



UNIVERSIDADE
LUSÓFONA

Aplicação de Inteligência Artificial para estudo de prática clínica inovadora

Trabalho Final de curso

Relatório Intercalar 2º Semestre

Nome do Aluno: Pedro Rodrigues

Nome do Orientador: Iolanda Velho

Nome do Coorientador: Maria Almeida Silva

Trabalho Final de Curso | Licenciatura de Informática de Gestão | 30/06/2023

www.ulusofona.pt

Direitos de cópia

Aplicação de Inteligência Artificial para estudo de prática clínica inovadora, Copyright de Pedro Emanuel Macedo Rodrigues, ULHT.

A Escola de Comunicação, Arquitetura, Artes e Tecnologias da Informação (ECATI) e a Universidade Lusófona de Humanidades e Tecnologias (ULHT) têm o direito, perpétuo e sem limites geográficos, de arquivar e publicar esta dissertação através de exemplares impressos reproduzidos em papel ou de forma digital, ou por qualquer outro meio conhecido ou que venha a ser inventado, e de a divulgar através de repositórios científicos e de admitir a sua cópia e distribuição com objetivos educacionais ou de investigação, não comerciais, desde que seja dado crédito ao autor e editor.

Resumo

A espasticidade é uma perturbação frequente nas lesões congénitas, que resulta de uma lesão no sistema nervoso central, podendo originar dificuldades funcionais, deformidades e/ou dor.

Em Portugal, anualmente, ocorrem 25 mil episódios de Acidentes Vasculares Cerebrais (AVC), sendo a espasticidade uma sequela que ocorre entre 17% a 43% dos casos. Esta é das mais frequentes sequelas do AVC, sendo que muitas vezes pode levar a problemas do foro emocional e a perturbações comportamentais. A compreensão do paciente perante uma condição que anteriormente não predominava o seu quotidiano pode causar um enorme transtorno sobre o mesmo. Para apoiar tais situações, instituições compostas de exímios profissionais das diversas áreas da Medicina de Reabilitação, prestam um serviço de apoio tanto físico como psicológico aos pacientes cuja vida mudou.

O Centro de Medicina e Reabilitação de Alcoitão é uma das instituições que tem como missão valorizar e potenciar as capacidades de cada indivíduo, apoiando-o no refazer do seu projeto de vida. Focada em ser uma linha da frente no tratamento dos pacientes com espasticidade, o Centro de Medicina e Reabilitação de Alcoitão tem sido líder no campo da investigação clínica aplicada e conquistou o reconhecimento mundial na área do tratamento da espasticidade com toxina botulínica.

O centro visa, em parceria com a Universidade Lusófona de Humanidades e Tecnologias, realizar um projeto com o intuito de estudar as variáveis de prognóstico utilizadas no momento de avaliação/reavaliação das condições do paciente para a aplicação de toxina botulínica no tratamento da espasticidade, bem como informatizar todo o sistema, atualmente manual, do registo dos questionários e armazenamento dos mesmos.

Este projeto será desenvolvido no âmbito do trabalho final de curso (TFC), a fim de demonstrar todo o conhecimento adquirido nas unidades curriculares referentes à Licenciatura de Informática de Gestão (LIG), com a orientação das professoras Iolanda Velho e Maria Almeida Silva, docentes da Universidade Lusófona de Humanidades e Tecnologia (ULHT).

Abstract

Spasticity is a frequent disruption in congenital injuries driven by central nervous trauma and may originate in functional difficulties, deformities, or pain.

Portugal has annually, twenty-five thousand new cases of Cerebrovascular Accident (CVA), and between 17% to 43% of the cases maintain spasticity. As indicated, this is one of the most frequent CVA sequels, sometimes leading to emotional and behavioral problems.

The assimilation of a patient to his new condition may cause some frustration and a need for help. That is why rehabilitation institutions gather the most qualified professionals from many rehabilitation areas, trying to provide the best support possible both physically and mentally. Alcoitão Rehabilitation Medicine Center is one of those institutions that embrace the mission of valorizing and thriving the individual capacities by supporting him in reshaping his life. Focussed on being the frontline in this type of treatment, the center has been recognized as a leader in the applied clinical investigation and worldwide recognition for the treatment of spasticity with botulin toxin.

The center aims, in partnership with the Universidade Lusófona de Humanidades e Tecnologias, to develop a project to study the variables of diagnosis upon doing check-ups in a patient, as well as, computerizing the questionnaire used alongside the check-ups and saving the information gathered from those.

This project will be developed, under the subject of the Final Course Work (TFC, in Portuguese), to demonstrate all the knowledge gathered from the classes lectured in the degree, with the supervision and coordination of Iolanda Velho and Maria Almeida Silva, both teachers from the Universidade Lusófona de Humanidades e Tecnologias.

Índice

Resumo.....	iii
Abstract	iv
Índice.....	v
Lista de Figuras.....	vi
Lista de Tabelas	vii
1 Identificação do Problema	1
2 Viabilidade e Pertinência.....	4
3 Benchmarking.....	5
4 Engenharia.....	8
4.1 Levantamento e Análise de Requisitos	8
4.2 Conceitos teóricos.....	8
4.3 Descrição de cenários de aplicação.....	12
5 Solução Desenvolvida.....	13
5.1 Introdução.....	13
5.2 Tecnologias e Ferramentas Utilizadas.....	14
5.3 1ª fase – Desenvolvimento de formulário digital.....	15
5.4 2ª fase – Análise dos dados.....	17
5.4.1 Descrição dos dados.....	17
5.4.2 Pré-processamento dos dados.....	18
5.4.3 Análise exploratória dos dados	20
5.4.4 Aplicação de PCA e Resultados	25
5.5 Parecer clínico & <i>Feedback</i>	30
5.6 Abrangência	31
6 Método e Planeamento	32
6.1 Planeamento inicial.....	32
6.2 Análise crítica ao planeamento.....	33
Bibliografia	35
Anexos.....	39
Lista de Acrónimos	54

Lista de Figuras

Figura 1 - Centro de Medicina de Reabilitação de Alcoitão (2 de julho de 1966)	1
Figura 2 - Formulário Consulta de Toxina Botulínica (Frente)	2
Figura 3 - Formulário Consulta de Toxina Botulínica (Verso)	2
Figura 4 - Ciclo manual de introdução de dados (Fase formulário)	2
Figura 5 - Processo para análise estatística realizado unicamente enquanto existiam fundos (Fase análise de dados)	3
Figura 6 – Método de cálculo das PC (Fonte: Slides aulas data mining)	9
Figura 7 – Exemplo de PCA aplicado a reconhecimento facial	10
Figura 8 - Representação gráfica para encontrar o K ótimo (Elbow Method).....	10
Figura 9 - Caso único do preenchimento da ficha no decorrer da aplicação da toxina botulínica	12
Figura 10 - A realidade vs. Os Resultados.....	12
Figura 11 - Plano de execução da solução.....	13
Figura 12 – Stack Tecnológico usado para Machine Learning	14
Figura 13 - Campos de identificação do formulário	15
Figura 14 - Secção de escala de dor do formulário.....	15
Figura 15 - Secção de força muscular do formulário	16
Figura 16 - Página de sucesso após inserção de formulário	16
Figura 17 – 6 passos para a limpeza e estruturação de um dataset	18
Figura 18 – Resumo do pré-processamento de dados.....	19
Figura 19 - Nº Pacientes por Género	20
Figura 20 – Idade dos pacientes por registo.....	20
Figura 21 - Idade dos pacientes por número de Pacientes	20
Figura 22 - Média de idade por género e objetivo do tratamento	22
Figura 23 – Causa mais comum por diagnóstico.....	23
Figura 24 - Taxa de sucesso de acordo com o objetivo do paciente	24
Figura 25 - Variância explicada cumulativa (%)	25
Figura 26 - Obtenção das magnitudes de cada variável	26
Figura 27 - Validação das datas (Formulário digital)	30
Figura 28 - Work Breakdown Structure	32
Figura 29 - Diagrama de Gantt Semanal (Realizado através do software Project Libre)	34
Figura 30 - Diagrama de Gantt anterior semanal (Realizado através do software Project Libre)	46
Figura 31 - Mapa de correlação entre as variáveis originais.....	47
Figura 32 - Mapa de correlação (Apenas com variáveis com correlação superior a 0.5 em valor absoluto)	48
Figura 33 – Mapa de correlação das variáveis originais com as componentes principais (correlação superior a 0.4 em valor absoluto).....	49

Lista de Tabelas

<i>Tabela 1 - Ferramentas utilizadas no formulário digital</i>	14
<i>Tabela 2 - Loadings (em valor absoluto) das variáveis mais impactantes na PC1</i>	26
<i>Tabela 3 – Top 6 variáveis com maior magnitude na variação total</i>	27
<i>Tabela 4 - 45 principais variáveis resultantes do estudo de PCA</i>	28
<i>Tabela 5 - Exemplos de ferramentas de avaliação de pacientes no mercado</i>	39
<i>Tabela 6 - Tabela de requisitos</i>	40
<i>Tabela 7 – Requisitos complementares e pedidos de alterações/modificações</i>	41
<i>Tabela 8 – Variáveis removidas devido à elevada multicolinearidade</i>	42
<i>Tabela 9 – Loadings mais impactantes por PC (PC2- PC20)</i>	43
<i>Tabela 10 - Tabela Milestones e Tarefas</i>	44
<i>Tabela 11 – “Novo paciente” criado com base na média de valores do modelo PCA</i>	50

1 Identificação do Problema

Conhecida usualmente pelo seu uso dermocosmético, a Toxina Botulínica, renomada por Botox, é uma neurotoxina produzida pela bactéria *Clostridium botulinum*. Inicialmente estudada em meados de 1800 por *Justinus Kerner*, esta constituía uma opção terapêutica para o tratamento de espasmos musculares. Apesar disso, apenas 200 anos depois, Dra. *Jean Carruthers*, ao aplicar a neurotoxina numa paciente que sofria de espasmos nas pálpebras, suspeitou que a mesma poderia ser utilizada para fins estéticos.

Bastante reconhecida no mundo científico, esta toxina tem vindo a ganhar dimensão pelos seus feitos originais: tratamento de espasticidade.

Estima-se que por ano cerca de 12,2 milhões de pessoas globalmente sofram um AVC, pelo que entre 17% e 43% podem prevalecer com espasticidade, ou seja, cerca de 1,9% da população mundial. Recorrentemente nas primeiras 6 semanas após o incidente, 25% dos pacientes com AVC sofrem de espasticidade, sendo esta uma sequela bastante frequente [1][2][3][4].

Atualmente, o Centro de Medicina e Reabilitação de Alcoitão (CMRA) (Figura 1) dispõe de 150 camas, sendo 16 exclusivas à reabilitação de pacientes pediátricos. O centro conta ainda com profissionais de saúde especializados e empenhados em investigação e desenvolvimento científico, sendo que cada utente tem à sua disposição uma equipa multiprofissional e multidisciplinar que proporciona uma reabilitação orientada por objetivos e focada no paciente, tanto em regime interno como ambulatório [5].

