

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE FÍSICA



Data Mining applied to Neurorehabilitation Data

Maria Salomé Coimbra Carmelo

Mestrado Integrado em Engenharia Biomédica e Biofísica
Perfil de Engenharia Clínica e Instrumentação Médica

Dissertação orientada por:

Paloma Chausa, Grupo de Bioingeniería y Telemedicina, Universidad Politécnica de
Madrid, Madrid, España
Prof. Dr. Nuno Matela, Instituto de Biofísica e Engenharia Biomédica, Universidade de
Lisboa, Lisboa, Portugal

2017

*“A felicidade pode ser encontrada, mesmo nos momentos mais sombrios, se alguém se lembrar
apenas de acender a luz.”*

*“Happiness can be found, even in the darkest of times, if one only remembers to turn on the
light.”*

J.K. Rowling

Acknowledgments

I would first like to thank to my external supervisor, Paloma Chausa. This report is being delivered in the first place because you accepted me to work with you. I felt a huge support during the whole internship and I am very grateful for everything.

Ao meu supervisor interno, o Professor Nuno Matela, queria agradecer não só pelo apoio, pela disponibilidade e pela compreensão (especialmente na fase final) como por, principalmente, ter aceitado ser meu orientador e mostrar logo de início entusiasmo pelo tema em que ia fazer a minha investigação. Um gigante obrigada.

También quería dar las gracias a todas las otras personas de GBT que me han acompañado los siete meses que estuve haciendo las practicas. A Profesor Enrique y a Profesora Elena les agradezco por su apoyo. A mis compañeros en la oficina, a los dos Joses, a Gemma, a Nacho, a Sandra, a Ana, a Estefanía, a Elisa: gracias por las charlas en los desayunos, pelas risas en las comidas, por todo el apoyo, que ha sido fundamental.

I want to thank to Prof. Dra. Ernestina Menasalvas for her availability and to Juan Tuñas for showing me the wonders of SPSS.

Quería hacer un agradecimiento muy especial, de todo corazón, a mis dos jefes de proyecto en Abassy, Marta y Leuvs. Si no fuera por vuestra bondad, nunca habría entregado esta tesis dentro del plazo.

A las dos personas que más me hicieron sentir que había elegido bien el sitio para hacer mis prácticas, Clara y Nacho, muchas gracias por vuestra amistad y por haberme integrado con tanta facilidad.

Às minhas duas meninas, Rita e Margarida, bem-dito seja o dia em que perguntei se alguém queria vir comigo passear. Obrigada pelas escapadinhas, pelas palhaçadas e por me terem feito aproveitar muito melhor o Erasmus. Tenho imensas saudades vossas! Também ao grupinho de portugueses que conheci logo no início da minha estadia: à Vânia, ao Paulo e à Joana, deram-me a motivação que precisava para começar bem a minha viagem, obrigada.

Às doze pessoas que me fizeram sentir em casa todo o tempo em que estive fora: Mariana, Inês, Ritas, Tatiana, Tomás, Marta, Francisca, Filipe, Tânia, obrigada por tudo. Um obrigada extra às três que me deram abrigo nas suas respetivas cidades para que eu pudesse aproveitar para cumprir este sonho que tenho de viajar e conhecer coisas novas.

Quería agradecer, com especial carinho, a toda a minha família próxima pelas palavras de apoio constantes, incluindo ao Acácio, pelo carinho desde o primeiro dia e por tomar tão bem conta de um dos diamantes da minha vida. Claro que esta família não poderia deixar de incluir também os meus três amigos de quatro patas que com as suas brincadeiras me impediram de passar um único dia sem rir.

À minha irmã, Madalena, obrigada por não me deixares esquecer que o amor deve ultrapassar todas as barreiras. Por nunca abdicares das coisas em que acreditas e por teres sido e continuares a ser o primeiro número na minha lista de contactos.

Aos meus pais, quero não só agradecer-vos como dedicar-vos este trabalho. Tudo o que sou, tanto a nível pessoal como profissional, é graças a vocês e só espero um dia poder recompensar-vos. Mãe, o teu apoio sempre incondicional, a tua alegria imensa, a forma linda como encaras a vida, tudo isso me deu forças para nunca desistir. És o meu ídolo a seguir. Pai, obrigada por nunca deixares de acreditar que eu sou capaz, por estares sempre presente quando preciso e por teres sempre uma palavra para me dar. És o meu porto de abrigo. Nunca existirão palavras suficientes neste mundo para vos agradecer.

Resumo

Apesar de não serem a principal causa de morte no Mundo, as lesões cerebrais são talvez a principal razão de existirem tantos casos de pessoas que veem a sua vida quotidiana afetada. Tal acontece devido a grandes dificuldades cognitivas que podem ser derivadas de um acidente de automóvel, de uma queda, da presença de um tumor, de um acidente vascular cerebral, da exposição a substâncias tóxicas ou de uma outra qualquer situação que tenha envolvido uma lesão do cérebro. De entre este tipo de lesões podem considerar-se aquelas que são provenientes de traumas por forças externas, ou seja, as chamadas lesões cerebrais traumáticas ou traumatismos crânio-encefálicos. É precisamente em pessoas que sofreram uma lesão desse tipo que se foca este estudo. Em pessoas que, depois dessas lesões, foram sujeitas a um tratamento de neuro reabilitação. Este tratamento, baseado na realização de tarefas especialmente desenhadas para estimular a reorganização das ligações neuronais, permite que os doentes tenham a possibilidade de voltar a conseguir realizar tarefas do dia-a-dia com a menor dificuldade possível. O objetivo da realização destas tarefas é a estimulação da capacidade de plasticidade cerebral, responsável pelo desenvolvimento das conexões sinápticas desde o nascimento e que permite ao cérebro voltar a estabelecer o seu funcionamento normal depois de uma lesão. Naturalmente, o grau de afetação de uma pessoa depende do tipo de lesão e tem uma grande influência não só no tempo de recuperação física e mental, como também no seu estado final.

O estudo documentado neste relatório de estágio constitui um meio para atingir um objetivo comum a outros trabalhos de investigação nesta área; pretende-se que os tratamentos de neuro reabilitação possam vir a ser personalizados para cada paciente, para que a sua recuperação seja otimizada. A ideia é que, conhecendo alguns dos dados pessoais de um doente, considerando informação sobre o seu estado inicial e através dos resultados de testes realizados, seja possível associá-lo a um determinado perfil disfuncional, de características bastante específicas, para o terapeuta poder adaptar o seu tratamento.

O Institut Guttmann, em Barcelona, foi o primeiro hospital espanhol a prestar cuidados a doentes de lesões medulares. Hoje em dia, um dos seus muitos projetos chama-se GNPT Guttmann NeuroPersonalTrainer e leva a casa dos seus doentes uma plataforma que lhes permite realizar as tarefas definidas pelos terapeutas, no âmbito dos seus tratamentos de neuro reabilitação. Dados desses doentes, incluindo informação dêmica e resultados de testes realizados antes e depois dos tratamentos, foram cedidos pelo Institut Guttmann ao Grupo de Biomédica e Telemedicina (GBT) sob a forma de bases de dados. Através da sua análise e utilizando ferramentas de Data Mining foi possível obter perfis gerais de disfunção cognitiva e descrever a evolução desses perfis, o principal objetivo desta dissertação.

Encontrar padrões em grandes volumes de dados é a principal função de um processo de Data Mining, tratando o assunto de forma muito genérica. Na verdade, é este o conceito utilizado quando são abordados temas de extração de conhecimento a partir de grandes quantidades de dados. Há diversas técnicas que o permitem fazer, que utilizam algoritmos baseados em funções estatísticas e redes neuronais e que têm vindo a ser melhoradas ao longo dos últimos anos, desde que surgiu a primeira necessidade de lidar com grandes conjuntos de elementos. O propósito é sempre o mesmo: que a análise feita a partir destas técnicas permita converter a informação oculta dos dados em informação que pode ser depois utilizada para caracterizar populações, tomar decisões ou para validar resultados. Neste caso, foram utilizados algoritmos de Clustering, um método de Data Mining que permite obter grupos de elementos semelhantes entre si, os clusters, considerando as características de cada um destes elementos.

Dados de 698 doentes que sofreram um traumatismo craniano e cuja informação disponível nas bases de dados fornecidas pelo Institut Guttmann satisfazia todas as condições necessárias para serem considerados no estudo, foram integrados num Data Warehouse - um depósito de armazenamento de

dados - e depois estruturados. A partir de funções criadas em SQL - a principal linguagem de consultas e organização de bases de dados relacionais - foram obtidas as pontuações correspondentes aos testes realizados pelos doentes, antes do início do tratamento e depois de este ser terminado. Estes testes visaram avaliar, utilizando cinco diferentes níveis de pontuação correspondentes a cada grau de afetação (0 para sem afetação, 1 para afetação suave, 2 para afetação moderada, 3 para afetação severa e 4 para afetação aguda), três funções estritamente relacionadas com o nível cognitivo, a atenção, a memória e algumas funções executivas. As pontuações obtidas para cada uma das funções constituem uma média ponderada da pontuação cada uma das subfunções (atenção dividida, atenção seletiva, memória de trabalho, entre outras), calculadas por pelo menos um dos 24 itens de avaliação a que cada pessoa foi sujeita. De seguida, foram determinados os grupos iniciais e finais, recorrendo a uma ferramenta muito útil para encontrar correlações em grandes conjuntos de dados, o software SPSS. Para determinar a constituição dos clusters iniciais foi aplicado um algoritmo de Clustering designado K-means e, para os finais, um outro denominado TwoStep. A principal característica desta técnica descritiva de Data Mining é a utilização da distância como medida de verificação da proximidade entre dois elementos de um cluster. Os seus algoritmos diferem no tipo de dados a que se aplicam e também na forma como calculam os agrupamentos de elementos. Para cada um dos clusters, e de acordo com cada uma das funções, foi observada a distribuição das pontuações, através de gráficos de barras. Foram também confrontados ambos os conjuntos de clusters para se poder interpretar a relação entre eles.

Os clusters, que neste contexto correspondem a perfis de afetação cognitiva, foram validados, e concluiu-se que permitem descrever bem a população em estudo. Por um lado, os seis clusters iniciais determinados representam de uma forma fiel, e com muito sentido do ponto de vista clínico, os conjuntos de pessoas com características suficientemente definidas que os distinguem entre si. Já os três clusters finais, usados para retratar a população no final do tratamento e analisar as evoluções dos pacientes, retratam perfis bastante opostos, o que permitiu, de certa forma interpretar com maior facilidade para que pacientes o efeito da neuro-reabilitação foi mais ou menos positivo.

Alguns estudos citados no estado de arte revelaram que algumas variáveis são suscetíveis de influenciar o estado final de um doente. Aproveitando a existência de dados suficientes para tal, foi observado se, tendo em conta os clusters finais, se poderia fazer alguma inferência sobre o efeito de algumas das variáveis – incluindo a idade, o nível de estudos, o intervalo de tempo entre a lesão e o início do tratamento e a sua duração – em cada um destes. No final, considerando apenas as pontuações dos testes em cada função, antes e depois dos tratamentos, foram analisados e interpretados, recorrendo a gráficos, os desenvolvimentos e a evolução global de cada doente. Como desenvolvimentos possíveis, foram tidos em conta os casos em que houve melhorias, agravamentos e também os casos em que os doentes mantiveram o seu estado. Fazendo uso da informação sobre a forma como evoluíram os pacientes, foi possível verificar se, de facto, utilizando apenas os valores das pontuações obtidas nos testes, se poderia ou não confirmar que outras variáveis poderiam ter efeitos na determinação do estado final de um paciente. Os gráficos obtidos demonstraram que há diferenças muito subtis considerando algumas das variáveis, principalmente entre os dos doentes que melhoraram e os dos doentes que viram a sua condição agravada. Concluiu-se que o facto de os clusters agruparem pessoas com tipos de evolução diferentes levou a que o efeito de outras variáveis se mostrasse muito disperso.

O tipo de investigação sugerido para futuros desenvolvimentos inclui: (i) o estudo das outras hipóteses de perfis apresentados pelo software usado (SPSS); (ii) considerar os diferentes aspetos das funções avaliadas a um nível mais detalhado; (iii) ter em conta outras variáveis com possíveis efeitos no estado final de um doente.

PALAVRAS-CHAVE: Data Mining – Neuro reabilitação – Clusters – Perfis Disfuncionais

Abstract

Although they are not the leading cause of death in the world, brain injuries are perhaps the main reason why there are so many cases of people who see their daily lives affected. This is due to the major cognitive difficulties that appear after brain lesion. Brain injuries include those that are derived from traumas due to external forces – the traumatic brain injuries. This study is focused in people who, after these injuries, were subjected to a neuro rehabilitation treatment. The treatment, based on tasks specially designed to stimulate the reorganization of neural connections, allows patients to regain their abilities to perform their everyday tasks with the least possible difficulty. These tasks aim to stimulate the brain plasticity capacity, responsible for the development of synaptic connections which allows the brain to re-establish its normal functioning after an injury.

The study documented in this internship report constitutes another step for a major goal, common to other studies in this area: that neuro rehabilitation treatments can be personalized for each patient, so that their recovery is optimized. Knowing some of the personal data of a patient, considering information about their initial state and through the results of tests performed, it is possible to assign a person to a certain dysfunctional profile, with specific characteristics and for the therapist to adapt treatment.

One of his many projects of the Institut Guttmann (IG) is called GNPT Guttman NeuroPersonalTrainer and brings into its patients' home a platform that allows them to perform the tasks set by the therapists in the context of their neurorehabilitation treatments. Data from these patients, including clinical information and test results performed before and after the treatment, were provided by the IG to the Biomedical and Telemedicine Group (GBT) as databases. Through its analysis and using Data Mining techniques it was possible to obtain general profiles of cognitive dysfunction and to characterize the evolution of these profiles, the objective of this work.

Finding patterns and extracting knowledge from large volumes of data are the main functions of a Data Mining process. An analysis performed using these techniques enables the conversion of information hidden in data into information that can later be used to make decisions or to validate results. In this case, Clustering algorithms, which build groups of elements with the similar characteristics called clusters, were used. Also, data from 698 patients who suffered brain trauma and whose information available in the databases provided by the IG satisfied all the conditions considered necessary were integrated into a Data Warehouse and then structured. The scores corresponding to the tests performed before and after the treatment were calculated, for each patient. These tests aimed to evaluate, using five different punctuation levels corresponding to each degree of affectation, three functions strictly related to cognitive level: attention, memory and some executive functions (cognitive processes necessary for the cognitive control of behavior).

The initial and final clusters, representing patients' profiles, were determined, using the SPSS software. The distribution of the scores over the clusters was observed through bar graphs. Both groups of clusters were also confronted to interpret the relationship between them. The clusters, which in this context correspond to profiles of cognitive affectation, were validated, and it was concluded that, at this moment, they represent well the state of patients under study. As some variables, like age and study level, are likely to influence the final state of a patient, it was observed if, given the final clusters, some inference could be made about the effect of those variables. No valuable conclusions were taken from this part. Also, considering the tests scores, patients' evolution was identified as improvements, aggravations and cases where the conditions is maintained. Using that information, conclusions were extracted, regarding the population and the variables effect. The plots obtained allowed us to correctly describe the patients' evolution and also to see if the variables considered were good descriptors of that evolution. A simple interpretation from of the facts allows to conclude that the calculated are good general, but not perfect descriptors of the population. The type of research suggested for future developments includes: (i) the study of the other hypothesis of profiles presented by the Data Mining software; (ii) consider the different aspects of the functions evaluated at a more detailed level; (iii) take into account other variables with possible effects on describing the final state of a patient.

KEYWORDS Data Mining – Neurorehabilitation – Clusters – Dysfunctional Profiles

Contents

Acknowledgments	i
Resumo.....	ii
Abstract.....	iv
List of Figures.....	vii
List of Tables	ix
List of Abbreviations	x
Preface.....	1
1 Introduction.....	2
2 Background	3
2.1 Brain Injury	3
2.2 Neurorehabilitation	4
2.2.1 Cognitive Functions	4
2.3 The Cognitio Project	5
2.4 The Guttman, NeuroPersonalTrainer® (GNPT).....	6
2.5 The Intelligent Therapy Assistant (ITA) algorithm	7
2.6 Data Mining applied to Neurorehabilitation	8
2.6.1 Clustering techniques	10
2.6.1.1 Clustering algorithms	11
2.6.1.2 Clustering and segmentation	13
2.7 Related work: DM techniques in Brain Injury outcome prediction.....	13
2.7.1 Influence of other variables in BI outcome prediction	16
2.7.2 Data Mining techniques used for other diseases or conditions	16
3 Materials and Methods.....	19
3.1 Data Sources	19
3.1.1 Guttman Clinical Database	19
3.1.2 Guttman NeuroPersonalTrainer Database.....	19
3.2 Software (MySQL, R, SPSS)	20
3.3 CRISP-DM Methodology	20
3.3.1 Data Understanding	21
3.3.1.1 The Clinical Process	21
3.3.1.2 Neuropsychological Assessment	21
3.3.2 Data Preparation	23
3.3.2.1 ETL: Integration Procedure	24
3.3.2.2 The Data Warehouse	24
3.3.2.3 Getting the Pre and Post test scores	25
3.3.3 Data Analysis	27
3.3.3.1 Finding Cognitive Profiles	27
4 Results	32
4.1 Finding the Initial and Final Cognitive Profiles	32
4.2 Distribution of tests scores by the Initial and Final Profiles	42
4.3 Relation between the Initial and Final Profiles	45
4.4 Influence of other variables in the evolution of a patient	47
4.5 Improvement, Worsening and Maintenance	50
4.5.1 Factors that may influence the type of evolution	52
5 Discussion	56

6	Conclusion and Future Work	60
7	References	61
	Appendix	65

List of Figures

Figure 2.1 - The Rehabilitation Process followed by the Guttman, NeuroPersonalTrainer, taken from [15]	7
Figure 2.2 - The process of attribution of a cognitive profile to a patient. Taken from [15]	8
Figure 2.3 – Examples of partitioned Clustering (on the left) and hierarchical Clustering (on the right) [22]	10
Figure 2.4 - Schematic representation of the steps taken by a k-Means algorithm	11
Figure 2.5 – Schematic representation of a Cluster Feature Tree, taken from [27]	12
Figure 2.6 - Schematic representation of the steps taken by an agglomerative hierarchical clustering method.	12
Figure 2.7 - Predictive decision tree model used in [36] showing GCS as the best predictor for 1 month after injury	15
Figure 2.8 - Algorithm used for learning the conditional probabilities and for the prediction of the patient's label. Figure taken from [37]	15
Figure 2.9 - Comparison of the accuracy between three classifier methods: Naïve Bayes, Decision Trees and Average k-Nearest Neighbor for a different number of attributes (12 and 13). Graphic taken from [42]	17
Figure 2.10 - The Maitreya Framework used in the study carried out in [46]	18
Figure 3.1 - Schematic representation of a common CRISP-DM flow	20
Figure 3.2 - Schematic representation of the clinical process flow which ends with the computation of the global improvement level.	21
Figure 3.3 - Screenshot of the MySQL Workbench with all the tables displayed on the left and showing the kind of data contained in one table (administrative_data)	25
Figure 3.4 - Screenshot of an Excel file containing all the scores, for each treatment, for each pre and post function.	27
Figure 3.5 – Screenshot of the IBM SPSS Modeler's initial Stream	28
Figure 3.6 - Example of how to start building a flow in IBM SPSS Modeler with a Source and Type node	28
Figure 3.7 - Screenshot of the box where it is possible to edit all the parameters from the source data	29
Figure 3.8 - Screenshot of the box where it is possible to edit all the parameters related to the type and format of the source data.	29
Figure 3.9 – Screenshot of the box to select the models to be used by SPSS to determine the best clusters for the source's data	30
Figure 3.10 – Screenshot of the box used to set more specific parameters for the Kohonen node, like the size of the two-dimensional output map.	30
Figure 4.1 - Screenshot of the box used to select which one of the calculated SPSS models shall be used to build the initial clusters, containing information about each one of them.	33
Figure 4.2 - Graphics showing the characteristics of the chosen k-Means model with 6 clusters	33
Figure 4.3 - Information about each one of the six clusters regarding the Predictor importance for every input	34
Figure 4.4 - Example of a cell distribution plot where it is possible to look at the distribution of the scores over the Attention function and compare it with the distribution over the entire cluster. .	35
Figure 4.5 - Screenshot of a table provided by SPSS that shows the assignation of the initial clusters for each treatment, according to each function.	35
Figure 4.6 - Cell distribution plots displaying the distribution of the patients over the scores for Attention (on top), Memory (in the middle) and Execute Functions (bottom for each cluster). ..	36
Figure 4.7 - Screenshot of the box used to select which one of the calculated SPSS models shall be used to build the final clusters, containing information about each one of them.	38
Figure 4.8 - Graphics showing the characteristics of the chosen TwoStep model with 3 clusters	38
Figure 4.9 - Information about each one of the three clusters regarding the Predictor importance for every input	39

Figure 4.10 - Table provided by SPSS that shows the assignation of the final clusters for each treatment, according to each function	39
Figure 4.11 - Cell distribution plots displaying the distribution of the scores for Attention (on top), Memory (in the middle) and Execute Functions (bottom).....	40
Figure 4.12 - Resultant table from the exclusive selection of the patients belonging to initial cluster 1.	41
Figure 4.13 - SPSS flow used to obtain the histograms for each function, considering only the patients from the initial cluster 1	42
Figure 4.14 - Histograms showing the number of patients for each score and Post function, for cluster 1	42
Figure 4.15 - Distribution of the scores obtained in the Attention Post function over the initial clusters	43
Figure 4.16 - Distribution of the scores obtained in the Memory Post function over the initial clusters	43
Figure 4.17 - Distribution of the scores obtained in the E.F Post function over the initial clusters	44
Figure 4.18 - Distribution of the scores obtained in the Attention Pre function over the final clusters	44
Figure 4.19 - Distribution of the scores obtained in the Memory Pre function over the final clusters	44
Figure 4.20 - Distribution of the scores obtained in the Executive Pre functions over the final clusters	45
Figure 4.21 - Distribution graphic of the final profiles (3) by the initial profiles (6).....	46
Figure 4.22 - Distribution of the gender, considering the initial (left) and final (right) clusters.....	47
Figure 4.23 - Distribution of the age classes of the patients over the final clusters	48
Figure 4.24 - Distribution of the studies of the patients over the final cluster	48
Figure 4.25 - Distribution of the break classes of the patients over the final clusters	49
Figure 4.26 - Distribution of the number of tasks of performed patients over the final clusters.....	49
Figure 4.27 - Distribution of the duration of the patient's treatment over the final clusters	50
Figure 4.28 - Excel file showing by colors, the patients from cluster 1 who improved, who got worse or did not evolve, in each of the functions.	50
Figure 4.29 - Excel file showing some of the patients who improved and their correspondent initial and final scores and clusters.....	51
Figure 4.30 - Excel file showing some of the patients who got worse and their correspondent initial and final scores and clusters.....	51
Figure 4.31 - Excel file showing some of the patients who did not show evolution and their correspondent initial and final scores and clusters.....	52
Figure 4.32 - Histograms crossing the variables Age, Study level and Gender, using data from the patients who had a global improvement of 1.	52
Figure 4.33 - Histograms crossing the variables Age, Study level and Gender, using data from the patients who had a global improvement of 2.	53
Figure 4.34 - Histograms crossing the variables Age, Study level and Gender, using data from the patients who had a global improvement of 3.	53
Figure 4.35 - Histograms showing the distribution of people according to the time they took to start the treatment after the lesion (in days), for each of the three evolution types.	54
Figure 4.36 - Histograms showing the distribution of people according to duration of their treatment (in days), for each of the three evolution types.....	54
Figure 4.37 - Histograms showing the distribution of people according to the number of tasks they performed in their treatment, for each of the three evolution types.....	55
Figure 5.1 – Schematic representation of the initial clusters and final clusters and the relation between all of them	57

List of Tables

Table 3.1 - Representation of the possible results for the global evolution of a patient considering the evolution in each one of the functions.	23
Table 3.2 – Simple representation of the table called “tests_results”, with information about the treatment, the test and the normalized result obtained.....	26
Table 4.1 - Schematic colored representation of the score’s distribution over each one of the 6 profiles, for each function	36
Table 4.2 - Schematic colored representation of the score’s distribution over each one of the 6 profiles, for each function	40

List of Abbreviations

ABI	Acquired Brain Injury
AD	Alzheimer's Disease
AKNN	Average k-Nearest Neighbor
BI	Brain Injury
BN	Bayesian Network
BM	Bayesian Method
BB.DD	Database
CART	Classification and Regression Tree
CFT	Cluster Feature Tree
CHCT	Canadian head CT
CLARA	Clustering LARge Applications
CLARANS	Clustering LARge ApplicatioNS
CPXR	Contrast Pattern Aided Regression
CT	Computerized Tomography
DM	Data Mining
DW	Data Warehouse
DT	Decision Trees
EM	Expectation-Maximization
ESKD	End Stage Kidney Disease
ETL	Extract, Transform, Load
ICF	International Classification of Functioning, Disability and Health
IG	Institut Guttmann
ITA	Intelligent Therapy Assistant
GA	Genetic Algorithm
GCS	Glasgow Coma Scale
GNPT	Guttmann NeuroPersonalTrainer
KDD	Knowledge Discovery in Databases
MEG	Magnetoencephalography
MRI	Magnetic Resonance Imaging
NN	Neural Networks
PAM	Partitioning Around Medoids
QEEG	Quantitative Analysis of the Electroencephalography
RF	Random Forest
SFP	Simple Feature Picker
SVM	Support Vector Machine
TIRP	Time Intervals Related Patterns
TBI	Traumatic Brain Injury

Preface

The Grupo de Bioingeniería y Telemedicina, or GBT, as an abbreviation, is a research group of the Universidad Politécnica de Madrid (UPM) founded in 1983 and it is the largest telemedicine research center in Spain. The main common aim of the research works carried out in this group is the technological development regarding bioengineering, what includes the application of information and communication technologies in the biomedicine field. This group is also responsible for the coordination of the Biomedical Engineering Degree at ETSIT (Escuela Técnica Superior de Ingenieros de Telecomunicación), the school from UPM where it is headquartered, in Ciudad Universitaria.

The GBT staff is constituted by over 50 members that include not only professors but doctors, pre-doctoral candidates, assistant researchers and graduate students as well. Almost all of them have a strong background in telecommunication engineering. The group itself maintains partnerships with several sectors, hospitals, clinical institutions and other groups in the area of information and communication technology applied to medicine, what benefits the research.

The GBT's research lines encompass biomedical imaging, diabetes technology, telemedicine and intelligent devices, surgical simulation, planning and image guided surgery, knowledge management and data mining and neurorehabilitation engineering. The proposed work of this dissertation, done in this institution, involves mainly the last two research lines. [1]

The external supervisor of this project was Paloma Chausa, an investigator from the GBT group and the internal supervisor was Prof. Dr. Nuno Matela from Instituto de Biofísica e Engenharia Biomédica, Universidade de Lisboa, Portugal. For purposes of the data-mining application techniques using SPSS, the help of two researchers from the CBT (Centro de Tecnología Biomédica), who are also part of the MiDaS (Minería de Datos y Simulación) research group: Prof. Dra Ernestina Menasalvas and his student in Data Analytics, Juan Tuñas.



ETSIT (Escuela Técnica Superior de Ingenieros de Telecomunicación)

1 Introduction

Acquired brain injuries are very often a major cause of death. They can happen due to traumatic (a car accident, a fall, a very hard concussion) or non-traumatic events (stroke, anoxia). When the person does not die after such kind of an event, it is very much possible that that person will end up with very severe consequences, mental and physical that will affect a big part of his or her life. People who have suffered from these events are expected to go under a neurorehabilitation treatment, to do the possible to go back to their healthiest state. Most commonly, this includes performing tasks that will require the activation and exercise of their brain functions. Some people are not too much affected but there might be people who completely lose their cognitive functions, including their Attention or Memory levels which may decrease significantly or even almost disappear. A person's affectation degree after an accident depends mainly on its severity level. On the other hand, a better or worse response to the treatment depends on that person's age, gender, study level, antecessors, amongst others. Even the time that passes between the day of the injury and starting the treatment or between the day the treatment starts and its ending might have an influence on the final state of the person. It is because of this that it makes sense that the treatment should be personalized in order to obtain the best improvement possible.

This brings us to the main objective of this study: to obtain dysfunctional profiles to characterize the population at the beginning and at the end of the treatment. This is done by extracting information from tests that evaluate three main cognitive functions (Attention, Memory and five Executive Functions), performed before the treatment starts and after the treatment ends. A comparison between the distributions of the two populations allows an analysis of the patient's evolution. After this process is clinically validated, the profiles can be used to customize the therapy. A second analysis was designed to assess the effects of other variables on the evolution and final state of a patient. At the end, the global improvement of the patients, considering only their scores in the assessments, is to be calculated.

