Le protocole expérimental pour la création d'un pipeline de sous-typage de l'AVC ischémique impliquait une approche globale utilisant plusieurs algorithmes de clustering, méthodes de sélection de caractéristiques et métriques d'évaluation. Les composantes clés du protocole sont les suivantes :

Ensemble de données et prétraitement

L'étude a utilisé un ensemble de données de 121 patients atteints d'AVC ischémique suivis longitudinalement sur 60 mois. La collecte des données s'est déroulée à quatre moments : initial (J7), court terme (M6) et long terme (M36, M60). La cohorte STROKDEM* utilisée contient un total de 1 340 caractéristiques par patient.

Les caractéristiques incluaient :

- Informations démographiques (âge, niveau d'éducation)
- Mesures cliniques
- Scores cognitifs (MMSE et MoCA)
- Données IRM à 72 heures et M6 post-AVC

Les étapes de prétraitement comprenaient :

- Gestion des valeurs manquantes
- Standardisation numérique
- Stratégies d'imputation multiple (MICE, KNN, médiane, remplissage vers l'avant)

Nous avons défini plusieurs ensembles de caractéristiques à utiliser dans l'analyse de sous-typage de l'AVC :

Caractéristiques catégorielles

Une liste de 18 caractéristiques catégorielles a été définie, incluant :

- Informations démographiques : 'Genre'
- Conditions médicales : 'Patho Coro', 'Insuf Card', 'Arteriopathie', etc.
- Facteurs de risque : 'Frisquvasc Hta', 'Frisquvasc Diab', 'Frisquvasc Hypchol', etc.
- Facteurs de style de vie : 'Facteur Alcool', 'Act Phys'

Caractéristiques cliniques et démographiques

Neuf ensembles de caractéristiques différents ont été définis, chacun contenant une combinaison des variables suivantes :

- Démographiques : 'Age', 'Nb An Scol' (Années de scolarité)
- Mesures cliniques : 'Iq Code J0', 'Poids', 'Taille'
- Scores cognitifs : 'Mmse J7', 'Moca J7'
- Sévérité de l'AVC : 'Nihss Score J7'

Ces ensembles varient dans leur composition, certains n'incluant que des mesures démographiques et cliniques de base, tandis que d'autres incorporent des scores cognitifs et la sévérité de l'AVC.

Ensemble de caractéristiques avec IRM

Un ensemble plus complet de caractéristiques a été défini, incluant :

- Mesures démographiques et cliniques de base
- Score cognitif (Moca J7)

Le protocole expérimental incluait deux ensembles de caractéristiques dérivées de l'IRM :

1. Un ensemble restreint de 17 caractéristiques IRM sélectionnées, ciblant des régions cérébrales spécifiques (notamment le putamen et l'hippocampe) et utilisant diverses techniques d'analyse d'image (telles que GLCM, GLDM et GLSZM).
2. Un ensemble élargi comprenant toutes les caractéristiques IRM disponibles.

Ces deux ensembles ont été utilisés dans des tests distincts afin d'évaluer l'impact de l'ajout de caractéristiques IRM supplémentaires sur les performances du clustering. Cette approche a permis une analyse comparative de l'efficacité des modèles avec un nombre limité de biomarqueurs IRM ciblés par rapport à une utilisation plus exhaustive des données d'imagerie.

```
3   categorical_features = ['Genre', 'Patho Coro', 'Insuf Card', 'Arteriopathie',
    'Synd Apnee Somm', 'Thromb Vx Prof', 'Embol Pulm', 'Tb Rythme Card',
    'Depression', 'Epilepsie', 'Cancer', 'Frisquvasc Hta', 'Frisquvasc Diab',
    'Frisquvasc Hypchol', 'Frisquvasc Hyptri', 'Frisquvasc Tabac', 'Facteur
    Alcool', 'Act Phys']
4
5   feature_sets = [
6     ['Age', 'Nb An Scol', 'Iq Code J0', 'Poids', 'Taille', 'Moca J7'],
7     ['Age', 'Nb An Scol', 'Iq Code J0', 'Poids', 'Taille', 'Mmse J7'],
8     ['Age', 'Nb An Scol', 'Iq Code J0', 'Poids', 'Taille', 'Mmse J7', 'Moca J7'],
9     ['Age', 'Nb An Scol', 'Iq Code J0', 'Poids', 'Taille', 'Moca J7', 'Nihss
      Score J7'],
10    ['Age', 'Nb An Scol', 'Iq Code J0', 'Poids', 'Taille', 'Mmse J7', 'Nihss
      Score J7'],
11    ['Age', 'Nb An Scol', 'Iq Code J0', 'Poids', 'Taille', 'Mmse J7', 'Moca J7',
      'Nihss Score J7'],
12    ['Age', 'Nb An Scol', 'Iq Code J0', 'Poids', 'Taille', 'Nihss Score J7'],
13  ]
14
15  features_set_with_IRM = ['Age', 'Nb An Scol', 'Iq Code J0', 'Poids', 'Taille',
    'Moca J7',
16      'Putamright 72h Original Glcm Imc1',
17      'Ceright 72h Original Gldm Largedependencehighgraylevelemphasis',
18      'Ceright 72h Original Glcm Jointenergy',
19      'Ceright 72h Original Gldm Largedependenceemphasis',
20      'Putamleft 72h Original Glszm Largeareahighgraylevelemphasis',
21      'Hippright 72h Original Glrlm Runpercentage',
22      'Hippleft 72h Original Firstorder Variance',
23      'Hippright 72h Original Gldm Graylevelnonuniformity',
24      'Ceright 72h Original Gldm Largedependencelowgraylevelemphasis',
25      'Hippright 72h Original Gldm Largedependencelowgraylevelemphasis',
26      'Hippright 72h Original Glrlm Longrunemphasis',
27      'Hippright 72h Original Firstorder Meanabsolutedeviation',
28      'Hippright 72h Original Glszm Largearealowgraylevelemphasis',
29      'Hippright 72h Original Firstorder Variance',
30      'Putamleft 72h Original Gldm Largedependencehighgraylevelemphasis',
31      'Hippright 72h Original Gldm Graylevelvariance',
32      'Hippright 72h Original Glszm Largeareahighgraylevelemphasis']
```

