UNIVERSITAT DE BARCELONA

FUNDAMENTAL PRINCIPLES OF DATA SCIENCE MASTER'S THESIS

Patient Profiling and Task Assignment in Neuropsychological Rehabilitation: A Data-Driven Analysis of GNPT's Decision Support System

---

Author:
Ilaria Curzi

Supervisor:
Dr. Javier Solana Sánchez and
Dr. Laura Igual

A thesis submitted in partial fulfillment of the requirements for the degree of MSc in Fundamental Principles of Data Science

in the

Facultat de Matemàtiques i Informàtica

June 30, 2025

UNIVERSITAT DE BARCELONA

Abstract

Facultat de Matemàtiques i Informàtica

MSc

Patient Profiling and Task Assignment in Neuropsychological Rehabilitation: A Data-Driven Analysis of GNPT's Decision Support System

by Ilaria Curzi

Personalizing neuropsychological rehabilitation remains a key challenge in clinical practice. This thesis presents a critical analysis and extension of the decision support system integrated into the Guttmann NeuroPersonal Trainer (GNPT) platform for neuropsychological rehabilitation, known as the Intelligent Therapy Assistant (ITA). The ITA is designed to personalize treatment for patients with cognitive impairments by identifying patients with similar neuropsychological profiles and recommending appropriate cognitive training tasks.

To enhance the ITA's adaptability and clinical impact, the work develops a new clustering model that integrates a wider range of cognitive, demographic, and clinical variables that better capture the diversity of patient profiles. These profiles are then used to reassess the ITA task allocation logic and to propose alternative, data-driven criteria to improve the alignment between recommended tasks and individual patient needs. [1]

[1] All code is available at: `https://github.com/Ilaria125/Master-Thesis-DS`.

Acknowledgements

I would like to express my heartfelt gratitude to my supervisor, Dr. Jordi Solana, for giving me the opportunity to step into the world of clinical research. Despite its complexity, this field has proven to be deeply inspiring and has shown me how meaningful it can be to work on projects that have a real impact on people's lives.

A special thank you goes to Laura for her constant guidance and support, not only academically but also emotionally. Her availability whenever I needed guidance and encouragement were essential in navigating the most challenging phases of this project.

The deepest and most profound gratitude goes to my parents. Thank you for every sacrifice you have made to allow me to pursue my dreams, even when it meant being far from home. Everything I am and everything I will become, I owe to you.

Contents

Chapter 1

Introduction

1.1 Clinical Context and Motivation

Acquired Brain Injury (ABI) often leads to cognitive difficulties that can affect a person's ability to live independently, return to work, or maintain social relationships. These impairments commonly involve attention, memory, and executive functions and require carefully adapted rehabilitation plans for each individual.

Designing personalized interventions can be challenging for clinicians, especially when dealing with large patient populations and diverse cognitive profiles. To meet this growing demand, scalable digital tools have become essential to help therapists offer more efficient, flexible, and evidence-based treatment strategies (Câmara, Geraldo, et al., 2024).

The *Guttmann NeuroPersonal Trainer* (GNPT)[1] was developed precisely with this goal in mind. It is a digital platform designed to assist professionals in the planning, monitoring, and dynamic adjustment of cognitive therapy programs for individuals with neurological disorders, particularly those with ABI (Solana, Rius, Tormos, et al., 2014).

To facilitate decision-making and optimize the rehabilitation process, GNPT integrates a Decision Support System (DSS) that automatically assigns cognitive tasks to patients based on their neuropsychological profiles. This system, known as the Intelligent Therapy Assistant (ITA), is designed to personalize rehabilitation plans through a two-step data-driven process.

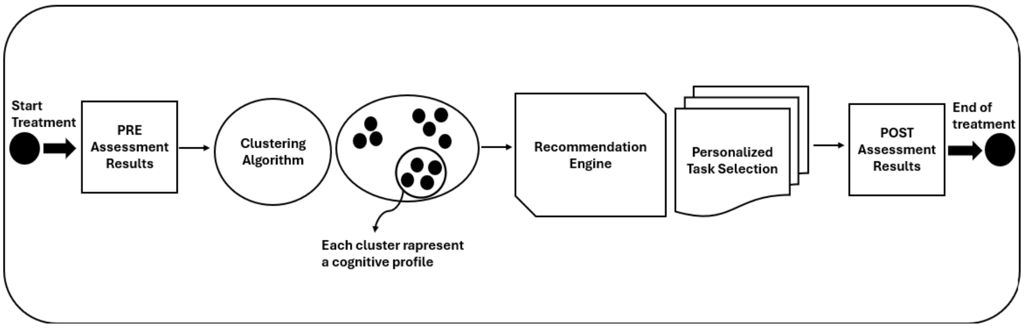

FIGURE 1.1: Overview of the ITA workflow from patient assessment to personalized task selection and post-treatment evaluation.

As shown in Figure 1.1, the process begins with the patient's neuropsychological evaluation results, referred to as the PRE phase. During this phase, the patient completes a series of tests designed to evaluate various cognitive functions. Each

[1]Developed by Institut Guttmann, https://barcelona.guttmann.com/en.

function is rated on a scale from 0 (no impairment) to 4 (very severe impairment), based on the patient's performance.

These scores are then used as input for a clustering algorithm that groups patients into cognitive profiles based on their patterns of impairment. Based on the assigned profile, the system uses a recommendation engine to identify rehabilitation tasks that have been proven effective in similar cases. After the intervention is completed, a second assessment (POST) is performed to measure cognitive improvement and guide further therapy planning.

While this workflow provides a structured approach to therapy, the current implementation relies exclusively on neuropsychological test scores and predefined clinical rules. For this reason, a thorough review is conducted to explore potential improvements. This project investigates whether the clustering process can be enhanced by adopting more modern unsupervised learning algorithms and by incorporating additional features, including demographic and clinical variables.

In parallel, historical task assignment data is analyzed to evaluate the current scoring logic and explore improvements for making recommendations more consistent and better adapted to each patient profile.

1.2 Thesis Objectives

The general objective of this thesis is to investigate how data-driven methods can enhance decision-making in neuropsychological rehabilitation within the GNPT platform. Specifically, the work focuses on improving two core components of the current Decision Support System (DSS):

- **Patient Profiling:** A new patient profiling model was developed using the K-Modes algorithm, selected for its suitability with categorical and ordinal data and its transparency in clinical contexts. Unlike the existing approach, this model incorporates a broader set of variables, including demographic and clinical information (e.g., age, etiology, education level) together with the results of neuropsychological tests that evaluate a wider range of cognitive functions beyond the standard domains used in the GNPT system. The aim is to better capture patient heterogeneity and produce clinically meaningful groupings that reflect both cognitive status and individual background. The model was evaluated using the Elbow Method and silhouette scores based on Gower distance, and the resulting clusters were interpreted through domain-level visualizations, PCA projections, and demographic analyses to ensure clinical relevance and interpretability.

- **Task Prioritization:** Historical task assignment data within the GNPT was analyzed to evaluate and refine the current scoring logic used to recommend cognitive rehabilitation tasks. A revised task prioritization framework was developed to improve how cognitive rehabilitation tasks are recommended within the GNPT system. Historical task assignment data were analyzed to evaluate and refine the current scoring logic using statistical transformation techniques, and the scoring was further adjusted at the cluster level. This approach ensures that tasks better match the needs of each patient profile. The updated system was reviewed with clinical experts to validate its practical relevance and to ensure it supports personalized and evidence-based rehabilitation planning.

Through these components, the thesis aims to enhance the transparency, personalization, and clinical effectiveness of the DSS in cognitive rehabilitation, aligning algorithmic recommendations with the diverse profiles and therapeutic needs of patients within the GNPT system.

1.3 Structure of the Thesis

The following chapters are organized as follows:

- **Chapter 2** presents background knowledge and previous work on patient clustering and task assignment within the GNPT platform, reviewing the existing methodology and highlighting the motivations for improvement.
- **Chapter 3** describes the data sources used in this study and details the preprocessing, integration, and methodological pipeline for both clustering and task scoring.
- **Chapter 4** presents the results of the clustering analysis and the evaluation of task assignment using the revised scoring methodology.
- **Chapter 5** summarizes the conclusions of the study, discusses its limitations, and outlines potential directions for future research.

Chapter 2

Background and ITA's system Overview

2.1 Clustering and Patient Profiling in Clinical Settings

Clustering techniques are widely used in clinical research to group patients with similar characteristics (Miotto et al., 2018), especially when the aim is to tailor interventions based on data-driven profiles. In the field of neuropsychological rehabilitation, clustering can help define cognitive profiles that reflect different levels and patterns of impairment, allowing for more personalized and potentially more effective therapeutic strategies.

Within the GNPT platform, an initial implementation of patient clustering was developed by Solana, Rius, Tormos, et al. (2014) as part of a broader effort to incorporate data mining tools into clinical decision-making. The authors used the Expectation-Maximization (EM) algorithm, implemented via the Weka toolkit, to group patients with Acquired Brain Injury (ABI) based on the results of their initial neuropsychological assessments (PRE). Each patient's profile was constructed using scores from 11 cognitive subfunctions, rated from 0 (no impairment) to 4 (very severe impairment).

Table 2.1 shows the original cognitive profiles obtained through GNPT's clustering module, based on PRE neuropsychological scores. Each cluster is characterized by the average impairment level across three key domains: attention, memory, and executive functions. These profiles were used by the system to guide rehabilitation task selection.

Cluster	N. Patients	Attention	Memory	Executive Functions
0	47	1.31	1.40	0.75
1	257	4.00	3.99	4.00
2	121	2.86	2.46	3.35
3	55	1.36	0.97	1.11
4	44	2.28	1.72	2.72
5	37	3.93	1.60	3.59
6	76	2.29	1.12	2.31
7	53	3.57	1.51	2.72
8	124	3.90	2.68	3.85
9	65	1.95	1.60	2.04

TABLE 2.1: Cognitive profiles obtained using the original GNPT clustering module (Solana, Rius, Tormos, et al., 2014). Scores represent the average impairment in each domain, from 0 (no impairment) to 4 (very severe).

While this approach represents an important step towards data-driven personalization, it relies exclusively on neuropsychological test results, without incorporating additional clinical or demographic features that may influence treatment outcomes.

Recent studies have explored alternative clustering methods more suited to clinical data structures. For example, García-Rudolph et al. (García-Rudolph et al., 2021) applied the *k*-prototypes algorithm to group ABI patients in GNPT based on both categorical and numerical variables, while other works have used *k*-modes clustering to identify patient profiles in web-based neurorehabilitation settings for traumatic brain injury (Zhang et al., 2020).