To reach our main objective, Data Mining (DM) was used. This tool is used whenever one wants to extract knowledge in a large set of data. As DM is used in this study, a technical objective had to be defined. This brings us to the definition of our DM purpose: to get clusters of patients based in the level of the score tests in three cognitive functions that are considered to be the most important when assessing the patient's state after a brain injury. By using DM, it was also possible to extend the study to access what could be the influence of other variables (age, gender, study level, interval between injury and starting the treatment, duration of the treatment and the number of tasks) in the evolution of a patient. The data from the patients used in this study had to be prepared and shaped so that data mining techniques could be applied and also for the extraction of the final graphics and results. In this document, it is described, step-by-step, how every stage was taken care of. The work here documented is based on data provided by the Institut Guttmann and on the platform they built, together with the GBT, called GNPT Guttmann, NeuropersonalTrainer (described later in this report). Although this dissertation focus on the application of data mining techniques in Neurorehabilitation data, these techniques have also been applied to other diseases, with the same or different approaches. In the Background section, a brief summary of the investigation which has been done in all of these fields is presented, in order to clarify a little bit more the purpose of this Master dissertation work.

2 Background

2.1 Brain Injury

Acquired Brain Injury (ABI) is the global expression used to define one of the most common neurological disorders and it includes all types of brain injuries occurred after birth. ABI may cause multiple temporary or permanent impairments which can be cognitive, physical, emotional or behavioral, in that way leading to a decrease in life quality of the people who suffer from this condition or even to death. It can be caused by either traumatic or non-traumatic events. In the first group, traumatic forces to the head are the main reasons while in the second group the factors are mostly related to illness. These include infections (meningitis or encephalitis triggered by inflammations of the brain), anoxia or hypoxia (due to smoke or carbon monoxide inhalation, exposure to high altitudes, asphyxiation, near drowning, drugs or alcohol abuse, exposure to toxic substances, anaesthetic mishaps and severe asthma or heart attacks), stroke and brain aneurisms, hemorrhages or tumors. [2] [3] [4] ABI is not a genetic or congenital condition and cannot be considered as a degenerative brain condition like Alzheimer's or Parkinson's Disease or even Multiple Sclerosis although these diseases may eventually lead to ABI.

Traumatic Brain Injury (TBI) is a type of acquired brain injury and is a consequence of the action of external physical forces [5]. TBI is associated with several symptoms and disabilities that might be experienced right after the injury or only appear weeks after the event and that may lead to long term incapacities or even death. The predominant causes for this condition have a strong relation with age. For young people, it is known that the most effective causes are motor vehicle crashes (which include autos, trucks, motorcycles, bicycles and pedestrians hit by vehicles) and for older people, falls are the main reason. Other causes include sports, firearms and cutting objects. Studies showed that the range of people that are most susceptible to suffer from TBI is between 15 and 24 years old or people with more than 65 years. [6]

Traumatic injuries can occur through two different mechanisms: open head injuries and closed head injuries. Open head injuries consider damages caused by the perforation of the skull by an object. The object may end up getting in contact with the brain directly or cause a piece of bone from the skull to penetrate it. Closed head injuries are the most frequent ones: they do not involve the crushing of the skull and include acceleration followed by deceleration injuries, where the brain still has a forward momentum when it is forced to stop, then hitting the inner surface of the skull. [7] This happens in motor vehicle accidents, falls and in any cases where there is a violent shaking of the head. [8] What aggravates even more the closed head injury picture is that in these cases it is possible that a bleeding or a swelling may occur which increases the intracranial pressure and promotes an injury of the brain cells.

TBI, in almost all the cases, affects abilities, namely cognitive functions that include attention, memory, communication, processing and understanding of information and many others. In a TBI event it is possible that only one region get affected but it is also possible that the injury is spread to other areas of the brain and most times, the nature of the injury is indefinite. Depending on the region of the injury, the brain functions affected may vary. For example, frontal lobes are responsible for consciousness and emotions and so, if the injury happens in that area, there will possibly be problems regarding attention, feelings control and language expression. Also, occipital lobes are totally related to vision, so if the injury happens to be there, the person might as well have problems with locating objects, writing and reading or even episodes of hallucinations. Although sometimes an injury only affects one step of an

activity which takes place in a determined region in the brain, as there is an interrelation between brain functions, it is not always possible to identify the nature of the injury and what is going to be the accurate prognosis for several months or years. [9]

2.2 Neurorehabilitation

If a person who suffers a brain injury does not end up with the worst prognostic, at least an initial treatment is needed. There are some cases when it needs to be surgical, but there are also cases where the treatment serves to help the person restore their daily life. Neurorehabilitation is a medical process which helps the recovery from a nervous system injury and to minimize the side effects from that same injury. This includes cognitive rehabilitation, in which the goal is to reduce the cognitive deficits associated with a brain injury event. The severity degree of a TBI is a signal of short and long-term prognosis. After injuries, in a primary phase the brain might restore the connections by itself – it undergoes through spontaneous recovery. However, there are connections that are not able to go back to their normal state and there is almost always a partial or complete loss of functions. In case of partial loss of brain functions, a training induced rehabilitation is necessary and it has an optimized effect if it has an intensive character, if it is done in the right time after the injury and if the exercises chosen by the therapist are the right ones to stimulate the specific harmed regions of the brain. [10] In rehabilitation, health service professionals work together with brain injury patients to circumvent the fact that there is still no surgical or pharmacological treatment to restore the lost functions and still try to provide them a better life quality. This is only achievable due to an intrinsic property of the nervous system called plasticity. Due to this plastic nature, the brain is able to generate new connections or to modify the existing ones. [11] The reorganization of the neural pathways also lays on the same basic neurobiological processes responsible for the initial behaviors acquired and that means those rearrangements may produce adaptive but also maladaptive responses, which can happen with a higher probability than in a normal brain. Besides that, the great number of heterogeneous injuries and deficits makes it necessary to invest in a customized neurorehabilitation treatment.

2.2.1 Cognitive Functions

Similar to what has been considered in other studies, there are three main functions that Guttmann Institute considers when performing the assessment of a patient's improvement level. Those include *Attention* that can be sustained, selective or divided, *Memory* that can be either verbal/visual or working memory and the *Executive Functions*, divided in five subfunctions, being that planification, inhibition, flexibility, sequencing and categorization. For the final purposes of this work, only the three main functions were the ones considered, although the next step will be to perform the studies considering also the subfunctions. The fact is that each one of the subfunctions considers a different aspect of the main ones and enables a more personalized study of the affectation levels.

Concerning Attention, the sustained attention is about keeping a stable response during an unstopping and repetitive ability, which means, keep the focus in a task for a continuous period without being distracted, for example, keeping focused during a long meeting. Selective attention refers to selecting which tasks to pay attention to. Conscientious act of focusing and avoid distractions from stimulus not only external, as noises, but also unnecessary thoughts. Example: focus on the teacher's voice in a room full of noisy people. Divided attention requires a simultaneous answering to multiple tasks. When one

needs to process two or more responses or react to two or more different requests at the same time, divided attention is used. For example: check email while participating on a meeting and drive while listening to music or talking and so on. Truth is, what a person does is rapidly alternate between tasks, given the impression of doing them simultaneously.

Having a good visual or verbal memory implies making an immediate recall of the characteristics of a given object (or form) or verbal terms (words), respectively. After being presented to visual or verbal information, the person remembers that specific information in short or long term. For example: telling someone about a conversation call that happened just a few seconds before or remembering a movie. Working memory is the kind of memory used when it is needed to store a determined amount of information, related to a certain topic, for as long as it is needed. This includes writing a master dissertation, studying for a test or remembering to take out the garbage in the morning.

Regarding the Executive functions, there are a lot of theories about which are the ones that should be considered. In this study five were contemplated and they include Planification, Inhibition, Flexibility, Sequencing and Categorization. Planification, highly connected with organization, is the capacity of coordinating thoughts in order to achieve a goal. An example would be to plan a vacation week. Inhibition relates to the ability of resisting to impulses or controlling responses. This happens, for example, when a person tries to stop biting the nails in a stressful situation. Having cognitive Flexibility, which also has a relation with creativity, means to be willing to accept different ideas. For example, when composing music. Sequencing is about the capacity of following or arranging a set of tasks in order. This may happen when someone is learning a new idiom, one of the main tasks is to learn the order of the spoken language. At last, Categorization is the aptitude to find similarities and differences in things, putting them in different groups accordingly to those disparities. For instance, babies are highly motivated to place different toys, considering their shape, into boxes.

2.3 The Cognitio Project

An already concluded project, conducted by the GBT center, the Instituto de Investigación de Inteligencia Artificial del CSIC and the Hospital de Neurorehabilitación Institut Guttmann (IG) in Barcelona, called the COGNITIO project, was focused on the optimization of cognitive rehabilitation in traumatic brain injury. The research works done were based in the fact that a patient-personalized treatment, with constant monitoring and based on clinical assessments, has better effects than a general treatment. The four main objectives of this project included research regarding the physio-pathological mechanisms involved in ABI cognitive rehabilitation, development of new neuro-image techniques to classify and categorize structural injuries from ABI and automatic learning techniques, and to also use knowledge inference, data mining and multi-parametric analysis for optimization of treatments. This work is included in the last objective. Also, it is very common for cognitive and structural disability profiles to be missing. For that reason, one of the first aims has been to propose a cognitive dysfunctional profile based on neuropsychological knowledge and medical imaging studies. Thus, in [12] an ABI dysfunctional profile containing theoretical, structural and neuropsychological information taken from neuroanatomical structures, cognitive functions, neuropsychological assessment data and medical imaging is generated. The authors present a conceptual framework to define the profile and use a KERM (Knowledge Elicitation and Representation Model) system to gather patients' information based on theoretical models and *a priori* knowledge regarding cognitive processes. Both the framework and this system favor the improvement of personalized treatments and, they also contribute to provide a better body of ABI neurorehabilitation knowledge. The COGNITIO project was carried out from 2013 to 2015

and it was highly involved with the development of the Guttman, NeuroPersonalTrainer® platform, presented in the next section.

2.4 The Guttman, NeuroPersonalTrainer® (GNPT)

The Guttman, NeuroPersonalTrainer® (GNPT) was developed by a research team from the Guttman Institute and the GBT and it is the second generation of an already existing tele-rehabilitation platform called PREVIRNEC. GNPT is, by definition, a telemedicine cognitive rehabilitation platform that provides neurorehabilitation services to people who suffered from ABI. It comprises a web application for therapists, who use it to plan and configure the treatments and another application, for patients, who are this way capable of executing the tasks assigned by the therapists and then send back the results to the server once they have finished the set [13]. It was built based on the patients and therapists needs and it came out as a result of a translational research and transfer of technology process in the ICT (Information and Communication Technology) field [14]. The traditional rehabilitation process depends on several factors, namely, it requires a neuropsychologist to supervise the procedures continuously and the patients are needed to go to the clinical center and perform the tasks and tests. The characteristics of the platform are exposed in [11] where the platform developers also established a comparison between this platform and the traditional rehabilitation system. Thus, although the target of this platform are the people with cognitive impairment after ABI, GNPT has also been used in patients with cognitive impairment due to aging or dementia or even in children with disorders of cognitive development. The platform helps the therapists to configure and schedule personalized rehabilitation sessions, to continuously monitor the patient's performance in tasks and to take decisions having into account the information gathered by a decision support system, this way improving treatment's effectiveness. Even better, the fact that it is a telemedicine platform, allows the therapists to readjust the treatments even if the patients are at home or at daily centers and not only in neurorehabilitation centers. The patients can even perform all the tasks requested by their therapists in an asynchronously way. In other words, there is not a need for the therapist to be present when the patient performs the exercises. The results are recorded by the platform and the therapist can have access to them later.

The GNPT starts with the assignment of a therapist to a patient. The therapist is responsible for performing a battery of initial tests to assess three cognitive functions: attention, memory and executive functions. The tests items are normalized onto the International Classification of Functioning, Disability and Health (ICF) which allows comparisons with other results and are then stored in the system as Pre tests. The patient's cognitive profile is computed using these tests, recurring to Data Mining clustering and there is a normalization process, in which some demographic information, including the person's age and study level, is used. The personalization of the treatment is based on the input parameters that the therapist chooses when stipulating daily tasks for the patient to execute considering his or her cognitive profile. The results of the tasks are sent back to the therapist, who decides to adapt the difficulty level for the next tasks according to the performance of the patient. The tasks assigned are always designed with the aim to improve the cognitive functions and autonomy and their suitability is evaluated in terms of ranges of scores between 0 (not suitable) and 100 (completely suitable). The platform uses three different score intervals that include an infra-therapeutic range (below 65% of right answers), a therapeutic range (between 65% and 85%) and a supra-therapeutic range (above 85%), where the difficulty level is assessed to be too high, appropriate and too low, respectively. This allows the system to readjust the difficulty level for the following task, automatically. After this rehabilitation treatment based on tasks, a final battery of tests is performed to evaluate the patient's improvement

(Figure 2.1). These tests are called the Post tests and are compared to the Pre tests to assess the improvement of the patient.

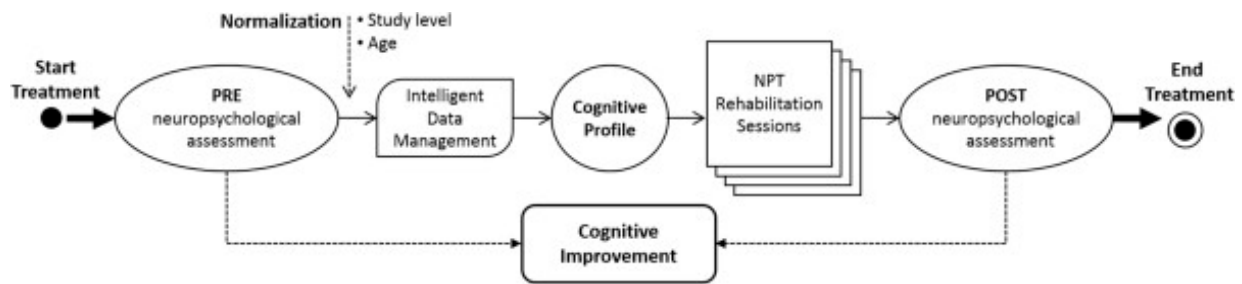


Figure 2.1 - The Rehabilitation Process followed by the Guttman, NeuroPersonalTrainer, taken from [15]

The efficiency of this platform was evaluated after three years after its implementation in Institute Guttman, both in terms of usability and efficiency. The score for usability for each group of users (therapists, patients and administrators) is higher than 70 out of 100 in the SUS (System Usability Scale) and the efficient ratio, in terms of costs and time is 1 to 20 (this rehabilitation platform is cheaper than the traditional face-to-face rehabilitation).

2.5 The Intelligent Therapy Assistant (ITA) algorithm

When using the GNPT platform, as described in the last section, therapists can easily define input parameters to adjust the patients' treatment accordingly to their level of performance. What the ITA, an integrated functionality, allows is an automatization of that parameter selection. It also helps therapists to plan rehabilitation sessions by selecting the best fitting tasks. The ITA algorithm, described and studied in [15], was developed to consider all the information stored in databases and to use it to compute and find the parameters that make the treatment as suitable to the patient as possible. When evaluating the technical viability of this algorithm, the aim was that the outcomes using ITA would be as good as the ones obtained from the manual planification method. Data Mining is used to generate clusters considering data from groups of patients with similar characteristics (see Figure 2.2). This information is useful to compare treatments and describe the evolution of ABI patients. The PRE tests set is composed of 27 tests in total. After normalization of the data, there is a scoring process of the patient's performance in each subfunction with a level between 0 (normality) and 4 (very severe impairment). The system then combines the information about the impairment level calculated and information from previous tasks results and assigns the patient to a certain cluster representing a specific cognitive profile. For the automatic assignation of the tasks, the ITA evaluates all of them (from 0 to 100, as previously said, in function of their usage, improvement and clinical scores), in order to select the most appropriate ones for a specific patient. After that, the system groups the tasks in quartiles, from most suitable (SQ1) to less suitable (SQ4). The ITA is set to schedule sessions in sets of 10 tasks at a time and each set is computed taking into account the previous set results. The system picks tasks randomly from all four quartiles until it is time to end the rehabilitation session. In the same way, there are also Difficulty quartiles, built upon the attribution of a weight to each parameter value in each task, from less difficulty (0) to a level n of difficulty. This is to adjust between easy and more demanding tasks. The suggestions given can always be modified by the therapist anytime. The evaluation of the whole system showed that, by comparing the number of times a task was chosen over a different task using the traditional way or the ITA way, there are no significant differences between the improvement values obtained with both

methods. This is, of course, considering that patients improve if at least one main cognitive function performance has increased and they did not get worse in any of the other functions.

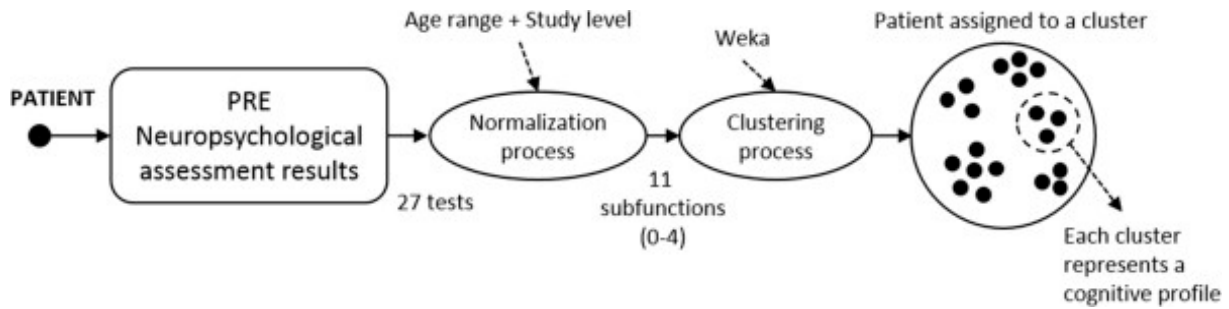


Figure 2.2 - The process of attribution of a cognitive profile to a patient. Taken from [15]

2.6 Data Mining applied to Neurorehabilitation

Data mining (DM), as a powerful instrument to extract information from the data, has been widely used since mid-1990 and surged as a solution for the urgent need to manage the great amounts of existing data. By that time, the advent of the World Wide Web required prevailing tools to mine data in a correct way and to obtain the right information from a big set of sources and databases [19]. Throughout the years, it has been applied to many fields including health care, banking, sales and marketing, education, amongst many others. Work has been done regarding different fronts of data mining application not only in neurorehabilitation but also in other fields. New techniques have been applied and some of the existing ones have been improved. The improvements have several goals, being one of them the reduction of the time needed to compute the data, that way making the process of obtaining results faster. Another improvement in the algorithms used is also, for example, the introduction of the so called “latent” variables that help to increase the accuracy of the descriptive models. In the specific case of TBI, the main goals have been to perform different analysis with different techniques to make good diagnostics and to predict the outcome in patients. Some common classification techniques not only include Classification, as the k-Nearest Neighbor, Decision Trees (DT), Support Vector Machines (SVM), Neural Networks (NN) or Bayesian Methods (BM), but also Regressions, Clustering, Association Rules and Sequential Patterns.

DM is widely used to find tendency lines or to extract patterns in big amounts of data. In the neurorehabilitation field, specifically, it can be used to prove the effectiveness of treatments, to describe disability or to predict the recovery of patients with a brain injury. It can also be used to find the variables that are better descriptors of those patients’ condition or better predictors of their recovery. There is an important difference between descriptive or predictive Data Mining. As the terms themselves indicate, a descriptive DM technique uses data from the past for the analysis (Clustering, Association Rule Discovery), while a predictive DM technique tries to determine future results (Regression, Classification). Usually, all the studies start describing the population and then proceed to the prediction part. That prediction part is always performed with the help of a classifier. It is very common to consider DM as one of the five phases of the Knowledge Discovery in Databases (KDD) process, which is a powerful tool for knowledge extraction in big databases. The Data-Mining phase is always preceded by data preparation and processing and followed by data evaluation, interpretation and implementation. [16] Within the scope of COGNITIO’s project, the multidisciplinary research team responsible for the GNPT has published several papers and conferences regarding the application of data mining to ABI

patients' data in order to predict cognitive outcomes, using machine learning techniques, with the aim of increasing knowledge in the theory of cognitive rehabilitation field. [16], [17], [18])

Many studies have been carried out using all different types of DM techniques. For example, in [17], the authors' aim was to assess if a set of trained classifiers would correctly predict the improvement in all memory, attention and executive functioning areas and to also evaluate the relevance of certain variables using a feature selection method. It was concluded that it is possible to predict with significant accuracy the outcome of a brain injury using variables like age at injury, etiology or neuropsychological evaluation scores, which are all data registered before the treatment. Having information about diagnosis and treatment is also thought to be enough to know the outcome of a patient, considering other previous patient's data. For this study, four machine learning techniques were exploited to prove the predictive value of the selected features: a CART (Classification and Regression Tree) method, a k-Nearest Neighbor method, a Naïve Bayes Classifier and a Support Vector Machine.

In [16], it is used the PREVIRNEC platform database (containing demographic, neuropsychological and tasks executions data) to obtain scores from neuropsychological assessments of the memory subfunctions, executed before and after the treatments in the scale from 0 to 4. Through comparison of those pre and post-rehabilitation scores it is estimated if the patient improved or not. Three DM techniques, named Decision Tree (DT), Multilayer Perceptron (MP) and general Regression Neural Network (RNN) are applied to build prediction models for the outcome in patients with ABI. The DM techniques are then validated with a 10-fold cross validation method. The analysis, based on values of specificity, sensitivity, accuracy and on a confusion matrix as well, showed that the first technique (DT) has the best prediction rates. In [18], other data mining techniques are used to predict the outcome: an Artificial Metaplasticity on Multilayer Perceptron (AMMLP), a Backpropagation Neural Network (BNN) and a C4.5 Decision Tree. The patients studied are also from the PREVIRNEC platform database and the variables considered are the cognitive affectation profile, the rehabilitation tasks result and the patient's outcome after 3 to 5 months of treatment.

As previously mentioned, the work to be developed in this dissertation comes a little bit in the wake of what has already been done and described in this section. In fact, what sets our work apart from what has been done so far is the fact that our intention is to customize therapies to achieve better patient recovery. To do so, it is necessary to characterize the patient very well and take into account how is his/her evolution according to his/her specific therapy. This could be thought as a predictive model, but it has a differential element: it includes information about the therapy, something that is not usually seen in any of the works mentioned in the state of the art. Most of the predictive models only consider the initial state of the patient and the accident data to predict their evolution – leaving the therapy in the background, not giving due importance to the treatment. In our case, the predictive model has not been made yet. Although there are publications release by the IG and the GBT in which predictive models are presented [16][17][18], those were to test algorithms and not to be applied in a real way in the clinic. As a matter of fact, the phase presented in this document is all about correctly describing the data.

Using the patients' database provided by the Guttmann, NeuroPersonalTrainer platform, with Guttmann Institute patients and other external patients, the main objective was to describe the population, to obtain and define a set of dysfunctional profiles. Due to the available GNPT data that includes information not only about the injury but also about the rehabilitation treatments, several different kinds of analysis could be applied. In this case, different Clustering techniques were applied to obtain dysfunctional profiles.

2.6.1 Clustering techniques

As previously mentioned, clusters are groups of elements that have great similarities between them. Thus, clustering methods are ways to ensure that these groups are well constructed: that the elements that share the same features are in the same group and that elements with huge differences remain apart from one another.

Usually, when trying to solve a problem, there is a principle of “divide and conquer”. Clustering techniques rely on that same basis. This means that the goal using these techniques is to break the big data set into smaller pieces that can be explained in a much simpler way and in which patterns can be found more easily. Clustering can, at least, be classified as either an unsupervised or a supervised machine learning approach [20] (there are authors who also consider semi-supervised and weakly supervised systems). In the first case, which is the most common one, the final number of categories, or clusters, is not known. This is because although there are pre-defined clusters (the initial groups that the user determines), the clustering techniques usually need to adapt to the situation and have the aim to find which is the most appropriate structure and number of clusters. Supervised clustering systems have already set at the beginning which are the characteristics of the final clusters, including their number, their type, amongst other features and evolve towards that objective [21]. The following presented techniques are all unsupervised techniques.

While there are a lot of clustering techniques more focused in certain data sets, there are two basic ones that are most commonly used: Partitioned clustering (Figure 2.3 on the left) and Hierarchical clustering (Figure 2.3 on the right).

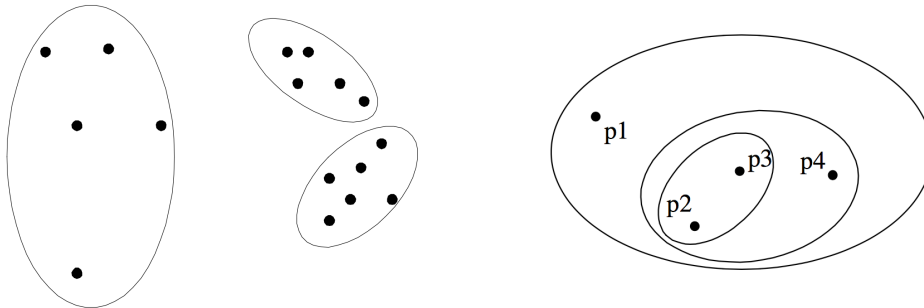


Figure 2.3 – Examples of partitioned Clustering (on the left) and hierarchical Clustering (on the right) [22]

About Partitioned clustering, an element corresponds to exactly one cluster, which means for example that elements 1,2,3 and 4 might belong to cluster 1 but not to any other cluster. In this clustering method, each cluster is indicated to optimize a specific clustering condition independently, being that all of them can optimize the same condition (for example, in k-Means, each cluster tries to minimize the value of the Squared Error Objective Function). On the other hand, in Hierarchical clustering [23], an element can be part of a cluster that is nested in a bigger one, as in a hierarchical tree structure. This means that, for instance, two elements 1 and 2 can be part of one cluster A. That cluster A can also be part of cluster B. Cluster B, besides containing the elements from cluster A, may also include elements 3 and 4. This technique includes two types of algorithms: the agglomerative ones that consider each element as an individual cluster and at each step, join the two clusters that are closest to one another, until only one or k clusters are left and the divisive ones that consider only one single cluster that includes all the members, and that at each step, divides that cluster until each cluster contains an element or until there are k clusters.

2.6.1.1 Clustering algorithms

A clustering algorithm might not apply in the same way to all the data. In fact, there are four parameters that, in a general way, justify the reasons why a certain technique may not apply correctly to a group of data: its size, its dimensionality, the objective function and the structure used [24]. Essentially, it is known that there is a set of characteristics that must be present in a good clustering technique, and those include having the ability to perform well in large data sets with high dimensionality, requiring as minimum information from the user as possible, being able to analyze both single and different attribute types, having the capacity to recognize the best shape for the clusters and eliminating the noise data, the aptitude to ignore the order by which the input records are introduced. At last and ideally, a good algorithm builds clusters that can be used in practical terms and that can be easily interpreted in the end.

Inside the Partitional techniques there are four main algorithms: k-Means, PAM (*Partitioning Around Medoids*), CLARA (*Clustering LARge Applications*) and CLARANS (*Clustering LARge ApplicationNS*). From this group, only the first one was used for the clustering performed in this research. PAM algorithm is an extension of the k-Means, CLARA was developed as a solution for the cost of the PAM algorithm regarding the high number of objects and clusters and CLARANS is nothing more than a combination of the PAM algorithm with sampling methods.

The k-Means algorithm (Figure 2.4) is an iterative clustering procedure that aims to find a local maximum value in each iteration. All the procedure can be described using only four steps: first, the user needs to determine the preferred number of clusters, called as k to assign a n number of elements to. Those k clusters are considered as the initial groups and each one of them is defined by its centroid, which is basically the weighted average (or mean m_i) of its elements. Being E the Squared-Error Objective Function, the main goal of this algorithm is to minimize its value when creating a set of k clusters from n elements. Analyzing the expression, it is observed that the E is calculated by using the distance function to calculate the space between an element x_i and the centroid of a cluster c_j , for all elements and for all initial clusters.

$$E = \sum_{i=1}^k \sum_{j=1}^n \|x_i - c_j\|^2 \quad (2.1)$$

The k-Means procedure starts by randomly assigning each data element to a cluster k . Then, it computes the cluster centroids and according to the results, it re-assigns each point to the closest cluster centroid. In the end, it calculates again the centroids for the new improved clusters. The last two steps are repeated

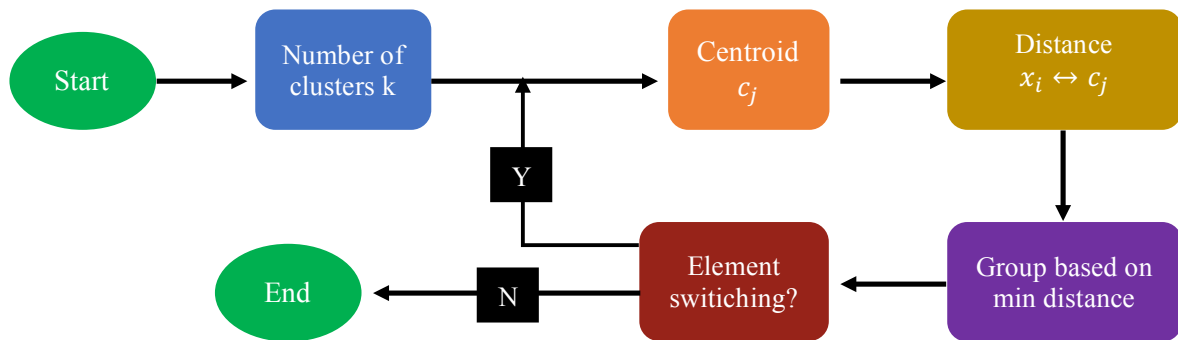


Figure 2.4 - Schematic representation of the steps taken by a k-Means algorithm

until there are no more ways to improve the results. In technical terms, if a result is optimal, there will not be any more elements swapping between two clusters for two consecutive attempts and the value of E will be as small as possible. It is estimated that there will always be more elements (n) than clusters (k). This method is easy to implement and it does not take into consideration the order by which the elements are inputted, what can be thought as advantages. Sadly, there is a high dependency on the initial number of clusters k and the presence of outliers might have a bigger influence on the results than what would be expected.