No centro são acompanhadas diferentes patologias, sendo as mais frequentes doenças ou sequelas de doenças neurológicas, sequelas de politraumatismos graves, amputações de membros, deficiências congénitas e doenças ou perturbações do desenvolvimento [6].

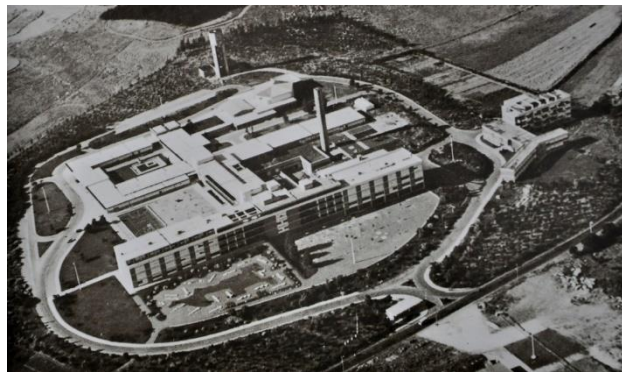


Figura 1 - Centro de Medicina de Reabilitação de Alcoitão (2 de julho de 1966) (fonte: <http://cmra.pt/centro/historia/>)

Cerca de 75% dos pacientes no CMRA padeceram de um AVC, pelo que, contrastando com os valores anteriores de forma bastante rudimentar compreende-se que $\pm 40\%$ dos 75% dos pacientes presentes no CMRA sofrem de espasticidade. Atendendo ao facto de ser um dos Centros da Europa com mais pacientes, é gerado assim um volume de dados avultado. Consequentemente, os profissionais de saúde investem o seu conhecimento e tempo a agregar informação para a possibilidade de realizar estudos científicos a fim de poder compreender certos padrões existentes neste tipo de sequelas [7]. Para tal, médicos e especialistas de saúde beneficiariam do uso de algoritmos e estudo da correlação dos dados para identificar possíveis padrões existentes, sendo que trará valor e conhecimento.

No caso concreto do CMRA, os profissionais de saúde do centro não dispõem, nos dias correntes, de um formulário digital que possa ser preenchido com os dados do paciente. Tal ainda é realizado através de impressos guardados em arquivos, para que mais tarde alguém os transponha para uma folha de Excel (Figura 2, Figura 3). Este problema, resulta na inutilização de alguns dados recolhidos, uma vez que para estudos científicos são utilizados dados que até à data se encontram em formato digital. A ausência de digitalização deve-se não só à falta de informatização como também à falta de recursos logísticos e financeiros.

CONSULTA DE TOXINA BOTULÍNICA
Registo de Intervenção, Definição de Objectivos e Avaliação

FOLHA DE REAVALIAÇÃO

IDENTIFICAÇÃO: _____ DATA: ____/____/____ REAV: ____/____/____

ESCALA DE ASHTWOTH MODIFICADA

M. Superior	D. C. P/D			M. Inferior		
	0	1	2	0	1	2
1						
2						
3						
4						

ESCALA VISUAL ANALÓGICA (EVA) - DOR

10 metros de marcha: Tempo(s) _____ Velocidade _____

Reacção associada: Tonturas _____ Náuseas _____ Velocidade _____

Caregiver Burden Scale

TOXINA BOTULÍNICA: Dose: _____

LOCALIZAÇÃO: Palpação Injeção Furo

Figura 2 - Formulário Consulta de Toxina Botulínica (Frente)

MEMBRO SUPERIOR

Classe	0	1	2
Classe I			
Classe II			
Classe III			
Classe IV			
Classe V			
Classe VI			
Classe VII			
Classe VIII			
Classe IX			
Classe X			
Classe XI			
Classe XII			
Classe XIII			
Classe XIV			
Classe XV			
Classe XVI			
Classe XVII			
Classe XVIII			
Classe XIX			
Classe XX			

MEMBRO INFERIOR

Classe	0	1	2
Classe I			
Classe II			
Classe III			
Classe IV			
Classe V			
Classe VI			
Classe VII			
Classe VIII			
Classe IX			
Classe X			
Classe XI			
Classe XII			
Classe XIII			
Classe XIV			
Classe XV			
Classe XVI			
Classe XVII			
Classe XVIII			
Classe XIX			
Classe XX			

OUTRAS MODALIDADES TERAPÊUTICAS: Físio-terapia Fono-terapia Terapia ocupacional Terapia psicológica Outra

INTERFERÊNCIAS ASSOCIADAS AO TRATAMENTO: Sim Não

OBJECTIVOS DO TRATAMENTO COM TOXINA BOTULÍNICA(S)

OBJECTIVO PRIMÁRIO

Classe	0	1	2
Classe I			
Classe II			
Classe III			
Classe IV			
Classe V			
Classe VI			
Classe VII			
Classe VIII			
Classe IX			
Classe X			
Classe XI			
Classe XII			
Classe XIII			
Classe XIV			
Classe XV			
Classe XVI			
Classe XVII			
Classe XVIII			
Classe XIX			
Classe XX			

OBJECTIVOS SECUNDÁRIOS

CLASSIFICAÇÃO: Score Botulínica: _____ Score Custom: _____

GRUPO DE SATISFAÇÃO DO PACIENTE

GRUPO DE SATISFAÇÃO DO PROFISSIONAL

OBSERVAÇÕES: _____

Figura 3 - Formulário Consulta de Toxina Botulínica (Verso)

Inconvenientemente, os profissionais de saúde do CMRA têm de preencher formulários manualmente, sendo que, o centro atualmente comporta em média mais de 700 consultas anuais da aplicação da toxina botulínica. Consequentemente, o uso de, pelo menos, um impresso por cada consulta implica a utilização de 700 folhas de papel, sem atender aos tinteiros, desgaste da impressora e consumo energético, fora, a frequente acumulação de trabalho (Figura 4). Foi-nos dado um exemplo: 6 meses de dados estagnados em impressos de papel por falta de tempo para transpor para o ficheiro Excel.

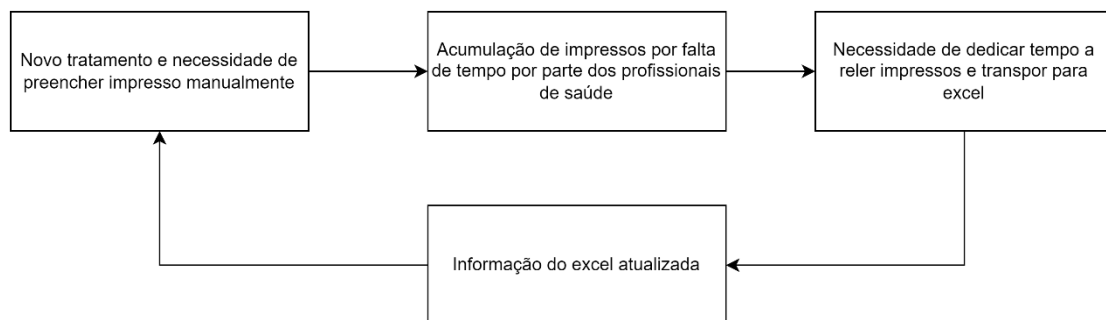


Figura 4 - Ciclo manual de introdução de dados (Fase formulário)

A criação de um formulário, e posterior armazenamento numa base de dados, em SQL, NO-SQL ou Excel (atendendo ao enquadramento para o CMRA e profissionais), será proveitosa para os profissionais de saúde, que terão os seus trabalhos facilitados, atendendo que neste momento os dados não são armazenados em formato digital, pelo que neste momento os dados teriam de ser transpostos folha a folha, para formato digital, para posterior análise. Após a informatização, a recolha e estudo dos dados pode originar questões que anteriormente não tinham sido feitas, e melhorar o diagnóstico realizado aos pacientes.

Relativamente à análise dos dados existentes, até ao momento, o CMRA efetuou unicamente um pedido a uma entidade externa para a realização de uma análise estatística sobre os dados recolhidos (*Figura 5*).

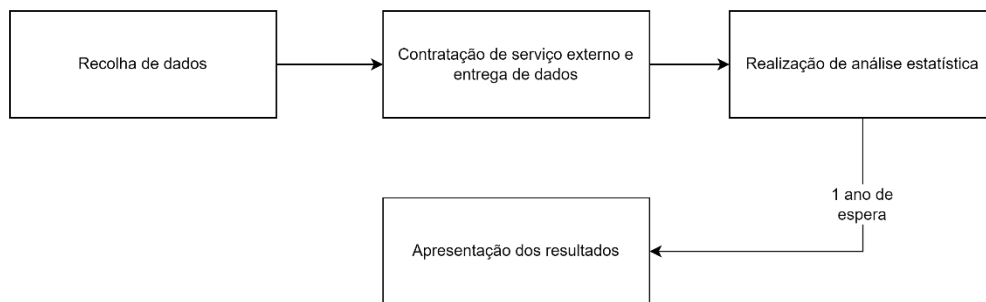


Figura 5 - Processo para análise estatística realizado unicamente enquanto existiam fundos (Fase análise de dados)

Os demais profissionais e diretores do CMRA reconhecem que, até ao momento, pouco sabem sobre os dados que dispõem e que seria benéfico, tanto para a comunidade científica, como para o auxílio aos pacientes, identificar quais as componentes importantes na realização do diagnóstico.

Em suma, o TFC irá focar-se na informatização do questionário sobre a aplicação da toxina botulínica em pacientes com espasticidade e no posterior estudo da correlação e importância das variáveis envolvidas nesta prática clínica.

2 Viabilidade e Pertinência

O CMRA atualmente é o líder a nível nacional no tratamento da espasticidade com a toxina botulínica. Contudo, as aplicações da toxina congregam um custo elevado, sendo necessária a aplicação de 2 a 4 vezes por ano para manter os resultados [7].

A título demonstrativo, à data corrente, a Santa Casa da Misericórdia de Lisboa, detentora do CMRA, emitiu um contrato público para a aquisição de Toxina Botulínica tipo A 100 UI AMP IM/IGL e tipo A 100 UI AMP IM/ID para o CMRA, no valor total de 36,000.00€. ¹

É do nosso conhecimento que cada aplicação acarreta um custo económico elevado, tanto para o paciente, como para o sistema de saúde ou para Seguradora de Saúde do paciente, que acarreta uma quota parte do valor.

Através do CMRA, surgiu a possibilidade de colaboração com a Universidade Lusófona de Humanidades e Tecnologias a fim de realizar uma parceria e ter apoio em dois aspetos. Em primeiro lugar, na informatização de alguns dos seus procedimentos, como é o caso do questionário aos pacientes. Em segundo lugar, para a realização de estudos através da aplicação de algoritmos de Machine Learning (ML) aos dados que têm para conseguir perceber quais as variáveis/questões importantes de colocar a um paciente, aquando da realização de uma consulta para aplicação da toxina.