Inspired by these contributions, this work adopts the *k*-modes algorithm to generate patient profiles, using a richer and more diverse feature set than previous approaches. These methodological extensions are presented and evaluated in Section 3.1.

2.2 Task Assignment Logic within the ITA

Once a patient is assigned to a cognitive profile through clustering, the GNPT platform uses an automated module (ITA) to configure individualized rehabilitation sessions.

The ITA computes a Global Suitability Score (GSS) for each of the available cognitive tasks, using a composite function based on three main components:

1. **Usage Score**: favors tasks that have been used more frequently with patients who belong to the same cognitive profile. The more a task has been used, the higher its score.

2. **Improvement Score**: rewards tasks associated with greater cognitive improvements in similar patients. If most patients improved after doing a task, that task receives a higher score.

3. **Clinical Score**: rewards tasks that train functions in which the patient is more impaired or that are clinically useful for improving other related functions.

These components and their computation are described in detail in Section 3.2.

In the original implementation, each component contributes equally to the final score, without any explicit weighting. The GSS is computed as a simple sum:

$$\text{GSS} = \text{Usage Score} + \text{Improvement Score} + \text{Clinical Score} \tag{2.1}$$

Based on their Global Suitability Score, tasks are ranked and divided into four quartiles (QI_1 to QI_4), with QI_1 containing the most suitable tasks. A therapy session is then created by combining tasks from each quartile in a fixed proportion, until the total duration of one hour is completed.

Figure 2.1 provides a visual summary of this process. It provides an overview of the combination of the three scoring components into a Global Suitability Score (GSS), the sorting of tasks into quartiles, and the process used to select tasks for a complete rehabilitation session.

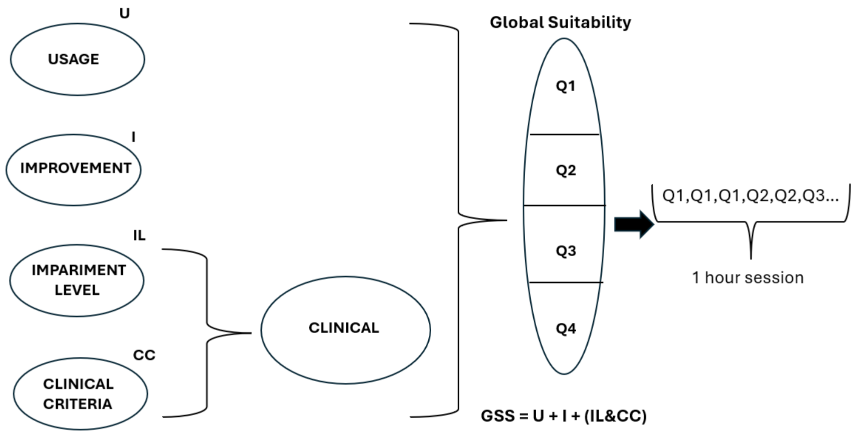

FIGURE 2.1: Diagram illustrating the scoring criteria and selection phases used to determine the most suitable tasks based on the patient's profile. The task selection follows a predefined scheme: 3 tasks from QI_1, 2 from QI_2, 2 from QI_3, and 1 from QI_4, repeated as needed until the session duration is reached (typically one hour).

While this framework offers a structured and replicable method for assigning tasks, the scoring logic relies on fixed rules and equal weights across all components. This ensures simplicity and transparency, but it does not account for potential differences in the relative importance or reliability of each component across patient profiles.

In this work, attempts are made to enhance the statistical soundness and degree of personalization of the original scoring model. Alternative weighting strategies are explored, and normalization techniques are introduced to better account for variability in task frequency and observed cognitive improvements.

These methodological extensions are presented and evaluated in Section 3.2.

Chapter 3

Methodology

The core of this study is based on structured clinical data extracted from the *GNPT*. The platform has been operational since 2014 and includes thousands of treatment records for patients with cognitive impairments.

Given the complexity and the wide range of the data, this chapter is organized into two main parts, each addressing a specific objective of the study:

- **Section 3.1 – Patient Clustering Methodology:** describes the construction of a patient-treatment level dataset, advanced handling of missing and non-evaluable values, feature selection with domain-specific weighting, and the application of the K-Modes clustering algorithm to identify clinically meaningful cognitive profiles using both cognitive and demographic variables.
- **Section 3.2 – Task Scoring Methodology:** describes the creation of a detailed task-level dataset and the development of an updated prioritization framework (GIp) that refines the task assignment logic within the GNPT platform by combining usage, improvement, and clinical relevance scores with normalization and cluster-specific weighting.

Each section details the methodology for dataset construction, preprocessing steps, and exploratory analyses conducted to guide methodological decisions, providing a clear and systematic overview aligned with the study's objectives.

3.1 Patient Clustering Methodology

3.1.1 Dataset Construction

For the purposes of clustering and patient profiling, the primary unit of analysis was the treatment (`idTreatment`), rather than the patient (`idPatient`), since a single patient may receive multiple treatments associated with different clinical conditions.

The dataset was constructed by merging three core tables from the GNPT database:

- **FUNCTION_LEVELS**: cognitive evaluations per treatment and function
- **DEMOGRAPHIC_DATA**: demographic attributes
- **ADMINISTRATIVE_DATA**: treatment metadata

For each treatment, only the earliest available evaluation per cognitive function was retained, resulting in a single record per `idFunction` within each `idTreatment`. This process produced a large dataset, initially containing approximately 356,000 treatment-function records, reflecting the extensive nature of the GNPT database. The final merged dataset consisted of the variables reported in Table 3.1.

Variable	Type	Description
idTreatment	Identifier	Unit of analysis
idFunction	Categorical	Cognitive function evaluated
resultNormalized	Ordinal (0–4, -2)	Level of cognitive impairment (-2 = non-evaluable)
age	Numeric	Patient's age at treatment start
gender	Categorical (0–1)	Gender of the patient (1=male)
studies	Ordinal (0–3)	Education level
ethiology	Categorical	Ethiology of the impairment

TABLE 3.1: Variables included in the merged dataset prior to clustering.

3.1.2 Handling of Missing and Non-Evaluable Values

Missing data was minimal and addressed with standard procedures strategies. Observations missing key fields such as age or education were removed due to their rarity, while missing gender values were imputed based on the observed distribution to preserve demographic balance.

The value `-2` in `resultNormalized` required careful interpretation, as it did not represent a technical missing value. Instead, it indicated that a cognitive function could not be evaluated, either because the function was not part of the treatment plan or because the patient was too compromised to complete the evaluation. Figure 3.1 shows the distribution of the first recorded affectation levels (`resultNormalized`) across all treatments. The plot highlights that the non-evaluable value (`-2`) is the most frequent, while mild to moderate impairment levels (0, 1, 2) occur with similar frequencies, and severe impairment levels (3 and 4) are less common but still substantial. This distribution underscores the importance of correctly interpreting the value `-2` and considering the full range of cognitive impairment levels in subsequent analyses.

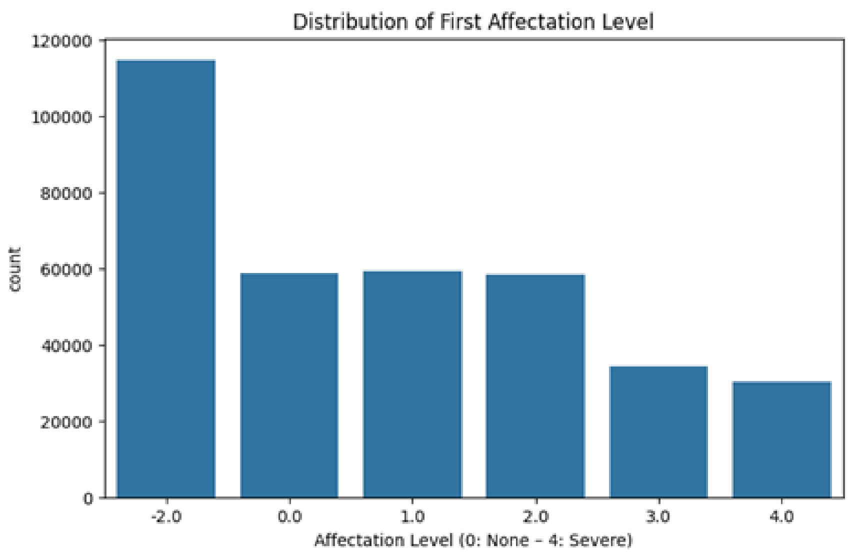

FIGURE 3.1: Distribution of affectation levels (`resultNormalized`) across the most frequently evaluated cognitive functions.

To operationalize this distinction, a clinically guided strategy was developed based on observed patterns in the data. Treatments were considered clinically severe when at least 30% of their evaluated cognitive functions scored 3 or higher.

In such cases, a consistent pattern emerged: specific functions, particularly those belonging to the domains of language, gnosias, and calculation, were sistematically marked as `-2`. These domains are known to require complex cognitive processing and are often the first to become non-assessable in patients with severe cognitive impairment, adding clinical credibility to interpreting `-2` as an indicator of very severe impairment, rather than missing data, in these specific contexts (Lezak et al., 2012).

Based on this, a two-step strategy was applied. First, all `-2` values in severe treatments were recoded as 4 (high cognitive impairment), while in all other cases they were treated as missing. Subsequently, for each treatment and cognitive macro-domain, if at least one subfunction had a valid score, missing values in other subfunctions within the same domain were imputed using the mode of the available scores. This ensured that imputations remained consistent with the patient's clinical profile. The strategy was reviewed and validated by a clinical expert.

idTreatment	**Domain**	**idFunction**	**Subfunction**	**resultNormalized**
250	Executive Functions	11	Planning	3
250	Executive Functions	12	Inhibition	3
250	Executive Functions	13	Flexibility	NaN
250	Executive Functions	14	Sequencing	2
250	Executive Functions	15	Categorization	NaN
			Imputed value (mode within domain)	3

TABLE 3.2: Example of missing value imputation using the mode within a cognitive domain.

An illustrative example is shown in Table 3.2, where a treatment includes five subfunctions from the *Funcionsexecutives* domain. In this case, three subfunctions have valid scores (3, 3, and 2), while two others are missing. Since at least one value is available, the missing entries are imputed using the mode (3), reflecting the most frequent level of impairment within that domain.

This approach allowed us to retain valuable information while minimizing bias introduced by arbitrary imputation, reducing the proportion of missing values from to 17.91% after domain-wise imputation.