Regarding hierarchical clustering, the main algorithm is the Two Step. It is a scalable hierarchical type clustering algorithm specially used in data sets with a lot of elements [25]. As the name implies, this technique consists of only two phases: the pre-clustering and the clustering. The first step is to pre-cluster the elements into small groups of data [26]. That means that, in this step, elements are checked one by one in order to be merged or not with the previously formed clusters. In case they are not merged, then they become the first element of a new cluster. This decision is based on the distance criterion. To build the pre-clusters and to review the records, it uses Cluster Feature Tree (CFT).

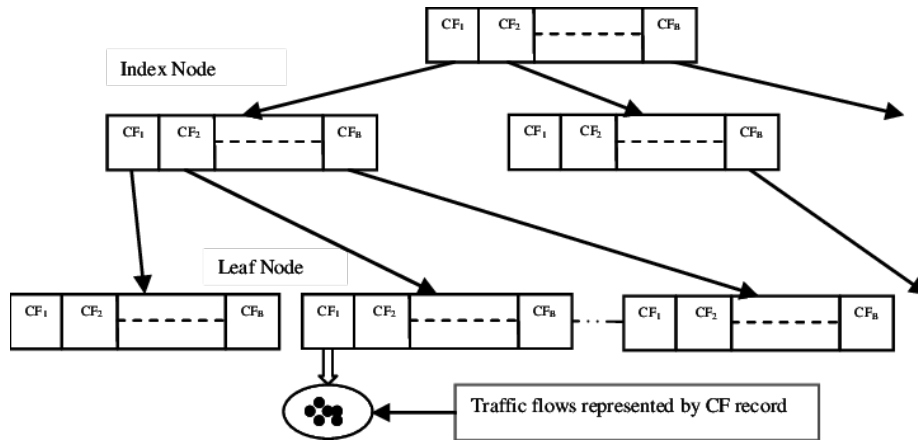


Figure 2.5 – Schematic representation of a Cluster Feature Tree, taken from [27].

A CFT (Figure 2.5) is a tree whose nodes with several entries (CF_1, CF_2, \dots, CF_n , each CF being the number of elements in an entry) are placed in several levels. The leaf nodes, containing the leaf entries, are the sub-clusters. All the other nodes, including the root node (the first one) are used to lead the new elements into their correspondent leaf levels. The basic idea is that when a new element enters through the root node, the closest entry in it guides that element to the closest child node that leads it to the its

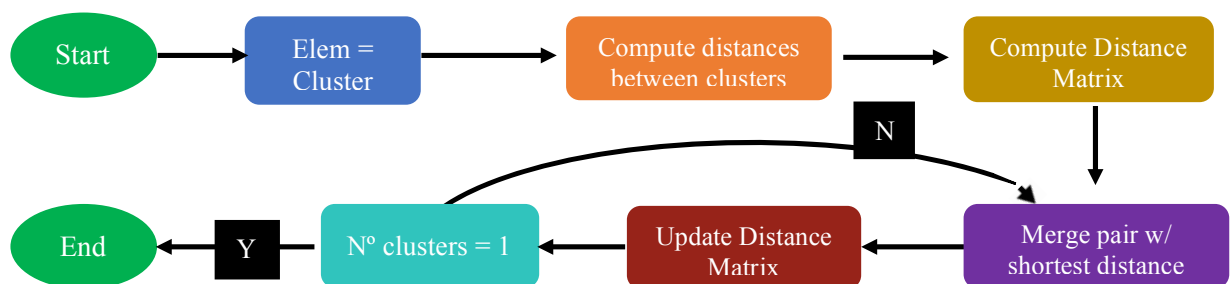


Figure 2.6 - Schematic representation of the steps taken by an agglomerative hierarchical clustering method.

next closest child node and so on, until it reaches the correct leaf entry, which would also be the closest one [28]. The existence of thresholds in this method reassures that no element ends up in a leaf node that is too far from where it should be. Every time a new element is placed in a leaf node, the CF of that node is updated.

The second step of this algorithm is to cluster the small groups formed in the pre-clustering into the required number of clusters. To do this, traditional methods can be used, since the number of clusters to be considered is always smaller than the initial number of elements to be clustered. Nonetheless, the software used in this research made use of the agglomerative hierarchical clustering method (Figure 2.6).

In this method, explained in [29], each cluster is initially constituted by only one element. Then, all the distances between clusters are measured and updated into a Distance Matrix. The pair with the shortest distance between the two points is merged into a single cluster and the Distance Matrix is then updated. This step is repeated over and over until there is only one cluster present in the Distance Matrix. This leaves us with a tree where in the leaf nodes are the clusters with the biggest distances between them.

2.6.1.2 Clustering and segmentation

Segmentation techniques are widely used in the fields of Biomedical Engineering, particularly in the Imaging fields, with the information taken from that process being massively used for several purposes, including diagnostic techniques. One could think that segmentation and clustering are the same, since they both consider an initial set of data and separate those data into smaller groups. Nonetheless, as written in [30] “Segmenting is the process of putting elements into groups based on similarities, and clustering is the process of finding similarities in elements so that they can be grouped, and therefore segmented”. Indeed, this establishes a very thin line between these two concepts but despite being two very similar terms, they are clearly not the same thing. Image segmentation relies on a set of eliminating characteristics to define if a pixel belongs to one group or another. For example, when segmenting the image of a bone, features like the color, the intensity or the texture are considered when assigning or not a pixel to a segment. The pixels that do not have that certain color are not allocated in that segment, the ones that have a different intensity of the desired intensity are not part of it either, and so on. Basically, there is a better definition of the segments if there are more characteristics to differentiate between two pixels that are together. On the other hand, what Clustering does is to find what are the relations between the different types of data, using Machine Learning and algorithms, so that new partitions (segments) depending on those relations, can be created. Following the previous example, clustering understands what the relations are between two pixels by their colors, intensities and textures and if they are similar, they stay in the same segment, if not, then they are assigned into different places. [31]

2.7 Related work: DM techniques in Brain Injury outcome prediction

After an extensive descriptive analysis of the population in cause, predicting the outcomes after a brain injury is considered to be a major goal by the therapists who need to make decisions to plan the neurorehabilitation treatments. Therefore, it is necessary to perform intensive analysis to assess whether the currently used techniques are efficient enough to make good descriptive and predictive models and to improve the techniques that are not yet optimized. In this sense, it is also important to select the features that constitute, first better descriptors and then, better predictors.

Molaei et al. [32] present a new algorithm and a set of features that therapists can use to evaluate if a certain TBI patient needs to perform a CT scan or not. CT scans are frequently used to make conclusions about the injuries after a brain injury event but only in 9% of the cases the findings are positive and justify the costs and the exposure of the patient to the ionizing radiation. The new algorithm that makes use of an ensemble learning random forest technique, is compared with the already existing Canadian head CT (CHCT) algorithm, which uses different rules and features, in terms of diagnostic accuracy. To build the algorithm, feature selection based in literature review was used and the features found to be the best predictors included, amongst others, age and amnesia episodes. The results showed that the proposed algorithm is better than the CHCT one.

In their research paper, Prichep et al. [33] rely on algorithms that provide multi-class classification. More specifically, they use an “*informed data reduction*” method based on age-regressed quantitative features extracted from quantitative analysis of the electroencephalography (QEEG), to classify and evaluate TBI patients. From the obtained 1536 features, the best ones are selected using two feature selection methods that increase performance of the classifier functions at each iteration: the genetic algorithm (GA) and a deterministic feature selection method called Simple Feature Picker (SFP).

At a conference, V. Taslimitehrani and G. Dong [34] proposed to provide accurate prognostic models for TBI patients, while different groups of patients require different prediction models and to improve an existing regression method called Contrast Pattern Aided Regression (CPXR), by considering a logistic regression instead of a linear regression. Their new method is intended to be used in general binary outcome prediction and it is called CPXR(Log). In this study, it is considered the Glasgow Coma Scale (GCS) as the outcome result for TBI. The GCS is the system that classifies brain injury patients into different categories accordingly to the level of recovery they need and the scale range goes from GCS 1 (dead) to GCS 5 (good recovery). The data set used comprises information taken from 15 variables: basic variables that include the age of the patient and the GCS score, variables with values obtained by CT scans like hypoxia and hypertension and also variables which are measured in a laboratory, as glycose and hemoglobin. The results of the CPXR(Log) method on prediction of 6-month outcome after TBI have proved that it has better accuracy than the standard logistic regression method.

Another outcome prediction analysis after 6 months was performed, in these specific conditions, for the first time by van der Ploeg et al. [35] in which five statistical modelling techniques (Logistic Regression, Classification and Regression Trees, Random Forest, Support Vector Machines and Neural Networks) were used and externally validated for varying complexity predictor sets. The fact that 11.026 TBI patients from 15 different studies were considered in this study sets it as a Big Data analysis. Similar to what happens in the previous mentioned study, the predictor sets were grouped into three different categories, here defined as Core, Core + CT and Core + CT + Laboratory. The results demonstrate that Logistic Regression methods stand out comparing with the other four methods studied.

In [36], Hyun Soo OH and Wha Sook SEO studied the variables related to recovery at one month after the brain injury event and developed a prediction model using a Decision Tree technique that takes into account certain cut off values for the variables considered, to predict if the recovery is good or poor. The GCS was found to be the most significant predictor, followed by age and blood glycose level – all values taken at the admission time at the hospital. Figure 2.7 shows the prediction tree model obtained after the analysis and detection of consistent patterns in the data, using the CART (Classification and Regression Trees) algorithm. Once again, the authors deliberate that one of the limitations of this type

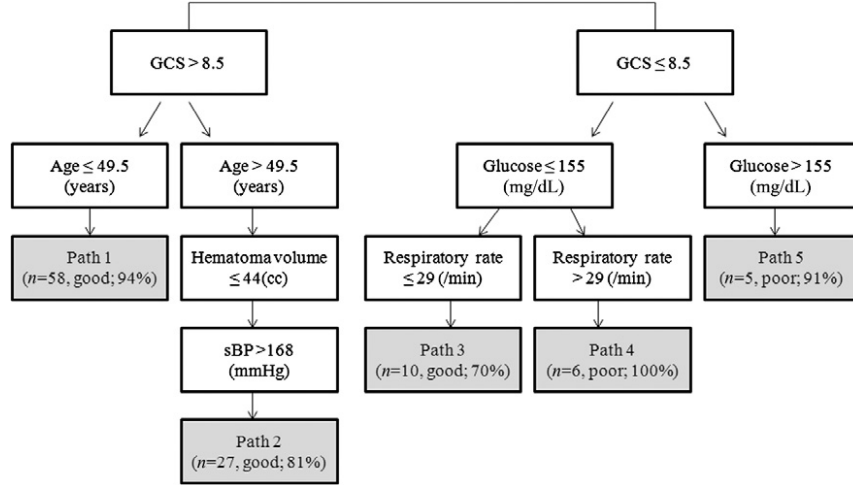


Figure 2.7 - Predictive decision tree model used in [36] showing GCS as the best predictor for 1 month after injury

of study is that the best predictors will always depend on the type of brain injury and this model is perhaps too general since it can be applied to various brain injury types.

In a closer approach of what was firstly intended to be one of the main aims of the work to be done in this project, Siddiqui et al. [37] try to understand what is the effect of a certain rehabilitation treatment in a patient. Medical tests performed before and after the treatment in 15 patients and 14 controls are considered by means of the scores that patients achieved in each type of test. Several methods are studied in order to model patients' evolution and final outcome. To do that, a supervised evolutionary label prediction method called EvoLabelPred is presented. This method is also able to learn an evolutionary model from unsupervised data and it is an improved version of the already existing EvolutionPred algorithm, developed by the same team. The difference between both methods is that, while in EvolutionPred each patient is projected into a future moment, EvoLabelPred predicts the patient's recovery label. Thus, the new method uses labelled longitudinal data as input and learns two models: a clustering model for each pre and post time point and also a cluster-based transition model used to predict the labels of the patients. To make this prediction, the algorithm (Figure 2.8) learns conditional probabilities over each one of the clusters. The mining workflow proposed with both algorithms clusters all the patients based on the similarity between their both pre-and post-assessments and then it accompanies the evolution of each cluster, by building a cluster evolution graphic. EvolutionPred

Algorithm 2: EvoLabelPred	
Input : $z_{\mathcal{X}}$ sampled patients, $z'_{\mathcal{X}}$ non-sampled patients, K number of clusters	
1	$\zeta_{pre} \leftarrow$ Learn clustering over $z_{\mathcal{X}}$ for t_{pre} instances.
2	$\zeta_{post} \leftarrow$ Learn clustering over $z_{\mathcal{X}}$ for t_{post} instances.
3	$\mathcal{G} \leftarrow$ Learn cluster transition graph over ζ_{pre} and ζ_{post} .
4	Learn conditional probabilities based on labels from t_{pre} for each cluster in ζ_{pre} .
5	foreach $x \in z'_{\mathcal{X}}$ do
6	$\hat{l}_{post} \leftarrow$ Predict label for x given l_{pre} and the nearest cluster $c \in \zeta_{pre}$ (Eq. 5)
7	end

Figure 2.8 - Algorithm used for learning the conditional probabilities and for the prediction of the patient's label. Figure taken from [37]

projected values have shown to be almost identical to the real ones, while in EvoLabelPred the models were found to correctly predict only part of the data. It is said that one of the following steps would be to incorporate image information from MagnetoEncephaloGraphy (MEG) into the workflow.

2.7.1 Influence of other variables in BI outcome prediction

The idea that demographic variables may have an effect on the evolution of a Brain Injury patient has many years now. In his book, published in 2008, James P. Tsai [38] wrote an entire section about the influential patient variables that need to be taken to consideration. He mentioned, among others, the age and the education level. Age because, for example, younger people are very likely to have their cognitive abilities still not totally developed and education due to its influence in the level of cognitive cerebral organization – usually people with a low level in studies or in cultural experience tend to have more difficulty in understating test instructions. The author also suggests that the handedness could also have an effect, mentioning studies that proved that right handed people's left hemisphere have a higher dominance in language and that left handed and ambidextrous are more inclined to have a variable linguistic-cognitive hemisphere dominance. In regards to gender, he brings up the fact that women show a better control in the verbal fluency and men in visual or special tasks. Polyglottism is also mentioned as an important variable, as people who know more languages deal with particular cognitive-cerebral processing mechanisms.

These and other several variables have been studied and are already documented. For example, in [39], Sherrill-Pattison and his research team considered variables like the age of the patient, the gender, ethnicity, the education level, the injury circumstances, the time that passed since the day of the lesion and also some IQ values. They started with the hypothesis that the age and education level of the patient would explain the existing variance in two tests that measure frontal lobe deficits, the “*Category Test*” and the “*Trail Making Test*”, but that the gender would have no influence on that. They confirmed that, in fact, their hypothesis was correct. Besides that, they have also concluded that the presence of coma for, at least, a day, was an affecting factor the performance on the tests. Another study, carried out by C. M. de la Plata et. al [40] intended to assess the specific influence of the age on long-term recovery from TBI. The conclusion taken was that older patients present more decline after 5 years after the injury than younger patients. A research regarding the affectation of the functional mobility, led by Haffejee et al. [41] also allowed the extraction of some important variables to have in consideration, also in cognitive terms. The results showed that a person with a younger age, from male gender, with a secondary education level, not smoking and drinking, and having occupational therapy sessions would have a better evolution in the mobility functions after a brain injury event. Because of what was mentioned in this section, part of this study was conducted to assess the influence that some demographic variables existing in the patient databases used, might have in their recovery from a brain injury.

2.7.2 Data Mining techniques used for other diseases or conditions

Improvements in data mining techniques have been done also regarding other diseases or conditions. It is more than possible that these improvements are able to be applied to Neurorehabilitation data, so they must be considered as well. Then, recent work in data mining is also presented here. For example, in [42], C. Kalaiselvi established the goal to improve the supervised k-nearest neighbor algorithm in terms of the timing of computation. This study is focused on heart disease prediction and presents a new learning algorithm called average k-Nearest Neighbor (AKNN), used for classification and prediction.

In this method, a super sample is created for each class and the AKNN searches the sample data to locate, by measuring the distance between neighbors, the closest one to the input. The efficacy of this method is limited by the number of chosen clusters but it is reinforced by the reduced number of samples used for training, which makes the algorithm faster when compared to the standard kNN and more

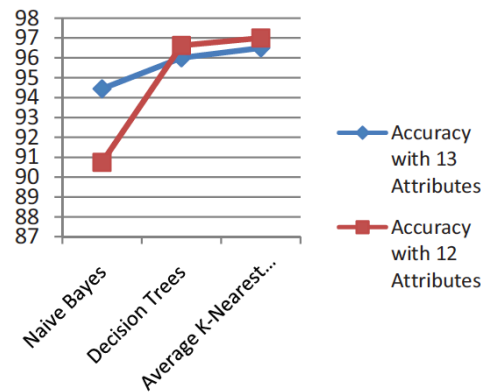


Figure 2.9 - Comparison of the accuracy between three classifier methods: Naïve Bayes, Decision Trees and Average k-Nearest Neighbor for a different number of attributes (12 and 13). Graphic taken from [42]

accurate than other classifier methods like Naïve Bayes and decision trees (Figure 2.9). Also, related to heart disease, studies conducted by S. Radhimeenakshi and published in a conference paper [43] make use of two well-known data mining techniques: a SVM and artificial neural networks (ANN) to predict heart disease risk. The data used was from two different databases and the results proved SVM to have the best accuracy prediction levels.

A research work regarding feature selection was carried out by Moretti et al. in [44]. In this study, the aim was to extract hemiparesis strongly related features and use them as input samples for data mining analysis. The inheritance of the features was evaluated through accuracy rates of machine learning algorithms and the best features were selected. Results of this KDD techniques application have showed that force attributes are the most inherent features to the hemiparetic sides. Having labelled data and a reduced number of attributes, the next step was to select a data mining technique and four were considered: a decision tree (J48), a random forest, a k-nearest neighbor and a multilayer perceptron neural network. The last one was found to be the best to perform an accurate distinction of hemiparetic sides.

Regarding data mining applied to the Immunoglobulin (IgA) nephropathy condition (inflammation of the glomeruli of the kidney), the main goal of Diciolla et al.'s work [45] was to make predictions regarding whether the patient will reach the ESKD (End Stage Kidney Disease) state or not, which is done by using the first classifier and, if this verifies, if the patient will reach it within 5 years, or not, which can be done using a second classifier. There were six inputs of the classification models that included age, gender, histological grade, amongst others and four data mining methods were trained with the available data: artificial MLP neural networks, neuro fuzzy systems, support vector machines and decision trees. The first one presented the best results with an accuracy percentage greater than 90, which means that it can be used as a decision support system for outcome prediction in IgA nephropathy patients and, possibly, for other conditions (which include TBI).

Maitreya, a framework used to predict outcomes considering symbolic time intervals is presented in a paper written by Moskovitch et al. [46]. The learning models are based in temporal patterns found in the clinical records. These temporal patterns are the prognostic markers used to train the predictive models. In this framework, the data is divided in three folds and mining iterations are performed in all of them.

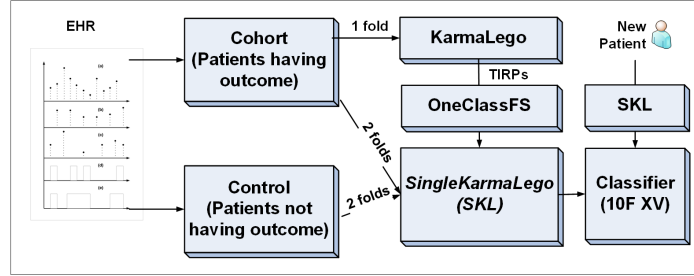


Figure 2.10 - The Maitreya Framework used in the study carried out in [46]

In the first fold, with data obtained only from patients with a n outcome, KarmaLego algorithm is used to discover the temporal patterns or the “Time Intervals Related Patterns” (TIRP). The TIRPs discovered in this phase undergo through a one class feature selection process and are found in the other two folds, with data from patients with an outcome and also control patients, using another algorithm called SingleKarmaLego (see Figure 2.10). The temporal data mining performed in this study constitutes an improvement to standard data mining techniques, since it takes into account additional information for prediction models.

Still regarding longitudinal data, but considering neuroimaging Alzheimer disease (AD) patients’ information, Huang et al. [47] used a nonlinear supervised sparse regression-based random forest (RF) framework to make predictions about longitudinal AD clinical scores with information from baseline scores and MRI-derived features. In this study, the random forest method used in the framework was also availed to estimate missing longitudinal scores and those estimated scores were then used at previous time points to predict the scores at future time points. Due to how the RF method is constructed, it can deal with nonlinear features and with a great amount of training data unlike, for example, linear regression.

Regarding missing scores, new methodologies have already been proposed. In [48], B. Yet and his work team introduce a new way to process the concept of latent variables. If so far the missing variables would lead to a partial elimination of the data, now it is possible to estimate those variables by inferring them from the other part of the data, observed data, using prediction models. For example, many diseases or conditions can only be diagnosed in an indirect way, by taking information from symptoms and tests on the patients. Thus, in this research work, a Bayesian Network (BN) model was developed to predict these latent variables and to do so, a case study of acute traumatic coagulopathy was considered. An iterative algorithm called Expectation-Maximization (EM) was taken for learning the Bayesian network parameters from the data set with the missing values, by computing the maximum likelihood estimate of its complete data. The results support the evidence that there is, in fact, an improvement in predictions using all the knowledge provided by using latent variables.

3 Materials and Methods

3.1 Data Sources

Two different databases were provided to the GBT. One of them, the Guttmann Clinical database, was called “curso_clinico” and the other, Guttmann NeuroPersonalTrainer was named “npt_produccion”. Every few months, both were updated and re-sent by the clinicians.

3.1.1 Guttmann Clinical Database

The Guttmann Clinical Database was the Hospital Guttmann database with administrative and demographic information. This database was used to get that data that was missing from the Guttmann NeuroPersonalTrainer Database and to double check the information available.

3.1.2 Guttmann NeuroPersonalTrainer Database

The Guttmann NeuroPersonalTrainer database (BB.DD) was entirely extracted from the GNPT platform and so, it contains data from patients both in Guttmann Institute and external patients from other rehabilitation centers, who were subjected to a neurorehabilitation treatment after a brain injury, and performed the tasks and tests that were previously planned by their therapists. Among the patients, test patients can also be encountered. This type of records are essentially fictional patients and not real persons, with false names, created by the therapists to test the platform. Even though they appear in the initial database, a manual selection was done to exclude them from the integrations.

The initial sum of patients in both databases, in February 2017, was 3299. Alongside with the selection of test patients, other filters were applied to make sure that after the integration, the set of patients had very well-known features. From the initial number of patients, 1563 were from the Guttmann Institute and 1736 were external patients. However, from this number, 65 were immediately excluded because they were being subjected to more than one treatment. As the aim was to take information about patients that had to undergo only one treatment, the database started with 3234 patients. This means that the data integration approached the selection and exclusion of some patients: patients with more than one treatment, prove patients and of some patients with errors in the data. Personal and clinical information from all the individuals was provided in the databases. This information included:

- | | |
|--|--|
| ● Name and Surname | ● Level of studies |
| ● Birth | ● Date when the injury occurred; |
| ● Spanish ID and Health Record Number | ● Category of the injury; |
| ● Place of Birth | ● Login used by the person in the platform; |
| ● Address (City, Province and Country) | ● Number of tasks performed in the treatment; |
| ● Phone number | ● Number and type of tests performed; |
| ● Email account | ● Number of sessions attended; |
| ● Gender | ● Dates of the tests (Pre and Post treatment); |
| ● Postal Code | ● Center where the treatment was done |
| ● Etiology | |

3.2 Software (MySQL, R, SPSS)

For this research three softwares were mainly used. MySQL Workbench (v. 6.3.7 build 1199 CE for 64bits) was used to explore and manage the database and to perform the integrations and queries as well. IBM SPSS Statistics 24 was used to obtain the clustering results and the histograms presented in the Results section. The R software was used as supporting tool for obtaining a statistic analysis of the variables considered.

3.3 CRISP-DM Methodology

CRISP-DM stands for Cross-Industry Standard Process in Data Mining. It is a very useful methodology to plan a Data Mining project, providing a strong and structured approach. It is a process most commonly composed of six stages, all of them considered in a very well-organized temporal flow.

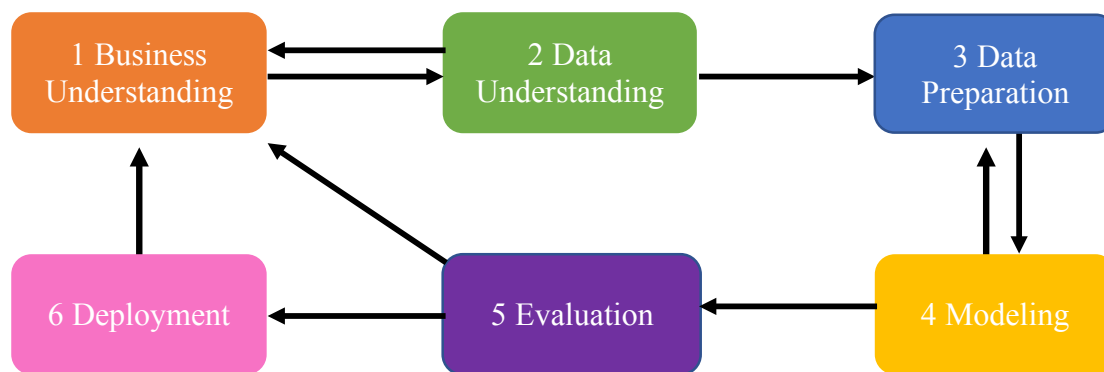


Figure 3.1 - Schematic representation of a common CRISP-DM flow

Only the first five stages were completed in this study. The Crisp-DM can be considered as a cyclic process because it can be performed as many times as a person wants – it depends on the DM objective.

The first step must always be to determine what the goals of the Data-Mining project are, considering the context in which it is being applied. This also includes building a planification of all the tasks to do. In our case, the general and the Data Mining objectives were established since the beginning. The second stage, Data Understanding, involves not only collecting and familiarizing with the data, but also describing it, exploring it and at, the end, verifying its quality. As our purpose was to get dysfunctional profiles to correctly describe the population, this was probably the most important Crisp-DM step in this study. In Data Preparation, as the name explains for itself, it is necessary to select the important data, clean it and model so that in can be used in future steps. At this point, data was filtered and integrated so that it could be used to obtain the clusters and the histograms. The following parts are already regarding the construction of predictive models. They include modeling, which means, selecting the techniques that better apply, building the model and assessing its value, the evaluation part, where the results obtained are evaluated and the process reviewed, and the last part, Deployment that makes a retrospective on the process, to understand if it worked properly and, if so, to plan some developments.

The Data Analysis stage was added in our case, considering that the objective was not to proceed, at least for now, with the prediction model. The analysis part is where the obtained data is interpreted and crossed over by means of graphics and tables, to extract valuable conclusions from it.

3.3.1 Data Understanding

3.3.1.1 The Clinical Process

The neurorehabilitation process followed by the Guttmann Institute starts with an initial neuropsychological assessment (Pre Tests). Once the cognitive affectation is obtained, the patient starts the treatment, which is constituted by several sessions, not necessarily with uniform frequency – the intervals between sessions may vary. After the last session, the patient executes another neuropsychological tests (Post Tests). The results of the initial and final assessments are compared and the level of improvement is assessed. In addition to the results coming from the initial assessment, the therapists collect another type of data in order to better define the treatment as personal information (age, gender) and structural data (type of damage, time passed since the injury happened, etc.). The entire mechanism is explained in the Figure 3.2.