Deste modo, seria interessante e relevante melhorar procedimentos realizados pelos médicos, tais como: 1) redução de tempo de preenchimento do formulário, o que por sua vez iria reduzir o tempo por consulta; 2) atualização das questões realizadas nos formulários; bem como de otimizar recursos através da criação de base de dados informatizada, reduzindo a necessidade de contratação de entidades externas, entre muitos outros.

Qualquer tipo de estudo ou informatização realizada no centro irá beneficiar diretamente os pacientes que, ao realizarem este tratamento, poderão ter na sua presença um médico com ferramentas eficientes para lhe proporcionar uma consulta de melhor qualidade, com questões otimizadas e fulcrais para o seu tratamento. Além disso, permitirá também uma melhoria contínua do centro relativamente ao estudo da aplicação da toxina.

De salientar que, na primeira visita realizada ao centro, foi apresentado que o volume de visitas e a complexidade de gestão de recursos no centro é elevada, pelo que qualquer tipo de auxílio na melhoria de processos que gere eficiência, terá um grande impacto no centro.

Em suma, a chave do sucesso para a entrega de valor e conhecimento, passa também pela melhoria de procedimentos, ou seja, mais eficiência dos recursos que o centro tem disponíveis, sendo estes através da realização de tarefas como: o preenchimento de um questionário mais curto e objetivo; ou simplesmente a possibilidade de poder ter os dados informatizados sem necessidade de vasculhar impressos em papel.

¹ Disponível em: <https://dre.tretas.org/dre/5106681/anuncio-de-procedimento-13855-2022-de-28-de-outubro>

3 Benchmarking

Atualmente, e pelo testemunho de profissionais de saúde do CMRA, todos os locais por onde realizaram e realizam trabalhos, dispõem de um sistema de introdução de fichas da aplicação de tratamentos, exceto no Centro. Como tal, os profissionais consideram obsoleto o facto de não disporem de um sistema de software para gestão de registos médicos de pacientes para este tipo de prática clínica. Este *feedback* advém de uma “necessidade” deste tipo de ferramentas, a fim de se tornar o *workflow* mais eficiente e eficaz.

Muito frequentemente, como o médico não tem capacidade de dar resposta a tanto trabalho, acaba por não transpor de imediato os questionários físicos para a folha de Excel, pelo que, o trabalho acaba sendo acumulado. O próprio relata que muito frequentemente apenas consegue atualizar a ficha de Excel de 6 em 6 meses tendo, da última vez, pago a externos para realizar esta introdução de dados.

No mercado não existe nenhum formulário igual ao do CMRA, pelo que, será necessário desenvolver um formulário digital exclusivo para o centro, uma vez que se trata de algo muito específico e concreto. Existe também, a nível da vertente tecnológica, a necessidade de utilizar, de preferência, tecnologias de cariz gratuito para manter o formulário digital, uma vez que o CMRA possui um *budget* limitado e todo o tipo de redução de custos será bem recebido. Deste modo, a criação de um formulário digital com HTML, o processo de tratamento e validação das informações através de PHP e o armazenamento dos dados em ficheiro CSV, será para o CMRA uma ferramenta simples, sem custos adicionais e que irá auxiliar na diminuição de tempos de espera na introdução de dados, redução do uso de recursos e aumento da satisfação dos profissionais de saúde, bem como a criação de uma base de dados preparada para futuras análises. Atendendo a uma visão de longevidade, na eventualidade do CMRA propor-se em armazenar os dados num formato diferente, por exemplo, uma base de dados, tal pode ser feito pelo departamento informático do centro a partir do ficheiro CSV.

Através de plataformas como a *Google Scholar*, a *b-On - Biblioteca do Conhecimento* e a *web of science* foi possível realizar pesquisas bibliográficas sobre as práticas realizadas na área da saúde. Contudo, e mais concretamente sobre o tema abordado da aplicação da toxina botulínica a pacientes com espasticidade, de acordo com a pesquisa realizada, até à data, não foi possível encontrar nenhum documento da aplicação de técnicas da ciência de dados neste caso específico. No entanto, foi possível encontrar algumas publicações sobre a evolução da prática clínica e algumas aplicações de algoritmos de ML (com aplicação da toxina botulínica) referentes a: previsão do tratamento do paciente [8], classificação de Gait [9], disfonia espasmódica adutora [10], retrospectiva de análise da utilização de recurso e custos em pacientes com espasticidade pós-AVC [11].

A fusão cada vez maior da ciência de dados com a área de saúde torna-se inevitável, uma vez que os benefícios como o conhecimento que pode advir é preponderante para o tratamento de um paciente a fim de antecipar possíveis problemas irreversíveis [12].

Deste modo, conferências realizadas globalmente publicaram um resumo de algumas práticas clínicas realizadas aonde se descreve qual o autor, qual a doença estudada e quais os algoritmos/técnicas utilizadas [12][13]. Ao analisar essas referências, é bastante perceptível que muitos destes autores recorreram frequentemente a técnicas como *K-Means* ou *K-nearest neighbors (KNN)*, contudo outros preferiram uma abordagem diferente utilizando *Support Vector Machines (SVM)*.

A investigação científica e os artigos que têm vindo a ser publicados demonstram que a ciência de dados tem um papel essencial nos dias correntes na área da saúde, mas particularmente na previsão de doenças e sintomas [12]. Além do mais, estas publicações demonstram que não existe apenas uma técnica que prevalece sobre as restantes, mas podem ser utilizadas diversas técnicas até se conseguir chegar a um resultado satisfatório para o estudo que se encontra a ser realizado.

Na área de saúde é bastante frequente a aplicação de PCA para a identificação de componentes que expliquem a maioria da variação total dos dados. Alguns exemplos desta aplicação são: a Associação entre componentes inflamatórios e forma física no estudo de saúde, envelhecimento e composição corporal [14], a Aplicação de PCA aos sinais de Eletrocardiograma para diagnóstico automatizado da arritmia cardíaca [15] e a Aplicação de PCA relativamente à entrevista de diagnóstico de autismo [16].

Em [16] encontram-se semelhanças ao trabalho que se pretende realizar neste TFC. O estudo do questionário utilizado e a análise da correlação das perguntas são pontos fulcrais para este estudo. No caso do diagnóstico de autismo existem 98 questões divididas por 6 critérios sendo que estes critérios foram analisados para perceber se efetivamente a correlação que existe entre eles é elevada ao ponto de ser redundante. Tal como descrito na discussão do estudo, os autores referem o paradigma que existe entre diversos autores se por exemplo, pelo facto dos critérios de interação social e comunicação estarem intimamente relacionados, estes devem ser considerados um único critério.

Também foi possível encontrar informação da evolução das práticas clínicas [17], bem como a história da toxina botulínica, desde o primeiro registo de informação (886 AC), até aos dias correntes [18].

As ferramentas de avaliação utilizadas pelos profissionais do CMRA em consultas de avaliação geram informações como: saber qual a necessidade de ajuda do paciente, qual o nível de dor dos membros, qual a frequência dos espasmos, entre outros. Até à data, de acordo com a pesquisa realizada, nunca foi realizada uma publicação sobre a testagem da viabilidade e correlação que as ferramentas e perguntas realizadas aos pacientes possuem. Este estudo pode levar a uma melhoria da informação gerada, podendo originar a futura criação de modelos de previsão. Por exemplo, supondo um paciente com determinadas características numa avaliação inicial, o médico consegue avaliar a probabilidade de agravamento do seu caso e, deste modo, prever se o mesmo necessita apenas de tratamento fisioterapêutico, cirurgia ou fisioterapia com aplicação da toxina.

Alguns exemplos de ferramentas de avaliação de pacientes utilizadas no mercado e no centro são (*Tabela 5 – Anexos*):

- Escala de *Ashworth* Modificada
- Escala Visual Analógica
- *Caregiver Burden Scale*
- Escala de frequência de espasmos
- Tempos de marcha
- Tempos de reação

4 Engenharia

4.1 Levantamento e Análise de Requisitos

Atendendo que o projeto incorpora uma vertente mais orientada para o âmbito científico, a concretização do levantamento e análise de requisitos remete-se a uma abordagem mais generalizada uma vez que existem diferentes práticas passíveis de serem aplicadas no estudo de dados. Embora o TFC incorpore duas fases, sendo a primeira passível de enumerar requisitos concretos fornecidos pelos médicos especialistas, a segunda fase remete para uma componente de estudo pela qual não se consegue levantar requisitos do cliente.

Como previamente referido, o projeto encontra-se dividido em duas fases, sendo a primeira fase a informatização de um formulário disponibilizado pelo CMRA, utilizado quando um paciente tem uma consulta de (re)avaliação sobre a sua condição de espasticidade. Este formulário, para além de ter disponibilizado com a devida antecedência, foi posteriormente possível reunir com o Médico responsável pela prática clínica em questão a fim de aferir quais os requisitos que o mesmo pretende. O mesmo indicou diversos requisitos, tais como:

- Adição de novos parâmetros
- Remoção de parâmetros obsoletos presentes no formulário manual
- Alteração de nomes de parâmetros
- Indicação de quais os campos de cariz obrigatório
- Indicação de quais as opções disponíveis para certo tipo de campos
- Indicação de requisitos de usabilidade (ex: uso de *sliders* por causa de pacientes)

Foi possível após a reunião proceder à realização de uma tabela indicativa de todos os requisitos imprescindíveis (*Tabela 6 - Anexos*).

4.2 Conceitos teóricos

Na segunda fase do projeto pretende-se estudar a importância das variáveis do formulário. Para tal, serão aplicadas técnicas como Análise de Componentes Principais (PCA) que têm como objetivo a redução da dimensionalidade de um *dataset*. As variáveis originais são transformadas em componentes principais, que resultam de combinações lineares das variáveis, e que permitem comportar uma elevada percentagem da informação original.

Neste TFC, poderemos aplicar correlação de variáveis, também conhecido como *Variable Correlation*, a fim de poder expressar a sua correlação linear entre variáveis. É frequentemente utilizada para demonstração de “relações” simples sem qualquer tipo de justificação sobre a causa e efeito.

A análise de componentes principais (PCA) é um método matemático utilizado para a redução da dimensionalidade. Implica encontrar as melhores combinações lineares de características originais para criar variáveis designadas por componentes principais (Z_i).

Como se pode verificar na *Figura 6*, estas componentes principais são calculadas através da combinação linear de variáveis sendo os coeficientes associados designados por *loadings* (a_{ij}).

Estas componentes anulam a multicolinearidade e são classificadas com base na sua importância para explicar a variância dos dados. As componentes principais captam a informação mais significativa das características originais, são ortogonais entre si e estão ordenadas da mais para a menos importante.

PCA

👉 PCA = finding the **best linear combination** of features...

$$Z_1 = a_{11}X_1 + a_{12}X_2 + a_{13}X_3$$

$$Z_2 = a_{21}X_1 + a_{22}X_2 + a_{23}X_3$$

$$Z_3 = a_{31}X_1 + a_{32}X_2 + a_{33}X_3$$

- **Cancelling all multicollinearity**
- **Ranking** new features **ZZ in the most unequal way** from most to least "important"

Z_i are called the **Principal Components (PC)**

Figura 6 – Método de cálculo das PC (Fonte: Slides aulas data mining)

Quando aplicado PCA, primeiramente, é aconselhável a padronização dos valores correspondentes a cada variável, no caso de existirem variáveis com valores de magnitudes diferentes. Uma variável que apresente valores numa escala elevada, por exemplo entre 0 e 100, ao ser equiparada a uma variável cuja escala seja mais reduzida, por exemplo entre 0 e 1, pode levar a resultados tendenciosos devido unicamente à magnitude dos valores e não à relevância da variável [19].