3.1.3 Feature Selection and Preparation

To prepare the cognitive input for clustering, an initial analysis was performed on the full set of 187 cognitive functions (`idFunction`) evaluated across treatments. This exploratory step revealed a highly imbalanced distribution: only a small subset of functions was consistently evaluated in all patients, while many appeared in only a limited number of treatments.

To better illustrate this distribution, Figure 3.2 shows a histogram of the number of treatments per function, ordered by frequency. The X-axis labels were omitted to improve clarity.

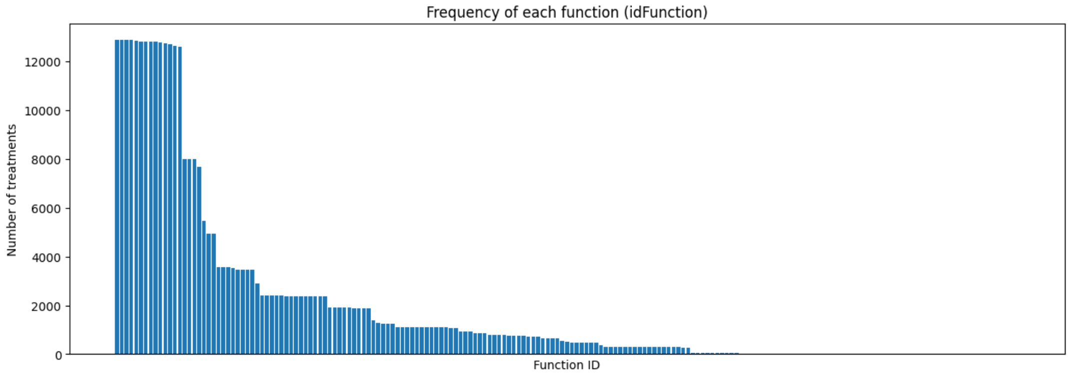

FIGURE 3.2: Histogram showing the frequency of evaluations for each cognitive function (`idFunction`), ordered by frequency.

Additionally, Figure 3.3 presents the cumulative distribution of functions by evaluation frequency, revealing a clear long-tail pattern.

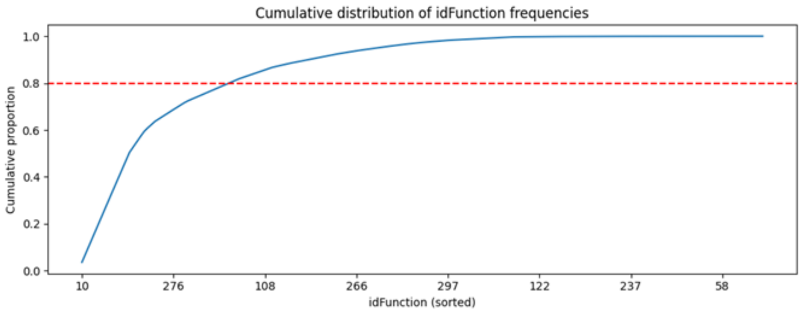

FIGURE 3.3: Cumulative distribution of cognitive functions (`idFunction`) by evaluation frequency.

As shown in Figure 3.3, approximately 20–25% of the functions accounted for over 80% of all evaluations. To reduce data sparsity and focus the analysis on consistently evaluated domains, only the most frequently evaluated functions were retained, specifically those that covered at least 80% of all treatments. The selected functions span several cognitive domains, including attention, memory, executive functions, orientation, calculation, and object and face recognition (gnosias).

To prepare the data for clustering, the dataset was reshaped from long to wide format: each row represented a single treatment, and each selected cognitive function became a separate column. In this structure, the value in each cell corresponds to the `resultNormalized` score for that specific function and treatment, treated as a categorical indicator of impairment level.

Figure 3.4 shows the distribution of affectation levels (`resultNormalized`) across the selected functions. As illustrated, macro-domains are represented by different numbers of cognitive functions. For instance, the *executive functions* domain is represented by five subfunctions (Planning, Inhibition, Flexibility, Sequencing, Categorization), whereas domains such as *Calculation* are represented by a single function. To account for this imbalance, domain-informed weights were applied during the clustering process, ensuring a more balanced influence across cognitive areas. (see Tables A.9)

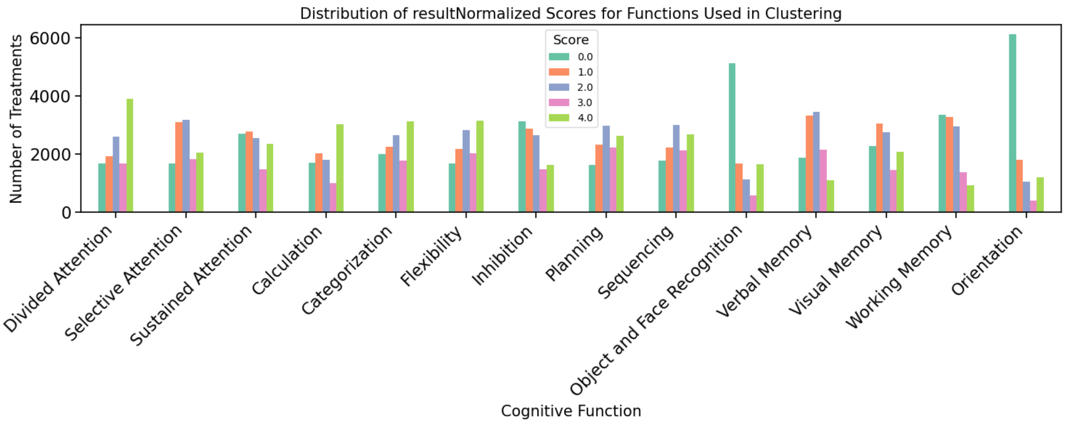

FIGURE 3.4: Distribution of `resultNormalized` scores across cognitive functions used in the clustering input.

Demographic and clinical variables were also included and duplicated to balance their influence against the more numerous cognitive features. To improve interpretability and ensure adequate representation across categories, `etiology` was recoded into six clinically meaningful groups (e.g., *Vascular*, *Neurodegenerative*, *Psychiatric*, etc.). Similarly, `age` was discretized into five categories to better capture relevant differences in cognitive profiles.

Despite recoding, some categories—such as *Vascular*—remained overrepresented, while others were comparatively rare. To mitigate this imbalance and prevent dominant categories from disproportionately influencing the clustering results, a weighting scheme was applied during preprocessing (see Tables A.10). The final distribution of demographic and clinical variables, after recoding and sparsity filtering, is illustrated in Figure 3.5.

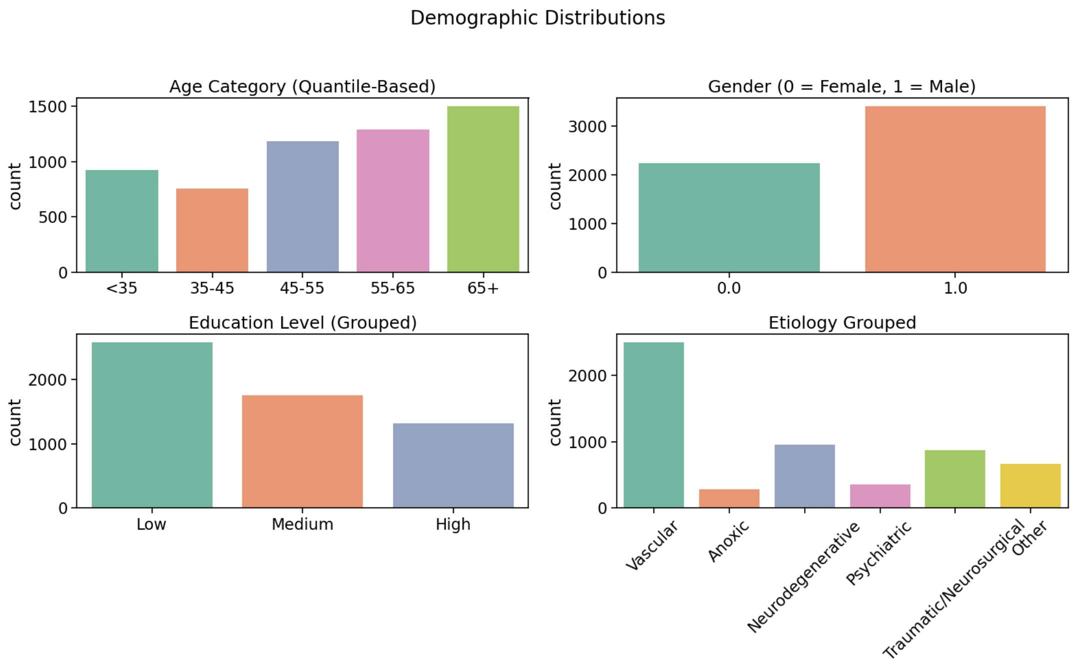

FIGURE 3.5: Distribution of demographic and clinical variables.

3.1.4 Clustering Method: K-Modes

To identify distinct cognitive profiles among patients, the K-Modes algorithm was selected as the most suitable clustering method (Huang, 1997). This choice was guided not only by its effectiveness, but also by its transparency and explainability, both of which are crucial in clinical contexts where results need to be communicated clearly to healthcare professionals who may not be familiar with complex machine learning techniques. This aligns with ethical principles in clinical AI, helping to build trust and ensuring that algorithmic decisions impacting individuals are made in a fair and understandable way (Vitrià, 2025).

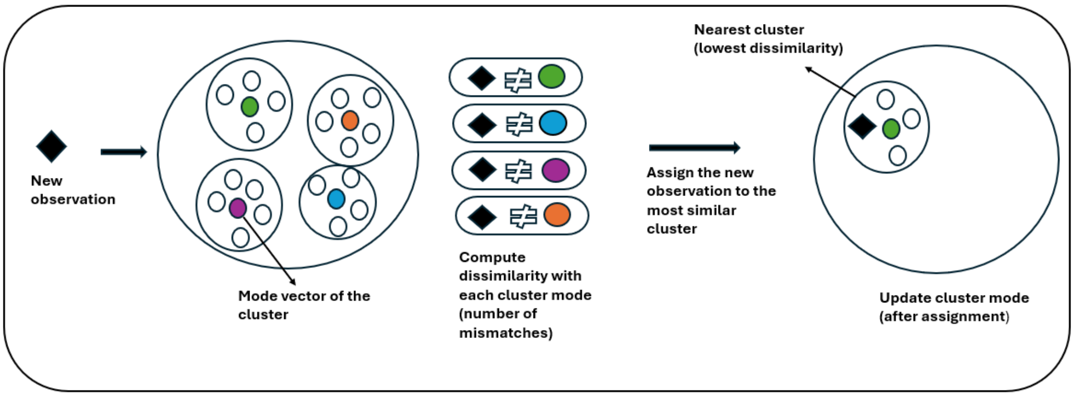

FIGURE 3.6: Schematic representation of the K-Modes algorithm.