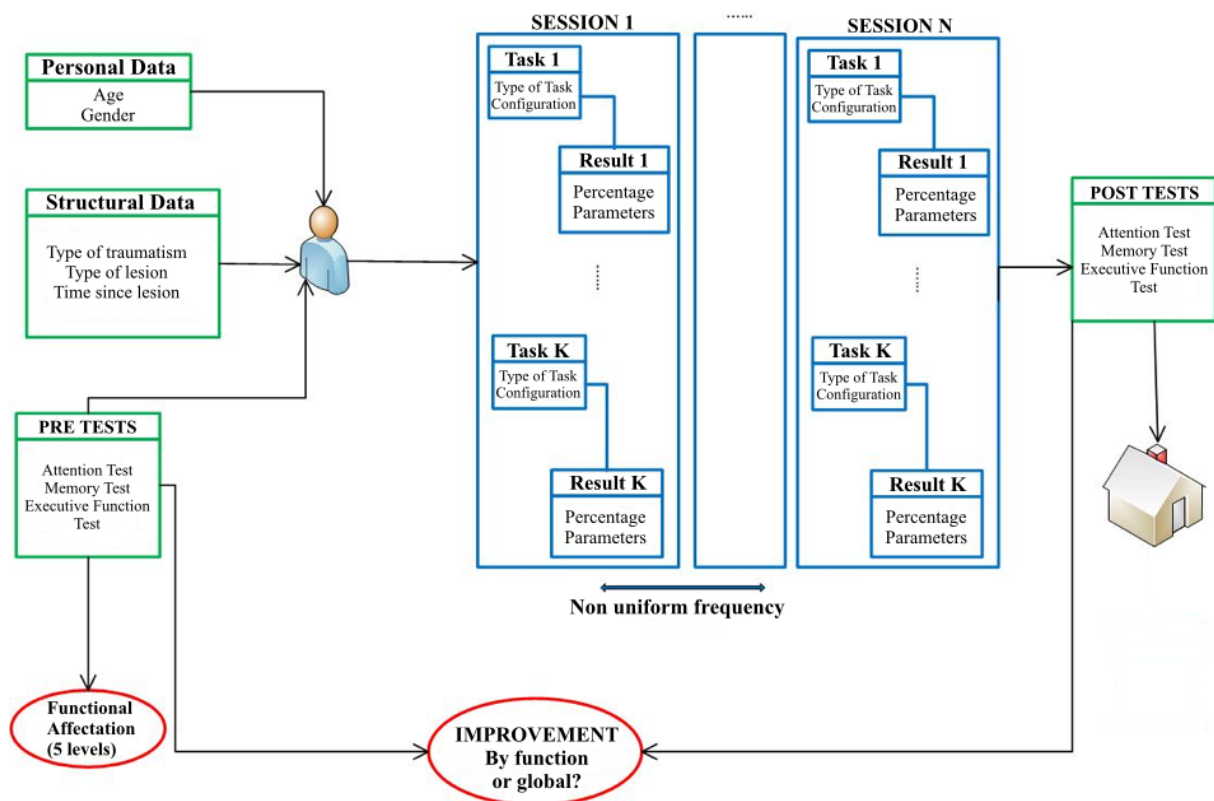


Figure 3.2 - Schematic representation of the clinical process flow which ends with the computation of the global improvement level.

3.3.1.2 Neuropsychological Assessment

After the injury and before starting the treatment, patients were subjected to pre-assessment tests. In total, each patient performed 27 tests: a battery of 24 tests and 3 CPT (Continuous Performance Test). In this research, only the battery was considered since the data from the CPT tests was not statistically sufficient to be contemplated – not all the patients had this information. From the 24 tests, only 17 were used in this study to evaluate the cognitive and functional capacity of the patients in the three functions that were considered the most important ones when assessing the state of a person who suffered a brain

injury. Thus, each function and/or subfunction is evaluated according to one or more specific tests and one test only might provide information about more than one function and/or subfunction. (Appendix Table A. 1). The range of calculated scores is very large, so a normalization was needed. This was done using the International Classification of Functioning, Disability and Health scale, with 5 levels of affectation: 0 for no affectation at all, 1 for mild affectation, 2 for moderate affectation, 3 for severe affectation and 4 for acute affectation. To perform the normalization, the age and the study level are considered since the tests performance most commonly depend on these variables. For example, it is reasonable to think that an older person without any studies will get worse results in certain tests, when comparing to a young person with a graduation level, even if the older person does not have the cognitive functions affected. Thus, the scores considered to be normal in these tests may vary due to the age and the study level of the patient. There is a file of all the Profiles that were contemplated until now which contain all the expected scores in all the tests for each function and subfunction, considering the affectation levels, the age and studies of the patients (see Table A.2 and Table A.3).

3.3.1.3 Criteria for evolution grouping

Still in the field of Data Understanding, in the validation part of the study - that aimed to get information among the groups of patients who improved, maintained or got worse, to compare it with the final dysfunctional profiles - some criteria had to be used to define these groups. Thus, some conditions had to be pre-established, so that the global improvement level of the patient could be assessed. The global improvement level parameter was calculated for each one of the patients according to the improvement levels in each of the functions. For example, if a patient had a level of 0 in the Pre Test of Attention and then a level 1 in the Post test, it was considered that the patient got worse in that function. If he had a 2 in Memory in the Pre and then another 2 in the Post test that would be considered a maintenance. If the patient had a 4 in the Pre test and then a 2 in the Post test that would be an improvement. The global improvement, considering these evolutions in all the three functions, was calculated by Excel, following the criteria:

- There is an *Improvement*, associated with a global improvement of 1, if the patient improves in, at least, one cognitive function, without getting worse in the rest of the functions;
- There is a *Worsening*, associated with a global improvement of 2, if the patient gets worse in, at least, one cognitive functions without improving in the rest of the functions;
- There is a *non-significant evolution of the global cognitive function (Maintaining)*, associated with a global improvement of 3, when none of the criteria of worsening or improvement for the two previous conditions is verified;

The maintaining criteria considers four options. First, the patient improves in one function, gets worse in another one and maintains his/her state in the third. Second, the patient maintains the scores in all the functions. Third, the patient improves in exactly two functions and gets worse in one. Fourth, he/she gets worse in exactly two functions and improves in one. In Table 3.1, all the possible combinations are presented, in a generic way, with no specific functions. A “+” represents a function whose score improved, a “-” represents a function whose score got worse and the “0” was used to represent a function that did not change its score.

Table 3.1 - Representation of the possible results for the global evolution of a patient considering the evolution in each one of the functions.

+	0	0	improves
+	+	0	improves
+	+	+	improves
-	0	0	worsens
-	-	0	worsens
-	-	-	worsens
+	0	-	maintains
0	0	0	maintains
+	+	-	maintains
-	-	+	maintains

3.3.2 Data Preparation

Throughout the internship that this dissertation documents, several different databases were provided by the IG, always including more patients than the previous version. The results here documented were obtained using the last database received. Every time, after receiving both files from the IG, the first step was to restore both of them in the MySQL Workbench platform. The most important part for achieving a correct integration, when copying the tables from the Guttman NeuroPersonalTrainer database that contains the data used in that integration, is to maintain the original database structure. Since this part was done manually, a special care was needed. After this, the complete integration procedure included three parts: preparation of the data, integration of the data and finally, its validation. In the first part, a small script was executed in order to delete all the previously created tables and to reset all the existing data. Then, the integration procedure was executed, for the integration of all the patients. This was the longest task since it usually takes more or less 2 or 3 minutes to integrate each patient. Each integration line for each patient is written as in the expression:

call treatment_integration ('login_of_the_patient');

The *treatment_integration* stored procedure starts by verifying all the information available from all the existing patients, using their login string or number. If a person only has one treatment, the procedure will check if he/she has already been integrated or not. If there is already a registration with that login, it confirms if the corresponding patient's information has been processed and if it is complete. If so and if there are not new tests or information, the program will interpret that there is no data to be migrated. On the other hand, if the patient is completely new, his/her login is registered and inserted into two tables with information about the treatment and the tests, called "check_treatment" and "check_clustering", respectively. The first table has a great part of administrative and demographic data about the patient and the second registers the existence or no existence of information about the birth, the study level and the pre and post tests associated with the login. After this, the "treatment_integration" checks if the patient is a language patient (if he has language problems) and if he/she is an active patient, the software will start looking for the demographic data and the migration of tests results. After everything is verified, the patient is marked as "Complete".

After the integration, it was possible to filter the information to know the data. A validation script that had previously been prepared was used to verify the quality of the data and to filter the patients by key parameters. To do this, several SELECT queries were used in SQL. The purpose was to find and eliminate some patients, depending on some factors, or for example, to determine the number of patients with a certain number of tests performed. In this study, patients with less than 10 tasks performed were immediately set aside. The reason for this is because the target patients are the ones who have performed enough tasks for the possibility of considering the therapy's effect on the evolution. Also, patients of language (with language problems), children under 16 and patients without demographic data were not taken in consideration. In the integration done for all of the patients (both Guttmann and external), the total number in the clinical course was 3234. Due to having less than 10 tasks performed, 525 were excluded, leaving the cohort with 2709 patients. From those, only 2104 were not language patients. By only keeping the patients with all the demographic data complete, this number goes down to 1813 and after selecting only the ones older than 16 years old and having both Pre and Post assessment tests, the final number was 929.

From the final number of patients, the main objective was to analyze those who had at least one test in Attention, two tests in Memory and two tests in Executive Functions, both in Pre and Post versions of the tests. An important thing to highlight is that, when calculating the affectation levels, the tests within each function were not always from the same item. For example, if a patient had a Pre test in one Attention item (TMTA) but did not perform that same item in the Post test, and yet it had performed another item from Attention (Interference), then this patient would be taken into account (see Table A.1 in the Appendix section). Using this filtering, the number of patients included in the sheet to be used for the clustering process was 698.

3.3.2.1 ETL: Integration Procedure

All the new information was subject to an ETL (Extract, Transform, Load) process. The ETL programming tool is used for extracting selected information from a certain data source, transforming it and converting it through calculations or through the application of processes and rules and then loading the result into a new or already existing database. It takes to consideration all the connections between the data and crosses all the information between matching elements. To access all the databases and to perform all the queries, MySQL Workbench was used. SQL stands for Structured Query Language and it is the most common method of accessing and transforming data within a database. MySQL is an open-source relational database management system that makes use of the SQL language.

Usually, the final aim of an ETL process is to build a Data Warehouse (DW). A DW is an aggregation of databases or sources, in which all the information is to be processed and cleaned into a consolidated structure, where queries can be done.

3.3.2.2 The Data Warehouse

The DW used to perform all the preparation and integration of the data, was called "CognitioDW" and it is displayed in the Appendix section, Figure A.1. Two different databases were brought into the final Data Warehouse: "curso_clinico" was the name of the Hospital Guttmann database with administrative

The screenshot displays the MySQL Workbench environment. The top menu bar includes File, Edit, View, Query, Database, Server, Tools, Scripting, and Help. The left sidebar shows the 'Navigator' pane with a tree view of the 'cognitidw' database schema, including tables, columns, indexes, foreign keys, triggers, and various other tables like 'areas', 'block_tasks', 'blocks', 'categories', etc. The main window shows 'Query 1' with the SQL statement: `SELECT * FROM cognitidw.administrative_data;`. The 'Result Grid' tab is active, displaying a table with 12 columns: idArea, idCategory, idRegime, therapeuticDays, sessionsPaid, literacy, ethiology, other, lessonDate, startTreatment, endTreatment, startPrevinec, globalImprovement, and complete. The data is presented in a grid format with 15 rows of results. The bottom status bar shows 'Apply' and 'Revert' buttons.

3.3.2.3 Getting the Pre and Post test scores

25

In fact, having all this information makes it easier to create a function to obtain the scores for a unique function, from the database tables. For that, SQL functions were generated. Following the example of obtaining the Pre test scores in Memory: a subfunction called “getMemoriaTrabajo” was designed to bring all the “resultNormalized” values from the “tests_results” table where the “idTest” was 15 or 16, into variable called “mdigits” and “mlletres”, because those are the tests who are used to evaluate the Working Memory cognitive subfunction. This function returns a rounded mean of all the tests that are not null or equal to -1. Another subfunction, called “getMemoriaVisualVerbal”, was designed to do and return the same thing, but only for the “idTest” values of 17, 18 and 19 and putting them into variables called “mravlt075”, “mravlt015” and “mravlt015r”, respectively. This was done for each subfunction of each main function.

tests_results			
idTreatment	...	idTest	normalizedResult
1		15	2
1	...	16	4
...
...
...
3	...	15	0
3	...	16	2
...
1	...	17	0
1	...	18	3
1	...	19	3
...
3	...	18	1
3	...	19	3

Table 3.2 – Simple representation of the table called “tests_results”, with information about the treatment, the test and the normalized result obtained.

Using the example presented in Table 3.2, the calculations should be:

1. For idTreatment 1,

$$getMemoriaTrabajo = \frac{(2+4)}{2} = 2 \quad (3.1)$$

and

$$getMemoriaVisualVerbal = \frac{(0+3+3)}{3} = 2 \quad (3.2)$$

2. For idTreatment 3,

$$getMemoriaTrabajo = \frac{(0+2)}{2} = 1 \quad (3.3)$$

and

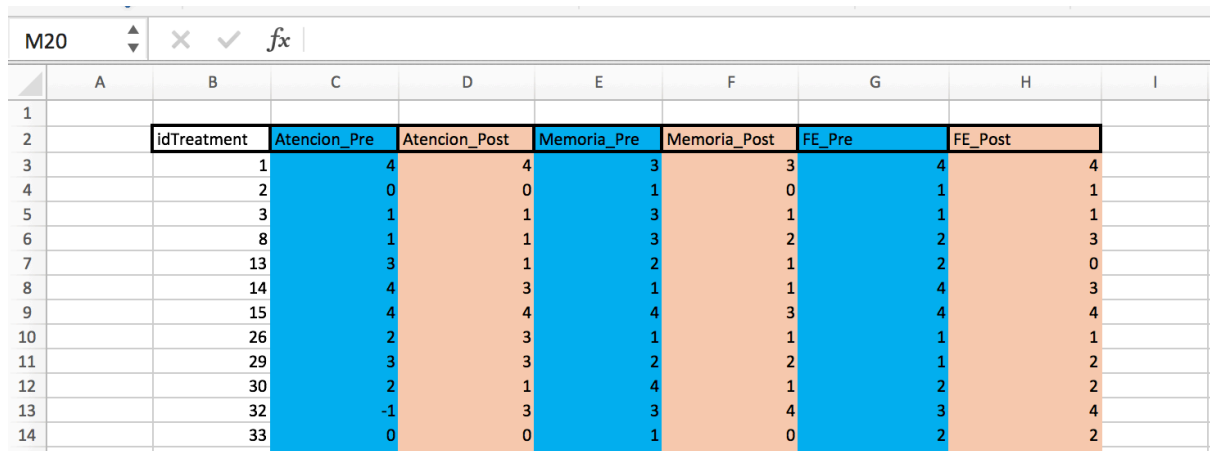
$$getMemoriaVisualVerbal = \frac{(1+3)}{2} = 2 \quad (3.4)$$

Then, a bigger function, called “insertFunctionLevels” was created to compute the values for each function and update them into a table called “function_levels”. In case of Memory, the value obtained

by executing this function, and brought into the table as the values for the “idFunction” number 2 (1 for Attention, 2 for Memory and 3 for E.F.), for the patient with the idTreatment 1, it would also be calculated as in calculation 3.5:

$$insertFunctionLevels(.) = \frac{getMemoriaTrabajo(.) + getMemoriaVisualVerbal(.)}{2} = \frac{(2+2)}{2} = 2 \quad (3.5)$$

A third function, named getMemory(), was built to get all the values (or normalized scores) from the table “function_levels”, from all idTreatments that have a 2 as their “idFunction”. Those values were copied into a column with the corresponding name in an Excel file (Figure 3.4).



	A	B	C	D	E	F	G	H	I
1									
2		idTreatment	Atencion_Pre	Atencion_Post	Memoria_Pre	Memoria_Post	FE_Pre	FE_Post	
3		1	4	4	3	3	4	4	
4		2	0	0	1	0	1	1	
5		3	1	1	3	1	1	1	
6		8	1	1	3	2	2	3	
7		13	3	1	2	1	2	0	
8		14	4	3	1	1	4	3	
9		15	4	4	4	3	4	4	
10		26	2	3	1	1	1	1	
11		29	3	3	2	2	1	2	
12		30	2	1	4	1	2	2	
13		32	-1	3	3	4	3	4	
14		33	0	0	1	0	2	2	

Figure 3.4 - Screenshot of an Excel file containing all the scores, for each treatment, for each pre and post function.

The normalized scores, as previously mentioned, take values in a range (0-4) in which 0 is no affectation, 1 is mild affectation, 2 is moderate affectation, 3 is severe affectation and 4 is acute affectation (the worst possible score for the patient).

3.3.3 Data Analysis

3.3.3.1 Finding Cognitive Profiles

For research purposes, it was considered that it would be interesting to evaluate the progress of a patient, considering his/her test score evolution. To do so, every single patient had to be assigned to an initial cognitive profile, considering his/her initial test performance, and to a final profile, considering his/her final test performance. The viable solution found for that assignation was to create a structure where patients would be grouped in clusters according to their test score levels in each of the three functions. This grouping, or clustering, was thus done using a software called IBM SPSS Modeler and the initial and final clusters were called, in practical terms, initial and final profiles since a cluster is, as a matter of fact, a cognitive profile.

For this next step, *IBM SPSS Modeler Version 18.0* was used. The basic environment of this software is called a Stream (Figure 3.5). A new stream can be created every time the user needs to build a new predictive model. Initially, its desktop is completely empty, and it is then filled with the icons (nodes) that the user wants to bring in the middle to start building a new flow. A flow is the path that needs to be followed from the moment that data is updated until the predictive model is built (see Appendix Figure A.2) The available nodes are displayed in the row that stands below in the screen, categorized by each of their functions. For example, there is a tab to select the Sources, another for Record Ops (operations that can be applied on the records), Field Ops (to operate on the fields), Graphs (which include plots, histograms from data), Modeling (to apply the DM algorithms) and then two for getting

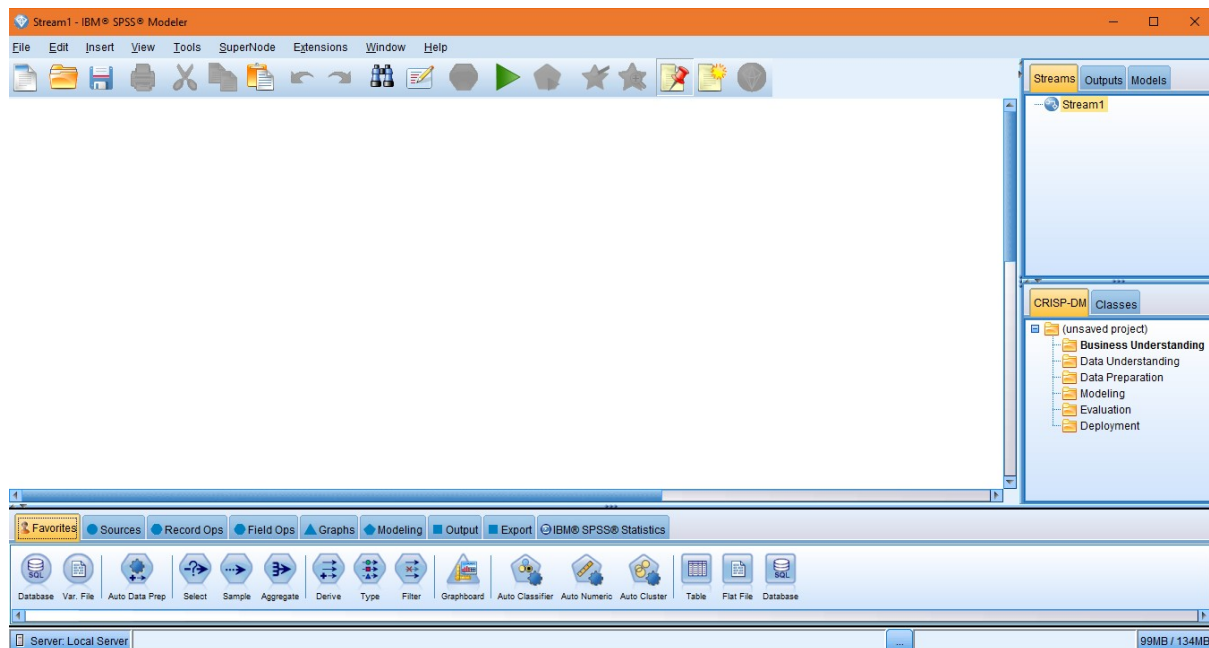


Figure 3.5 – Screenshot of the IBM SPSS Modeler's initial Stream

the results, Output and Export (data can be exported to other applications). In the right area of the stream, there is a tab called CRISP-DM that organizes all the data mining work. There, the user can put all the files that will be needed to build the predictive model, for example.

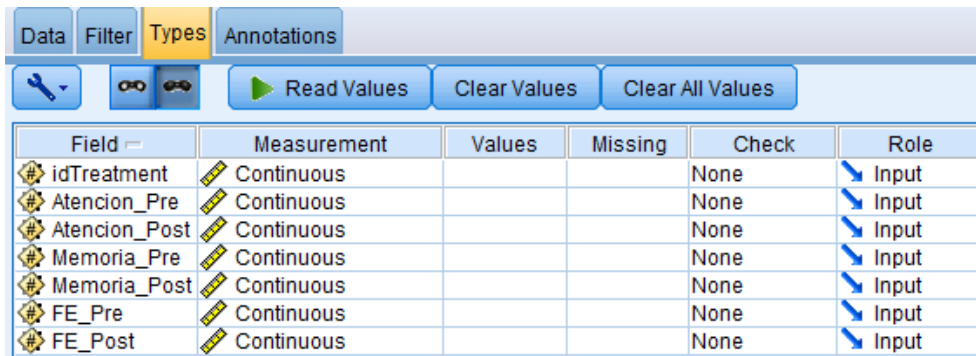
The first elements to put in a Stream are most commonly the external sources the user wants to bring with the data and the node that represents these elements has the same name. Here the source node was named “Results3functions”. The second element is usually the type (Figure 3.6). It is always possible to edit the information that relies on a node by opening it.



Figure 3.6 - Example of how to start building a flow in IBM SPSS Modeler with a Source and Type node

The source file can come from several types. In this case, the file used was the previously created Excel file containing the normalized results/values of the tests of the 3 functions (A, M, EF), both Pre and

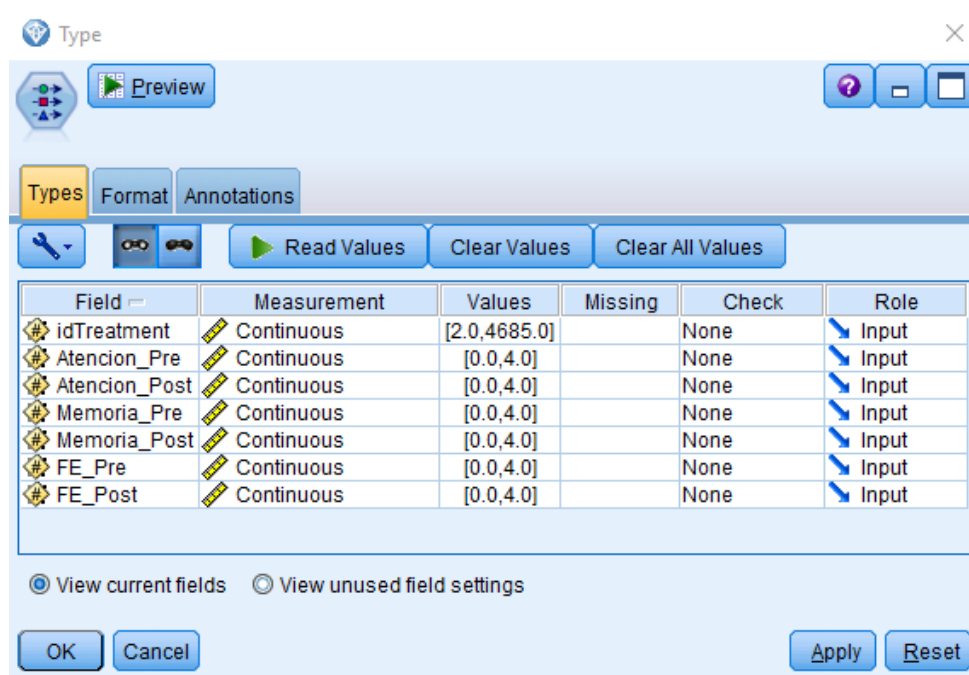
Post, for each one of the 698 idTreatments (Figure 3.4). By choosing one of the several tabs and sections, the user is allowed to view, classify or filter the data (Figure 3.7).



Field	Measurement	Values	Missing	Check	Role
# idTreatment	Continuous			None	Input
# Atencion_Pre	Continuous			None	Input
# Atencion_Post	Continuous			None	Input
# Memoria_Pre	Continuous			None	Input
# Memoria_Post	Continuous			None	Input
# FE_Pre	Continuous			None	Input
# FE_Post	Continuous			None	Input

Figure 3.7 - Screenshot of the box where it is possible to edit all the parameters from the source data

For example, the “Filter” section allows to select all the important fields and to eliminate the ones that are not needed or the ones that have no meaning in the context. By selecting “Preview”, one can see the preview of the data, which means, if everything is correct, the same as what is seen by opening the Excel file normally. The SPSS automatically attributes a type to the variables, although all of them can be changed by the user. The variables can be: continuous, categorical, flag (binary), ordinal, nominal or tapeless. Here, the variables can all be considered as Continuous, although accurately, they would be taken into account as discrete variables. The values of the tests vary between 0 and 4 and the variable idTreatment has values between 2 and 4685.



Field	Measurement	Values	Missing	Check	Role
# idTreatment	Continuous	[2.0,4685.0]		None	Input
# Atencion_Pre	Continuous	[0.0,4.0]		None	Input
# Atencion_Post	Continuous	[0.0,4.0]		None	Input
# Memoria_Pre	Continuous	[0.0,4.0]		None	Input
# Memoria_Post	Continuous	[0.0,4.0]		None	Input
# FE_Pre	Continuous	[0.0,4.0]		None	Input
# FE_Post	Continuous	[0.0,4.0]		None	Input

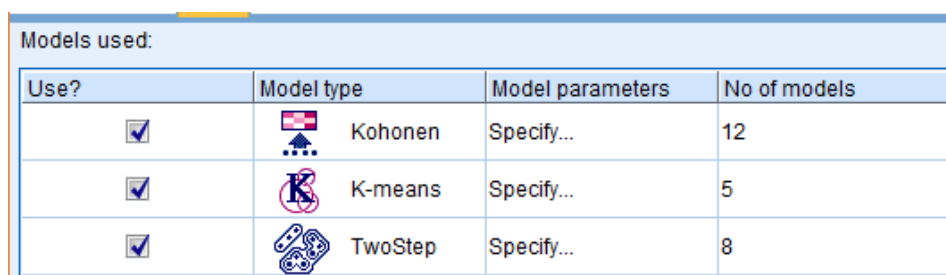
☒ View current fields ☐ View unused field settings

OK Cancel Apply Reset

Figure 3.8 - Screenshot of the box where it is possible to edit all the parameters related to the type and format of the source data.

The Type node can be found in the Field Ops tab. Selecting this node opens a box where the types for the variables are verified, as well as the measurements, the ranges of the values and their role (Figure 3.8). This part is important so that the software considers the data exactly as it is.

