les scores de validation croisée avec cross\_val\_score. On a les utiliser pour évaluer la performance moyenne du modèle sur plusieurs plis de données (5)

Résultats et Analyses :

Pour évaluation de modèle de régression logistique :

* Accuracy :

OUTPUT : 8

Accuracy: 0.9402390438247012

* Classification Report :
* precision recall f1-score support
* 0 0.91 0.98 0.95 400
* 1 0.98 0.89 0.93 353
* accuracy 0.94 753
* macro avg 0.95 0.94 0.94 753

weighted avg 0.94 0.94 0.94 753

Précision "Precision": Pour la classe 0 "non malade", une précision de 0.91 signifie que 91% des personnes prédites comme non malades par le modèle le sont réellement. Pour la classe 1 "malade", une précision de 0.98 signifie que 98% des personnes prédites comme malades par le modèle le sont réellement.

Rappel "Recall" : Pour la classe 0, un rappel de 0.98 signifie que le modèle a correctement identifié 98% de toutes les personnes non malades. Pour la classe 1, un rappel de 0.89 signifie que le modèle a correctement identifié 89% de toutes les personnes malades.

F1-score : Pour la classe 0, un F1-score de 0.95 indique une combinaison équilibrée de précision et de rappel pour les personnes non malades. Pour la classe 1, un F1-score de 0.93 indique une combinaison équilibrée de précision et de rappel pour les personnes malades.

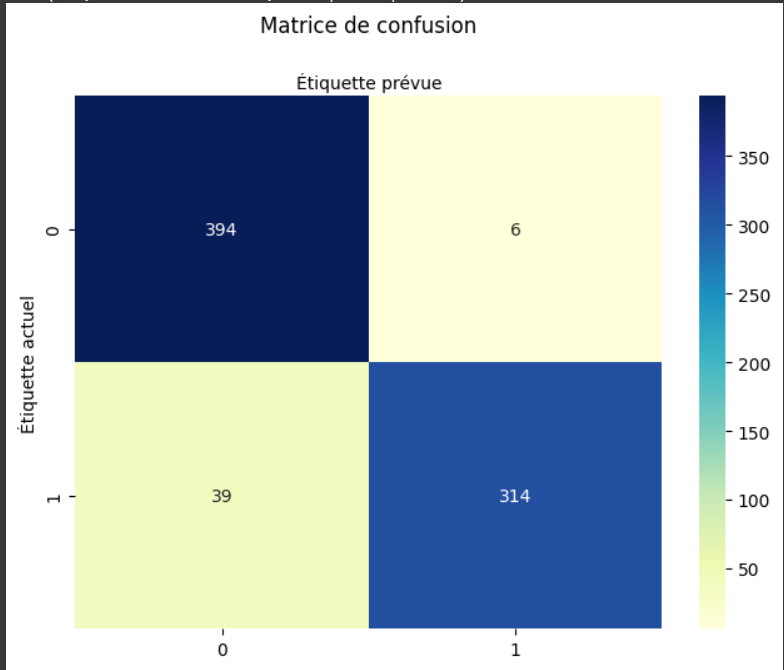
Support : Il s'agit du nombre d'occurrences de chaque classe dans l'ensemble de test, c'est-à-dire le nombre total de personnes non malades et de personnes malades dans vos données de test.

Pour 753 personnes, se trouve 400 personnes qui sont classées comme non malades et 353 personnes qui sont classées comme malades 9

Accuracy (Précision globale) : C'est la proportion de prédictions correctes parmi toutes les prédictions effectuées par le modèle. Dans ce cas, l'accuracy est de 0.94, ce qui signifie que le modèle a correctement prédit 94% de toutes les personnes (malades et non malades) dans l'ensemble de test.

Accuracy: 0.9402390438247012

* Matrice de Confusion :



True Positives (TP) | False Positives (FP)

--------------------------------------------------------

False Negatives (FN) | True Negatives (TN) 10

True Positive (Vrai Positif) : C'est le nombre de cas où le modèle a correctement prédit que les personnes étaient malades, et qu'elles étaient effectivement malades selon les données réelles.

False Positive (Faux Positif) : C'est le nombre de cas où le modèle a incorrectement prédit que les personnes étaient malades, alors qu'elles ne l'étaient pas en réalité.

False Negative (Faux Négatif) : C'est le nombre de cas où le modèle a incorrectement prédit que les personnes n'étaient pas malades, alors qu'elles l'étaient en réalité.

True Negative (Vrai Négatif) : C'est le nombre de cas où le modèle a correctement prédit que les personnes n'étaient pas malades, et qu'elles ne l'étaient effectivement pas selon les données réelles.

Note : La matrice de confusion est limitée par un nombre maximal de valeurs.

Pour évaluation de modèle Classificateur d’arbres :

* Scores de Validation Croisée : La validation croisée fournit un score fixe en raison de l'évaluation multiple sur les données. Plus le nombre de plis est élevé, plus l'évaluation est précise. Pour notre cas on a utilisé 5 plis pour l’évaluation.

OUTPUT :

Evaluation des cinq plis et la moyenne :

Scores de validation croisée: [0.98671096 0.99003322 0.98172757 0.9833887 0.98336106]

Moyenne des scores de validation croisée: 0.9850443060016252

11

7. Déploiement et Usage Clinique

Intégration dans la Pratique Clinique :

Ces modèles, une fois entraînés sur un ensemble de données pertinents et validés, peuvent être déployés dans des applications cliniques pour aider à la prise de décision et à la gestion des patients.

Le modèle de régression logistique pourrait être utilisé pour prédire le risque de développer certaines conditions médicales en fonction des caractéristiques individuelles du patient, le modèle d'arbre de décision pourrait être utilisé pour identifier les facteurs les plus importants contribuant à un diagnostic particulier, pourtant que on doit assurer la mise a jour des données utiliser pour l’entrainement des modelés.

Considérations Éthiques :

Cette utilisation de ces modèles peut influencer les résultats des patients. C’est pourquoi il est essentiel d’assurer le développement et la mise à jour de ces modèles pour obtenir des résultats plus précis et plus transparents.

8. Conclusion

Résumer :

Pour ce projet, nous avons utilisé une base de données tabulaire contenant des informations sur les patients. C'est pourquoi nous avons réalisé un prétraitement des données avant de les utiliser. Après la validation des données, nous avons construit deux modèles : une régression logistique et un arbre de décision. Bien qu'il ait été suggéré d'utiliser l'un ou l'autre des deux modèles, nous avons décidé de les mettre en œuvre tous les deux afin d'analyser les évaluations de différents types et les résultats obtenus.

12

En conclusion, ce projet a été une expérience d’application des deux

Modèles de prédiction Nous sommes impatients de poursuivre notre apprentissage et de l’améliorer par la découverte d’autre modèles comme le Réseaux de neurones convolutifs (CNN) et d'appliquer ces connaissances pour contribuer à l'amélioration des soins de santé.

Cordialement,

Ouerdi Ilyas, Es-sahli Soufiane

13