In K-Modes, each observation is represented as a vector of categorical values (e.g., [Male, Vascular, HighSchool]). For each cluster, a centroid is defined by computing the most frequent category (mode) for each variable among the observations it contains. The dissimilarity between a new observation and a cluster centroid is computed as the number of attribute mismatches (i.e., positions where the categories differ). The algorithm then assigns the observation to the cluster with the lowest dissimilarity and recomputes the centroid based on the updated members. This process is repeated iteratively until the cluster assignments stabilize, as illustrated in Figure 3.6.

This approach is particularly appropriate for our data, where all the variables are categorical or ordinal with non-linear relationships. By avoiding assumptions about numeric distances, K-Modes preserves clinical interpretability and avoids introducing bias from artificial scaling. Moreover, it ensures that the clustering results remain clear and coherent, as discussed in Chapter 4.

Cluster Evaluation Metrics

To determine the optimal number of clusters, the Elbow Method was used. This method evaluates the clustering cost, defined as the within-cluster dissimilarity (or inertia), as a function of the number of clusters k. For each k, the algorithm calculates the average dissimilarity of observations to their respective cluster centroids, which typically decreases as the number of clusters increases. However, adding more clusters beyond a certain point results in only marginal gains in compactness while increasing model complexity. The Elbow Method identifies this point of diminishing returns, known as the 'elbow', as the optimal value of k that balances compactness with simplicity (Kodinariya and Makwana, 2013).

Although K-Modes effectively identifies clusters in categorical data, it does not generate a distance matrix usable for traditional cluster validation metrics. Therefore, to evaluate the cohesion and separation of the resulting clusters, the silhouette score was calculated using the Gower distance. The Gower distance is specifically designed to compute dissimilarities between observations containing categorical variables, assigning a value of 0 if the categories match and 1 if they differ for each variable, with the overall distance calculated as the average across all variables (Choi, Kim, and Lee, 2019). This approach ensures that the evaluation of cluster quality remains consistent with the structure of the input data while providing a meaningful measure of clustering performance (Hennig et al., 2015).

3.2 Task Scoring Methodology

3.2.1 Dataset Construction

For the purposes of the task assignment analysis, the primary unit of analysis was the cognitive task (`idTask`) assigned within each treatment (`idTreatment`). The goal was to evaluate the suitability of 208 tasks specifically related to the rehabilitation of Acquired Brain Injury (ABI) by calculating the Global Indicator of Priority (`GIp`), which combines usage, improvement, and clinical relevance scores. The rationale for this prioritization framework was introduced in Section 2.2.

The dataset was constructed by merging the following core tables from the GNPT database:

- **BLOCK_TASKS**, **BLOCKS**, and **SESSIONS**: to associate each task with its corresponding treatment
- **ADMINISTRATIVE_DATA**: to retrieve metadata including treatment category
- **PATIENT_CLUSTERS**: to include the final patient cluster assignments
- **TASKS_TO_BE_SCORED**: to filter the 208 cognitive tasks under evaluation
- **FUNCTION_IMPROVEMENT**: to compute task-related improvement scores
- **FUNCTION_LEVELS** and **TASKS_SUITABILITY**: to calculate clinical relevance scores

After preprocessing, the final dataset included 396,897 rows, each representing a specific cognitive task assigned to a treatment, along with the patient cluster associated with that treatment. This structure allowed us to compute and rank tasks by priority within each patient cluster, forming the basis for the scoring and prioritization analysis presented in the following sections.

Table 3.3 summarizes the variables included in the task-level dataset used for the prioritization analysis.

Variable	Type	Description
`idTask`	Identifier	Cognitive task under evaluation
`idTreatment`	Identifier	Treatment associated with the task
`cluster_k5`	Categorical (0–4)	Final patient cluster assignment
`category_id`	Categorical	Treatment category (filtered for DCA)
`idFunction`	Categorical	Main cognitive function associated with the task
`n_uses`	Numeric	Number of times the task was used within each cluster
`result`	Ordinal	Improvement value in cognitive functions
`resultNormalized`	Ordinal (0–4, -2)	Level of cognitive impairment
`suitability`	Numeric	Task suitability weights for related functions
`usage_score`	Scaled (1–4)	Normalized frequency of task usage per cluster
`improvement_score`	Integer (0–4)	Improvement score calculated per task-cluster
`clinical_score`	Scaled (0–4)	Clinical relevance score per task-cluster
`GIp`	Numeric	Combined prioritization score per task-cluster

TABLE 3.3: Variables included in the task-level dataset for the prioritization analysis.

3.2.2 Usage Score

The first component of the `GIp` score is the Usage Score, which reflects how frequently a task has been used in past interventions within each patient cluster (`cluster_k5`).

Historical Quartile-Based Scoring. Initially, task usage was discretized into quartiles following the original GNPT DSS logic. In this method, tasks were ranked by their usage counts within each cluster and divided into four equally sized groups, assigning scores from 1 (lowest usage group) to 4 (highest usage group).

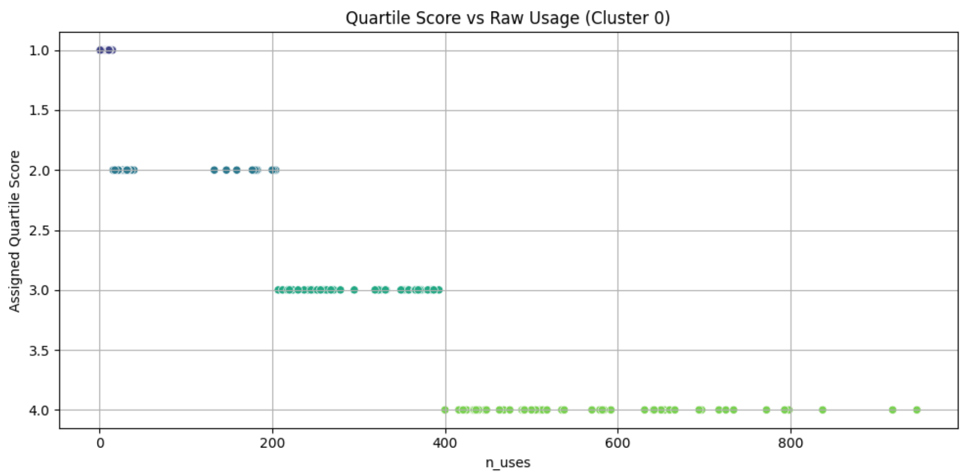

FIGURE 3.7: Quartile score assignment vs raw task usage (Cluster 0, representative example).

However, this approach relies solely on the relative position of a task within the distribution, without considering the actual differences in usage frequency between tasks. As shown in Figure 3.7, this method proved inadequate in the presence of highly skewed task usage distributions. For example, tasks used around 400 times and tasks used over 800 times ended up in the same highest quartile, despite significant differences in actual usage. Similarly, tasks with zero usage were assigned to different quartiles alongside tasks with very low usage, reflecting the positional nature of quartile assignment. These issues reduced the method's ability to meaningfully differentiate between low and high-usage tasks, limiting its informativeness for prioritization.

Z-score Normalization Approach. To address these limitations, a z-score based normalization approach was adopted. This method calculates the standardized distance of each task's usage count from the cluster mean, capturing how much a task deviates from typical usage within the cluster. The resulting z-scores were then rescaled to the [1, 4] interval using MinMax scaling to align with the scoring range of the other `GIp` components while preserving the relative differences in usage frequency.

FIGURE 3.8: Z-score normalized usage scores scaled to [1, 4] (Cluster 0). The continuous distribution highlights differences across the entire range of usage frequencies.

As shown in Figure 3.8, this approach produces a smooth, continuous distribution of scores across tasks within Cluster 0, ensuring that tasks with higher usage receive proportionally higher scores while maintaining differentiation among tasks with low or moderate usage. This approach improves the overall fairness and interpretability of the Usage Score, ensuring that frequently used tasks are appropriately prioritized while maintaining sensitivity to differences among less frequently used tasks. The consistency of this pattern across clusters confirms the general applicability of the z-score approach for this analysis.

3.2.3 Improvement Score

The second component of the `GIp` score is the Improvement Score, which measures how much each cognitive task is linked to improvements in patient performance within each cluster.

Calculation Method. Following the original GNPT DSS logic, improvement was evaluated using four criteria based on the proportion of patients showing progress in specific cognitive domains after receiving a task:

- **S (Subfunction):** Improvement in the specific cognitive subfunction targeted by the task.
- **F (Function):** Improvement in other subfunctions within the same main cognitive domain.
- **M1 and M2 (Other Main Domains):** Improvement in each of the two remaining main cognitive domains.

For each task-cluster combination, the set of patients who received the task was identified, and the proportion of patients showing improvement in each criterion was computed. A criterion was considered fulfilled if more than 50% of the patients in the cluster demonstrated improvement in that dimension. The Improvement Score was then calculated as the sum of the fulfilled criteria, resulting in an integer score ranging from 0 to 4.

Rationale and Threshold Criteria. This method ensured that the Improvement Score captured whether a task was generally associated with improvements across different cognitive domains within each cluster. To ensure that the computed scores remained robust and representative, the analysis was restricted to tasks performed by at least 10 patients within each cluster, excluding those with lower frequencies from the analysis.

3.2.4 Clinical Score

The third component of the `GIp` score is the Clinical Score, which captures the clinical relevance of each cognitive task within each patient cluster based on the levels of cognitive impairment observed in the targeted domains.

Calculation Method. The Clinical Score was computed using two components: the average impairment level in the specific subfunction targeted by the task (**NA**) and the weighted average impairment across other related subfunctions based on their clinical suitability (**CC**). This approach ensured that the Clinical Score reflected both the direct clinical relevance of a task to the cluster's primary impairments and its broader alignment with the cognitive profiles within each cluster.

The resulting Clinical Score for each task-cluster was defined as the weighted sum of the NA, doubled to emphasize primary impairment, and the CC components.

Normalization and Rationale. However, following the original GNPT DSS logic, these components had incompatible scales: while the Usage and Improvement Scores were bounded between 0 and 4, the Clinical Score ranged from 6 to over 60 points. As shown in Figure 3.9, this imbalance caused the Clinical Score to dominate the `GIp` calculation, reducing the contribution of the other components and limiting the effectiveness of the prioritization framework.