It is also possible to edit the format of the variables like, for example, the number of decimal places. The third element already depends on the Data Mining objective. As in this case the aim was to obtain clusters, the node was taken from the Modeling tab. Since the goal was to perform some research on the best number of clusters to characterize the population of patients, other than specify the number of clusters, the Auto Cluster node was used - the most useful in these situations. In the first step, the interest






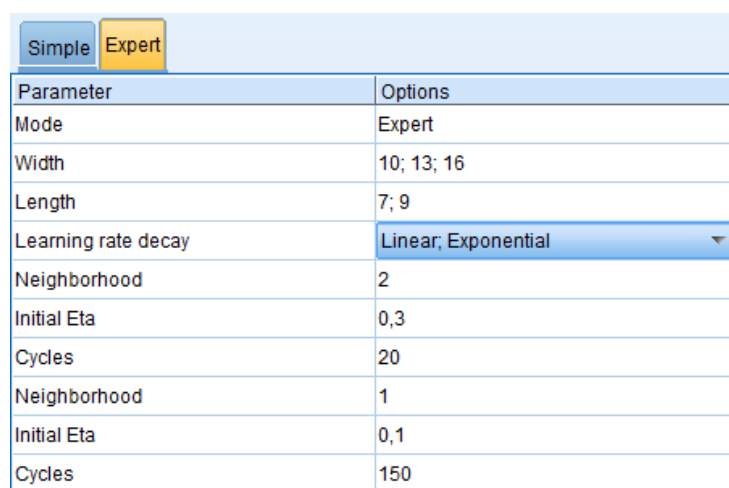
Models used:			
Use?	Model type	Model parameters	No of models
<input checked="" type="checkbox"/>	 Kohonen	Specify...	12
<input checked="" type="checkbox"/>	 K-means	Specify...	5
<input checked="" type="checkbox"/>	 TwoStep	Specify...	8

Figure 3.9 – Screenshot of the box to select the models to be used by SPSS to determine the best clusters for the source's data

was to first assess what were the initial profiles. This means, it would be useful to know what the clusters formed by the Pre tests are. This way, in the tab Fields, the option “*Use custom fields assignment*” was selected, and all the 3 functions in Pre were selected (Atencion_Pre, Memoria_Pre and FE_Pre) as inputs. For the evaluation, the chosen parameter does not have a special relevance at this point because the main interest is to have a Post reference and not a specific function to evaluate the clusters, so one of the three was randomly selected (Memoria Post). The parameters to be selected in the Model test are: “*use partitioned data*” – which means to use a training and a validation set, rank models by their “*silhouette*” and using test partition. A selection of which are the type of models that shall be the most adequate to be used (Figure 3.9) is also allowed. It is very useful to select all the models and not exclude any options to a general view of all the types of clusters that can be created. Here, there is the Kohonen model, the K-means model and the TwoStep model. For example, in the Kohonen networks model parameters' box, in the “Simple” tab, the “*Repeatable partition assignment*” was set to false and the



Simple Expert	
Parameter	Options
Mode	Expert
Width	10; 13; 16
Length	7; 9
Learning rate decay	Linear; Exponential
Neighborhood	2
Initial Eta	0,3
Cycles	20
Neighborhood	1
Initial Eta	0,1
Cycles	150

Figure 3.10 – Screenshot of the box used to set more specific parameters for the Kohonen node, like the size of the two-dimensional output map.

memory was set to be optimized. In the Expert tab (Figure 3.10) some parameters were set so that the execution could be as fast as possible.

In this case, it would not be very useful to assess the situation where there are more than 7 clusters because due to the number of variables, it will result in clusters way too small or clusters with huge size differences between them. The number of models with these parameters is 5. To initiate the calculations, the user needed to press “Run” on the element, which was called Pre. The execution results in another element that has the name of the evaluation column selected. In this case, the label “Memoria_Post” was applied, although it has nothing to do with the Memory function or the Post functions. The results of the calculations are presented in the next section.

4 Results

As previously mentioned, clustering can be thought as a process of organizing and grouping elements according to the similar characteristics between them. In this study, the first objective was to get enough information that would able an accurate description about the initial and final dysfunctional profiles from a patient, by obtaining the scores of the tests performed before and after the treatment, respectively, for each main function. This was achieved when clusters were obtained using SPSS Modeler. In this section, the tables and graphics presented first show the characteristics of the descriptive clustering model calculated using the Pre and Post test scores. Then, an analysis was made in order to better describe each one of the clusters in the model and also to understand the influence of the calculated initial profiles in the determination of the final ones. For a better perception of the effect of each function in the calculated profiles, graphics exhibiting the distribution of the Post functions scores through the initial profiles and the distribution of the Pre functions scores over the final profiles, are also displayed. This part was meant to illustrate how in each of these cases, the functions have an effect in the formation of the respective clusters.

As the second aim was to assess the effect of some variables on the evolution of the patient (age, study level, the interval of time that has passed between the injury and the start of the treatment, the number of tasks performed by the patient and the duration of the treatment), the graphics presented illustrate the distribution of those variables in the initial and final clusters.

The last part of this section first includes tables that are representative of the different types of evolutions of the patients according to their initial and final scores in the functions (not the clusters). Knowing the number of patients who achieved a certain evolution type and comparing it to the number of patients in the final profiles allowed us the extraction of conclusions and a first validation of the clusters. Besides that, it was also possible to observe through plotted histograms, if the previously studied variables could have had an influence on the type of evolution of the patient, especially the age, study level and gender considering only the initial and final test scores. This was the second validation method used for the computed initial and final clusters.

This section was accomplished by interpreting SPSS Modeler graphics, histograms and density plots.

4.1 Finding the Initial and Final Cognitive Profiles

To acomplish the first objective and after trying to find the best relation between the elements to build the clusters, SPSS Modeler presents the table shown below (Figure 4.1) with all the calculated models considering the input values defined by the user. For future considerations, clusters represent cognitive profiles. After an interpretation of the table, the selection the model that is found to be a best choice is required. This decision is most commonly made basing on two important parameter values. One of these important parameters is the silhouette, as mentioned before.

This parameter is a measure of consistency within the clusters, since it considers both cluster cohesion and cluster separation. In the cluster cohesion approach, models with tightly cohesive clusters are preferred while in cluster separation, the chosen models are the ones that include clusters which have a relatively large distance between them. The silhouette measure can be used to evaluate not only individual objects but also clusters or models. It ranges from -1, if it is a very poor model, to 1, when

the model is very good. This definition means that in the model with bigger silhouette number, there is a better patient fit within the cluster he was assigned into, when compared to the others.

Us...	Graph	Model	Build Time	Silhouette	Number of Clusters	Smallest Cluster	Smallest Cluster (%)	Largest Cluster	Largest Cluster (%)	Smallest/Largest	Importance
<input checked="" type="checkbox"/>		K-means 4	< 1	0,445	6	75	10	169	24	0,444	0,202
<input type="checkbox"/>		K-means 5	< 1	0,397	7	29	4	170	24	0,171	0,188
<input type="checkbox"/>		TwoStep 4	< 1	0,378	2	45	6	653	93	0,069	0,020

Figure 4.1 - Screenshot of the box used to select which one of the calculated SPSS models shall be used to build the initial clusters, containing information about each one of them.

Another important parameter is the Predictor importance (corresponding to the Importance column in the table) which refers to the relative influence of a field in a predictive model. It measures, for both numeric or discrete, how well a variable can distinguish different clusters. This implies that for a larger measure of importance, it is less likely that the variation for a variable between clusters is explained by a random factor, and it is more likely that the variation is due to some unknown difference.

SPSS Modeler found and presented two K-means models with 6 clusters and 7 clusters and a TwoStep model with 2 clusters. From the three of them, the model with the 6 clusters was the one with the best values for the silhouette and importance measures. Even so, taking these parameters values into consideration was not the decisive factor. The most important thing when selecting a model is to choose the one whose interpretation brings more value for the objective of the study. Thus, the model with 6 clusters was chosen for determining the Initial Clusters. Comparing this model with the others, the clusters were more detailed than those in the Two Step model and also better descriptors than the ones from the other K-means model with 7 clusters.

When selecting a model, it is possible to see the characteristics of all the clusters presented in several types of plots (Figure 4.2). A summary of the model is displayed, alongside with a plot indicating the cluster quality based on the silhouette measure, and a plot showing the cluster sizes, using percentage values. Observing the Cluster Quality graphic, it is possible to infer about how well these clusters represent the data – in this case, one could say that the cluster quality is sufficiently good. In the Cluster

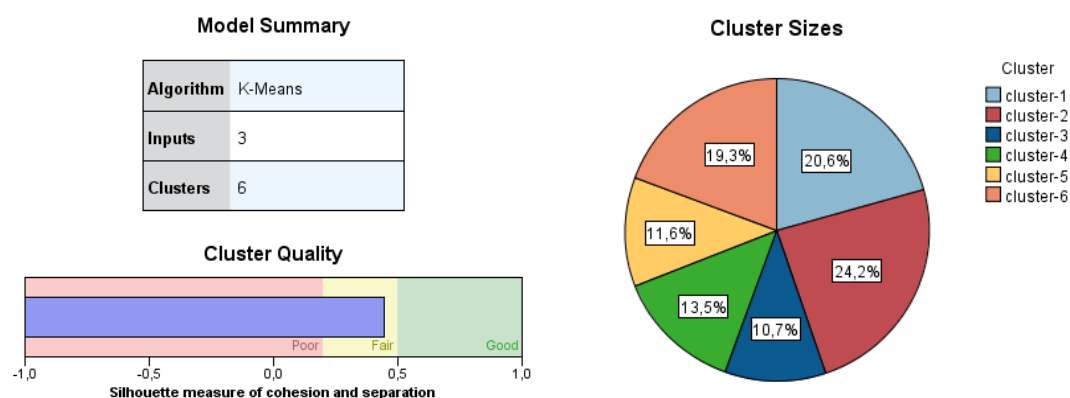


Figure 4.2 - Graphics showing the characteristics of the chosen k-Means model with 6 clusters

Sizes diagram, all the clusters seem to be relatively similar in terms of size: there is not a cluster or two that stand out from the others at first sight.

The Figure 4.3 also displayed, illustrates the Predictor's importance of each test - attributing a different shade of blue (the scale on the top right) - and the mean of normalized results, for each cluster and for each input. The clusters are sorted by size, from left to right.

Input (Predictor) Importance
 1,0 0,8 0,6 0,4 0,2 0,0

Cluster	cluster-2	cluster-1	cluster-6	cluster-4	cluster-5	cluster-3
Label						
Description						
Size	24,2% (169)	20,6% (144)	19,3% (135)	13,5% (94)	11,6% (81)	10,7% (75)
Inputs	Atencion_Pre 3,52	Atencion_Pre 0,80	Atencion_Pre 2,14	Atencion_Pre 1,41	Atencion_Pre 2,42	Atencion_Pre 3,31
	Memoria_Pre 3,02	Memoria_Pre 1,00	Memoria_Pre 1,56	Memoria_Pre 2,87	Memoria_Pre 2,22	Memoria_Pre 0,71
	FE_Pre 3,63	FE_Pre 1,36	FE_Pre 2,90	FE_Pre 3,13	FE_Pre 1,74	FE_Pre 2,47

Figure 4.3 - Information about each one of the six clusters regarding the Predictor importance for every input

Analyzing the table on this figure, on cluster number two (the biggest one, on the first column), the means are all around 3, which leads to a relatively high profile of affectation. The fact that the shade of blue in the Attention function line is darker (importance = 1.00) than in the other two functions, means that this specific function is likely to have a higher importance in the cluster's differentiation, when comparing to Memory (0.84) and Executive Functions (0.76).

If the Clusters View is selected, a different type of plot, called Cell Distribution appears, also for each cluster and for each input. This type of plots exemplifies the total distributed frequencies of the population of that specific function in the background (in a light red color) and it shows the frequencies for that specific cluster in the front (in a dark red color). This way, it is possible to make a comparison between clusters regarding the same function.

In the case presented in Figure 4.4, the cell distribution is for Attention Pre. Other features of this SPSS window allow different views, which include the sorting of inputs by within-cluster importance, by name or by label. The graphics can also show the cluster centers and the absolute or relative distributions.

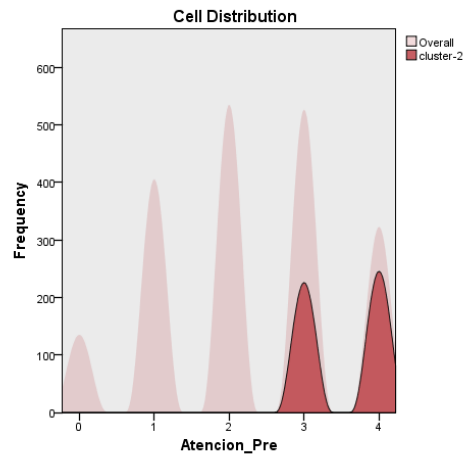


Figure 4.4 - Example of a cell distribution plot where it is possible to look at the distribution of the scores over the Attention function and compare it with the distribution over the entire cluster.

The next step was to interpret each one of the graphics and determine which model fitted best the purpose of the aims to reach. The following table (Figure 4.5) allowed an interpretation of the models for the Pre tests. It approaches a preview of the initial clusters that can be seen in the new column \$XC-Memoria_Post (or as it will from now on be called, the Initial Clusters column).

	idTreatment	Atencion_Pre	Atencion_Post	Memoria_Pre	Memoria_Post	FE_Pre	FE_Post	\$XC-Memoria_Post
1	2	0	0	1	0	1	1	cluster-1
2	3	1	1	3	1	1	1	cluster-5
3	8	1	1	3	2	2	3	cluster-4
4	13	3	1	2	1	2	0	cluster-5
5	14	4	3	1	1	4	3	cluster-3
6	15	4	4	4	3	4	4	cluster-2
7	26	2	3	1	1	1	1	cluster-1
8	29	3	3	2	2	1	2	cluster-5
9	30	2	1	4	1	2	2	cluster-4
10	33	0	0	1	0	2	2	cluster-1

Figure 4.5 - Screenshot of a table provided by SPSS that shows the assignation of the initial clusters for each treatment, according to each function.

In this last column, it is possible to understand to which initial cluster the patient belongs to, accordingly to the model selected, in this case, the k-Means model with 6 clusters. It is a fact that the cluster depends on the results for each of the functions, Pre and Post. This preview was done using the functions Pre as inputs, which means that, for example, since the patient with the idTreatment number 2 has a 0 in the Attention Pre test and a 1 in Memory Pre and FE Pre tests, the patient is very likely to belong to cluster 1 – the cluster whose patients are just slightly affected. On the other hand, the patient with the idTreatment number 15, has a score of 4 in all of the three Pre tests – one would think immediately that this patient is most likely to belong to a cluster where the patients have an higher affectation level. This preview shows us precisely that this patient is part of cluster 2, the cluster which has the biggest number of high affectation patients.

Next, the cell distribution graphics are analyzed (Figure 4.6) in order to understand how the patient's scores are distributed within each cluster, depending on the function. The background, or the column, shades, represent the total for the function considered. The colored columns represent the values for the

In cluster 2, focusing in attention, there is a very similar number of patients with a score of 3 and 4. No patient in this function has a normalized result of 0, 1 or 2. In memory, patients do not have any 0 or 1 value. The peak of patients in this cluster have scores of 3, but there are also some with a 2 or a 4. In this function, almost all patients with a 4 belong to this cluster. In executive functions, there are no patients of cluster 2 with scores lower than 3. This analysis leads to a profile of a patient with general **great affectation** in all of the cognitive functions.

Cluster number 1 is the second biggest cluster. In all three functions, the patients in this cluster only have normalized results of 0, 1 and 2. In both attention and memory, the peak of frequency is on the value 1. In EF, all patients with a 0 and a great part of patients with a 1 belong to this cluster. In attention, almost all patients with a 0 belong to this cluster. This analysis leads to a profile of a patient with general **low affectation** in the 3 functions.

Cluster number 6 is the third biggest cluster. In attention, patients have values of 1,2 and 3. The peak of patients has a value of 2. There are no patients with a 0 or a 4, which means there are no extreme values. In memory, the only values are 1 and 2 – values that tend to low affectation. On the contrary in executive functions, patients have values of 2, 3 and 4 – which tend to high affectation. This analysis leads to a profile with **medium affectation** in Attention and Memory and high affectation in EF.

Cluster 4 is the fourth biggest cluster. In attention, patients do not have any high values (only 0,1 and 2) and in memory and executive functions the patients do not have any low values (only 2,3,4 with a peak on 3). This is the case where a patient has a good level of attention but not a good performance in memory and in EF. This analysis leads to a profile with **low affectation** in Attention but with high affectation in Memory and in EF.

Cluster 5 is the fifth biggest cluster. In both attention and memory, patients have medium high values (on 2 and 3). In both functions, there is an almost null number of patients with a 1 or a 4. In EF, patient have medium low values (on 1 and 2) and an almost null number of patients with a 0. This analysis leads to a profile with **medium affectation** in Attention and Memory and medium low affectation in EF.

Cluster 3 is the smallest cluster. In attention, the patients do not have low values while in memory patients only have low values. In EF, there is only no record of a patient with a 0. This analysis leads to a profile with **high affectation** in Attention, low affectation in Memory and slightly high affectation in EF.

This general perspective suggests the existence of profiles of high (cluster 2) and low (cluster 1) affectation levels, independent from one another. The remain profiles differ in the level of affectation in the Attention function. One of the profiles has a high level of affectation (cluster 3), other has a low level (cluster 4) and the other two have a medium level. There is a considerable difference between the size of the clusters (the difference on the number of elements in the clusters is notable), which probably reflects, in a certain way, the importance of each cluster: it is conceivable that a cluster that contains a great number of patients describes better that population range that a cluster that has almost no patients in it. Another detail that needs to be mentioned in the analysis is the fact that there are about four clusters which are not very well defined, with very well distributed frequencies (clusters 3, 4, 5 and 6). Although their analysis does not allow a general description of the data, regarding the distribution of patients in the profiles considering only their Pre test scores, they certainly are important to evaluate the evaluation of the patients, considering the final cluster they will be assigned into. This could be easily explained: these four clusters represent profiles with very specific characteristics, and there is indeed a set of

patients that fall into those determined features. For example, there are around 135 patients (only 9 patients away from the second biggest cluster) that specifically have the biggest frequency of scores in the level 2 in Attention, the biggest frequency of scores in the level 3 in Executive Functions and that have the frequencies equally distributed only in the levels 1 and 2 in Memory. With this kind of information, more precise and trustable will be the comparison with the final profiles, allowing us to infer in a more personalized way about the evolution of the patient.

For the final clusters, the three Post functions were used as input. The model with the best silhouette was found by the SPSS Modeler to be the TwoStep with 3 clusters, with a value of 0.455 (that appears to be reasonable, or in a more technical word, fair), and a 1,0 of importance. The other two models found had lower values in both parameters, as showed in Figure 4.7:

Us...	Graph	Model	Build Time	Silhouette	Number of Clusters	Smallest Cluster	Smallest Cluster (%)	Largest Cluster	Largest Cluster (%)	Smallest/Largest	Importance
<input checked="" type="checkbox"/>		TwoStep 8	< 1	0,455	3	54	7	386	55	0,140	1,0
<input type="checkbox"/>		K-means 4	< 1	0,415	6	60	8	226	32	0,265	1,0
<input type="checkbox"/>		K-means 5	< 1	0,397	7	47	6	226	32	0,208	1,0

Figure 4.7 - Screenshot of the box used to select which one of the calculated SPSS models shall be used to build the final clusters, containing information about each one of them.

A fair value for the silhouette in the model of the 3 clusters means that this number of clusters is a sufficient estimation to represent the population. Nonetheless and once again, the model with three clusters was chosen not only for its good values in the parameters but also due the fact that its clusters represent very distinct and well-determined final profiles.

As it is seen in the Cluster Sizes diagram in Figure 4.8, representing the characteristics of the k-Means model with 3 clusters, there are two main groups (cluster 1 and 3, with percentages of 37.0% and 55.3%, respectively) and a small one that represents only 7.7% of the total number of patients.

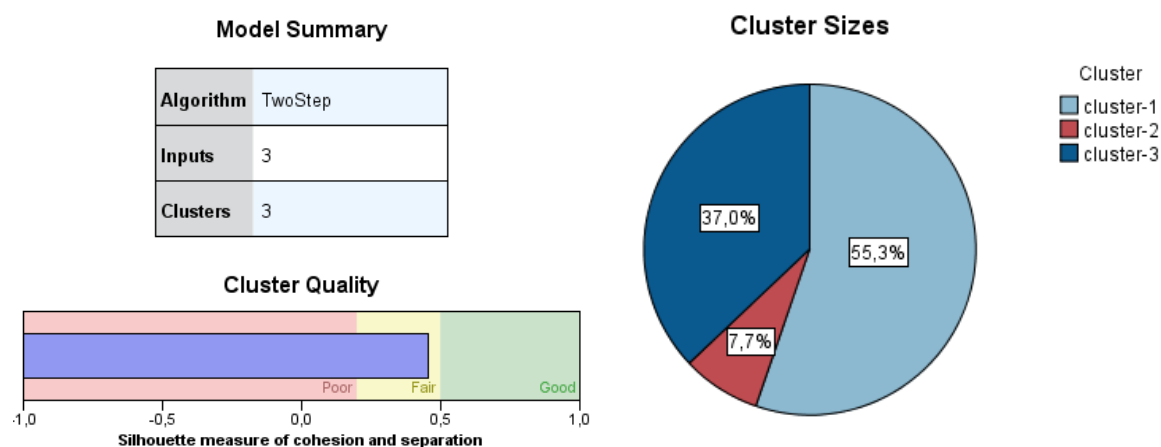


Figure 4.8 - Graphics showing the characteristics of the chosen TwoStep model with 3 clusters

In the following table (Figure 4.9) the percentages are again presented, but there is additional information regarding the real number of patients in each cluster. In cluster 1 there were 386 patients (55.3%), in cluster 2 there were 54 patients (7.7%) and in cluster 3 there were 258 (37.0%). Similar to what happened in the Initial Clusters table, here the shade in the Attention cells is darker than in the other functions. It makes sense that in both types of clusters this function is more important, as it is explained in the Discussion Section.

Size	55,3% (386)	37,0% (258)	7,7% (54)
Inputs	Atencion_Post 1,02	Atencion_Post 2,62	Atencion_Post 3,46
	FE_Post 1,35	FE_Post 2,98	FE_Post 1,87
	Memoria_Post 0,80	Memoria_Post 2,17	Memoria_Post 0,98
	1	3	2

Figure 4.9 - Information about each one of the three clusters regarding the Predictor importance for every input

The next step was to interpret the distribution of the patients score's in the final clusters, accordingly to their levels on the Post tests. The following preview (Figure 4.10) shows the results for the first idTreatments in the database. In this table, the final clusters appear in the column \$XC-Atencion_Post (or as they will be from now on called, Final Clusters). The tests used as inputs were all from the Post functions. In this last column, it is possible to find the cluster to which patient belongs to correspondingly to the model selected. In the same way as the Initial Clusters, the clusters in this case were calculated considering the values from the columns with the Attention, Memory and FE Post tests.

	idTreatment	Atencion_Pre	Atencion_Post	Memoria_Pre	Memoria_Post	FE_Pre	FE_Post	\$XC-Atencion_Post
1	2	0	0	1	0	1	1	1 cluster-1
2	3	1	1	3	1	1	1	1 cluster-1
3	8	1	1	3	2	2	2	3 cluster-3
4	13	3	1	2	1	2	0	0 cluster-1
5	14	4	3	1	1	4	3	3 cluster-3
6	15	4	4	4	3	4	4	4 cluster-3
7	26	2	3	1	1	1	1	1 cluster-2
8	29	3	3	2	2	1	2	2 cluster-3
9	30	2	1	4	1	2	2	2 cluster-1
10	33	0	0	1	0	2	2	2 cluster-1

Figure 4.10 - Table provided by SPSS that shows the assignation of the final clusters for each treatment, according to each function

Next, the cell distribution graphics were analyzed (Figure 4.11) in order to understand how the patient's Pre scores are distributed over each initial cluster, depending on the function. The representations for

Figure 1 displays a 3x3 grid of plots showing the evolution of probability distributions for three different cases (1, 2, 3) across three rows. Each plot has a y-axis from 0 to 600 and an x-axis from 0 to 4. The distributions are represented by red filled areas and light red outlines. The plots show how the distributions change over time or iterations, with peaks shifting and changing in height.

Also for this case, schematic representation of the profiles (Table 4.2) used during the analysis to have a more over-all perception of all the clusters. The table presented here summarizes all the three previous figures: a different color was assigned to each of the clusters and the transparency level of a color distinguishes the different frequencies within the same cluster. Each square resembles a score level in a specific function. The first row corresponds to Attention, the second row to Memory and the third row are the Execute Functions results.

[illegible]

In a general analysis, the colored cells in cluster 1 are only the ones on the left, while in cluster 3 they are only on the right area. In cluster 2, the cells in the Attention function are only on the right, the ones from EF are on the middle area and the ones from Memory are on the left part. This means the affectation levels are not very high in the Memory and EF, contrary to what happens in Attention, where the patients are severely affected. From the 3 clusters, one can easily distinguish that profile 1 is very characteristic of a patient with almost no affectation in all functions and that the profile 3 probably corresponds to a set of patients with signs of significant affectation in all functions (even though there are not so many patients with a score of 4 in the Post functions tests. From profile 2, where the real number of patients is low, the existing patients have their attention affected, the executive functions slightly affected and memory almost not affected at all.

Using the selected models, the next phase in this analysis was to plot histogram graphics for each cluster, to observe the distribution of patients through each one of the levels, for each function. The number of plots was, as expected, a multiple of three, since there was a plot for each main function and since this would happen as many times as the number of existing clusters. To consider only one cluster at a time, it was needed to tell the program to use just one cluster at a time. This was done by means of a condition.

The assignment of a condition in SPSS Modeler is accomplished by choosing the “Select” button from the “Record Ops” tab below and connecting it to the Initial Clusters icon. Starting with the selection of the patients in cluster 1, in the Settings section, the following script condition (Equation 4.1) was written:

$$'\$XC - Memoria_Post' = 'cluster - 1' \quad (4.1)$$

This condition tells the program which are the lines from the whole database to be selected (in the case below, it selects all the patients from the Initial Clusters that were assigned to cluster 1). This way, only the patients in cluster 1 were contemplated in the following analysis. The procedure explained here to obtain the histogram for this cluster was then used for all the six clusters.

The current preview of the data, in the following Figure 4.12, validates that the assignment was done correctly and that there are only patients from the selected cluster (there are only patients who belong to the initial cluster 1).

	idTreatment	Atencion_Pre	Atencion_Post	Memoria_Pre	Memoria_Post	FE_Pre	FE_Post	\$XC-Memoria_Post
1	2	0	0	1	0	1	1	cluster-1
2	26	2	3	1	1	1	1	cluster-1
3	33	0	0	1	0	2	2	cluster-1
4	61	1	1	1	0	2	1	cluster-1
5	98	1	1	1	0	1	0	cluster-1
6	118	0	0	1	0	1	0	cluster-1
7	122	1	0	1	0	1	2	cluster-1
8	157	1	2	2	0	2	2	cluster-1
9	188	1	1	2	1	1	2	cluster-1
10	232	0	1	1	1	3	1	cluster-1

Figure 4.12 - Resultant table from the exclusive selection of the patients belonging to initial cluster 1.

The “Select” node was then connected to a “Histogram” icon. These types of plots can be found in the “Graphs” tab. In this element, it is possible to choose the Field. In this case, three Fields were selected, matching each of the three Post functions, one at a time. For example, each “Select” node shall be connected to an Attention Post histogram, to a Memoria Post histogram and to an EF histogram, as it is shown in the next illustration (Figure 4.13):



Figure 4.13 - SPSS flow used to obtain the histograms for each function, considering only the patients from the initial cluster 1

Following the example of Cluster 1, the histogram graphics obtained by the flow exemplified in the figure above, for the three functions, were those presented in the following Figure 4.14:

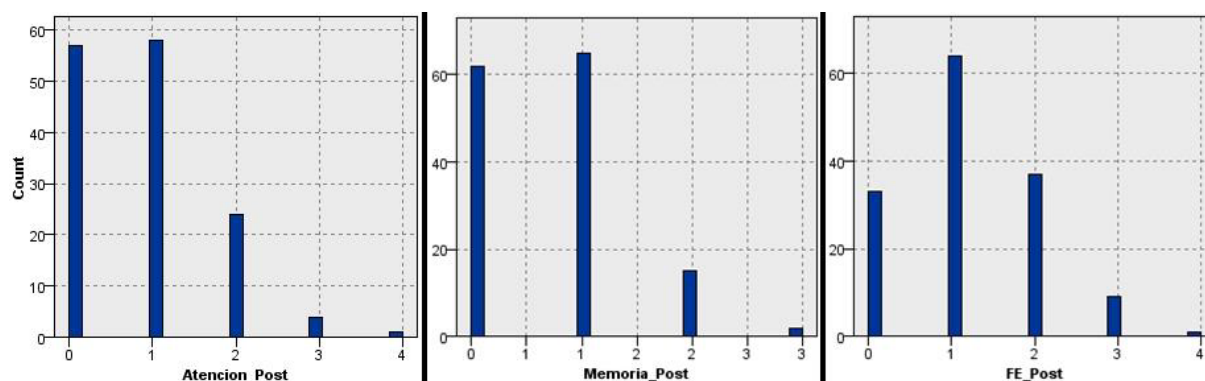


Figure 4.14 - Histograms showing the number of patients for each score and Post function, for cluster 1

The histograms reveal, in this case, for initial cluster 1, the number of patients that had presented one of the scores, for each Post function. The SPSS flow was built in order to provide histograms for the three functions, for all of the initial and final clusters. The information they contain is resumed in the graphics presented in the next section.

4.2 Distribution of tests scores by the Initial and Final Profiles

More important than having a global picture of the distribution of the Final Profiles by the Initial ones is to look at the scores in each Pre or Post functions and see how are they scattered in the profiles. Having already seen how are the Pre functions spread over the Initial clusters and how the Final clusters are formed with the Post functions scores, it was thought that it would also be interesting to observe it from a different point of view: how the initial clusters are defined in terms of Post functions results and how the final clusters are determined considering the scores in Pre functions. To do so, the crossings made were from Initial Profiles and Post functions and from Final Profiles and Pre functions and the results are showing in the next six graphics. In the following three graphics one can see the most important aspects on the distribution of the levels (normalized results) by Initial clusters for each function Post. In this whole process, SPSS Modeler is using the scores obtained in the Post functions tests as inputs to calculate the distribution of the patients through the six initial clusters, which allows us a backwards view.