Após a padronização encontram-se as direções nos dados com maior variabilidade, através do cálculo dos valores próprios e dos vetores próprios da matriz de covariância dos dados. Os vetores próprios, que permitem calcular as componentes principais, são ortogonais (perpendiculares) entre si e definem os novos eixos do espaço. Para reduzir a dimensionalidade de um conjunto de dados usando PCA, as componentes principais são ordenadas por ordem decrescente de percentagem de variabilidade explicada por cada uma e são mantidas apenas as componentes que explicam em conjunto um total de, por exemplo, 95% da variabilidade total dos dados. Isso resulta num novo conjunto de dados de menor dimensão que preserva o máximo possível da variação original [19].

Este tipo de prática é possível ser observado na *Figura 7*, em que ao ser aplicado a técnica de PCA, foi possível preservar a imagem com apenas 5 componentes (última fotografia da primeira linha) ao invés das 8 componentes (última fotografia).

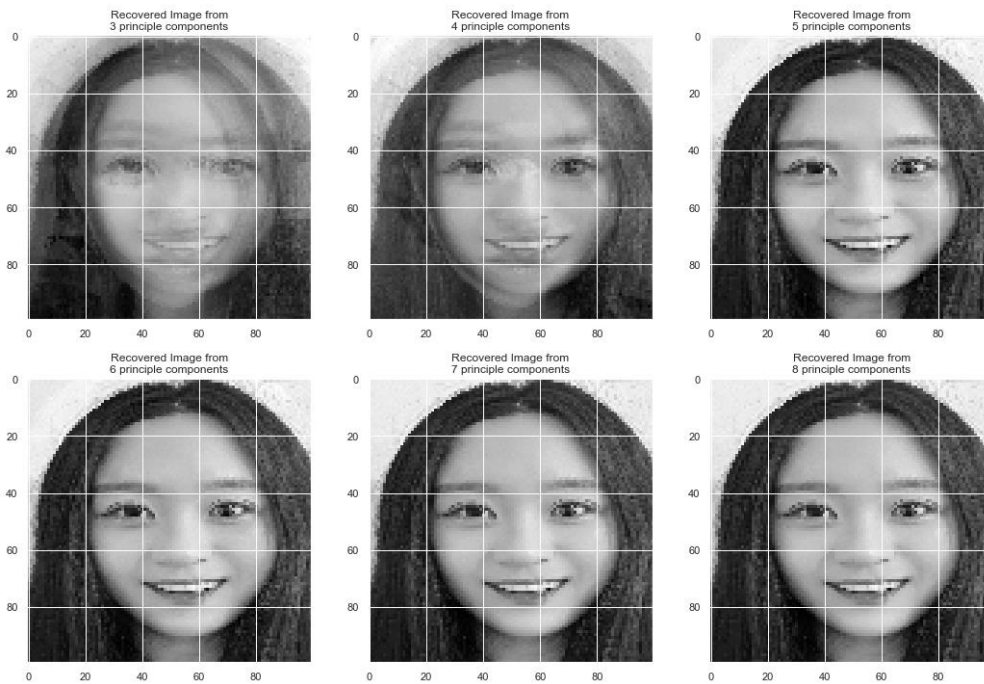


Figura 7 – Exemplo de PCA aplicado a reconhecimento facial

A escolha do número ótimo de componentes principais (K) a manter também pode ser feita pela aplicação do método do cotovelo, mais conhecido por “*Elbow Method*” (*Figura 8*).

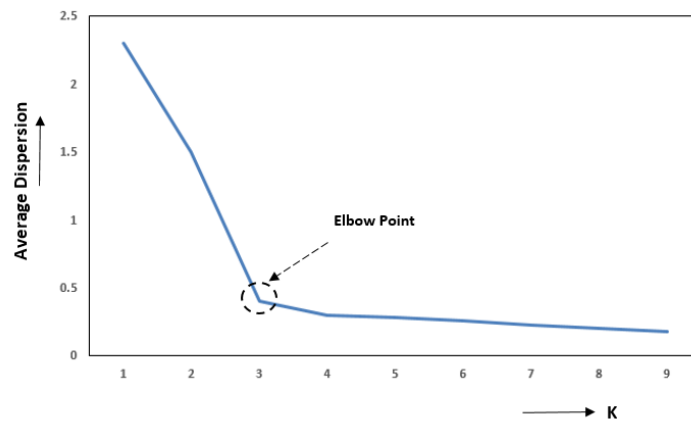


Figura 8 - Representação gráfica para encontrar o K ótimo (Elbow Method)

Em suma, para o sucesso na aplicação da segunda fase do projeto é necessário garantir um pré-processamento adequado e com espírito crítico assegure que se tem informação sobre todas as variáveis para todos os pacientes, os dados encontram-se limpos, isto é, a inexistência de valores em falta, o preenchimento adequado de todos os campos e valores coerentes e fidedignos, de modo a proceder a:

- Análise exploratória dos dados
- Padronização dos dados
- Cálculo da matriz de covariâncias (ou de correlações utilizando os dados originais)
- Cálculo dos valores e vetores próprios
- Construção das componentes principais
- Identificação das variáveis com maior importância.

4.3 Descrição de cenários de aplicação

Este TFC, tal como anunciado anteriormente, divide-se em duas fases sendo a primeira fase, a concretização da digitalização do formulário que o CMRA dispõe atualmente, necessitando de pequenas atualizações, todas de cariz obrigatório. O formulário em si não requer nenhuma componente estrutural tecnológica complexa, isto é, não precisa da criação de nenhuma base de dados ou plataforma, sendo apenas necessário disponibilizar aos profissionais de saúde um formulário onde possam digitar quais as informações obtidas aquando da realização da consulta com o paciente, submeter o mesmo e ficar com a informação armazenada em formato digital. Deste modo, apenas existe um cenário de aplicação, sendo o mesmo a abertura de um formulário e o respetivo preenchimento do mesmo no decorrer da consulta, bem como a submissão do formulário.

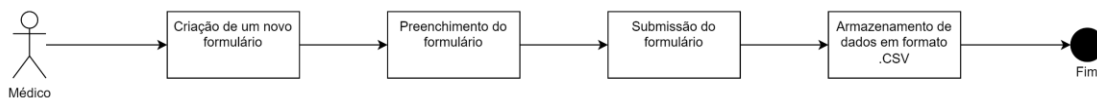


Figura 9 - Caso único do preenchimento da ficha no decorrer da aplicação da toxina botulínica

A segunda fase necessita da posterior avaliação clínica para efetivamente ser implementada em casos reais. Embora no TFC decorra o estudo das diversas variáveis e sejam aplicadas técnicas conhecidas na área de ciência de dados, o médico responsável pela prática clínica inovadora, deve dar o seu parecer para a implementação dos resultados obtidos no CMRA. Assim, poder-se-á perceber se efetivamente os resultados obtidos transcrevem a realidade ou se existem falhas entre a análise feita e a realidade (Figura 10). Caso estas perguntas tenham correlações umas com as outras pode significar que se esteja a obter a mesma informação e que na realidade as perguntas realizadas, embora diferentes, possam estar a responder à mesma questão [20].

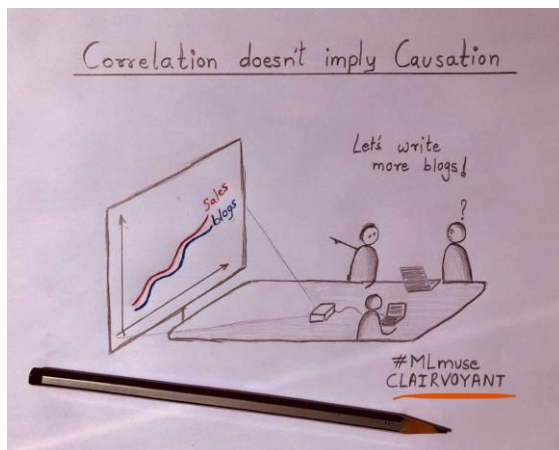


Figura 10 - A realidade vs. Os Resultados

5 Solução Desenvolvida

5.1 Introdução

A solução foi desenvolvida em duas fases (*Figura 11*):

1ª fase: Criação de um formulário em HTML com auxílio de PHP para poder facilitar a inserção dos dados por parte dos profissionais de saúde no ficheiro CSV para posterior análise. Esta proposta atende ao facto do CMRA necessitar de informatizar o processo de inserção de dados no ficheiro CSV e descontinuar a prática em papel.

2ª fase: Estudo da prática clínica, através do estudo dos dados e variáveis, aplicando diferentes técnicas de *Machine Learning*, tais como, *Feature Selection*, *Principal Component Analysis*, *Variable Correlations*, etc. Pretende-se que nesta fase se ganhe conhecimento sobre a prática e conclua quais as componentes mais importantes do questionário, a fim de apurar os dados indicadores de informação relevante e os que não trazem qualquer tipo de informação adicional. Serão utilizados os dados armazenados pelo CMRA em Excel dos últimos anos da prática clínica.

Pretende-se aplicar os conhecimentos adquiridos curricularmente, bem como demonstrar iniciativa na utilização de tecnologias e metodologias não lecionadas nas unidades curriculares da Licenciatura.