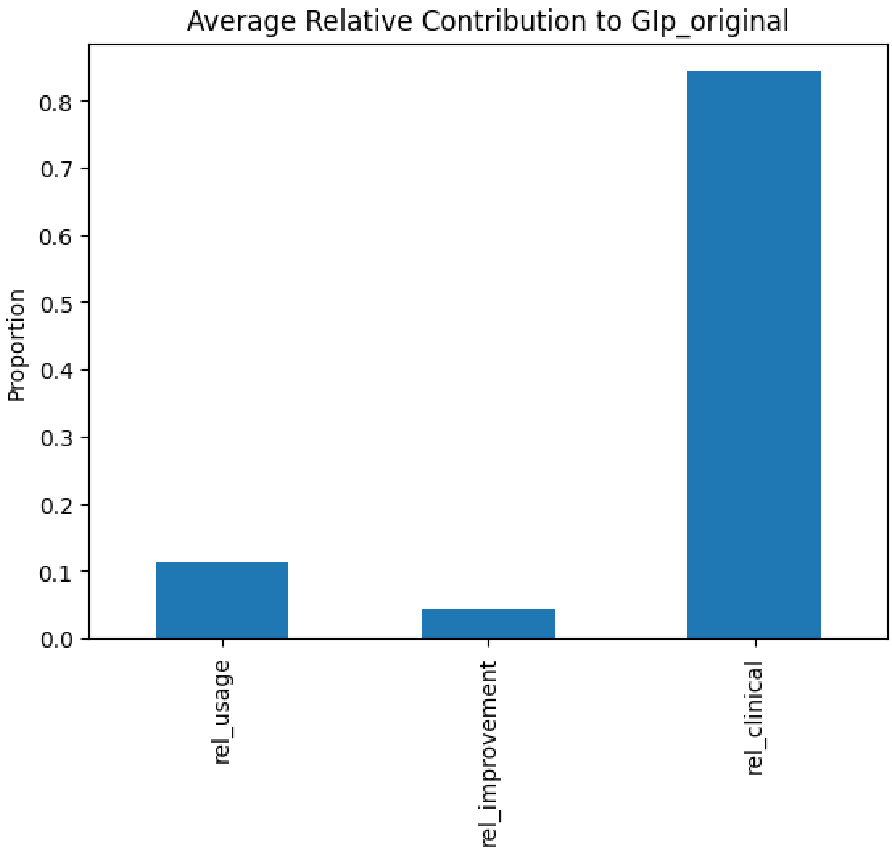

FIGURE 3.9: Average contribution of each component to the original GIp score. The Clinical Score accounts for more than 80% of the total value, dominating the prioritization.

To address this limitation, the Clinical Score was normalized within each cluster to the [0, 4] interval using MinMax scaling, aligning it with the scales of the other GIp components while preserving the relative differences in impairment levels across tasks within each cluster. As shown in Figure 3.10, this approach produced a balanced distribution of contributions across the Clinical, Usage, and Improvement Scores, ensuring that clinical relevance was appropriately represented without overpowering the other dimensions. This normalization improved the fairness and interpretability of the Clinical Score within the GIp framework, enabling consistent prioritization of tasks across clusters based on clinical needs while maintaining balance with usage frequency and observed effectiveness.

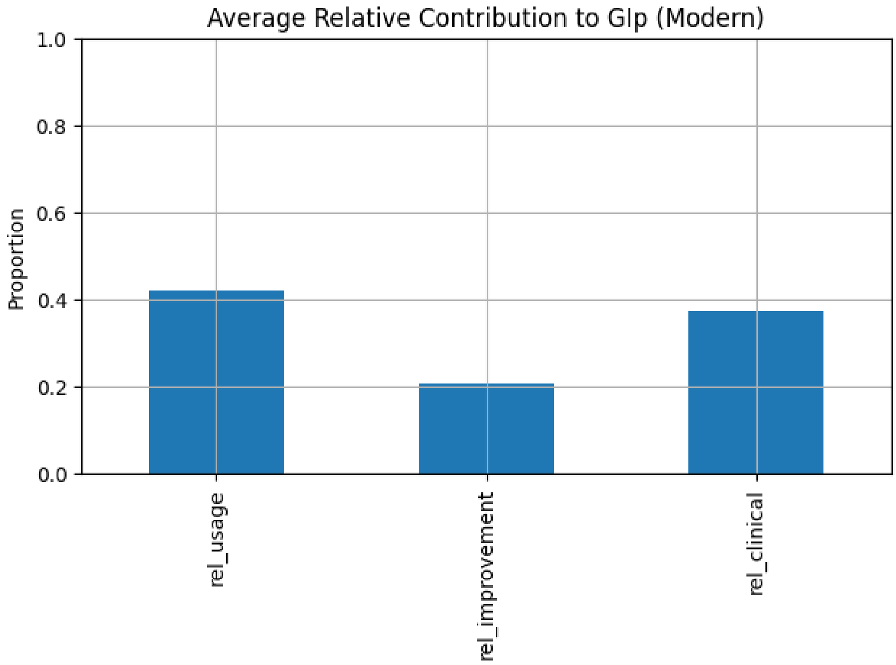

FIGURE 3.10: Average contribution of each component to the updated `GIp` score after normalization.

Figure 3.10 illustrates the balanced contribution of each component within the updated `GIp` framework.

In summary, the methodological steps described in this section prepared the individual Usage, Improvement, and Clinical Scores for integration into an updated prioritization framework within the GNPT system. The formulation of this unified metric and its application using cluster-specific weights are detailed and evaluated in Section 4.2.

Chapter 4

Results

4.1 Clustering Results

4.1.1 Cluster Structure and Elbow Method

The number of clusters was selected using the Elbow Method (see Section 3.1.4). Initially, increasing the number of clusters leads to a substantial reduction in dissimilarity, but after a certain point, the improvement becomes marginal. As shown in Figure 4.1, the curve displays a clear inflection point, or 'elbow', at $k = 5$, representing the balance between achieving sufficient compactness while avoiding unnecessary complexity.

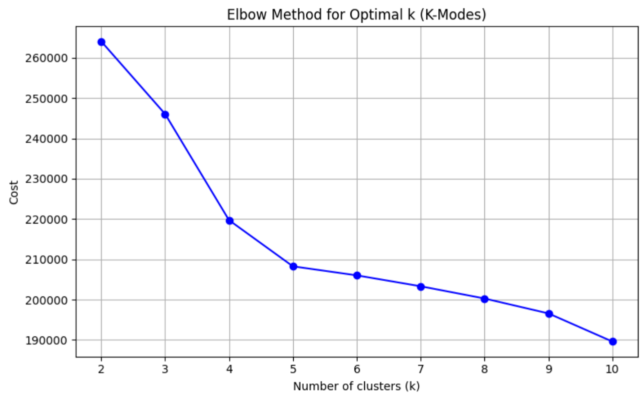

FIGURE 4.1: Elbow method showing the clustering cost for different values of k (K-Modes).

The clustering solution for $k = 5$ was further evaluated using the silhouette score, computed based on the Gower distance. The resulting score, 0.167, is considered low in general clustering contexts. However it is expected in a clinical dataset where preserving a certain degree of within-cluster variability is valuable to reflect the heterogeneity of patient conditions. For example, it is clinically meaningful for a cluster to include patients with severe memory impairment but preserved calculation abilities, as these nuanced profiles are essential for supporting personalized rehabilitation strategies. In this setting, interpretability and clinical relevance were prioritized over purely numerical separability.

4.1.2 Cognitive Profiles by Cluster

The five clusters identified through the K-Modes algorithm displayed distinct cognitive impairment patterns across major domains, as well as clear differences in size and demographic composition. The mean severity levels across cognitive macro-domains for each cluster are reported in Table 4.1, providing a high-level overview of the cognitive profiles identified. A detailed heatmap showing the severity distribution for each individual cognitive function is available in Appendix A.

Cluster	Number of Treatments
Cluster 0	2,173
Cluster 1	2,436
Cluster 2	1,656
Cluster 3	1,897
Cluster 4	2,578

TABLE 4.1: Distribution of treatments across the five clusters ($k = 5$).

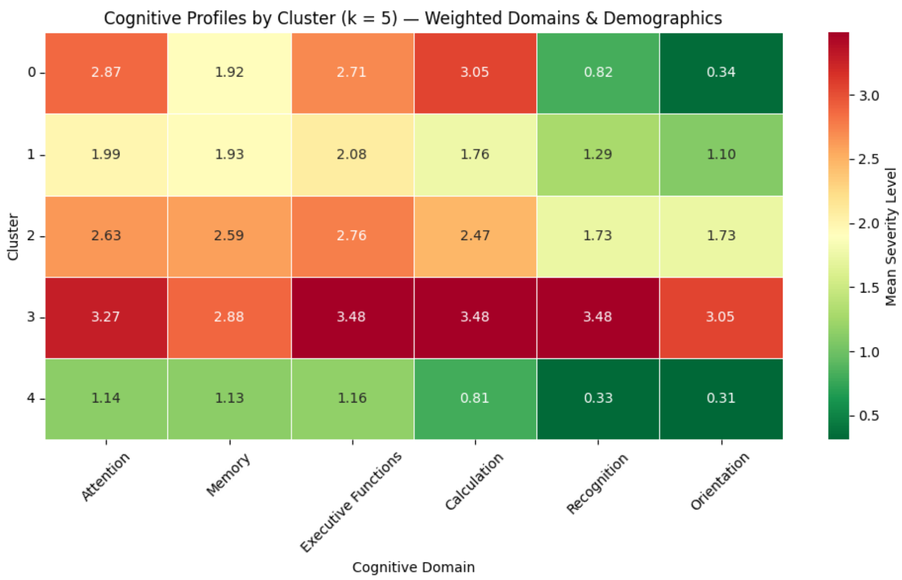

FIGURE 4.2: Mean severity levels across cognitive macro-domains for each cluster ($k = 5$). Higher values indicate greater impairment.

Figure 4.2 presents the average severity scores per cluster across six cognitive domains, using a scale from 0 (no impairment) to 4 (severe impairment). This visualization allows for a direct comparison of cognitive profiles among clusters, supporting the interpretation of clinical differences.