Starting with the first graphic (Figure 4.15), for example, there were a lot of patients from the initial cluster 1 with a 0 or 1 score in the Post test in Attention. Other way to see it, there were only a few patients with high scores in the Post tests of this function. The contrary happens when looking at cluster 2, where most patients did not obtain good scores in their final tests. Moreover, besides cluster 1, all the initial clusters had almost no patients with scores of 0. On the other hand, in clusters 1 and 6, there are almost no patients with a score of 4 and additionally, in the first one, there almost no patients with a score of 3. In a general point of view, the values are well distributed by the clusters.

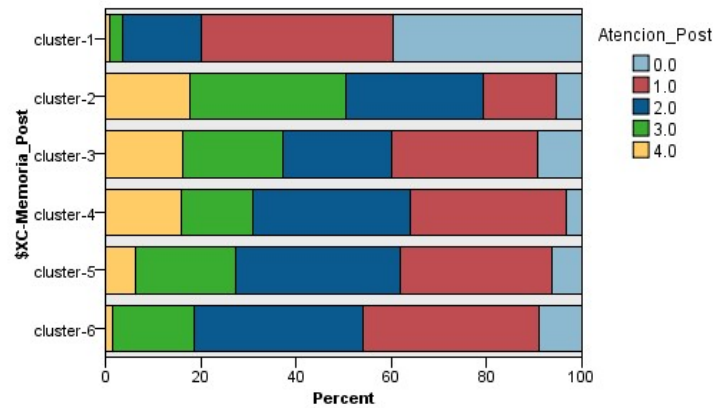


Figure 4.15 - Distribution of the scores obtained in the Attention Post function over the initial clusters

The same analysis can be done with Memory (Figure 4.16). Looking at the histogram, there are almost no patients with value 4: cluster 2 is the only exception with only a few patients with that score level. In cluster 3 there also almost no patients with values 2 or 3, therefore, it is basically composed by patients who got scores of either 0 or 1 in their final tests in this function. Something similar seems to happen also in cluster 1, where most patients who were assigned to that profile had a 0 or 1. In the remaining clusters (2, 4, 5 and 6), the biggest part of the patients has a value of 1 or 2 – the red and the dark blue colors are the most present.

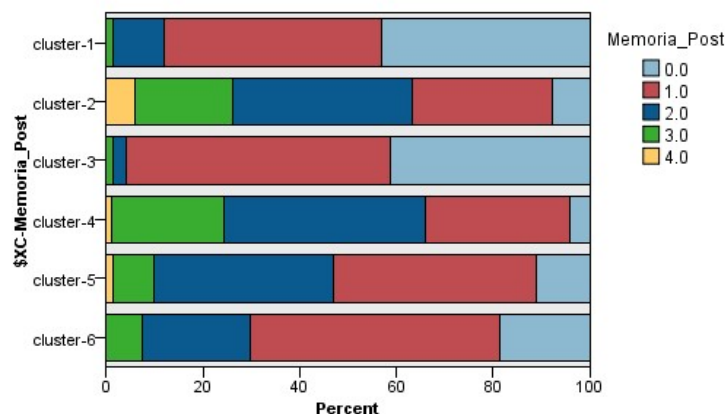


Figure 4.16 - Distribution of the scores obtained in the Memory Post function over the initial clusters

Regarding the Executive Functions (Figure 4.17), the profiles with the numbers 1, 3, 5 and 6 have almost no patients with a score of 4 in the Post tests, meaning that there are nearly no patients with severe affection in these functions. Clusters 2, 4 and 6 have almost no patients with value 0, which also means that in these profiles the patients did not improve in the best possible way.

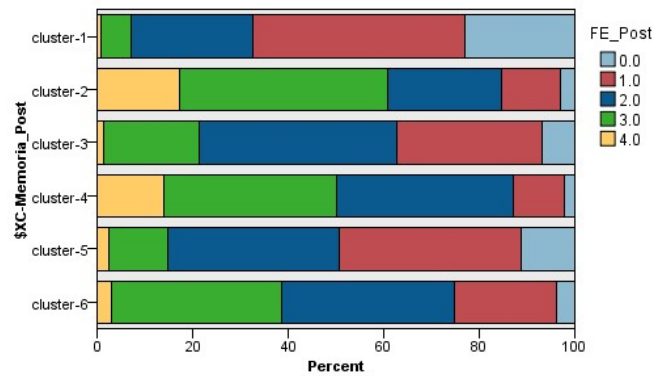


Figure 4.17 - Distribution of the scores obtained in the E.F Post function over the initial clusters

Another perspective is to think about the distribution of patients through the final profiles, considering their scores in the initial tests. For example, for Attention (Figure 4.18), only a few percentage of patients had a score 0 at the beginning. In addition, only some patients in profile 2 and 3 managed to reach a score of 1. This indicates that yellow, green and dark blue colors, that correspond to the “bad” scores (the scores associated with a great affectation profile), are the most present ones, which in practical terms, is not such a good result. Considering that the cluster 1 is a good profile and that the cluster 3 is a bad profile, it was supposed that the cluster 1 should be full of good scores (red and light blue) and that cluster 3 would be filled with yellows and greens.

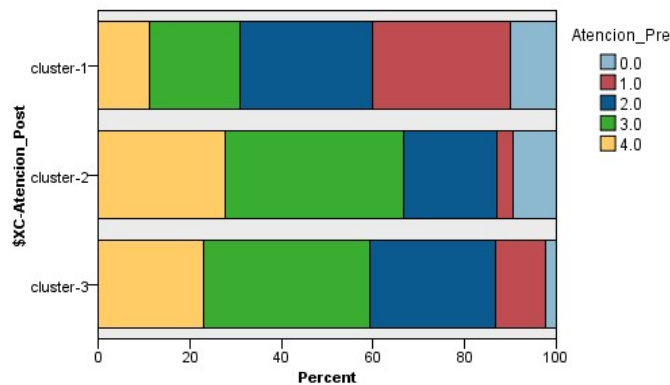


Figure 4.18 - Distribution of the scores obtained in the Attention Pre function over the final clusters

Cluster 3 has almost no patients with value 0, while cluster 2 has almost no patients with value 1. The rest of the values seem to be relatively well distributed over all three clusters. In Memory (Figure 4.19),

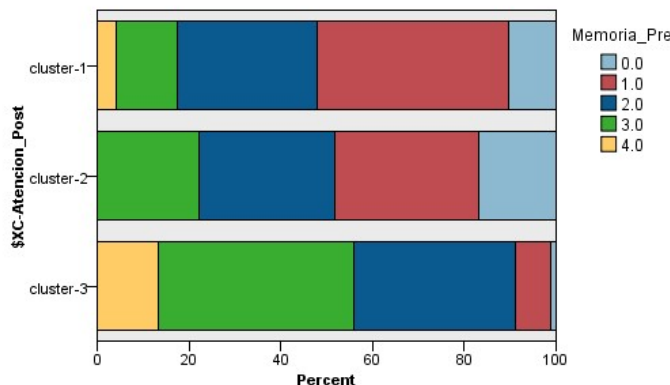


Figure 4.19 - Distribution of the scores obtained in the Memory Pre function over the final clusters

Cluster 1 only has a few patients with value 4 and cluster 2 has no patients at all with this value. Cluster 3 has almost no patients with value 0 and just a small number of patients with a 1. All the remaining values seem to be relatively well distributed over all three clusters. Comparing with the graphic from attention, the biggest difference relies on cluster 2, where the score of 1 has a much bigger percentage and where there are a lot of patients with a 4. The other two clusters have small differences. In regards to the Excecutive Functions (Figure 4.20), the amount of patients with scores of 0 is aproximately null and only a few had a score of 1. This is even more evident in cluster 3. The remaining values, represented in yellow, green and dark blue, seem to be relatively well distributed over the three clusters. In cluster 1 the biggest percentage of the patients had a 2, in cluster 2 there is a bigger quantity of score 3 and in the cluster 3 there is almost an equality between the scores 3 and 4 on the initial tests, which together occupy the biggest part of this cluster. Comparing with the other two previous functions, the most important aspect is that there are less patients with a 0 score on the 1st and 2nd profiles.

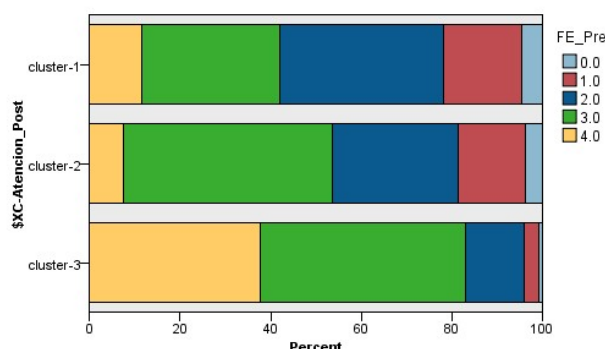


Figure 4.20 - Distribution of the scores obtained in the Executive Pre functions over the final clusters

In an overall view of the three graphics previously presented, it seems that in Attention and Executive functions, half of the graphics is occupied by the scores 3 and 4, which means, the scores that indicate a high level of affectation. In Memory this does not happen.

4.3 Relation between the Initial and Final Profiles

After selecting the models from the Pre and the Post tests, the results were combined, or merged, into a unique graphic. In this graphic one should be able to see to which Post cluster a patient from a certain initial profile was assigned to. This can be done by using the “Merge” element in the “Record Ops” tab and connect it to both Pre and Post product elements. In the “Keys for Merge” field, the idTreatment is selected, since this parameter is the primary key that connects both elements. Then, it is needed to connect the merge button to another graph, in this case, a Distribution graph (also available on the Graphs tab). In this graph, the field is set to be constituted by the Initial Clusters and that the Overlay (by color) corresponds to the Final Clusters. It can be normalized by color, which allows a general perspective of the distribution and a comparison between clusters. In case that in the Pre the model with 6 clusters is selected and that in the Post the model with 3 clusters is selected, the resulting distribution graphic is as follows (Figure 4.21). This is the graphic that satisfies one of the main objectives of this research work: observe and take conclusions about the evolution of a patient, considering the initial and final clusters that were assigned to him/her.

The analysis done in this graphic and in all the following graphics is based in a general view and not in exact percentage values. This way, the margins of error in the interpretation of the graphics might be influenced by some factors. Also, the results presented might be interpreted in many different ways. To

avoid an extensive interpretation of all the graphics a selection of what it seemed to be the most important details to noticed in each one was done.

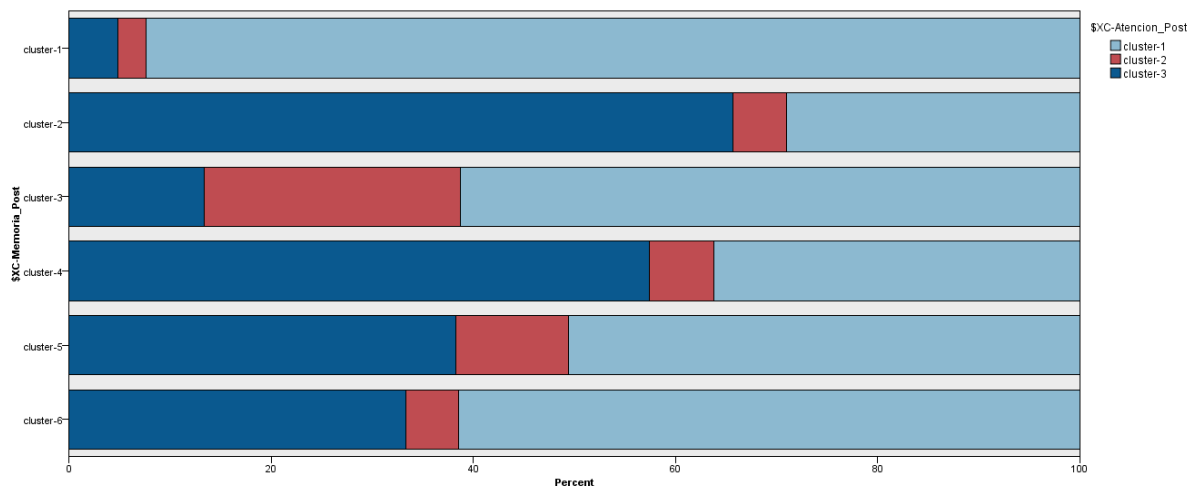


Figure 4.21 - Distribution graphic of the final profiles (3) by the initial profiles (6)

By analyzing and interpreting the graphic, valuable information can be taken. In an overall view, one conclusion can be taken: there are patients from all initial clusters evolving to all final clusters, which means, there are transitions of patients from initial clusters 1,2,3,4,5 or 6 to all the final clusters 1, 2 and 3. In practice terms this also means that there were some patients that went from a low affectation initial profile to a high affectation final profile and vice-versa.

For example, most of the patients from the **first initial** profile end up in the **first final** profile. This makes sense because if a patient has good scores before the treatment, it is also expected that he will not have a worst performance in the final tests. Indeed, it can happen for a patient that initially did not have a high affectation profile to get worse in the end, but it is not very common and that is the reason why there are not many patients going from the first initial profile to the second or third final profiles.

The biggest part of the patients from the **second initial** profile go to the **third final** profile. In an ideal situation, considering that the treatment was successful, patients belonging to an initial great affectation cluster would, somehow and admittedly, improve even almost insignificantly in their final scores. Truth is, only approximately one third of the patients from the **second initial profile** ended up in the **first** or **second final profiles**. This is because, in most cases, and specially when patients have acute and severe affectation levels in Attention they do not get better with the treatment in any of the functions. However, it must also be pointed out that even though the final cluster 3 also suggests a high affectation profile, it looks better than the initial cluster 2. A detailed analysis shows that the final cluster 3 also includes patients who obtained scores of 1 or 2 in the Attention and Memory functions - there were some patients that actually might have improved in those fields. Nonetheless, from all the initial profiles, this is the situation where there is a higher percentage of patients in the final cluster 3, the worst-case scenario of the final profiles.

Regarding the **third initial** profile, more than half of the patients evolve to the **first final** profile. Noticing that in the third initial profile, attention was bad and that executive functions were not much better, the fact that there was an improvement in those functions for a great part of the patients in the

initial cluster 3 is a great accomplishment. As memory was already not so affected, it was expected to stay the same.

Also, most patients from the **fourth initial** profile evolve to the **third final** profile. This can be considered the second worst situation in these results, since in the fourth initial profile the scores in the attention function tests were relatively good. It is true that Memory and Executive Functions were bad and basically, they remained bad in the biggest percentage of the cases. Only a little more than one third of the patients from this initial cluster evolved to the first or second profiles.

Patients from the **fifth initial** profile evolve approximately in the same way to the **first** as they do for the **second** and **third final** profiles together. In this initial profile there was no scores of 0, what can be considered as a good sign. There were patients who improved, especially in the Attention and Memory functions. There was also a significant percentage of patients that ended up in cluster 3, which means that they probably did not improve or even got worse.

Finally, the biggest part of patients from the **sixth initial** profile evolves to the **first final** profile. These were great results because in this initial profile there were not scores of 0 in any of the functions and there were inclusive patients with a score of 4 in Executive Functions. The biggest percentage of patients from cluster 6 going to cluster 1 means that the greatest part of the patients improved their test scores after treatment. The number of patients evolving to the **second final** profile is, in almost all cases, very low. The only exception is the set of patients from the **third initial** profile.

4.4 Influence of other variables in the evolution of a patient

In this second part of the study, the aim was, as it was indicated before, observe if some variables included in the database could possibly be descriptors of the improvement of a patient. That being said, the six available variables used were the **Gender** of the patient, the **Age** of the patient when the injury occurred, the **Study Level** of the patient, the **Break** - which means the time that passed between the day of the injury and the day when the patient started the treatment, the **Number of tasks** - the number of tasks performed by the patient during the treatment and the **Duration**, which stands for the duration of the treatment. The procedures to obtain the next plots were the same followed in the first part, although here the inputs used were the variables considered in each case. As all of the last three variables are continuous variables, some statistics were performed in order to obtain classes of values to better represent the data in graphics:

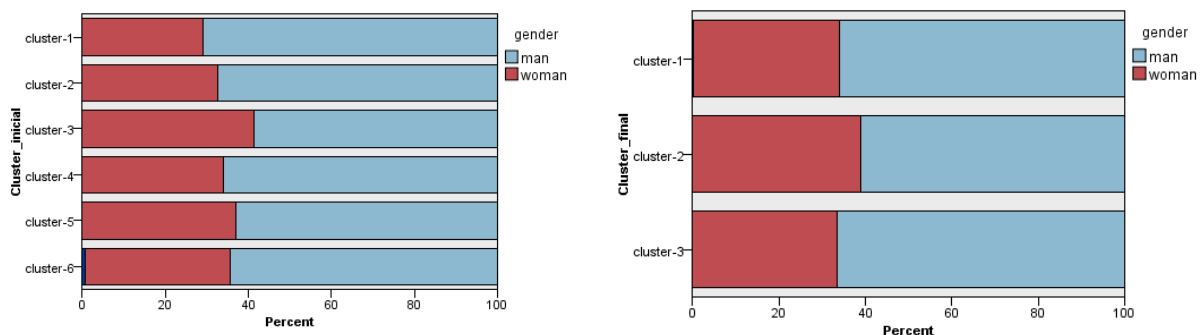


Figure 4.22 - Distribution of the gender, considering the initial (left) and final (right) clusters

A small analysis regarding the gender of the patients was also performed (Figure 4.22). Accordingly to the studies performed by Niemeier et al., women could have an advantage in post injury recovery, particularly in executive functions, when compared to men. Here, the graphics for both genders according to the initial and final clusters do not have considerable differences to the naked eye. After observing the graphic of the initial clusters, it can be assumed that in all of them, the distribution of men and women is very similar between clusters. The percentage of men is, in all clusters, clearly higher than the percentage of women but that could be explained by the fact that the population has more male than female patients. In the initial cluster 6 there are a few null records, of patients who do not have gender information available. Similar to what happens with the initial clusters, the distribution of men and women looks the same for all three final clusters. In all of them, the percentage of men is also higher than the percentage of women and in both initial and final high affectation profiles the percentage of women is bigger than in the both initial and final low affectation profiles.

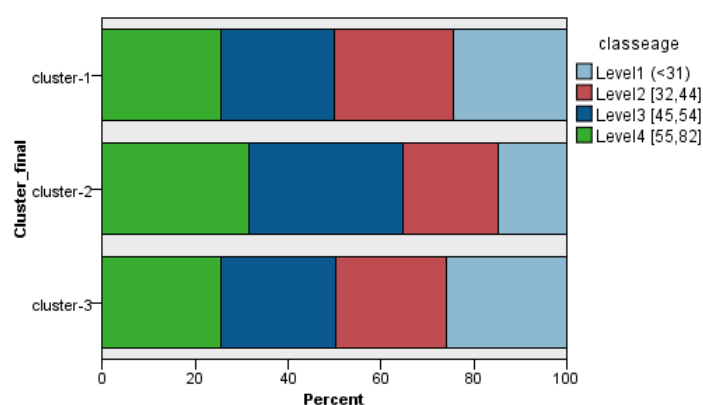


Figure 4.23 - Distribution of the age classes of the patients over the final clusters

The distribution of the age classes (in years), presented in Figure 4.23 over the final clusters seems to be very homogeneous. The class with biggest percentage of patients is different for each cluster. In the case of cluster 1 and 3, all four classes seem to have the same percentage. This means that the age does not have an influence on the assignment of the final profiles. If it had influence, e.g. if in cluster 1 there was a bigger percentage of patients aged 30 (or younger) than older patients (>55), it would mean that younger people could have a better final response to the treatment than the older ones. However, when observing the graphics, that does not happen. Moreover, percentages in final cluster 1 look almost equally distributed as in the final cluster 3. No valuable conclusions could be taken from this analysis.

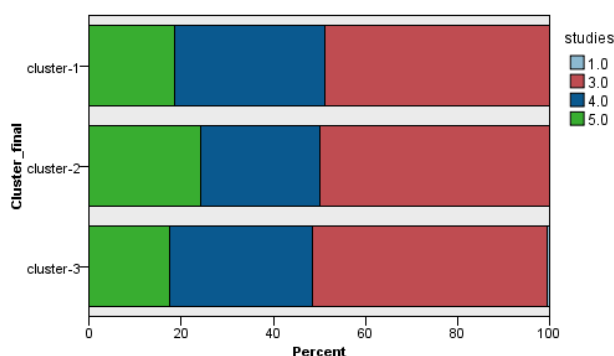


Figure 4.24 - Distribution of the studies of the patients over the final cluster

The distribution of study level per final cluster (Figure 4.24) is also very homogeneous. The only peculiarity to point out is that as there are almost no patients in the population with a study level of 1 or 2, the distribution of patients through the clusters accordingly to this variable one complains the levels of 3, 4 and 5. In all clusters, the biggest percentage of patients – almost half - has a level 3. Without being able to extract an exact conclusion, it looks like, in this dataset, there is no influence of the study level in the assignment of the final profiles to the patients.

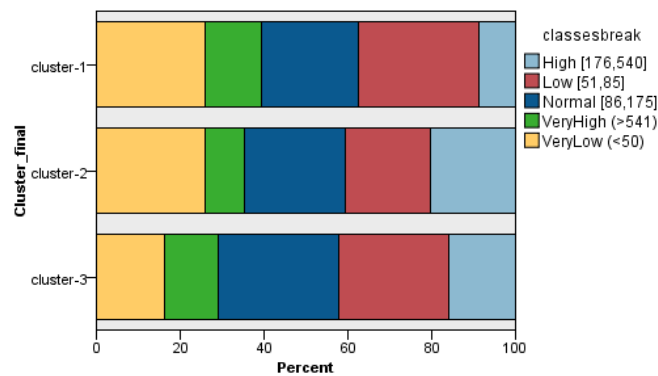


Figure 4.25 - Distribution of the break classes of the patients over the final clusters

Regarding the distribution of the Break classes by the final profiles (Figure 4.25), they were divided in Very High, High, Normal, Low and Very Low. In general, patients belonging to the different classes seem to be relatively well distributed over all the three final clusters. Although, there are some little differences that must be noticed. For example, considering that the colors green, dark blue and red are more or less equally distributed, and comparing only the two opposite clusters (1 and 3), in the first cluster, the color yellow - correspondent to a very low interval (less than 50 days) - occupies a larger space than it does in the third cluster. Also, the light blue occupies more space in cluster 3 than in cluster 1. This is indicative that in a profile of lower affection, the percentage of patients that took less time to start the treatment is bigger than the percentage of patients that took much more time. A conclusion to take from here would be that a smaller interval of time between the injury and the start of the treatment could lead to an improvement of the patient's state. Considering the distribution of the number of tasks levels by the final profiles, the results are presented in Figure 4.26.

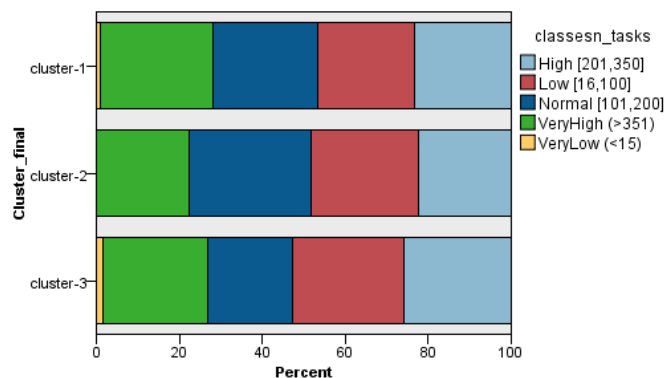


Figure 4.26 - Distribution of the number of tasks of performed patients over the final clusters

In all three clusters, there are almost no patients with less than 15 tasks. Regarding the other classes levels it seems that there is a very good and equal distribution of the patients over the clusters. It can be concluded that there is not a visible influence of the number of tasks in the evolution of a patient's state. The same happens when looking at the distribution of the duration of the treatment classes (in days) over the final clusters (Figure 4.27): all the classes are almost equally distributed. Once more, as far as a conclusion can be taken, there is not also a considerable influence of duration of a treatment on the evolution.

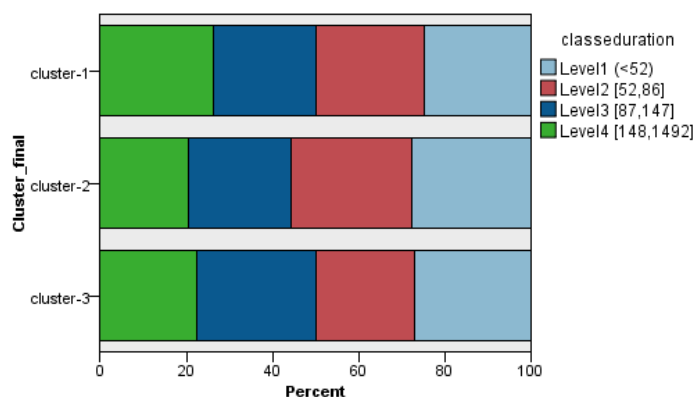


Figure 4.27 - Distribution of the duration of the patient's treatment over the final clusters

4.5 Improvement, Worsening and Maintenance

Regarding the scores and not only the clusters, an Excel file (Figure 4.28) was organized in a way that would make it possible to see which are the patients in each of the three common evolution types: Maintenance (M), Improvement (I) and Worsening (W). The values showing the variations in Attention, Memory and Executive Functions was determined by using simple Excel formulas.

	A	B	C	D	E	F	G	H	I	J	K	L
1	idTreatment	Initial Profile	Atencion_Pre	Atencion_Post	Δ Atencion	Memoria_Pre	Memoria_Post	Δ Memoria	FE_Pre	FE_Post	Δ FE	Final Profile
2	2.000	cluster-1	0	0	0	1	0	-1	1	1	0	cluster-1
3	33.000	cluster-1	0	0	0	1	0	-1	2	2	0	cluster-1
4	61.000	cluster-1	1	1	0	1	0	-1	2	1	-1	cluster-1
5	98.000	cluster-1	1	1	0	1	0	-1	1	0	-1	cluster-1
6	118.000	cluster-1	0	0	0	1	0	-1	1	0	-1	cluster-1
7	122.000	cluster-1	1	0	-1	1	0	-1	1	2	1	cluster-1
8	157.000	cluster-1	1	2	1	2	0	-2	2	2	0	cluster-1
9	188.000	cluster-1	1	1	0	2	1	-1	1	2	1	cluster-1
10	232.000	cluster-1	0	1	1	1	1	0	3	1	-2	cluster-1
11	272.000	cluster-1	1	1	0	1	1	0	1	2	1	cluster-1
12	316.000	cluster-1	1	1	0	2	1	-1	2	2	0	cluster-1
13	330.000	cluster-1	2	2	0	1	0	-1	1	2	1	cluster-1
14	367.000	cluster-1	1	0	-1	2	2	0	2	1	-1	cluster-1
15	408.000	cluster-1	1	1	0	2	1	-1	1	1	0	cluster-1
16	459.000	cluster-1	1	0	-1	1	1	0	1	1	0	cluster-1
17	494.000	cluster-1	2	2	0	0	1	1	1	2	1	cluster-1
18	507.000	cluster-1	1	1	0	2	1	-1	1	1	0	cluster-1
19	538.000	cluster-1	1	1	0	0	0	0	2	0	-2	cluster-1
20	570.000	cluster-1	1	1	0	1	0	-1	1	0	-1	cluster-1
21	616.000	cluster-1	1	0	-1	1	0	-1	2	1	-1	cluster-1
22	653.000	cluster-1	1	1	0	1	0	-1	1	1	0	cluster-1
23	793.000	cluster-1	1	1	0	2	2	0	2	2	0	cluster-1
24	814.000	cluster-1	1	1	0	1	1	0	2	1	-1	cluster-1
25	841.000	cluster-1	1	0	-1	2	1	-1	2	1	-1	cluster-1
26	845.000	cluster-1	1	1	0	1	0	-1	1	1	0	cluster-1
27	871.000	cluster-1	1	0	-1	1	0	-1	0	0	0	cluster-1
28	901.000	cluster-1	1	0	-1	1	1	0	2	1	-1	cluster-1
29	927.000	cluster-1	0	2	2	0	2	2	0	1	1	cluster-1
30	936.000	cluster-1	2	2	0	1	1	0	0	0	0	cluster-1
31	963.000	cluster-1	0	1	1	1	0	-1	1	2	1	cluster-1

Figure 4.28 - Excel file showing by colors, the patients from cluster 1 who improved, who got worse or did not evolve, in each of the functions.

In this table, there is an example for Cluster 1, on how the analysis was performed first in a general way: just by looking at the most prevalent colors in each of the functions. The Excel file contains the differences between the scores obtained in Pre and Post functions that were calculated with a simple difference formula. In the present screenshot, it shows for cluster 1 the evolutions for Attention, Memory and Executive Functions by using color green for the improvements (when the differences between scores is less than 0, grey for the maintenance cases (difference equal to zero) and red for the worsening cases (when the subtracting the Pre score to the Post score the value obtained was bigger than zero). For example, in this case, Memory is full of green cells that indicate a lot of improvements in this function and Attention has more grey cells than any other functions, this is, most patients maintained their score level.