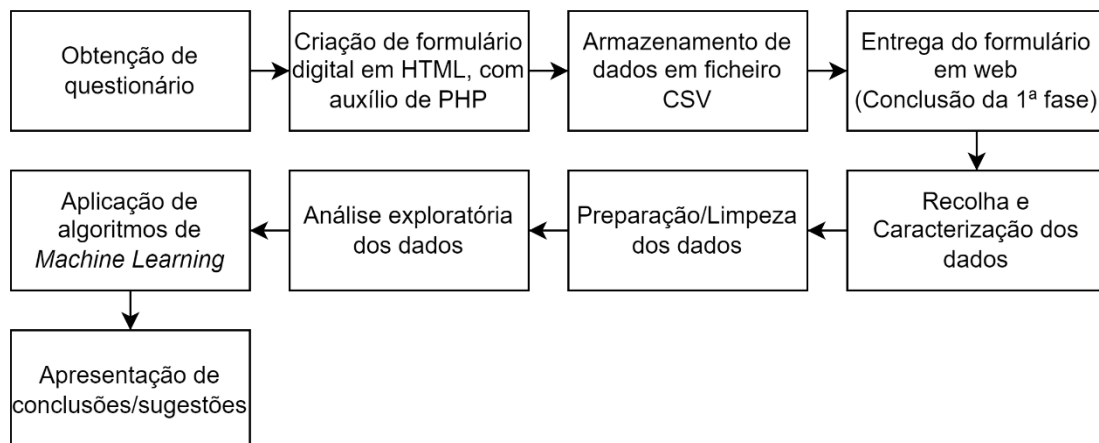


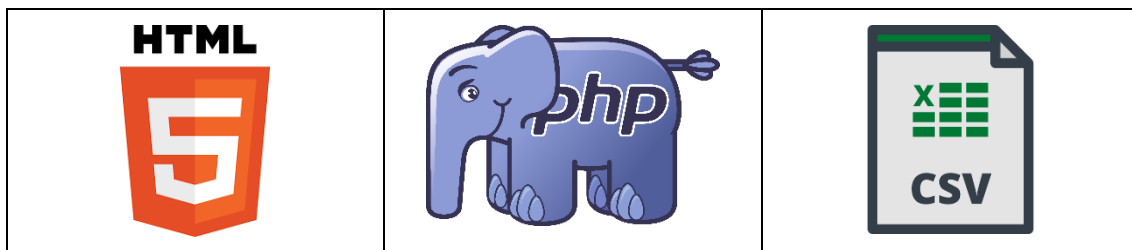
Figura 11 - Plano de execução da solução

NOTA: Por motivos de confidencialidade não será possível disponibilizar o *notebook* com o desenvolvimento realizado para o estudo dos dados, uma vez que compreendem dados sensíveis. Consequentemente, não será providenciado um vídeo demonstrativo uma vez que adjacientemente incorre no incumprimento de confidencialidade.

5.2 Tecnologias e Ferramentas Utilizadas

Para a 1ª fase, foi diagnosticado que não existe a necessidade de desenvolver um sistema complexo, com implementações de login, registo de utilizadores e leitura, atualização e eliminação de questionários. De acordo com responsáveis da prática clínica, apenas é necessário que o processo seja informatizado, simples e eficaz. É então proposto o desenvolvimento de uma página em *HyperText Markup Language* (HTML) com o auxílio de *Hypertext Preprocessor* (PHP) para transpor os dados para um ficheiro *Comma Separated Values* (CSV) (Tabela 1).

Tabela 1 - Ferramentas utilizadas no formulário digital



A 2ª fase requer a implementação de bibliotecas, ferramentas e processos (Figura 12) de *Machine Learning* (ML) todas elas orientadas a *Python*, sendo o principal objetivo correlacionar variáveis/features para poder, *á posteriori*, interpretar os resultados e proporcionar conclusões e/ou sugestões para o CMRA.



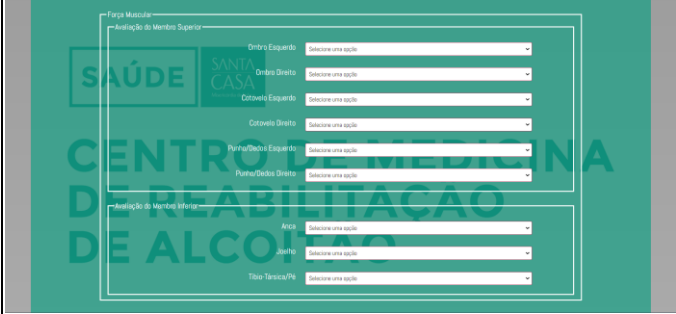
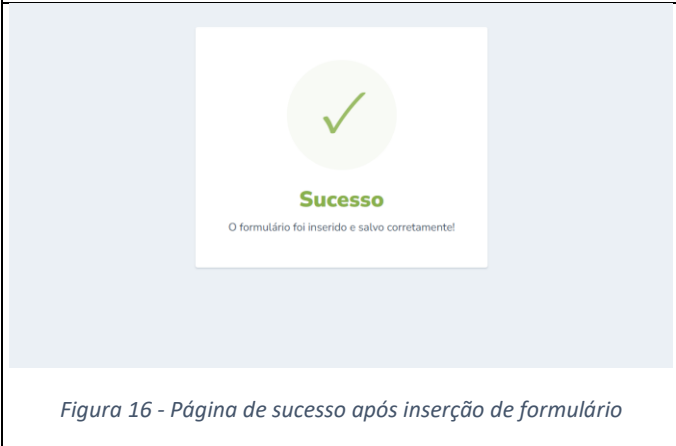
Figura 12 – Stack Tecnológico usado para Machine Learning

5.3 1ª fase – Desenvolvimento de formulário digital

Comporta-se na 1ª fase o desenvolvimento de um formulário digital que permita a clínicos realizar a inserção das consultas num formato digital, de modo a reduzir o gasto de papel e facilitar na preservação e estudo dos dados. Após reunião com o médico responsável, foi possível compreender que existia a falta de inserção de alguns campos requisitados pelos profissionais de saúde, pelo facto de a ficha digitalizada que fora disponibilizada se encontrar desatualizada dos pressupostos médicos atuais. Foi também pedido que, se possível, fossem feitos alguns cálculos e incorporadas algumas fórmulas no formulário a fim de evitar o erro humano aquando da inserção dos dados no formulário (Tabela 7 – Anexos).

A tabela abaixo pretende demonstrar com algumas imagens o desenvolvimento realizado na primeira fase.

<p>The screenshot shows the top section of a digital form. At the top, it says 'CONSULTA DE TOXINA BOTULÍNICA' and 'Registro de intervenção, Definição de Objetivos e Avaliação'. Below this, there are several input fields: 'Nome completo do Profissional', 'Nome do paciente', 'Médico responsável', 'Bairro', and 'Município'. The background features the logo of 'SAÚDE SAUVIA CASA' and 'CENTRO DE MEDICINA DE REABILITAÇÃO DE ALCOITÃO'.</p>	<p>Campo de identificação do paciente, bem como a inserção do nome do profissional que preencheu o formulário.</p>
<p>The screenshot shows a section titled 'Escala Visual Analógica (EVA) - ODI' for 'Avaliação de Membros Superior'. It lists several body parts with corresponding dropdown menus for pain level: 'Ombro Esquerdo', 'Ombro Direito', 'Cotovelo Esquerdo', 'Cotovelo Direito', 'Punho/Dirito Esquerdo', 'Punho/Dirito Direito', 'Axa Esquerda', 'Axa Direita', 'Joelho Esquerdo', 'Joelho Direito', 'Tubo Torácico/Pé esquerdo', and 'Tubo Torácico/Pé direito'. Each dropdown menu is currently set to 'Indicar nível de dor'.</p>	<p>Implementação da escala visual de dor para permitir a pacientes com espasticidade de indicar numa escala de 0-10 a dor que sentem.</p>

	<p>Secção de análise da força muscular do paciente dos membros superiores e inferiores.</p>
<p>Figura 15 - Secção de força muscular do formulário</p>	<p>Após a inserção do formulário preenchendo todos os campos devidamente e tendo todas as regras cumpridas (dose declarada igual à soma das dosagens atribuídas por músculo), será apresentada ao utilizador uma página que lhe indica o sucesso na inserção do formulário.</p>
	

O formulário incorpora as funcionalidades requisitadas em ambas as vertentes de campos necessários e validação/cálculos realizados pelo sistema. Através de mecanismos como *jQuery*, *JavaScript* e *PHP* foi possível agilizar todo o processo e disponibilizar ao CMRA um formulário consistente e fidedigno para a inserção dos dados nos ambientes computacionais no dia-a-dia dos profissionais.

Considera-se interessante a possibilidade de nova colaboração para dar continuidade ao desenvolvimento de novas funcionalidades, desenvolvimento de outros formulários utilizados em práticas clínicas ou sistemas de gestão documental.

Link GitHub: <https://github.com/DEISI-ULHT-TFC-2022-23/TFC-DEISI341-Aplicacao-de-AI-para-estudo-de-pratica-clinica-inovadora>

Link Youtube: <https://www.youtube.com/watch?v=vIrywvLe0il>

5.4 2ª fase – Análise dos dados

5.4.1 Descrição dos dados

No ficheiro disponibilizado, encontram-se 163 variáveis que representam informações relacionadas ao histórico clínico e tratamento de um paciente que sofreu um acidente vascular cerebral (AVC) e foi tratado com toxina botulínica tipo A (BoNTA). Algumas das informações incluídas são:

- Data de nascimento: Data em que o paciente nasceu;
- Género: Sexo do paciente;
- Diagnóstico: Diagnóstico clínico do paciente;
- Data do AVC: Data em que ocorreu o AVC;
- Primeira administração de BoNTA - sempre: Data da primeira administração da toxina botulínica;
- Data do tratamento com BoNTA;
- Data da avaliação: Data em que o paciente foi avaliado;
- Idade: Idade atual do paciente;
- Idade no momento do AVC: Idade do paciente quando sofreu o AVC;
- Intervalo entre AVC e primeira administração de BoNTA: Tempo decorrido entre o AVC e a primeira administração da toxina botulínica;
- Intervalo entre AVC e BoNTA: Tempo decorrido entre o AVC e o tratamento com BoNTA;
- Escala de espasticidade de Ashworth (MAS): Escala utilizada para avaliar a espasticidade nos músculos;
- Cadência (passos/min): Quantidade de passos por minuto;
- Comprimento do passo (metros): Comprimento do passo do paciente;
- Reações associadas (ARRS): Reações involuntárias durante a realização de uma tarefa;
- Escala de frequência de espasmos: Escala que avalia a frequência de espasmos musculares;
- Dose de Dysport/Xeomin/Botox: Quantidade de toxina botulínica administrada;
- Percentual de Dmax Dysport/Xeomin/Botox: Percentagem do máximo efeito da dose de toxina botulínica;
- Palpação/Neuroestimulação/EMG/Ultrassom: Métodos utilizados para guiar a injeção de toxina botulínica nos músculos afetados;
- Músculos tratados: Lista dos músculos tratados com toxina botulínica.

Estes são apenas alguns exemplos que podem também ser perceptíveis no questionário (*Figura 13, Figura 14*), uma vez que este é utilizado para realizar a anotação das diversas informações aquando da realização de uma (re)avaliação do tratamento. Posteriormente à realização da consulta, e como indicado anteriormente, estes dados são colocados no ficheiro que nos foi disponibilizado, como método de armazenamento em formato digital, contudo estes dados encontram-se bastante incompletos e nunca foram utilizados num estudo científico com práticas de *machine learning*.

5.4.2 Pré-processamento dos dados

Foi disponibilizado pelo médico responsável pela prática clínica do CMRA um ficheiro Excel com diversos registos de consultas a pacientes a quem foi aplicada a toxina. Ao analisar este ficheiro, foi possível verificar a ausência de informação em determinados campos, originando registos incompletos ou vazios, bem como colunas sem qualquer tipo de estrutura. Estas colunas são de extrema relevância, tal como indicado anteriormente, tanto na componente do formulário digital como no estudo dos dados, pelo que, é necessário ponderar a viabilidade da remoção da coluna.

O ficheiro disponibilizado possuía diversos tipos de informação com formatação e estruturas diferentes, mas contendo informações de extrema relevância para a medicina reabilitativa e para a prática clínica. Também, com base neste ficheiro, foi possível compreender de forma mais abrangente, o tipo de dados e a correlação entre alguns campos do formulário digital, que até ao momento causavam alguma dúvida e dificuldade de interpretação.