Cluster 3 exhibited the most severe impairments, with all domain scores exceeding 3.0. In contrast, Cluster 4 showed minimal impairment across all domains, with values close to or below 1.0, indicating a low level of cognitive compromise. Cluster 0 displayed the most heterogeneous profile, with high impairment levels in cognitive functions such as attention, executive functions, and calculation, low impairment in gnosis and orientation, and a moderate level of impairment in memory. Clusters 1 and 2 presented similar intermediate profiles, with Cluster 2 displaying slightly higher impairment across all domains, indicating a greater level of cognitive compromise compared to Cluster 1.

Overall, the mean values across domains are well stratified, suggesting good cognitive separability between clusters.

To further explore the distribution of patients across clusters, we conducted a Principal Component Analysis (PCA) to reduce dimensionality and enable 2D visualization. Figure 4.3 shows the projection of patients onto the first two principal components, which together explain 60.2% of the variance in the data.

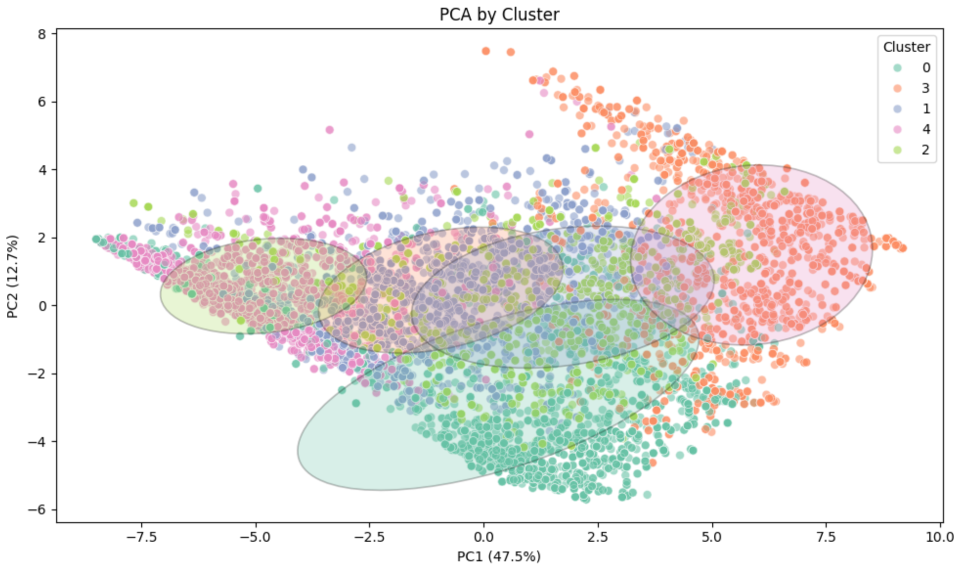

FIGURE 4.3: PCA projection of patients by cluster, based on cognitive and demographic features. Ellipses indicate the general spread of each group.

Notably, Clusters 3 and 4 are positioned at opposite ends of the PCA space, reflecting the clear contrast between the severe cognitive deficits in Cluster 3 and the minimal impairment in Cluster 4.

Cluster 0, which displayed the most heterogeneous cognitive profile, appears as the most numerous and heterogeneous group in the PCA.

Clusters 1 and 2, characterized by intermediate cognitive profiles, are located near the center of the PCA and partially overlap, indicating moderate levels of impairment and a higher degree of similarity between these groups.

Overall, while the central clusters exhibit some overlap, the PCA confirms meaningful differences across clusters—particularly at the extremes—and supports the interpretability of the profiling approach. A detailed analysis of the demographic and clinical characteristics associated with each cluster is provided in the following section.

4.1.3 Demographic and Clinical Interpretation

To interpret the cognitive clusters meaningfully, we first examined their etiological composition, as etiology emerged as a key driver of cognitive profiles when interpreting the impact of age and education on impairment severity. It is important to note that demographic separation across clusters was less pronounced than cognitive differentiation, reflecting the stronger influence of cognitive features in the clustering process while highlighting the complex interplay between demographic and clinical factors. This characteristic also helps to explain the silhouette score observed, as the inclusion of diverse feature types, cognitive, demographic, and clinical, naturally results in overlapping distributions across clusters in clinical datasets.

Etiology. Table 4.2 reports the distribution of etiologies across clusters. Cluster 3 (severely impaired) was predominantly traumatic (23.4%), Cluster 4 had a higher rate of neurodegenerative conditions (28.5%), and Cluster 0 was enriched in vascular (28.6%) and anoxic (23.6%) cases. Cluster 1 and Cluster 2 (mildly impaired) showed a higher proportion of psychiatric and other non-acute conditions (including epilepsy, paralysis, neurodevelopmental disorders, and tumors), which are typically associated with less aggressive cognitive impairment.

TABLE 4.2: Distribution (%) of etiologies across clusters ($k = 5$).

	Anoxic	Neurodeg.	Other	Psychiatric	Trauma/NS	Vascular
Cluster 0	23.6	13.2	11.7	3.5	19.5	28.6
Cluster 1	7.6	15.1	18.1	20.7	13.2	25.2
Cluster 2	7.7	13.5	28.1	17.4	15.5	17.8
Cluster 3	9.7	16.3	19.6	8.0	23.4	22.9
Cluster 4	5.7	28.5	17.8	15.0	12.2	20.8

The clear alignment between cognitive severity and etiology highlights the critical role of the underlying neurological condition in shaping impairment patterns within rehabilitation contexts. Considering etiology also clarifies the observed associations with demographic variables, avoiding misleading interpretations regarding the effects of age and education on cognitive profiles.

Age. Age did not directly correlate with the severity of cognitive impairment across clusters. Cluster 4, which exhibited lower impairment levels, included a higher proportion of older patients (49.6% aged 55+), while Cluster 3, the most severely impaired, contained a greater proportion of younger patients (26.8% under 35). Once etiology was considered, these patterns became clearer: Cluster 3, despite including younger patients, had more traumatic and vascular etiologies associated with acute, severe impairments, whereas Cluster 4, with older patients, had more neurodegenerative conditions that progress slowly, allowing milder impairment for longer periods. Full distributions of age across clusters are available in Appendix A (Table A.1).

Gender. A male predominance (62–67%) was observed across most clusters, reflecting the epidemiology of acquired brain injury, except for Cluster 4, the least impaired patient group, which interestingly had a higher proportion of females (53.5%). Full gender distributions are reported in Appendix A (Table A.2).

Education. Cluster 4, despite being the most cognitively impaired, also had the highest proportion of highly educated patients (29.5%). Although education is often linked to cognitive reserve, this does not fully protect against acute neurological damage. However, higher education may support better engagement in rehabilitation, even in severely impaired patients, aiding recovery trajectories. Full distributions of education levels across clusters are provided in Appendix A (Table A.3).

Summary and Interpretation

In conclusion, the clustering solution with $k = 5$ not only identified statistically distinct cognitive profiles, but also revealed coherent demographic and etiological patterns that enhance their clinical interpretability.

To further explore the role of demographic context, an additional clustering model with $k = 8$ was developed to match the current GNPT system configuration, focusing exclusively on the three core cognitive domains (attention, memory, and executive functions). This clustering solution, presented in Section A.2, revealed even sharper demographic differentiation across clusters and highlights the need to interpret demographic factors—such as age, education, and etiology—in combination, as their clinical impact is often interdependent

Taken together, these findings confirm the relevance of integrating demographic features into data-driven profiling, not only to improve cluster interpretability but also to move toward more personalized rehabilitation strategies.

4.2 Task Assignment Results

This section presents the results obtained by applying the revised `GIp` scoring framework, introduced in Section 3.2, to support task prioritization across patient clusters.

4.2.1 Global Indicator of Priority (GIp)

The Global Indicator of Priority (`GIp`) represents the final prioritization score used to rank cognitive tasks within each patient cluster, integrating information on task frequency, observed effectiveness, and clinical relevance in a unified metric.

The `GIp` score is computed as a weighted sum of the three components:

$$\texttt{GIp} = k_u \cdot \texttt{Usage} + k_c \cdot \texttt{Clinical} + k_m \cdot \texttt{Improvement}$$

where:

- k_u: weight for task usage,
- k_c: weight for clinical severity,
- k_m: weight for observed improvement.

Unlike the original GNPT implementation that used equal weights across components, this thesis introduced cluster-specific weights to better reflect the differing clinical priorities of each patient group. The applied weights are reported in Table 4.3.

TABLE 4.3: Cluster-specific weights used for the Global Suitability Score (GSS) calculation, indicating the importance assigned to usage (k_u), clinical relevance (k_c), and improvement (k_m) components for each cluster.

Cluster	k_u (Usage)	k_c (Clinical)	k_m (Improvement)
0	0.3	0.4	0.3
1	0.4	0.2	0.4
2	0.2	0.2	0.6
3	0.0	0.7	0.3
4	0.5	0.1	0.4

These weights reflect the specific clinical priorities of each patient group:

- **Cluster 3 (severely impaired):** prioritizes clinical severity ($k_c = 0.7$), with usage excluded.
- **Cluster 4 (preserved profile):** emphasizes profile consistency ($k_u = 0.5$).
- **Cluster 0 (executive function deficit):** higher clinical severity weight.
- **Cluster 1 (moderate generalized impairment):** balanced weights.
- **Cluster 2 (heterogeneous profile):** emphasis on improvement.

The adoption of differentiated weights was proposed in this thesis as an alternative to the uniform weighting scheme to align task prioritization with the therapeutic focus of each cluster. The specific numerical values used for each cluster (reported in Table 4.3) were defined to reflect the dominant clinical profiles identified through clustering while ensuring that the weights summed to one across components. This allocation was informed by iterative consultation with a clinical expert, who confirmed the coherence of this approach with practical therapeutic needs and its potential to enhance the clinical utility of the scoring model.

4.2.2 Distribution and Task Differentiation

Applying the updated `GIp` formulation enabled a clearer differentiation of tasks based on their suitability for each cognitive profile. The combination of normalized components and cluster-specific weights resulted in scores that aligned well with the clinical characteristics and therapeutic priorities of each group.