Cet ensemble (features_set_with_IRM) vise à incorporer des biomarqueurs de neuroimagerie avancés dans l'analyse de clustering, capturant potentiellement des différences plus subtiles dans la structure et la fonction cérébrale parmi les patients AVC.

Sélection de caractéristiques

Trois méthodes de sélection de caractéristiques ont été employées pour identifier les variables les plus pertinentes :

1. Seuil de variance : Élimine les caractéristiques à faible variance, aidant à éliminer les caractéristiques non informatives.
2. Analyse en composantes principales (ACP) : Réduit la dimensionnalité tout en préservant la variance maximale.
3. Agglomération de caractéristiques : Regroupe les caractéristiques similaires, utile pour gérer les données de haute dimension.

Algorithmes de clustering

Cinq algorithmes de clustering distincts ont été implémentés :

1. K-means : Efficace pour identifier des clusters sphériques.
2. Modèles de mélange gaussien (GMM) : Flexibles pour modéliser des distributions complexes.
3. Clustering spectral : Efficace pour les clusters non convexes, capturant potentiellement des relations complexes dans les données d'AVC.
4. Clustering hiérarchique : Révèle des structures imbriquées de sous-groupes de patients.
5. BIRCH : Gère efficacement les grands ensembles de données, potentiellement utile pour de futures études à plus grande échelle.

Chaque algorithme a été évalué avec des configurations k=2 et k=3 pour identifier le nombre optimal de clusters.

Méthodes d'imputation

Plusieurs méthodes d'imputation ont été implémentées pour gérer les valeurs manquantes :

- MICE (Multiple Imputation by Chained Equations)
- KNN (K-Nearest Neighbors)
- Médiane
- Remplissage vers l'avant (forward-fill)

Ces méthodes permettent de traiter les données manquantes de manière robuste, en utilisant différentes approches statistiques et basées sur les données pour estimer les valeurs manquantes. Cela aide à préserver la taille de l'échantillon et à réduire les biais potentiels.

Métriques d'évaluation

Les résultats du clustering ont été évalués à l'aide de plusieurs métriques complémentaires :

1. Score silhouette : Mesure la similarité d'un objet avec son propre cluster par rapport aux autres clusters.
2. Indice Calinski-Harabasz : Évalue la séparation des clusters basée sur le ratio de dispersion inter-cluster à intra-cluster.
3. Indice Davies-Bouldin : Calcule la similarité moyenne entre chaque cluster et son cluster le plus similaire.

De plus, une validation croisée a été effectuée pour évaluer la stabilité des clusters, avec des scores et des écarts-types rapportés.

Sélection de la variable cible

La variable cible principale était Moca M6 (score MoCA à 6 mois), choisie pour sa pertinence clinique dans l'évaluation de la fonction cognitive post-AVC.

Les deux autres variables cibles utilisées dans l'étude, en plus de "Moca M6", étaient "Synd Tb Cog M6" et "Moca M6 - Moca J7". Voici une explication de leur choix et de leur validité :

Synd Tb Cog M6

Cette variable représente la présence ou l'absence d'un syndrome de trouble cognitif à 6 mois post-AVC.

Justification du choix :

- Elle fournit une classification binaire des patients, ce qui peut simplifier l'interprétation clinique.
- Elle capture l'état cognitif global du patient à un moment crucial de la récupération post-AVC.

Validité :

- La validité de cette variable dépend de la définition précise et standardisée du syndrome de trouble cognitif utilisée dans l'étude.
- Elle peut être moins sensible aux changements subtils de la fonction cognitive comparée à une mesure continue comme le score MoCA.
- Le choix du seuil pour définir la présence du syndrome peut influencer significativement les résultats du clustering.