For each patient, associated with an idTreatment, the Global Improvement (GI) was determined, accordingly to the conditions set in section 3.2.3 from the Methods. The GI parameter considers the evolution levels in the three functions and defines a value that resumes the general state of the patient. It was considered that a 1 in GI was an Improvement, a 2 was a Worsening and a 3 a Maintenance. After the total counting, who maintained their state. Next, three tables from Excel are presented, each one corresponding to an evolution type, with the scores of tests performed before and after the treatments, the initial and final calculated profiles, for each idTreatment.

Considering the improvements (Figure 4.29), there was a total of 453 patients (64.9%) whose scores, in general, evolved in a very positive way.

	A	B	C	D	E	F	G	H	I	J	K	L	N
1	idTreatment	Initial Profile	Atencion_ Pre	Atencion_ Post	Δ Atencion	Memoria_ Pre	Memoria_ Post	Δ Memoria	FE_Pre	FE_Post	Δ FE	Final Profile	Global Improvement
2	2.000	cluster-1	0	0	0	1	0	-1	1	1	0	cluster-1	1
3	3.000	cluster-5	1	1	0	3	1	-2	1	1	0	cluster-1	1
4	13.000	cluster-5	3	1	-2	2	1	-1	2	0	-2	cluster-1	1
5	14.000	cluster-3	4	3	-1	1	1	0	4	3	-1	cluster-3	1
6	15.000	cluster-2	4	4	0	4	3	-1	4	4	0	cluster-3	1
7	30.000	cluster-4	2	1	-1	4	1	-3	2	2	0	cluster-1	1
8	33.000	cluster-1	0	0	0	1	0	-1	2	2	0	cluster-1	1
9	39.000	cluster-6	2	2	0	2	1	-1	4	3	-1	cluster-3	1
10	45.000	cluster-2	4	3	-1	3	2	-1	4	2	-2	cluster-3	1
11	53.000	cluster-5	3	1	-2	2	2	0	2	1	-1	cluster-1	1

Figure 4.29 - Excel file showing some of the patients who improved and their correspondent initial and final scores and clusters

Regarding the worsening cases (Figure 4.30), there were 76 (10.9%) patients who got worse or the same scores in the Post functions then they did in the Pre ones.

	A	B	C	D	E	F	G	H	I	J	K	L	N
1	idTreatment	Initial Profile	Atencion_ Pre	Atencion_ Post	Δ Atencion	Memoria_ Pre	Memoria_ Post	Δ Memoria	FE_Pre	FE_Post	Δ FE	Final Profile	Global Improvement
2	26.000	cluster-1	2	3	1	1	1	0	1	1	0	cluster-2	2
3	29.000	cluster-5	3	3	0	2	2	0	1	2	1	cluster-3	2
4	64.000	cluster-6	2	3	1	2	2	0	3	3	0	cluster-3	2
5	142.000	cluster-6	3	3	0	2	3	1	3	3	0	cluster-3	2
6	145.000	cluster-5	2	3	1	2	2	0	1	2	1	cluster-3	2
7	146.000	cluster-2	3	3	0	3	4	1	4	4	0	cluster-3	2
8	272.000	cluster-1	1	1	0	1	1	0	1	2	1	cluster-1	2
9	323.000	cluster-1	0	3	3	1	1	0	0	2	2	cluster-2	2
10	337.000	cluster-5	2	4	2	2	4	2	2	4	2	cluster-3	2
11	349.000	cluster-3	3	3	0	1	1	0	2	3	1	cluster-3	2

Figure 4.30 - Excel file showing some of the patients who got worse and their correspondent initial and final scores and clusters.

Concerning the people that maintained their scores (Figure 4.31) there were 169 patients (24.2%) who got worse, better or maintained their scores in some or all the functions.

	A	B	C	D	E	F	G	H	I	J	K	L	N
1	idTreatment	Initial Profiles	Atencion_Pre	Atencion_Post	Δ Atencion	Memoria_Pre	Memoria_Post	Δ Memoria	FE_Pre	FE_Post	Δ FE	Final Profiles	Global Improvement
2	8.000	cluster-4	1	1	0	3	2	-1	2	3	1	cluster-3	3
3	34.000	cluster-6	2	1	-1	1	2	1	2	2	0	cluster-1	3
4	56.000	cluster-4	0	4	4	2	2	0	3	2	-1	cluster-2	3
5	69.000	cluster-6	2	3	1	2	1	-1	3	2	-1	cluster-2	3
6	100.000	cluster-5	2	3	1	2	1	-1	2	3	1	cluster-3	3
7	115.000	cluster-6	2	3	1	2	1	-1	3	3	0	cluster-3	3
8	122.000	cluster-1	1	0	-1	1	0	-1	1	2	1	cluster-1	3
9	135.000	cluster-2	3	4	1	3	2	-1	3	3	0	cluster-3	3
10	144.000	cluster-3	3	4	1	1	0	-1	3	2	-1	cluster-2	3
11	148.000	cluster-4	2	3	1	4	3	-1	4	3	-1	cluster-3	3

Figure 4.31 - Excel file showing some of the patients who did not show evolution and their correspondent initial and final scores and clusters.

In the Improvements table (Figure 4.29), there are no red cells, which means there are no people who got worse in any of the functions. In the Worsening table (Figure 4.30), there are no greens in the entire table, which means that all the patients maintained or got worse in all the functions. In the Maintenance table (Figure 4.31), the variety is much bigger, for there are red, green and grey cells.

4.5.1 Factors that may influence the type of evolution

Having the information about the patients who improved, maintained their score levels or got worse ones, it was possible to observe if the variables previously considered could have influenced each one of this evolution types. In regards to age, study level and gender, the graphics presented in the next three figures show for the patients whose Global Improvement was 1, 2 or 3, respectively, the effects of these three features.

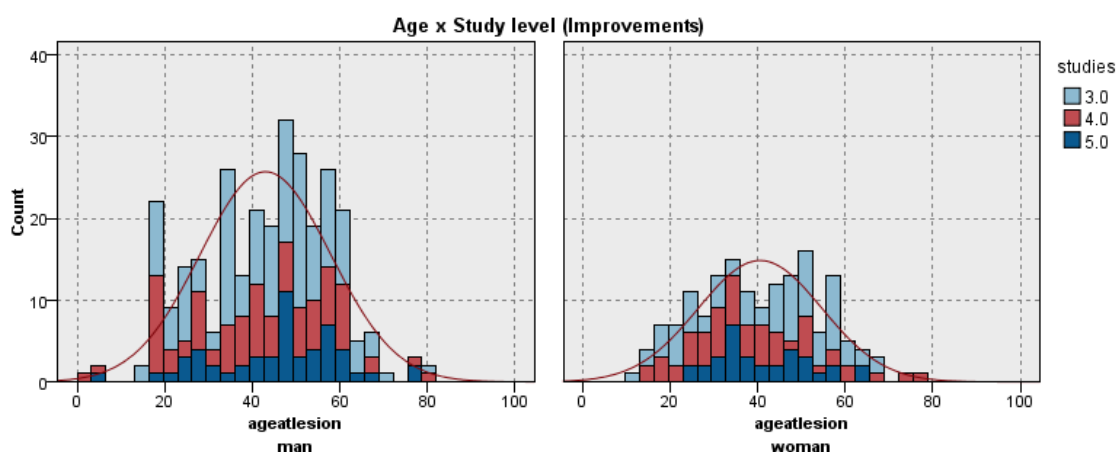


Figure 4.32 - Histograms crossing the variables Age, Study level and Gender, using data from the patients who had a global improvement of 1.

Regarding the variables assessed in the two histograms from Figure 4.32, the first thing to notice is the fact that there are more male patients than female patients. The second is that, in both cases, there are no great visible differences in relation with the study levels – for both genders it seems that, in general,

there are more people with a study level of 3 than with a 4 or 5. The peak of the age, which also corresponds to the peak of the normal curve is slightly more to the right for the male graphic.

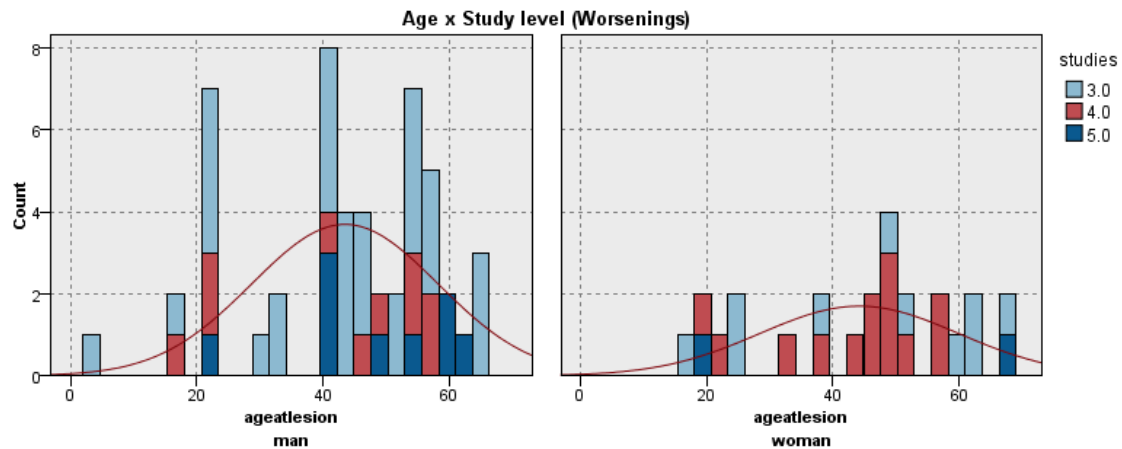


Figure 4.33 - Histograms crossing the variables Age, Study level and Gender, using data from the patients who had a global improvement of 2.

In the worse cases (Figure 4.33), the age ranges are much more scattered than in the improvements. Amongst the male patients, who also sum a bigger count than women, it seems that most of them are older than 40 years old, while women have better their age better distributed. Regarding the study level, there is a difference between genders: most part of the male patients who got worse have an education level of 3, while most part of the female patients in that same situation had a 4. Comparing these graphics with the improvement ones, besides the great difference on the number of patients, there is only a slight difference, considering that in the improvements population has a better distribution between the 20 and 60 years old and in the worsening evolution type, there are more old than young people (most are older than 40, for both genders). Regarding the study levels, there are very few patients with a study level of 5 in the worsening population, especially when comparing that number with the one from the improvements one.

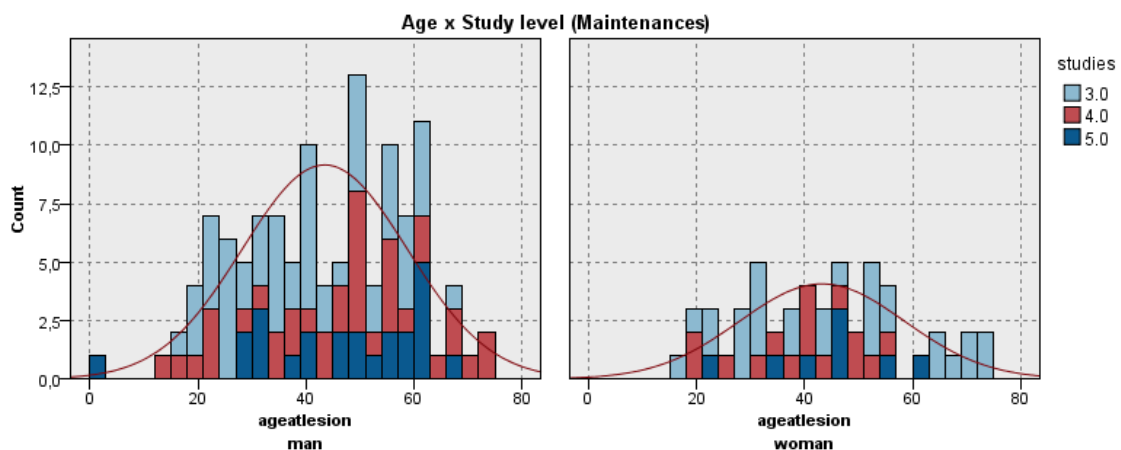


Figure 4.34 - Histograms crossing the variables Age, Study level and Gender, using data from the patients who had a global improvement of 3.

Concerning the people that maintained their level (Figure 4.34), men are, once again, in larger number and the ages are well-distributed. The study levels are also relatively well spread, although patients with a 5 are in less number than the ones with a 4 and in an even smaller number than the ones with a 3.

Comparing these two graphics with the previous ones, they have more similarities with the improvements one.

Propagating this study to the three other variables considered, the results for the improvement, worsening and maintenance were as shown in the following figures (Figure 4.35, Figure 4.36 and Figure 4.37). For the interval between the injury and the start of the treatment (Break), only patients who started their treatment less than 800 days (a little bit more than 2 years) after the injury were taken into account, since there were only a few patients who took much longer than that to start and that were considered as outliers for the purpose of this study. The differences between the three graphics are very small. The only thing to point out is that comparing the improvements and the worsening situations, in the second group a there is a bigger percentage of patients who took longer than 400 days.

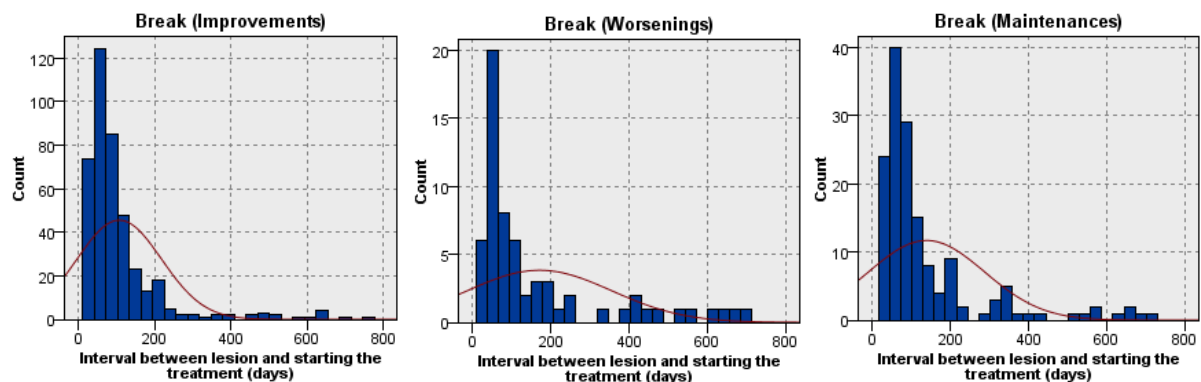


Figure 4.35 - Histograms showing the distribution of people according to the time they took to start the treatment after the lesion (in days), for each of the three evolution types.

For the duration of the treatment (Duration), the histograms show that there is not a big difference between the three evolution levels. In all of them, most of the patients' treatments lasted for less than 200 days. In the Worsening population, there was only one patient whose treatment lasted for more than 400 days and. When comparing this population with the improvements one, there are more patients beyond the line of the 200 days, which means, there are more patients whose treatments lasted longer.

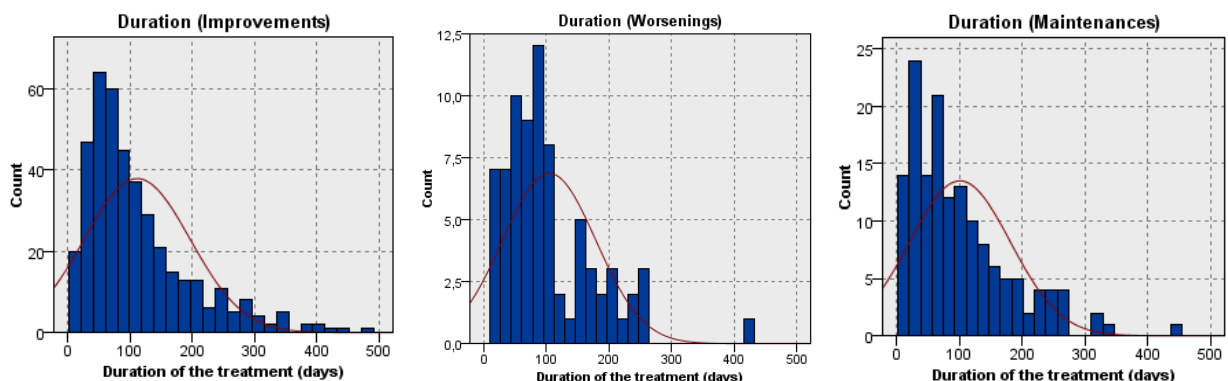


Figure 4.36 - Histograms showing the distribution of people according to duration of their treatment (in days), for each of the three evolution types.

Observing the number of tasks performed in the entire rehabilitation treatment (n_tasks) and comparing the three histograms, it appears that there are almost no variances, excepting some punctual ones.

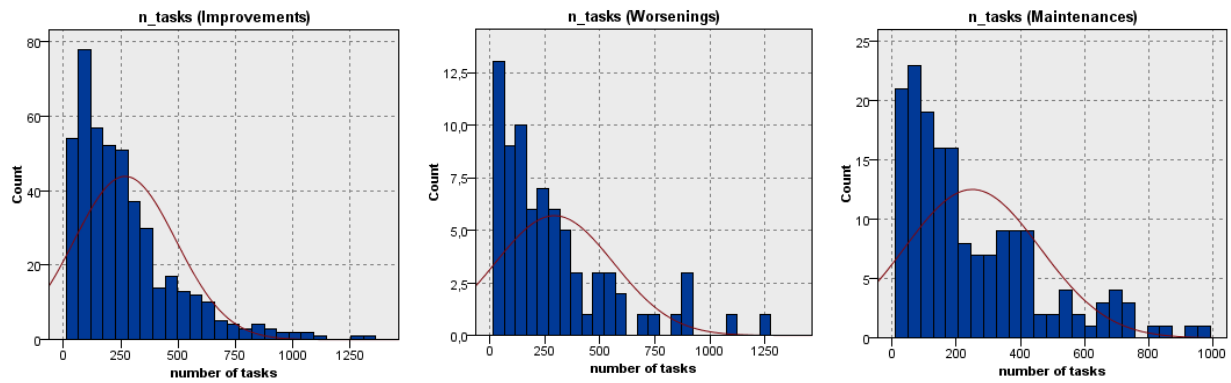


Figure 4.37 - Histograms showing the distribution of people according to the number of tasks they performed in their treatment, for each of the three evolution types.

The comparisons between the results obtained for the variable's influence taking into account the calculated clusters and considering only the scores obtained before and after the treatments constitute another way to validate the computed clusters and are established in the Discussion part.

5 Discussion

Before starting to interpret the results the first point to be considered is that the way the tests were performed by the patients can influence the scores obtained. This includes the psychological state of the patient in the moment: his/her motivation, his/her anxiety level or the interest he shows in performing the test. Other conditions that may affect the performance are the restless levels, suffering from a depression, experiencing the side effects from drugs' consumption or even the time of the day that the tests are executed. The validity of the tests scores used in this study is compromised by all of these factors.

Regarding the calculation of the affectation in the functions, it can be improved. For this study, the evolution of a patient in a function, like Attention, can be calculated even if the patient has only one test item in Pre and another in Post, with both not having much to do with each other (even if they are from the same function, they might evaluate different subfunctions). Truth is, although the clinicians confirmed that the two different items could still be used to make the estimation, logically that brings some noise into it. Ideally, there would be only patients who have the same Pre and Post test item for all functions (even if they only have one test per function). This might be considered as a future work suggestion.

For the initial profiles, the model with 6 clusters was chosen and for the final profiles the model with 3 clusters. Both were chosen after considering not only two important parameters (silhouette and importance level) but mostly after looking at the general panorama and try to understand from which ones more valuable information could be taken. A model that has good values of silhouette and importance but does not allow a good interpretation of the clusters and the features of its elements is not considered a good model for this specific situation. In this case, having a number of 6 initial clusters allowed us to consider not only patients with “good” or “bad” scores, but also other groups of people with specific characteristics. In an ideal situation, and in order to optimize the personalization of the treatments, there would be as many clusters as the number of groups of people with very well discriminated characteristics. Those clusters would not be too big or too small and would represent the reality in a correct way. Here, the chosen number of initial clusters, considering the differentiating characteristics of the elements that constitute them and their size, was considered a suitable one.

The fact that the model chosen for the final clusters only had three well determined groups allowed us to see if the patients had, in general, evolved in a good or in a bad way. Observing the initial clusters, it can be concluded that the biggest part of the population was placed in the two most differentiated clusters: the one with the best scores and the one with the worst scores, in all three functions. The smallest part of the data was placed in clusters not with features very well defined but very heterogeneous. While the two biggest clusters have data points that only have good scores or that only have bad scores, the smallest clusters have elements with, for example, good scores in Attention but bad scores in Memory. Although this is a very normal situation, it was more difficult to understand the evolution of the patients in these situations. Nonetheless, and once again, the fact that the final clusters had great differences between them, helped getting a better perception of the evolution of not only the patients in the well-defined initial clusters but also of the patients in the smallest ones.

Choosing a model with three clusters over a model with five or six clusters for the description of the data might bring the risk of having a cluster with too many elements. If that is the case, there is always

the possibility that the cluster is not a good descriptor of its elements, meaning there could be pairs of elements in that cluster that have great differences between them and those differences might not have been taken into account. A big cluster usually appears with a higher frequency when the study needs to be more general. For example, the existence of clusters with a higher number of elements is more common when those elements are supposed to be placed either in a positive or in a negative cluster, no other options are considered. In the case of the selected models, it happens that two out of three clusters are very well differentiated, opposite to one another.

In the first part of the results, a histogram establishing the relation between the initial and the final profiles was shown. That relation was calculated by using a clustering method to set the groups. For a better understanding of the histogram, the following illustration (Figure 5.1) was conceived. The colored icons represent the Pre tests and the black, white and grey, the Post functions. This is to demonstrate that each Post function is build up of members from all Pre clusters and that each element was evaluated in each one of the three functions.

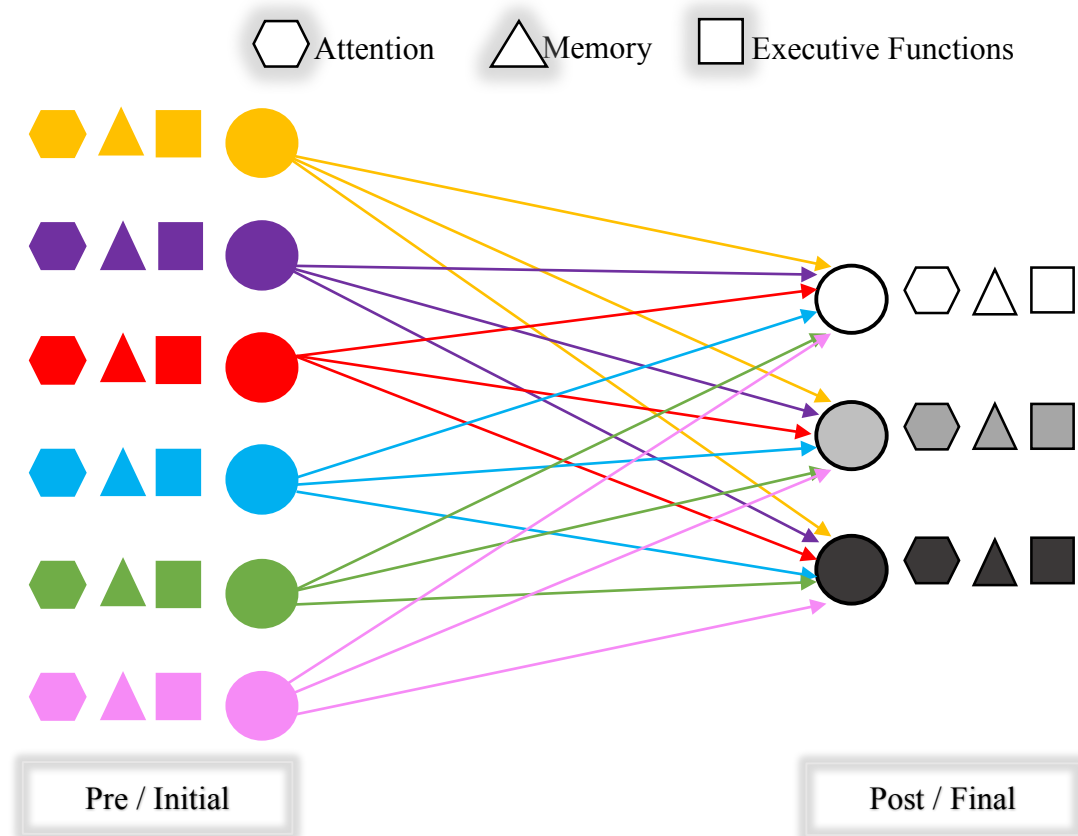


Figure 5.1 – Schematic representation of the initial clusters and final clusters and the relation between all of them

The Results section, shows that the population at the end, in general, has less affectation, as expected. Also, the migrations observed between the initial and final clusters seem to be very reasonable. Concerning the particular event of some elements from the “good” clusters, migrating to “bad” clusters, there is not an easy justification for it. It can be explained, for example, by the conditions described in the beginning of this section. On the other hand, it is very plausible that some observations go from the

“bad” initial cluster to the “bad” final clusters: if the patient has an acute or severe affectation level it might happen that the rehabilitation treatment is not sufficient for his/her recovery.

Looking at the number and type of final clusters and comparing it with the different type of evolution possible it seems there could be a correspondence between both. Nonetheless, finding relations between improvement, worsening and maintenances with the three clusters was not the objective of this study. As it was mentioned already, among the final clusters, there is one whose scores are very “good”, there is another that has very “bad” scores and there is a third one which includes patients with both very “good”, very “bad” and normal scores. Also, evolution types include improvements, characterized by their excellent final scores, people who worsened and got poor scores and who did not improve or got any worse in their tests, just maintained. At first sight, a relation could be established between the final cluster 1 and the patients who improved, between the final cluster 2 and the people who maintained and between the final cluster 3 and the ones who got worse but, indeed, it could be considered as coincidence. That correspondence would be considering that all the patients who improved were the ones who ended up in the final cluster one and that all the patients who got worse were the ones placed in the final cluster 3. Of course, this relation is not simple at all. Patients from the improvement population got better scores in, at least, one of the functions and did not get worse in any of the others. Nonetheless, there were patients from the initial cluster 2, formed by people with “bad” scores that ended up in cluster 2 and not in cluster 1, although they have improved. This is the reason why the relation between the final clusters and the evolution types is not entirely equivalent. In fact, when looking at the numbers, there is a great discrepancy, essentially between the patients in cluster 3 (represent 37.0% of the total) and the ones from the worsening cases (10.9%) and between the patients in cluster 2 (7.7%) and the ones from the maintenance (24.2%). Nonetheless, the percentages for the patients in cluster 1 and the ones in improvements only hold a difference of less than 10 percentage points between them, which means 67 patients. The only correct and possible way to establish a relation here and to reduce these differences, yet with some reservations to consider, would be to contemplate also the patients who evolved, for example, from the initial cluster 2 and ended up in the final cluster 2, as improvements, or the patients who migrated from the initial cluster 1 to the final cluster 2, getting worse as well.

The most important thing to retain when discussing the graphics and inferring conclusions is that the obtained clusters try to describe the population based on distances and similar characteristics between them. Knowing the evolution types after the description of the patients’ cognitive profiles in two different time points and observing how their migration from the initial to the final clusters helps us make a clinical validation of those clusters, and nothing else. As future work, this could be done for other time points during the treatment, in real time, to assess the patient’s evolution and to adapt the therapy immediately.

Regarding the influence of other variables, the part of the study that considered the clusters revealed that there was no special effect from any of the variables. In the case of the gender, the results were obtained for both initial and final clusters. For all the other variables, only the final profiles were taking into account. In any of the graphics for a specific variable was there a bigger number of patients for a determined cluster when comparing with the other clusters. In all the situations patients were equally distributed over the clusters. Thus, information that can be extracted from the analysis of the graphics could only mean that these variables (gender, age, study level, interval between the injury and the start of the treatment, the duration of the treatment and the number of tasks performed during the treatment) do not have effect on the final profile of a patient or, at least, with not very clear influence. Analyzing the graphics where the influence of these variables was found for each type of evolution, the conclusions are not the same. Indeed, looking at the first three figures (Figure 4.32, Figure 4.33 and Figure 4.34),

one for the improvements, other for the worsening situations and another one for the maintenances, there were several, but not too big, differences regarding the gender, the age and the education level. The fact that in the improvements population most of the people are well distributed between the 20 and 60 years old and that in the worsening evolution type, there are more old than young people (most are older than 40, for both genders) might reveal that older people tend to get worse instead of improving. Considering the study levels, there are very few patients with a study level of 5 in the worsening population, especially when comparing that number with the one from the improvements one. A rather simple conclusion, yet reasonable, would be that patients with higher education levels tend to improve their scores instead of getting worse. Regarding the gender, there are no differences between the evolution types: in all three situations, there are more men than women. This means that this specific variable does not appear to have an influence on the evolution of a patient. For the rest of the variables, there only seems to exist a slight influence of the time that passes from the injury and the start of the treatment and the duration of the treatment. The shorter the time that the patient takes to start the treatment, and longer the treatment, more likely he/she seems to show improvements. A basic conclusion that can be taken from this analysis is that the computed clusters are not completely able to be used to describe the influence of the other variables.