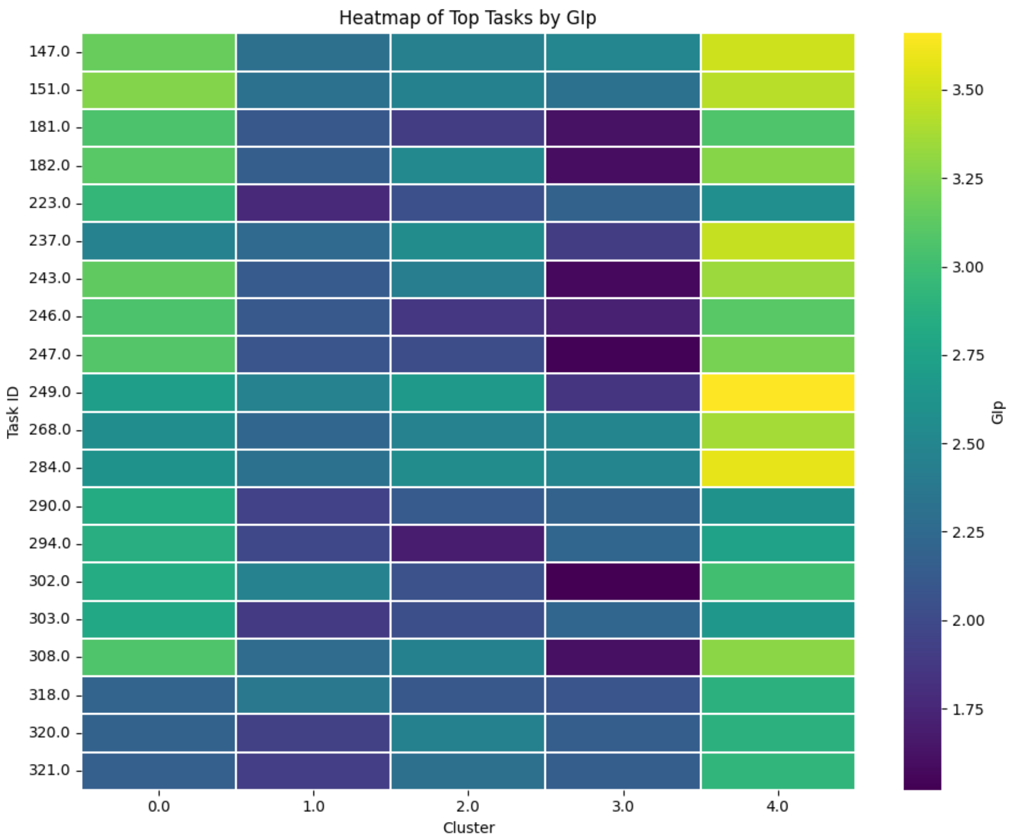

FIGURE 4.4: Top 20 tasks ranked by the revised `GIp` scores across clusters. Color intensities reflect task suitability for each cognitive profile.

Figure 4.4 shows the revised `GIp` (Global Improved Prioritization) scores of the top 20 cognitive tasks across the five patient clusters. These scores were computed using the cluster-specific weights described previously, combining the usage, improvement, and clinical components for each task within each cluster.

As a result, the `GIp` value of a given task can vary across clusters, reflecting its differential suitability depending on the cognitive and clinical profile of each patient group. Cluster 3 (severely impaired) tends to display lower `GIp` scores, as the system prioritizes clinical severity over historical usage, ensuring that only tasks with high clinical relevance are recommended for these patients. Conversely, Cluster 4 (preserved profiles) shows consistently higher `GIp` scores across many tasks, as the system emphasizes task usage to maintain cognitive activity variety in patients with minimal impairment.

In the heatmap, color intensities indicate the prioritization level of each task within each cluster, with higher scores (yellow) reflecting greater suitability for that specific profile. This visualization demonstrates how the revised `GIp` system dynamically adjusts task prioritization based on patient profiles, supporting personalized, ethically aligned, and clinically meaningful rehabilitation planning within the GNPT framework.

FIGURE 4.5: Distribution of `GIp` scores across clusters using the revised method.

This pattern is also reflected in Figure 4.5, which shows the distribution of revised `GIp` scores across clusters. Cluster 4 exhibits the highest median and wider spread of `GIp` scores, reflecting the higher suitability of a broader range of tasks due to the contribution of the usage score in preserved profiles. Conversely, Cluster 3 shows lower medians and a narrower spread, consistent with the system's emphasis on clinical severity over historical usage in severely impaired patients, resulting in lower prioritization of non-essential tasks. Overall, the distributions are compact and distinct across clusters, facilitating clearer interpretability of prioritization patterns and enhancing clinical decision support.

Summary and Interpretation. The application of the revised `GIp` scoring method enabled a clear and clinically meaningful differentiation of task priorities across patient clusters within the GNPT system. By incorporating cluster-specific weighting, the system adapts to the unique therapeutic needs of each cognitive profile, enhancing the alignment between task selection and patient characteristics. This contributes to a more transparent and systematic prioritization process that can support therapists in planning personalized interventions. While the results demonstrate the potential of this approach, further evaluation in real-world clinical workflows will be essential to determine its effectiveness in improving task assignment practices and patient outcomes in diverse rehabilitation settings.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

This thesis explored the potential of data-driven approaches to enhance decision support in neuropsychological rehabilitation. The main contributions can be summarized as follows:

- **Patient Clustering:** We applied the K-Modes algorithm to group patients based on cognitive test scores together with additional demographic and clinical variables. This integration allowed the identification of five clinically interpretable profiles that reflect distinct patterns of cognitive impairment while capturing relevant contextual factors, providing clearer and more meaningful distinctions among patient subgroups.

  Importantly, the inclusion of additional variables enabled a more comprehensive analysis of factors that may act as confounders in the interpretation of cognitive profiles. For example, the variable "etiology" emerged as a critical factor influencing both the cause and the degree of cognitive impairment, highlighting the necessity of considering underlying medical conditions when interpreting patterns of decline. Without accounting for etiology, there is a risk of drawing misleading conclusions, such as attributing lower cognitive impairment to advanced age, when in fact the type of neurological condition may be the primary driver of cognitive outcomes.

  This underlines the importance of carefully selecting and incorporating additional contextual variables to improve the explanatory power of clustering models in clinical settings.

- **Task Scoring Redesign:** We analyzed and restructured the logic for task assignment in GNPT by revisiting the computation of the `GIp` score. The original formulation tended to overweight the clinical adequacy component, limiting the ability to capture real-world task effectiveness and patient-specific needs. To address this imbalance, we introduced a revised scoring methodology in which each component—usage frequency, clinical relevance, and observed improvement—was independently normalized to a 0–4 scale, enabling fair comparison and aggregation.

  A key innovation was the introduction of cluster-specific weighting, allowing the prioritization logic to adapt to different cognitive profiles rather than applying a uniform rule across all patients. For instance, greater weight was assigned to clinical severity in more impaired profiles, while usage and profile consistency were emphasized for patients with preserved cognitive functioning. These weighting strategies were derived from the interpretation of clustering results and reviewed by a clinical expert to ensure alignment with therapeutic reasoning.

The revised scoring method was implemented and explored through descriptive analyses, including heatmaps and score distribution plots, to illustrate how task priorities are distributed across clusters in alignment with clinical reasoning. By moving beyond a one-size-fits-all approach, the redesigned `GIp` scoring method supports the generation of personalized and clinically relevant task recommendations, enhancing the potential of GNPT to deliver targeted and effective cognitive rehabilitation.

Overall, this work demonstrates how clustering and tailored scoring strategies can contribute to more effective and patient-centered rehabilitation planning in clinical decision support systems like GNPT. These contributions represent a step forward in making neuropsychological rehabilitation more adaptive, data-driven, and clinically grounded.

5.2 Limitations and Future Directions

While this thesis demonstrates the potential of data-driven methods to enhance decision support in neuropsychological rehabilitation, several limitations should be acknowledged.

First, the clustering analysis was based on a single snapshot of cognitive, demographic, and clinical data, without considering how patient profiles may evolve over time. Cognitive rehabilitation is inherently dynamic, and incorporating longitudinal data could improve the adaptiveness and clinical utility of patient profiling, particularly in long-term treatment plans. Additionally, although the revised `GIp` scoring method leverages historical data to refine task prioritization, it does not integrate therapist feedback, measures of patient engagement, or session-level adaptations, all of which may influence real-world outcomes.

Methodologically, while the K-Modes algorithm was chosen for its interpretability and suitability for categorical data, it has limitations, including sensitivity to initialization, the need to manually select the number of clusters, and the assumption of equal variable importance unless explicitly reweighted. These factors may affect the stability and flexibility of the clustering process, especially in heterogeneous clinical populations.

Finally, the dataset used in this study was sourced exclusively from the GNPT platform, reflecting specific clinical practices and protocols that may limit the generalizability of findings to other healthcare settings or rehabilitation systems.

Building on these considerations, several avenues remain open for further development:

- **Clinical Validation:** While clinical review supported the interpretability of the clusters and the updated task scoring logic, future work should evaluate their practical impact within real rehabilitation workflows, assessing whether they improve task assignment and patient outcomes in diverse clinical settings.

- **Dynamic Patient Profiling:** Incorporating longitudinal data and reassessing patient profiles over time could enhance the system's adaptiveness, allowing recommendations to evolve alongside patient progress during rehabilitation.

- **Extension to Multiple Pathologies:** Future studies could explore adapting and validating the current clustering methodology for other conditions beyond acquired brain injury, such as dementia and neurodevelopmental disorders, to ensure its effectiveness across diverse patient populations.
- **Exploration of Advanced DSS Extensions:** With a robust and interpretable patient profiling pipeline in place, future work could explore developing predictive models and recommendation systems to further refine rehabilitation personalization. For instance, supervised learning approaches could predict task suitability or session frequency based on patient histories, while optimization algorithms could support personalized scheduling that accounts for clinical priorities and patient evolution.

Advancing these directions will be essential to further enhance the personalization, transparency, and clinical relevance of data-driven cognitive rehabilitation. As clinical needs evolve and technologies advance, combining data science with practitioner expertise will be key to shaping the next generation of intelligent, ethical, and patient-centered rehabilitation platforms.

Appendix A

Supplementary Figures and Tables

This appendix reports supplementary figures and tables supporting the results described in Chapter 4. It first presents the full demographic and clinical distributions from the main clustering solution ($k = 5$), including a detailed heatmap showing the mean severity levels for each individual cognitive function across clusters. It then reports the detailed results of the second clustering model ($k = 8$). Additionally, it includes the weighting factors applied during clustering to balance the influence of cognitive, demographic, and clinical features in the analysis.

A.1 Main Clustering Model ($k = 5$)

Figure A.1 reports the detailed mean severity levels for each individual cognitive function across clusters, complementing the macro-domain analysis presented in the main text.