Moca M6 - Moca J7

Cette variable représente le changement du score MoCA entre 7 jours (J7) et 6 mois (M6) post-AVC.

Justification du choix :

- Elle capture la trajectoire de récupération cognitive plutôt qu'un état statique.
- Elle peut révéler des patterns de récupération distincts qui ne seraient pas apparents en examinant uniquement les scores à un moment donné.

Validité :

- Cette mesure de changement peut être plus informative sur la progression de la récupération cognitive.
- Elle contrôle indirectement les différences de niveau cognitif initial entre les patients.
- Cependant, elle peut être influencée par des effets plafond ou plancher du test MoCA, surtout pour les patients ayant des scores très élevés ou très bas à J7.

Comparaison avec Moca M6

L'utilisation de "Moca M6" comme variable cible principale semble avoir été le choix le plus judicieux pour plusieurs raisons :

1. Elle fournit une mesure directe et continue de la fonction cognitive à un moment cliniquement significatif (6 mois post-AVC).
2. Les résultats montrent des performances de clustering supérieures avec cette variable, notamment des scores de silhouette plus élevés et une meilleure stabilité en validation croisée.
3. Elle évite les potentiels problèmes de définition associés à "Synd Tb Cog M6" et les complications d'interprétation liées à une mesure de changement comme "Moca M6 - Moca J7".

En conclusion, bien que chaque variable cible apporte une perspective unique, "Moca M6" semble offrir le meilleur équilibre entre la pertinence clinique et la robustesse statistique pour le sous-typage des patients AVC dans cette étude.

Analyse des résultats

Les résultats ont été analysés en fonction de :

- Métriques de performance du clustering
- Tailles et équilibre des clusters
- Stabilité de la validation croisée
- Interprétabilité clinique des clusters résultants

L'algorithme K-means avec k=2 s'est révélé être la configuration optimale, démontrant :

- L'indice Calinski-Harabasz le plus élevé (35,558)
- Un équilibre raisonnable des clusters (52/68 patients)
- Une performance stable en validation croisée (0,193 ± 0,024)

Voici un tableau récapitulatif détaillé du pipeline de clustering pour le sous-typage des AVC :

Pipeline de Clustering pour le Sous-typage des AVC

Étape	Composants	Détails	Justification
Préparation des données	Cohorte	121 patients AVC ischémiques	Base statistique suffisante
	Points temporels	J7, M6, M36, M60	Suivi longitudinal complet

Étape	Composants	Détails	Justification
	Caractéristiques de base	Démographiques, cliniques, scores cognitifs	Approche multidimensionnelle
Prétraitement	Nettoyage des données	Gestion des valeurs manquantes, standardisation numérique	Optimisation de la qualité des données
	Méthodes d'imputation	MICE, KNN, médiane, remplissage vers l'avant	Traitement robuste des données manquantes
Caractéristiques	6 caractéristiques clés	Age, Nb An Scol, Iq Code J0, Poids, Taille, MoCA J7	Variables fondamentales
	Ensembles de données	+12 sets dont 7 principaux (cliniques + démographiques)	Exploration exhaustive
Sélection de caractéristiques	Méthodes	Seuil de variance, ACP, Agglomération	Réduction dimensionnelle optimale
Clustering	Algorithmes	K-means, GMM, Spectral, Hiérarchique, BIRCH	Diversité des approches
	Nombre de clusters	k=2 et k=3	Exploration des groupements
Variables cibles	Principales	Moca M6, Synd Tb Cog M6, Moca M6 - Moca J7	Évaluation multi-critères

Étape	Composants	Détails	Justification
Évaluation	Métriques internes	Score silhouette, Indice Calinski-Harabasz, Indice Davies-Bouldin	Validation quantitative
	Validation croisée	Évaluation de la stabilité des clusters	Robustesse des résultats
Visualisation	Graphiques	Trajectoires des scores cognitifs (MMSE, MoCA)	Interprétation clinique
Analyse	Tests statistiques	ANOVA pour variables continues	Validation statistique
	Comparaison clusters	Taille, caractéristiques, trajectoires cognitives	Interprétation clinique

Résultats Optimaux Obtenus

Algorithme	Configuration	Métriques	Distribution
K-means	k=2	Silhouette: 0.214, CH: 35.558, DB: 1.633	52/68 patients
Spectral	k=2	Silhouette: 0.400, CH: 14.070, DB: 0.795	116/4 patients

CH: Calinski-Harabasz, DB: Davies-Bouldin

Cette approche systématique et exhaustive permet une exploration approfondie des sous-types d'AVC, en combinant données cliniques, démographiques et cognitives avec des méthodes d'analyse avancées. La pertinence de ce protocole pour le sous-typage de l'AVC réside dans sa capacité à identifier des sous-groupes de patients cliniquement significatifs tout en tenant compte de la nature complexe et multidimensionnelle des données d'AVC. La combinaison de multiples algorithmes et métriques d'évaluation assure une approche robuste pour capturer des trajectoires de récupération cognitive distinctes.