6 Conclusion and Future Work

The conclusions that can be taken from the analysis of the results disclose that the objectives of this dissertation could be reached. The DM objective of getting clusters from patients based on the level of the 3 main cognitive functions (attention, memory and executive functions) was successfully achieved. Also, the main aim of getting dysfunctional profiles to personalize treatments based on the cognitive assessment performed before the treatment was well accomplished. The initial clusters obtained embody the possible initial cognitive profiles. What happens in the real-life situation is that there are great amounts people with a lot of dissimilar characteristics. In a generic approach, the clusters chosen for this study by representing six different groups of people, with very distinctive features, resemble reality. The final clusters, on the other hand, represent the final states of the patients according to their scores. They are supposed to have enough number and quality to differentiate the final possible states of the patients, considering that what the therapist wants to know is if the patient improves, gets worse or did not suffer any change with the treatment. Indeed, in this case, the final three clusters allowed us to distinguish between the patients with good scores, patients with bad scores and patients who did not seem to have a very well defined global improvement level. They allowed us to see that the population, at the end of the treatment and in general terms, had less affectionation (there were some clusters with not so good scores but, in global terms, the patients improved).

In summary, what was done in this dissertation was to get the initial dysfunctional profiles considering the test scores obtained in the tests performed before the treatment. Having calculated the initial profiles, the next step was to look for patients who had improved within that profile. This was done so that a similar therapy could be chosen (since it seems that therapy had been successful) to be applied to the other patients of that same initial profile. The final clusters were done to describe the population at the end of the treatment and also to analyze if the way patients evolved between an initial and final cluster made sense. In general, it did but, in fact the final clusters do not have to match exactly patients who improve, get worse, or maintain their status. This part could easily be calculated without the need to apply any clustering algorithm.

The final aim was to see if the calculated profiles could reveal any information about the influence of some demographic and lesion-related variables on the final status of a patient. The validation performed showed that there was no visible effect. However, the same study performed with the original scores indicated that that effect exists, but it is not very accentuated. This difference in the effect, when comparing clusters and scores might as well have sense: clusters group people who improve with people who do not improve in certain functions, and so, the effect is much more blurred.

For future work, the suggestion is to perform this same analysis for the 10 subfunction levels (Attention Sustained, Divided and Selective, Working and Visual/Verbal Memory, Planification, Flexibility, Sequencing, Inhibition and Categorization). Considering this level of detail makes it possible to calculate clusters with a much higher level of personalization.

7 References

- [1] Grupo de Bioingeniería y Telemedicina (2016) [Online]. Available: <http://www.gbt.tfo.upm.es>. [Accessed: 13-Nov-2016]
- [2] Traumatic Brain Injury vs Acquired Brain Injury (2016) [Online]. Available: <http://www.brainbehappy.com/brain-injury/tbi-vs-abi/>. [Accessed: 22-Nov-2016].
- [3] ABI Manual (2016) [Online]. Available: http://www.acquiredbraininjury.com/abi_manual. [Accessed: 28-Nov-2016].
- [4] Non-traumatic brain injury (2016) [Online]. Available: <http://www.braininjuryhub.co.uk/information-library/non-traumatic>. [Accessed: 28-Nov-2016].
- [5] Traumatic Brain Injury (2016) [Online]. Available: <https://www.uabmedicine.org/patient-care/conditions/traumatic-brain-injury> . [Accessed: 22-Nov-2016].
- [6] Brain Injury Facts (2016) [Online]. Available: <http://www.internationalbrain.org/brain-injury-facts/>. [Accessed: 20-Nov-2016].
- [7] What are the causes of TBI? (2016) [Online]. Available: <http://www.traumaticbraininjury.com/understanding-tbi/what-are-the-causes-of-tbi/> . [Accessed: 28-Nov-2016].
- [8] Traumatic Brain Injury (2016) [Online]. Available: http://www.abistafftraining.info/Content/3_Type_a.html . [Accessed: 29-Nov-2016].
- [9] Lehr Jr, R.P. Brain Functions (2016) [Online]. Available: <http://www.neuroskills.com/brain-injury/brain-function.php> . [Accessed: 6-Dec-2016].
- [10] Reyst, H. Neuroplasticity After Acquired Brain Injury (2016) [Online]. Available: <http://www.rainbowrehab.com/neuroplasticity-aquired-brain-injury/#attachment%20wp-att-6066/0/> . [Accessed: 6-Dec-2016].
- [11] Solana, J., Cáceres, C., García-Molina, A., Opisso, E., Roig, T., Tormos, J.M. and Gómez, E.J. Improving brain injury cognitive rehabilitation by personalized telerehabilitation services: Guttman neuropersonal trainer, *IEEE Journal of Biomedical and Health Informatics* **19(1)**, 124-31 (2015). DOI: 10.1109/JBHI.2014.2354537
- [12] Luna, M., Caballero, R., Chausa, P., García-Molina, A., Cáceres, C., Bernabeu, M., Roig, T., Tormos, J.M. and Gómez, E.J. Acquired Brain Injury Cognitive Dysfunctional Profile Based on Neuropsychological Knowledge and Medical Imaging Studies, *Biomedical and Health Informatics*, 260-263 (2014). DOI10.1109/BHI.2014.6864353.
- [13] Guttman, NeuroPersonalTrainer (2016) [Online]. Available: <https://www.gnpt.es/es/> . [Accessed: 29-Nov-2016].

- [14] Information and communication technologies applied to neurorehabilitation (2016) [Online]. Available: <http://www.guttmanninnova.com/en/guttmann-research/strategic-research-programs/information-and-communication-technologies-applied-to-neurorehabilitation.html> [Accessed: 29-Nov-2016].
- [15] Solana, J., Cáceres, C., García-Molina, A., Chausa, P., Opisso, E., Roig-Rovira, T., Menasalvas, E., Tormos-Muñoz, J.M. and Gómez, E.J. Intelligent Therapy Assistant (ITA) for cognitive rehabilitation in patients with acquired brain injury, *BMC Medical Informatics and Decision Making* **14**(1) (2014).
- [16] Cedeño, A.M., Chausa, P., García-Molina, A., Cáceres, C., Tormos, J.M., Gómez, E.J. Data mining applied to the cognitive rehabilitation of patients with acquired brain injury, *Expert Systems with Applications* **40**(4), 1054-1060, (2013).
- [17] Serrà, J. and Arcos, J. Cognitive prognosis of acquired brain injury patients using machine learning techniques, *The Fifth International Conference on Advanced Cognitive Technologies and Applications*, 108-113 (2013). ISBN: 978-1-61208-273-8.
- [18] Cedeño, A.M., Chausa, P., García-Molina, A., Cáceres, C., Tormos, J.M., Gómez, E.J. Artificial metaplasticity prediction model for cognitive rehabilitation outcome in acquired brain injury patients, *Artificial Intelligence in Medicine* **58**(2), 91-99 (2013).
- [19] Thuraisingham, B. Web Data Mining and Applications in Business Intelligence and Counter-Terrorism, *CRC Press* (2003). ISSN: 9780203499511
- [20] Wanner, L. Introduction to Clustering Techniques” (2017) [Online]. Available: <http://www.iula.upf.edu/materials/040701wanner.pdf> . [Accessed: 28-Nov-2016].
- [21] Biemann, C. Unsupervised and Knowledge-free Natural Language Processing in the Structure Discovery Paradigm, *Ausgezeichnete Informatikdissertationen*, 31-40 (2007). ISBN: 978-3-88579-412-7
- [22] Martin, A. What is Cluster Analysis? [Online]. Available: <http://slideplayer.com/slide/4706529/> . [Accessed: 9-Apr-2017].
- [23] Kaushik, S. An Introduction to Clustering and different methods of clustering [Online]. Available: <https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/> . [Accessed: 9-Apr-2017].
- [24] Andritsos, P. Data Clustering Techniques, *Qualifying Oral Examination Paper* (2012) [Online]. Available: <http://www.cs.toronto.edu/~periklis/pubs/depth.pdf> . [Accessed: 10-Apr-2017].
- [25] Technical Report: The SPSS TwoStep Cluster Component: A scalable component enabling more efficient customer segmentation (2001) [Online]. Available: https://www.spss.ch/upload/1122644952_The SPSS TwoStep Cluster Component.pdf . [Accessed: 11-Apr-2016].

- [26] Gamper, J., Kacimi, M. Data Mining: Clustering: Partitioning Methods (2012) [PowerPoint Slides] [Online]. Available: <http://www.inf.unibz.it/dis/teaching/DWDM/slides2010/lesson9-Clustering.pdf> . [Accessed: 9-Apr-2017].
- [27] Mahmood, A., Leckie, C., Islam, Md R., Tari, Z. Hierarchical summarization techniques for network traffic, *6th IEEE Conference on Industrial Electronics and Applications*, 2474-2479 (2011). DOI: 10.1109/ICIEA.2011.5976009.
- [28] IBM, Pre-Cluster (2012) [Online]. Available: https://www.ibm.com/support/knowledgecenter/SSLVMB_21.0.0/com.ibm.spss.statistics.help/alg_2step_precluster.htm . [Accessed: 5-May-2017].
- [29] Castillo, C. Algorithmic Methods of Data Mining: Hierarchical Clustering (2015) [PowerPoint Slides] [Online]. Available: <https://pt.slideshare.net/ChaToX/hierarchical-clustering-56364612> . [Accessed: 5-May-2017].
- [30] AgilOne, The Difference Between Segmentation and Clustering (2014) [Online]. Available: http://blog.agilone.com/blog/segmentation-vs-clustering_ . [Accessed: 8-May-2017].
- [31] Evgeniou, T. Cluster Analysis and Segmentation (2017) [Lecture] [Online]. Available: <http://inseaddataanalytics.github.io/INSEADAnalytics/CourseSessions/Sessions45/ClusterAnalysisReading.html> . [Accessed: 8-May-2017].
- [32] Molaei, S., Korley, F.K., Soroushmehr, S.M.R., Falk, H., Sair, H., Ward, K. and Najarian, K. A Machine Learning based approach for Identifying Traumatic Brain Injury Patients from Whom a Head CT Scan can be avoided, 2258-2261 (2016). DOI: 10.1109/EMBC.2016.7591179.
- [33] Prichep, L.S., Jacquin, A., Filipenko, J., Dastidar, S.G., Zabele, S., Vodenčarević, A. and Rothman, N.S. Classification of Traumatic Brain Injury Severity Using Informed Data Reduction in a Series of Binary Classifier Algorithms, *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **20(6)**, 806-822 (2012). DOI: 10.1109/TNSRE.2012.2206609.
- [34] Taslimitehrani, V. and Dong, G. A new CPXR Based Logistic Regression Method and Clinical Prognostic Modeling Results Using the Method on Traumatic Brain Injury, *IEEE 14th International Conference on Bioinformatics and Bioengineering*, 283-290 (2015). DOI: 10.1109/BIBE.2014.16
- [35] Van der Ploeg, T., Nieboerc, D. and Steyerberg, E. W. Modern modeling techniques had limited external validity in predicting mortality from traumatic brain injury, *Journal of Clinical Epidemiology* **78**, 83-89 (2016).
- [36] OH, H.S. and SEO, W.S. Development of a Decision Tree Analysis model that predicts recovery from acute brain injury, *Japan Journal of Nursing Science* **10**, 89–97 (2013).
- [37] Siddiqui, Z.F., Krempf, G., Spiliopoulou, M., Peña, J.M., Paul, N. and Maestu, F. Predicting the post-treatment recovery of patients suffering from traumatic brain injury (TBI), *Brain Informatics* **2**, 33–44 (2015).

- [38] Tsai, J.P. Leading-edge cognitive disorders research, *New York: Nova Science Publishers* (2008). ISBN: 9781600218439
- [39] Sherrill-Pattison, S., Donders, J., Thompson, E. Influence of demographic variables on neuropsychological test performance after traumatic brain injury, *Clin Neuropsychol* **14**(4), 496-503 (2000). DOI: 10.1076/clin.14.4.496.7196.
- [40] De la Plata, C., Hart, T., Hammond, F. M., Frol, A., Hudak, A., Harper, C. R., O'Neil-Pirozzi, T., Whyte, J., Carlile, M., Diaz-Arrastia, R. Impact of Age on Long-term Recovery From Traumatic Brain Injury, *Archives of Physical and Medical Rehabilitation* **89**(5), 896–903 (2008). DOI: 10.1016/j.apmr.2007.12.030.
- [41] Haffeejee, S., Ntsiea, V., Mudzi, W. Factors That Influence Functional Mobility Outcomes of Patients After Traumatic Brain Injury, *Hong Kong Journal of Occupational Therapy* **23**(1), 39-44 (2013). DOI: 10.1016/j.hkjot.2013.08.001.
- [42] Kalaiselvi, C. Diagnosing of Heart Diseases using Average k-Nearest Neighbor algorithm of Data Mining, *3rd International Conference on Computing for Sustainable Global Development*, (2016). Electronic ISBN: 978-9-3805-4421-2.
- [43] Radhimeenakshi, S. Classification and prediction of heart disease risk using data mining techniques of support vector machine and artificial neural network, *3rd International Conference on Computing for Sustainable Global Development*, 3107-3111, (2016). Electronic ISBN: 978-9-3805-4421-2.
- [44] Moretti, C.B., Joaquim, R.C., Terranova, T.T., Battistella, L.R., Mazzoleni, S. and Caurin, G.A.P. Knowledge Discovery strategy over patient performance data towards the extraction of hemiparesis-inherent features: A case study, *6th IEEE RAS/EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, 26-29 (2016). DOI: 10.1109/BIOROB.2016.7523711
- [45] Diciolla, M., Binetti, G., DiNoia, T., Pesce, F., Schena, F.P., Vågane, A.M., Bjørneklett, R., Suzuki, H., Tomino, Y. and Naso, D. Patient classification and outcome prediction in IgA nephropathy, *Computers in Biology and Medicine* **66**, 278–286 (2015).
- [46] Moskovitch, R., Choi, H., Hripsack, G. and Tatonetti, N. Prognosis of Clinical Outcomes with Temporal Patterns and Experiences with One Class Feature Selection, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2016). DOI: 10.1109/TCBB.2016.2591539.
- [47] Huang, L., Jin, Y., Gao, Y., Thung, K-H. and Shen, D. Longitudinal clinical score prediction in Alzheimer's disease with soft-split sparse regression based random forest, *Neurobiology of Aging* **46**, 180-191 (2016).
- [48] Yet, Z. Perkins, B., Fenton, N., Tai, N. and Marsh, W. Not just data: A method for improving prediction with knowledge, *Journal of Biomedical Informatics* **48**, 28–37 (2014).

Appendix

The tests performed by the patients were designed to evaluate certain types of capacities. In Table A. 1 all the 17 tests that were used to calculate the scores for the Attention, Memory and EF subfunctions, are presented, alongside with the number they were matched to. For example, in the case of this study, the patients considered had performed at least one complete (Pre and Post) test in Attention, two in Memory and another two in EF. A hypothetical patient, to be considered, could have a combination of a TMTA test, to evaluate the Sustained Attention a Midigits test, to assess the Working Memory performance, an Aspan test, for the Visual/Verbal Memory, and also a PMR planification test and a VCCUBS sequencing test, for example.

Table A. 1 - Table of tests used to evaluate each Attention, Memory and EF functions

ATTENTION	Sustained	TMTA		5
	Selective	Stroop	Palabra	6
			Color	7
			PalabraColor	8
			VPWALSS	9
	Divided		Interferencia	23
		TMTB	20	
MEMORY	Work		Mdigits	15
			Mlletres	16
	Visual/Verbal		Aspan	4
		RAVLT	Inmediato	17
			Diferido	18
			Reconocimiento	19
EXECUTIVE FUNCTIONS	Planification	VCCUBS		14
		PMR		24
	Inhibition	Interferencia		23
		Mlletres		16
	Flexibility	FEWCSTE		22
		PMR		24
	Sequencing	VCCUBS		14
		TMTB		20
	Categorization	FEWCSTC		21
		PMR		24

Next, Figure A. 1 presents the CognitioDW data warehouse with all the tables and relations between them.

The Cognitio project already has several tables that parametrize the results, or scores, according to the affectation levels, to the functions, the tests and specially the age and the study level (Table A. 2 and Table A. 3)

Table A. 2 - Profile of the affectation degree for a person aged between 17 and 30 and with a level of primary studies

GRADOS DE AFECTACIÓN PERFIL NPS			Rango de edad: 17 a 30 años.	Nivel de estudios: Primarios (EGB/ESO).				
FUNCIÓN	SUBFUNCIÓN	PRUEBA	variable	0	1	2	3	4
				Normalidad	afectación leve	afectación moderada	afectación grave	afectación muy grave
Orientación	B/o.persona	PIENC	O.persona	7	6	5-2	1	0
	B/o.espacio	PIENC	O.espacio	5	4	3-2	1	0
	B/o.tiempo	PIENC	O.tiempo	23	22-20	19-15	14-5	≤ 4
Gnosis	B/f.superpuestas	PIENC	F.superpuestas	20	19-18	17-15	14-9	≤ 8
Lenguaje	Repetición	PIENC	palabras	10	9-8	7-4	3-2	≤ 1
	Denominación	PIENC	visoverbal	14	13-11	10-7	6-4	≤ 3
	Comprensión	PIENC	órdenes	16-15	14-13	12-8	7-4	≤ 3
Atención	A. Sostenida	CPT	Omisiones	30,000-60,000	60,001-65,000	65,001-70,000	70,001-75,000	≥75,001
	A. Sostenida	CPT	Comisiones	30,000-60,000	60,001-65,000	65,001-70,000	70,001-75,000	≥75,001
	A. Sostenida	CPT	T. Reacción	30,000-60,000	60,001-65,000	65,001-70,000	70,001-75,000	≥75,001
	A.Selectiva	Stroop	APalabra	≥99	98-79	78-59	58-39	≤ 38
	A.Selectiva	Stroop	AColor	≥65	64-51	50-37	36-23	≤ 22
	A.Selectiva	Stroop	APalabra-color	≥39	38-28	27-17	16-6	≤ 5
	A.Dividida	Stroop	Interferencia	≥-6,35	de -15,44 a -6,351	de -24,45 a -15,45	-33,5 a -24,46	≤ -33,51
	A.Selectiva	TMT	TMT-A	≤ 38	39-49	50-60	61-71	≥72
	A. Dividida	TMT	TMT-B	≤ 63	64-74	75-85	86-96	≥ 97
Memoria	Inmediata	WAIS-III	Dígitos directos	≥6	5	4	3	2-1
	Inmediata	WAIS-III	Dígitos inversos	≥5	4	3	2	1
	M. trabajo	WAIS-III	Letras/números	≥10	9	8-7	6-5	≤ 4
	Verbal	RAVLT	Corto plazo	≥53	52-45	44-37	36-29	≤28
	Verbal	RAVLT	Largo plazo	≥11	10-9	8-7	6-5	≤4
	Verbal	RAVLT	Reconocimiento	≥13	12-11	10-9	8-7	≤6
	Flexibilidad	PMR	nº palabras	≥ 41	40-35	34-30	29-26	≤ 25
	Categorización	WCST	nº categorías	6	5	4	3	2-0
	Flexibilidad	WCST	e.perseverativos	≤ 17	18-22	23-33	34-50	≥ 51
Visoconstrucción		WAIS-III	Cubos	≥45	44-40	39-36	35-33	≤ 32
Velocidad de procesamiento		WAIS-III	Claves/números	≥76	75-68	67-63	62-57	≤ 56

Table A. 3 - Profile of the affectation degree for a person aged between 17 and 30 and with a level of medium studies

GRADOS DE AFECTACIÓN PERFIL NPS			Rango de edad: 17 a 30 años.	Nivel de estudios: Medios (BUP/COU/Bach./FP/C. Formativos).				
FUNCIÓN	SUBFUNCIÓN	PRUEBA	variable	0	1	2	3	4
				Normalidad	afectación leve	afectación moderada	afectación grave	afectación muy grave
Orientación	B/o.persona	PIENC	O.persona	7	6	5-2	1	0
	B/o.espacio	PIENC	O.espacio	5	4	3-2	1	0
	B/o.tiempo	PIENC	O.tiempo	23	22-20	19-15	14-5	≤ 4
Gnosis	B/f.superpuestas	PIENC	F.superpuestas	20	19-18	17-15	14-9	≤ 8
Lenguaje	Repetición	PIENC	palabras	10	9-8	7-4	3-2	≤ 1
	Denominación	PIENC	visoverbal	14	13-11	10-7	6-4	≤ 3
	Comprensión	PIENC	órdenes	16-15	14-13	12-8	7-4	≤ 3
Atención	A. Sostenida	CPT	Omisiones	30,000-60,000	60,001-65,000	65,001-70,000	70,001-75,000	≥75,001
	A. Sostenida	CPT	Comisiones	30,000-60,000	60,001-65,000	65,001-70,000	70,001-75,000	≥75,001
	A. Sostenida	CPT	T. Reacción	30,000-60,000	60,001-65,000	65,001-70,000	70,001-75,000	≥75,001
	A.Selectiva	Stroop	APalabra	≥99	98-79	78-59	58-39	≤ 38
	A.Selectiva	Stroop	AColor	≥65	64-51	50-37	36-23	≤ 22
	A.Selectiva	Stroop	APalabra-color	≥39	38-28	27-17	16-6	≤ 5
	A.Dividida	Stroop	Interferencia	≥-6,35	de -15,44 a -6,351	de -24,45 a -15,45	-33,5 a -24,46	≤ -33,51
	A.Selectiva	TMT	TMT-A	≤ 28	29-39	40-50	51-61	≥62
	A. Dividida	TMT	TMT-B	≤ 53	54-64	65-75	76-86	≥ 87
Memoria	Inmediata	WAIS-III	Dígitos directos	≥6	5	4	3	2-1
	Inmediata	WAIS-III	Dígitos inversos	≥5	4	3	2	1
	M. trabajo	WAIS-III	Letras/números	≥10	9	8-7	6-5	≤ 4
	Verbal	RAVLT	Corto plazo	≥53	52-45	44-37	36-29	≤28
	Verbal	RAVLT	Largo plazo	≥11	10-9	8-7	6-5	≤4
	Verbal	RAVLT	Reconocimiento	≥13	12-11	10-9	8-7	≤6
	Flexibilidad	PMR	nº palabras	≥ 46	45-41	40-35	34-30	≤ 29
	Categorización	WCST	nº categorías	6	5	4	3	2-0
	Flexibilidad	WCST	e.perseverativos	≤ 13	14-18	19-29	30-46	≥ 47
Visoconstrucción		WAIS-III	Cubos	≥45	44-40	39-36	35-33	≤ 32
Velocidad de procesamiento		WAIS-III	Claves/números	≥76	75-68	67-63	62-57	≤ 56

Next, in, it is presented the flow that was used for this research to obtain both the initial and final clusters and the histograms regarding the distribution of scores through these clusters.

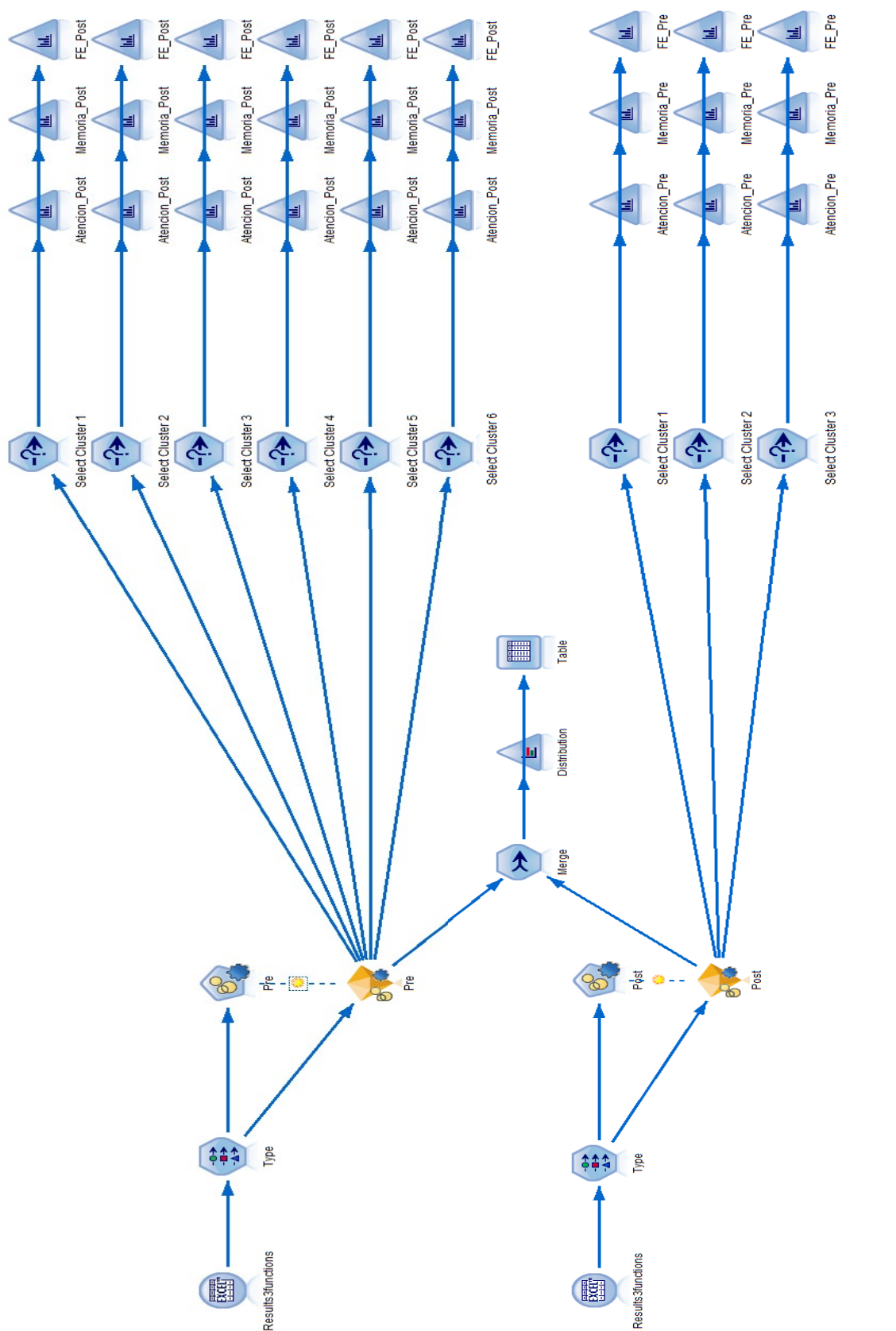


Figure A. 2 - Workflow to build a descriptive clustering model in SPSS

Then, another flow was created using as sources the files containing, separately, the patients who improved, who got worse and who did not evolve (Figure A. 3). The histograms showing the effects of the other variables were also calculated each one at a time.

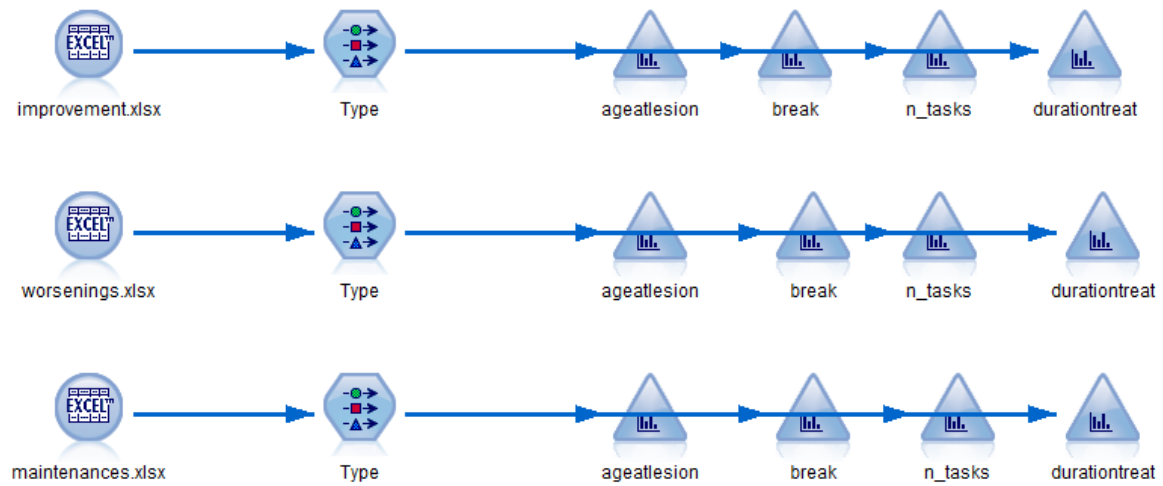


Figure A. 3 - SPSS flow to assess the influence of variables on the evolution type