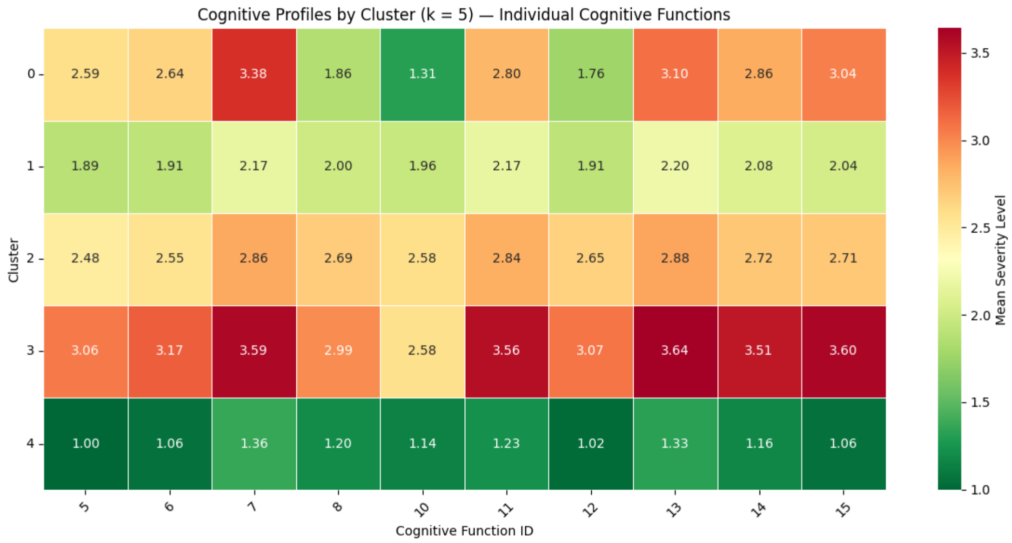

FIGURE A.1: Detailed heatmap showing the mean severity levels for each individual cognitive function across clusters ($k = 5$). Higher values indicate greater impairment.

TABLE A.1: Distribution (%) of age categories across clusters ($k = 5$).

	<35	35–45	45–55	55–65	65+
Cluster 0	14.9	14.5	30.3	18.0	22.2
Cluster 1	21.2	14.6	19.6	22.8	21.8
Cluster 2	31.5	13.9	14.1	18.4	22.1
Cluster 3	26.8	16.9	18.3	15.5	22.5
Cluster 4	18.1	12.4	19.8	25.2	24.4

TABLE A.2: Distribution (%) of gender across clusters ($k = 5$).

	Female	Male
Cluster 0	35.1	64.9
Cluster 1	32.8	67.2
Cluster 2	37.0	63.0
Cluster 3	37.8	62.2
Cluster 4	53.5	46.5

TABLE A.3: Distribution (%) of education levels across clusters ($k = 5$).

	Low	Medium	High
Cluster 0	47.9	29.7	22.4
Cluster 1	49.7	31.2	19.0
Cluster 2	50.7	29.2	20.2
Cluster 3	48.6	28.9	22.5
Cluster 4	35.7	34.9	29.5

TABLE A.4: Distribution (%) of etiologies across clusters ($k = 5$).

	Anoxic	Neurodeg.	Other	Psychiatric	Trauma/NS	Vascular
Cluster 0	23.6	13.2	11.7	3.5	19.5	28.6
Cluster 1	7.6	15.1	18.1	20.7	13.2	25.2
Cluster 2	7.7	13.5	28.1	17.4	15.5	17.8
Cluster 3	9.7	16.3	19.6	8.0	23.4	22.9
Cluster 4	5.7	28.5	17.8	15.0	12.2	20.8

A.2 Clustering Model ($k = 8$)

This section reports the detailed results of the exploratory clustering model ($k = 8$), which focused on the three core cognitive domains and is referenced at the end of Chapter 4.

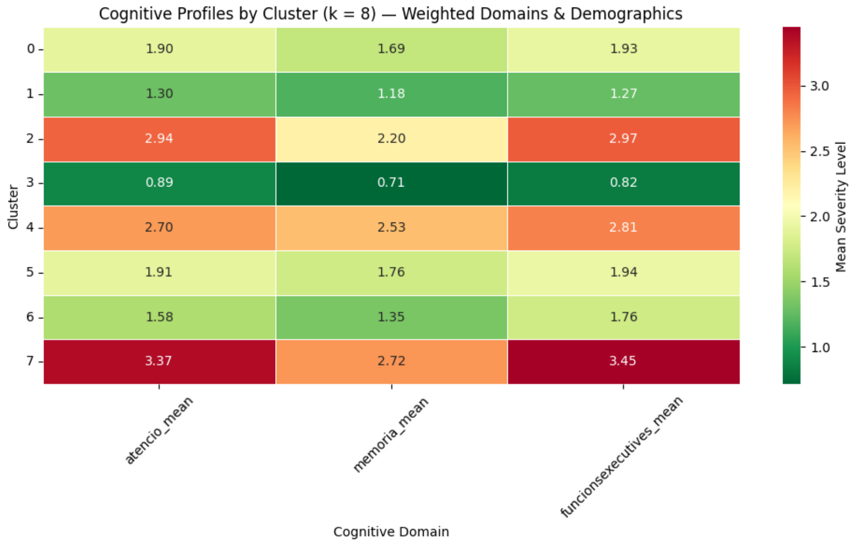

FIGURE A.2: Cluster distribution obtained from the exploratory clustering model ($k = 8$), focusing on attention, memory, and executive functions.

TABLE A.5: Distribution (%) of age categories across clusters ($k = 8$).

	<35	35–45	45–55	55–65	65+
Cluster 0	10.3	9.7	18.3	46.0	15.7
Cluster 1	10.9	11.4	20.6	18.2	38.9
Cluster 2	18.3	14.2	48.0	12.9	6.6
Cluster 3	31.9	15.7	19.7	17.7	15.0
Cluster 4	24.9	17.2	11.5	14.2	32.2
Cluster 5	63.9	10.2	14.5	5.1	6.3
Cluster 6	13.4	45.6	19.6	12.5	8.9
Cluster 7	13.9	13.9	13.0	15.6	43.6

TABLE A.6: Distribution (%) of gender across clusters ($k = 8$).

	Female	Male
Cluster 0	24.4	75.6
Cluster 1	69.4	30.6
Cluster 2	15.1	84.9
Cluster 3	68.9	31.1
Cluster 4	22.3	77.7
Cluster 5	27.9	72.1
Cluster 6	21.4	78.6
Cluster 7	73.2	26.8

TABLE A.7: Distribution (%) of education levels across clusters ($k = 8$).

	Low	Medium	High
Cluster 0	67.6	16.9	15.5
Cluster 1	28.0	50.2	21.8
Cluster 2	23.4	58.9	17.7
Cluster 3	17.3	22.9	59.8
Cluster 4	69.4	16.7	13.9
Cluster 5	12.5	78.5	9.0
Cluster 6	16.9	14.2	68.9
Cluster 7	76.5	9.2	14.2

TABLE A.8: Distribution (%) of etiologies across clusters ($k = 8$).

	Anoxic	Neurodeg.	Other	Psychiatric	Trauma/NS	Vascular
Cluster 0	9.2	13.2	13.8	10.0	9.9	43.9
Cluster 1	6.9	45.3	11.6	8.8	8.7	18.7
Cluster 2	17.5	9.2	12.2	3.5	43.1	14.6
Cluster 3	10.9	19.3	32.4	5.0	13.3	19.1
Cluster 4	10.4	14.5	16.3	14.8	29.3	14.8
Cluster 5	4.4	7.6	13.7	51.1	18.9	4.3
Cluster 6	12.8	10.9	39.0	5.8	17.1	14.5
Cluster 7	15.9	16.7	13.8	5.4	5.1	43.1

A.3 Weighting Factors for Clustering

To balance the influence of cognitive domains during clustering, domain-informed weights were assigned to each cognitive function based on the size of its macro-domain. Functions belonging to smaller domains received higher weights, ensuring a balanced contribution across cognitive areas.

TABLE A.9: Weights assigned to cognitive functions for clustering.

idFunction	Function	Weight
5	Sustained Attention	2
6	Selective Attention	2
7	Divided Attention	2
8	Verbal Memory	2
9	Visual Memory	2
10	Working Memory	2
11	Planning	1
12	Inhibition	1
13	Flexibility	1
14	Sequencing	1
15	Categorization	1
83	Calculation	6
85	Object and Face Recognition	6
87	Orientation	6

TABLE A.10: Weights assigned to etiology groups to balance their representation during clustering.

Etiology Group	Weight
Vascular	1
Traumatic/Neurosurgical	2
Neurodegenerative	2
Other	3
Anoxic	4
Psychiatric	4

Author's Note

This thesis used ChatGPT to improve language clarity and rephrase preliminary drafts. All data analysis, interpretation, and argumentation were conducted by the author.

Bibliography

Choi, Hojung, Sun Kim, and Duck Lee (2019). "A survey of distance metrics for data clustering". In: *Journal of the Korean Data and Information Science Society* 30.6, pp. 1321–1330.

Câmara, Joana, Andreia Geraldo, et al. (2024). "Remote delivered cognitive rehabilitation programs in Acquired Brain Injury: a systematic review of methods and outcomes". In: *Current Psychology* 43, pp. 31990–32026.

García-Rudolph, Alejandro et al. (2021). "A clinical decision support system to enhance the personalization of cognitive rehabilitation". In: *JMIR Medical Informatics* 9.2, e23675.

Hennig, Christian et al. (2015). *Handbook of Cluster Analysis*. CRC Press.

Huang, Zhexue (1997). "Clustering large data sets with mixed numeric and categorical values". In: *Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pp. 21–34.

Kodinariya, Trupti M. and Prashant R. Makwana (2013). "Review on determining the number of clusters in K-means clustering". In: *International Journal of Advance Research in Computer Science and Management Studies* 1.6, pp. 90–95.

Lezak, Muriel Deutsch et al. (2012). *Neuropsychological Assessment*. Oxford University Press.

Miotto, Riccardo et al. (2018). "Deep learning for healthcare: review, opportunities and challenges". In: *Briefings in Bioinformatics* 19.6, pp. 1236–1246.

Solana, Jordi, Anna Rius, Jose M Tormos, et al. (2014). "Adaptive cognitive rehabilitation: a data-driven approach". In: *Frontiers in Human Neuroscience* 8, p. 336.

Vitrià, Jordi (2025). *Transparency and Explainability*. Lecture slides, Ethical Data Science, MSc in Fundamental Principles of Data Science, Universitat de Barcelona.

Zhang, Y. et al. (2020). "Clustering analysis of TBI patients in a web-based cognitive rehabilitation platform". In: *Frontiers in Neurology* 11, p. 812.