Deste modo, foram efetuados diversos processos relacionados com *data wrangling* (Figura 17). Ao importar o ficheiro Excel para o *Jupyter notebook*, foi observado que inicialmente existiam 2985 linhas (registos) e 163 colunas (variáveis), sendo que nesta primeira fase não tinha ainda sido realizado qualquer tipo de limpeza aos dados. Primeiramente foram removidas todas as colunas que se encontravam sem cabeçalho, reduzindo assim de 163 para 149, de seguida foi analisada a percentagem de valores em falta em cada coluna. Atendendo às linhas, objetivou-se preservar o máximo de linhas possíveis removendo primeiramente os registos em branco, ficando assim com 2808.



Figura 17 – 6 passos para a limpeza e estruturação de um dataset

É prática comum, sempre que não seja possível obter dados fiáveis para o preenchimento de *missing values*, remover todas as colunas cuja percentagem destes valores seja igual ou superior a 30%, evitando assim enviesamentos no estudo. A referida técnica foi aplicada à base de dados, ficando assim um total de 95 colunas. Ao analisar em maior detalhe aplicou-se o mesmo critério aos registos com mais de 30% de *missing values*, reduzindo o número de registos de 2808 para 2661. Mesmo após todos estes procedimentos, foi possível verificar a existência de colunas com uma percentagem de valores em falta muito próximos a 30%, sendo que deste modo foi necessário verificar cada coluna. No decorrer da análise das colunas, verificou-se também a existência de somas incorretas das dosagens, problemas de inserção relativamente às áreas tratadas, cálculos de percentagens das dosagens incorretas, entre outros erros de preenchimento de registos. Perante tais circunstâncias, deliberou-se que o preenchimento dos *missing values* com a média ou moda da respetiva coluna não seria uma boa prática, uma vez que existem práticas clínicas e ações que não se deve “manipular” pois pode induzir todo o resultado do estudo em erro. Concluiu-se deste modo e após bastante análise que se deve remover todos os registos cuja informação não esteja totalmente completa. Face à impossibilidade de os corrigir, devido à inacessibilidade aos registos em papel, que poderão conter a restante informação, e sendo inúmeros os registos que se encontram incompletos, seria um trabalho que iria exceder o tempo disponível para realizar o TFC. Igualmente não se dispõe do conhecimento necessário para interpretar os dados presentes nesses registos em papel, com exceção dos responsáveis de saúde do CMRA. Deste modo, foi reavaliado todo o procedimento realizado até então, tendo ficado um total de 1331 registos e 94 colunas. Foi notada a existência de uma coluna cuja informação era irrelevante pois o valor era sempre igual, tal foi removida e em contrapartida adicionadas 3 outras, que outrora haveriam sido descartadas por não cumprirem os requisitos previamente indicados. Contudo, as mesmas são relevantes para cálculos de somatório e percepção do resultado obtido atendendo ao objetivo do paciente, ficando assim com 1331 registos e 97 colunas (Figura 18).

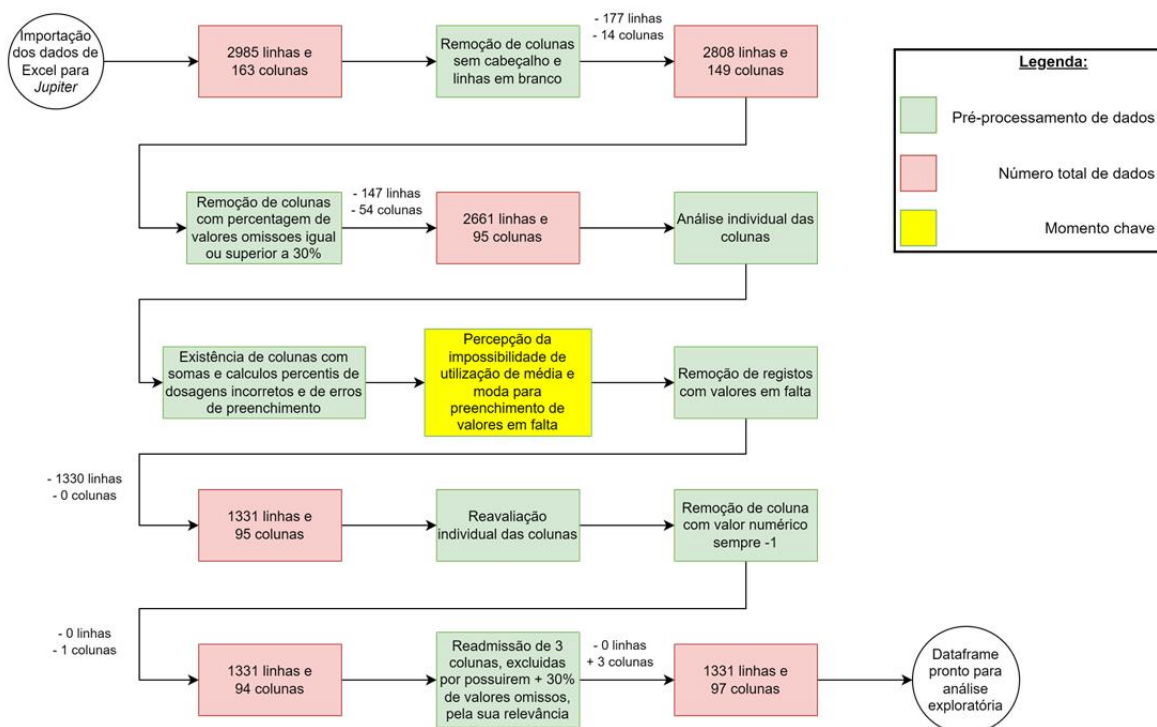


Figura 18 – Resumo do pré-processamento de dados

5.4.3 Análise exploratória dos dados

Compreende-se que a prática clínica é realizada em 127 pacientes do sexo masculino (57.21%) e 95 do sexo feminino (42.79%) (Figura 19) cuja idade se encontra compreendida entre os 20 e 86 anos. A idade média dos pacientes encontra-se nos 58.28 anos, sendo o registo do paciente mais novo com 19 anos e do paciente mais velho com 86 anos.

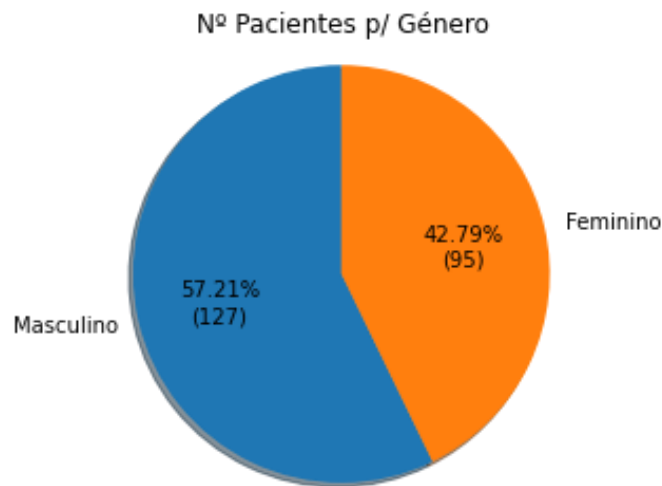


Figura 19 - Nº Pacientes por Género

Como tal, foi possível realizar uma análise comparativa entre a distribuição da idade por registo (Figura 20) e a distribuição da idade média por paciente (idade do paciente em cada tratamento aplicado) (Figura 21). A diferença entre as duas deve-se ao facto de um mesmo paciente poder ter vários registos associados a aplicações da toxina em diferentes momentos. Desta análise concluiu-se que, embora o número de registos em pacientes com idade compreendida entre os 50 e os 59 anos e entre os 60 e os 69 anos seja bastante próxima, o número de pacientes no primeiro intervalo de idades é inferior. Este indica um maior número de aplicações por paciente com idade entre os 50 e os 59 anos. Este resultado pode indicar que pacientes com idade inferior a 60 anos requerem mais aplicações possivelmente por motivos de objetivos pessoais ou maior desconforto.

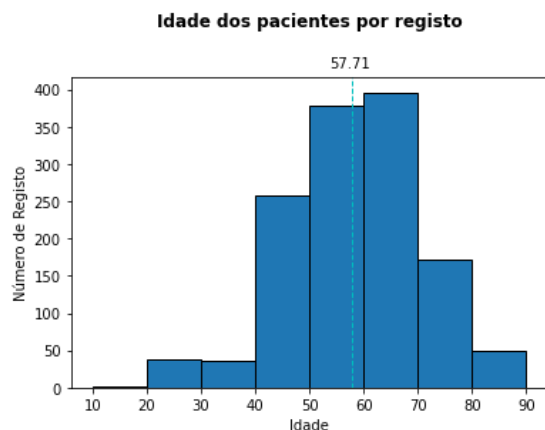


Figura 20 – Idade dos pacientes por registo

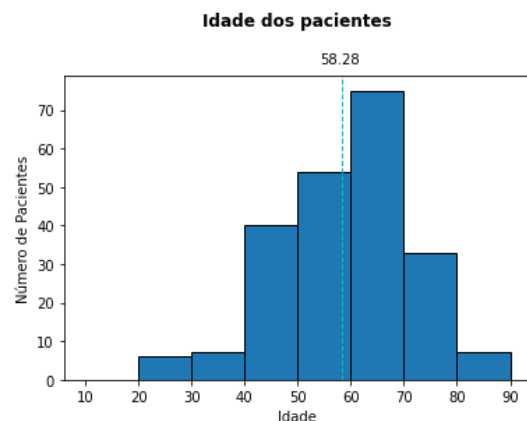


Figura 21 - Idade dos pacientes por número de Pacientes

No decorrer da caracterização e limpeza dos dados, foi possível perceber que existem diversas variáveis cuja informação considera-se pertinente para a análise exploratória. Por questão de facilidade em termos de análise da informação que uma determinada variável acarreta, foi gerado um mapa de correlação (*Figura 31, Figura 32 - Anexos*) que demonstra o quão correlacionadas as variáveis se encontram entre si (valores próximos de 1 (representadas na escala por cor mais branca) e -1 (representadas na escala por cor mais roxa) indicam uma forte correlação, positiva e negativa respetivamente, enquanto valores próximos de 0 indicam uma inexistência de correlação entre as variáveis). Graças a este mapa, foi possível verificar que existia uma coluna cujo valor encontrava-se sempre igual a -1, como tal, apresentava uma linha branca no gráfico. Esta variável é irrelevante para o estudo e podemos removê-la.

A correlação negativa, também conhecida como correlação inversa, é um tipo de correlação estatística que indica uma relação inversa entre duas variáveis. Ao analisar o mapa de correlação, encontram-se alguns casos tais como:

- *Impairment e Diagnosis* (-0.83)
- *Stroke date e Stroke-BonTA interval* (-0.77)
- *Diagnosis e Treated Side* (-0.71)
- *First BoNTA administration ever e Stroke-BonTA interval* (-0.6)
- *Stroke-First BonTA interval e Stroke Date* (-0.51)

A correlação positiva é um tipo de correlação estatística que indica uma relação direta entre duas variáveis. Em outras palavras, quando uma variável aumenta, a outra tende a aumentar, e quando uma variável diminui, a outra tende a diminuir. No mapa de correlação encontram-se inúmeros casos de correlação positiva:

- *Gastrocnemius lateralis e medialis* (0.95)
- *Fibularis Longus e Fibularis brevis* (0.92)
- *Sintomas/Défices e Principal Goal Subcategory* (0.88)
- *Treated Side e Impairment* (0.85)
- *Semimembranosus e Semitendinosus* (0.84)

As correlações indicadas anteriormente são alguns exemplos de casos que se encontram altamente correlacionados de forma negativa ou positiva. No entanto, existem casos cuja correlação, embora menor, pode acarretar maior significado para os profissionais de saúde, uma vez que podem conter indícios de outros músculos a ter em conta ou mesmo escalabilidade da situação do paciente. Alguns desses casos podem ser:

- *Supinator e Digitorum extensor* (0.32)
- *Biceps cruris e Psoas iliacus* (0.28)
- *Semimembranosus e Adductor muscles* (0.26)
- *Rectus anterior e Vastus internus* (0.26)
- *Flexor hallucis longus e Flexor digitorum longus* (0.33)

Nestes casos de menor correlação seria interessante perceber junto do médico responsável qual poderá ser o motivo para existir alguma correlação (embora pequena) relativamente ao *Supinator* e *Digitorum extensor* (0.32). No entanto, casos como *Gastrocnemius lateralis* e *medialis* (0.95) ou *Fibularis Longus* e *Fibularis brevis* (0.92), apresentam uma correlação muito elevada pelo que pode indiciar que quando tem de ser administrada ao paciente a sua dose num determinado músculo, muito possivelmente será também administrada no outro.

Posteriormente e em discussão com o médico responsável, o mesmo indicou que a idade do paciente e a quantidade de tratamentos poderá estar relacionada com os objetivos do paciente. Deste modo foi considerado pertinente averiguar as causas mais comuns, bem como o objetivo do paciente e a média da idade por género, a fim de tentar perceber se de facto o relatado se verificava nos dados.

		Medium Age
	Principal goal Subcategory	Gender
D1-Amplitude de movimento/prevenção contracturas	Female	61.32
	Male	56.66
D1-Dor/Desconforto	Female	59.46
	Male	60.09
D1-Movimentos involuntários	Female	52.36
	Male	57.67
D2- Facilitating therapy	Female	55.52
	Male	61.58
D2-Funções activas/atividade motora	Female	56.20
	Male	55.35
D2-Funções passivas/cuidados	Female	62.67
	Male	59.43
D2-Mobilidade(transf/vert/marcha)	Female	58.12
	Male	56.97

Figura 22 - Média de idade por género e objetivo do tratamento

A Figura 22 apresenta informações sobre a relação entre a média de idade e diferentes objetivos terapêuticos em duas subcategorias (D1 e D2), de acordo com o género (feminino e masculino). Observa-se que, na maioria dos casos, as mulheres têm uma idade média superior do que os homens em todas as subcategorias, exceto D1-Dor/Desconforto, D1-Movimentos involuntários e D2-Facilitating therapy. Isso sugere que, em média, as mulheres apresentam mais tardiamente necessidade de terapia em relação aos homens.

Essas informações podem ser úteis para orientar terapeutas na escolha de abordagens terapêuticas específicas para cada subcategoria, levando em consideração a idade e o género do paciente. Além disso, pode ser útil investigar as razões pelas quais as mulheres apresentam, em média, uma necessidade tardia de terapia em relação aos homens.

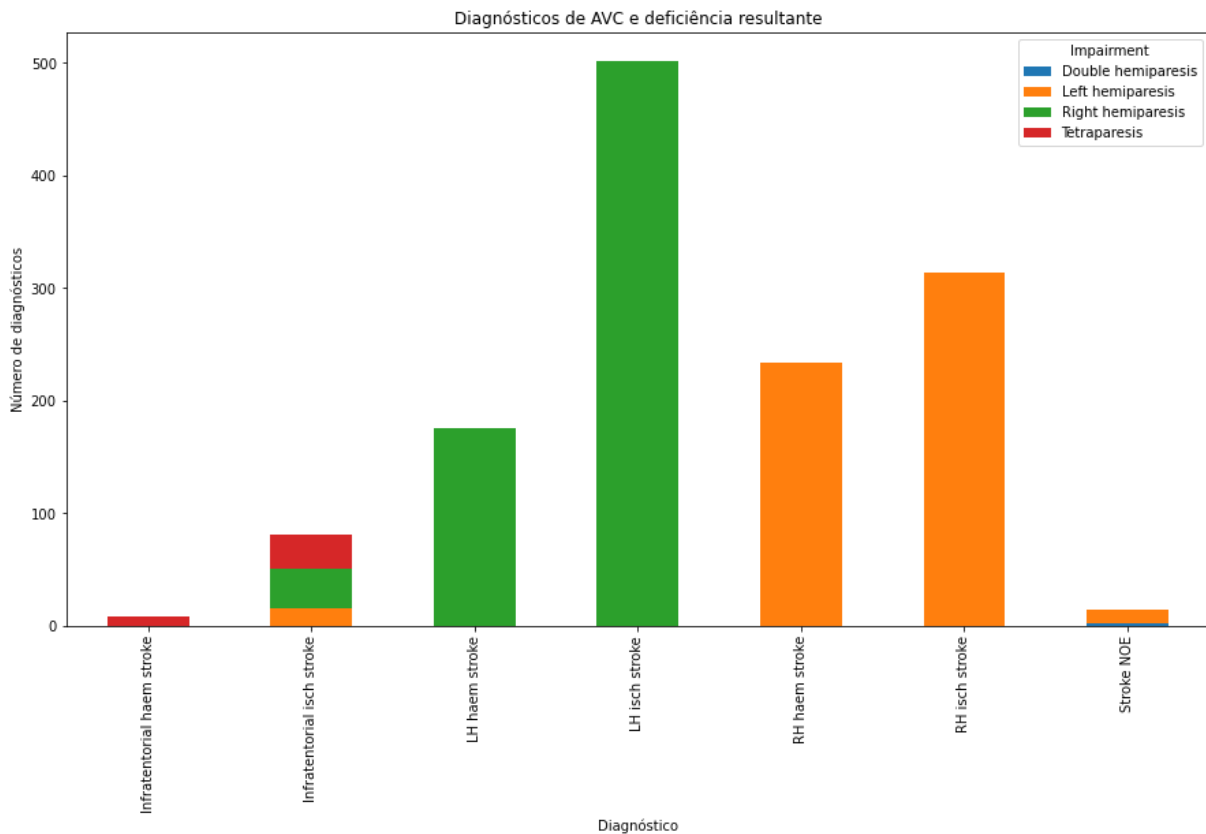


Figura 23 – Causa mais comum por diagnóstico

A *Figura 23* apresenta informações sobre diferentes diagnósticos de AVC (acidente vascular cerebral) e a deficiência resultante em termos de paralisia em membros (tetraparesia, hemiparesia direita, esquerda ou dupla). Observa-se que os diagnósticos de AVC associados a hemorragias cerebrais (haem) ou isquemia (isch), tendem a resultar em tetraparesia ou hemiparesia no lado oposto do corpo afetado.

Essa informação pode ser útil para terapeutas ao planejar abordagens terapêuticas específicas para cada tipo de debilitação resultante de um AVC. Por exemplo, em casos de hemiparesia, o terapeuta pode-se concentrar em melhorar a força e a amplitude de movimento nos membros afetados, enquanto em casos de tetraparesia, pode ser necessário um esforço maior na reabilitação dos membros inferiores para recuperar a capacidade de locomoção.

No entanto, é importante lembrar que cada paciente pode apresentar uma recuperação e necessidades de reabilitação diferentes, mesmo com o mesmo diagnóstico de AVC. Portanto, é importante que o plano terapêutico seja individualizado para cada paciente e que o terapeuta avalie continuamente a eficácia da intervenção e faça ajustes conforme necessário.

Para analisar as taxas de sucesso e insucesso recorreremos à variável GAS T Score (Figura 24). Esta é uma medida que avalia o grau de mudança no desempenho do paciente após o tratamento. A pontuação varia de -2 a +2, sendo que valores acima de -1 são considerados um sucesso e valores iguais ou abaixo de -1 são considerados insucesso.

De acordo com os dados fornecidos, podemos calcular a taxa de sucesso e insucesso da seguinte maneira:

- Taxa de sucesso: $(1109 + 59 + 9) / (1109 + 149 + 59 + 9 + 3) = 0,8856$ ou 88.56%
- Taxa de insucesso: $(149 + 3) / (1109 + 149 + 59 + 9 + 3) = 0.1144$ ou 11.44%

Podemos concluir que a taxa de sucesso é alta, com 88.56% dos pacientes apresentando uma pontuação acima de -1 na GAS T Score. A taxa de insucesso é relativamente baixa, com apenas 11.44% dos pacientes apresentando uma pontuação igual ou abaixo de -1.

Taxa de Sucesso e Insucesso por Tratamento

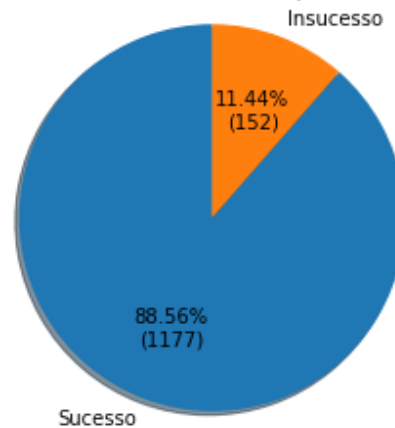


Figura 24 - Taxa de sucesso de acordo com o objetivo do paciente

Todas as informações fruto da aplicação de técnicas e práticas de *machine learning* no conjunto de dados, são de enorme ganho de conhecimento para profissionais de saúde especialistas em reabilitação. Estas informações são novas e nunca publicadas ou vistas uma vez que os dados não tinham ainda sido submetidos a nenhuma componente de inteligência artificial. Foi deste modo, desenvolvido um modelo capaz de nos indicar quais as variáveis que acarretam maior informação, perpetuando e reforçando a análise de modo a evidenciar o máximo de conhecimento possível a fim de oferecer um contributo sólido e enriquecido de informação.

A análise do *Variance Inflation Factor* (VIF) permite identificar multicolinearidades entre as variáveis presentes, devendo ser removidas as variáveis com VIF igual ou superior a 5, pois representam elevada multicolinearidade e poderão ser problemáticas para os modelos de ML. No caso dos dados resultantes do pré-processamento, averiguou-se que existiam 11 variáveis que devido à sua multicolinearidade e pouca importância numa ótica de analista de dados (Tabela 8 – Anexos), deviam ser removidas, reduzindo assim o número de variáveis para 84.

5.4.4 Aplicação de PCA e Resultados

Prosseguindo no estudo dos dados, após o respetivo pré-processamento e posterior análise exploratória, aplicamos o modelo de PCA.

Devido ao facto de os valores das diferentes variáveis serem bastante diversificados, isto é, existem valores de diferentes escalas, será necessário aplicar um *StandardScaler* de forma a “padronizar” ou “normalizar” os mesmos. O *StandardScaler* transforma os dados de forma que eles tenham média zero e desvio padrão igual a 1. Ao padronizar os dados, todas as características têm a mesma importância relativa, evitando que características com valores numéricos maiores dominem o modelo devido à sua escala.

Após a normalização dos dados, prossegue-se com a aplicação de PCA com o número de variáveis igual a 84, sendo necessário encontrar um equilíbrio entre reduzir a dimensionalidade e preservar a quantidade adequada de informação. Geralmente, quanto menor o número de componentes selecionados, maior será a perda de informação dos dados originais.

Uma abordagem prática é analisar o gráfico da variância explicada acumulada em relação ao número de componentes. O gráfico da *Figura 25*, relativo à aplicação do PCA nos nossos dados, mostra como a quantidade de informação preservada aumenta à medida que adicionamos mais componentes. É possível selecionar um número de componentes que mantenha uma percentagem adequada de variância explicada, como 80%, 90% ou 95%, dependendo do caso.

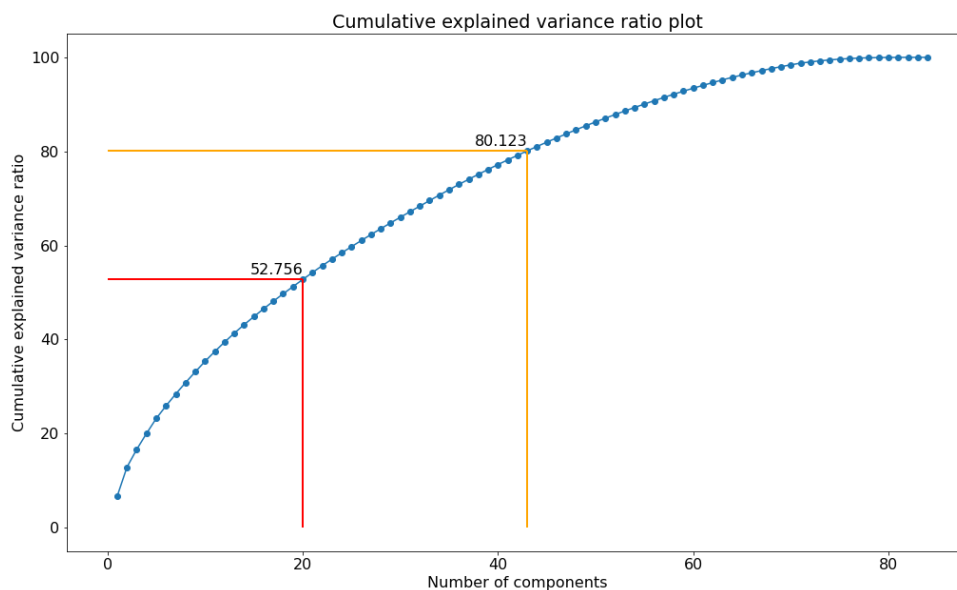


Figura 25 - Variância explicada cumulativa (%)

Pela análise da *Figura 25*, é possível preservar mais de 50% da informação com a utilização de 20 componentes, sendo necessárias 43 componentes para garantir 80% da variância dos dados explicada.

De seguida foram avaliadas quais são as principais variáveis em cada Componente Principal (PC), em termos absolutos (agilizando a ordenação do impacto que cada variável teve nas componentes), e como é que as mesmas traduzem importância na explicação da variância das variáveis. Este tipo de prática tem o intento de ganhar interpretabilidade das variáveis, uma vez que quando aplicada PCA se perde a mesma.

Tabela 2 - Loadings (em valor absoluto) das variáveis mais impactantes na PC1

UL Dose	0.735533
Dysport	0.709040
Gastrocnemius lateralis	0.592053
Gastrocnemius medialis	0.583871
LL Dose	0.579186
Flexor digitorum superficialis	0.573096

Embora as variáveis que comportam maior informação se encontrem nas principais componentes (*Tabela 2, Tabela 9 – Anexos*), essas mesmas variáveis podem não explicar a variação total. Enquanto cada variável associada a uma componente principal explica o quanto a mesma impacta a variância da componente, quando somamos a ponderação de cada variável ao longo das diversas PC, é perceptível que as que se encontravam nas primeiras PC na realidade não explicam a variação total.

```
np.abs (pca.components_.T * np.sqrt(pca.explained_variance_))
```

Figura 26 - Obtenção das magnitudes de cada variável

Outra forma de perceber a importância de cada variável em cada PC é através do cálculo das magnitudes de cada variável. A magnitude permite verificar a proporção da variância de cada variável explicada em cada PC, ou seja, a contribuição relativa de cada variável. Para obter estes valores, efetuou-se o produto dos *loadings* pelo desvio-padrão e colocou-se o resultado em valor absoluto (*Figura 26*), de modo a ignorar qualquer tipo de efeito negativo ou positivo que a variável possa ter. Assim, nesta primeira fase, apenas se tenta compreender qual o impacto que cada variável possui na explicação da variância total. Os resultados apresentados na *Tabela 3* foram obtidos através da soma das magnitudes associadas a cada variável em todas as PC.

Tabela 3 – Top 6 variáveis com maior magnitude na variação total

Flexor digitorum profundus	6.863647
Pectoralis major	6.806062
Systemic medication	6.778461
Orthosis	6.765528
Triceps brachii	6.749912
Pronator teres	6.741226

É importante ressaltar que embora as magnitudes da variação total possam auxiliar na explicação dos dados como um todo entende-se que matematicamente os cálculos realizados possam não demonstrar a sua real importância. Enquanto na aplicação de PCA obtemos as principais componentes em que estas indicam quais as variáveis principais, ao realizar a soma de todos os coeficientes de uma determinada variável, obtemos uma percepção ofuscada da realidade uma vez que os valores das primeiras PC comportam maior informação que as restantes (*Figura 25*). Este caso é perceptível ao analisar a matriz de correlação das variáveis originais com as componentes principais da PCA e verificar que as variáveis apresentadas na matriz são idênticas às que se encontravam presentes na *Tabela 2* e na *Tabela 9* (Anexos) uma vez que são as que incorporam maior importância e informação.

Deste modo, interpela-se a questão fulcral do trabalho: Quais as variáveis mais significativas em termos de informação para o questionário e possíveis futuros estudos, sendo que, após o pré-processamento dos dados, o *dataset* era composto por 98 variáveis? Concebe-se através deste estudo a possibilidade de revelar que, das 98 variáveis, considera-se sustentável dar ênfase a 45 das mesmas. Estas encontram-se destacadas na *Figura 33* dos Anexos, onde se apresenta a correlação anteriormente descrita entre as variáveis e as componentes principais. A *Tabela 4* são o resultado de aplicar um *threshold* de 0,4. Foi aplicado esse valor, pois quando analisados os resultados, conclui-se que existiam inúmeras variáveis cujos valores se encontravam entre 0,5-0,4. Deste modo, na renderização do gráfico, para que o mesmo não fique desproporcional, aplicamos o *threshold* de 0,4 e percebemos que aquelas seriam as variáveis com maior correlação com as PC, atingindo os 80% mínimos (o equivalente a 43 componentes principais) de informação preservada.

Tabela 4 - 45 principais variáveis resultantes do estudo de PCA

Diagnosis	Neurostimulation	Biceps brachii	Semimembranosus	Fibularis Longus
Impairment	Ultrasound	Supinator	Semitendinosus	Fibularis brevis
Stroke date	Upper limbs	Flexor carpi radialis	Gastrocnemius medialis	Flexor hallucis longus
Sheet date	Lower limbs	Carpi extensor	Gastrocnemius lateralis	LL Dose
Age	Upper+Lower limbs	Flexor digitorum superficialis	Soleus	Others
Age at Stroke	Treated side	Opponens pollicis	Flexor digitorum longus	TENS
Dysport	Supraspinatus	Lumbricoides	Extensor digitorum longus	Intercorrências
Xeomin	Infraspinatus	UL Dose	Extensor digitorum brevis	D1-Sintomas/Défices
Botox	Deltoideus	Psoas iliacus	Flexor digitorum brevis	Principal goal Subcategory

Com base nas 45 variáveis mais importantes mencionadas, é possível fazer alguns comentários sobre o número de componentes principais a serem escolhidos e a possibilidade de remover algumas variáveis menos importantes do formulário. A seleção de 45 variáveis indicia a redução da dimensionalidade dos dados, mantendo um número significativo de informações relevantes. No entanto, é importante considerar a interpretabilidade dos resultados e acautelar a não existência de ajuste excessivo (*overfitting*) ao escolher um número elevado de componentes. Recomenda-se realizar uma análise adicional para determinar se as 45 variáveis são de facto essenciais para o problema em questão.

Ao remover as variáveis menos importantes, será simplificada a colheita de dados e permitirá concentrar-se nas informações mais relevantes, resultando num formulário mais conciso e eficiente. No entanto, é sempre importante realizar uma avaliação e validar os resultados para garantir que a exclusão dessas variáveis não prejudica a qualidade da análise ou do modelo final.

A seleção cuidadosa das variáveis relevantes no meio clínico é crucial para otimizar o processo de diagnóstico e tratamento, permitindo uma abordagem mais precisa e eficiente. Ao identificar as variáveis mais importantes, os profissionais de saúde podem direcionar os seus esforços para os aspetos mais relevantes da condição do paciente, aprimorando a qualidade dos cuidados e os resultados clínicos.

Apresenta-se também, e reforçando a possibilidade de trabalhos futuros, a criação de um novo paciente “fictício” desenvolvido com base nos dados pós pré-processamento e com a média de valores obtidos do modelo PCA (*Tabela 11 – “Novo paciente” criado com base na média de valores do modelo PCA*). Este paciente apresenta características que podem ser consideradas mais comuns, uma vez que é um paciente criado com base na média dos existentes. O cálculo da injeção total de toxina botulínica bem como outros valores, foram posteriormente corrigidos através de somatórios e outros métodos uma vez que os valores das variáveis como UL Dose ou LL Dose são somatórios da quantidade total aplicada em cada músculo do membro superior e inferior respetivamente. Poder-se-á em trabalhos futuros, realizar um estudo sobre a classificação de pacientes ou a previsão de sucesso ou insucesso na aplicação do tratamento. Através do trabalho realizado, pretende-se abrir portas para outros estudos, a fim de realizar uma parceria duradoura, positiva e enriquecedora para o âmbito científico e o CMRA, oferecendo a discentes da Universidade Lusófona a possibilidade de empreender em projetos únicos e magnânicos.

5.5 Parecer clínico & Feedback

Após a finalização do TFC, realizou-se uma reunião junto da diretora de unidade do CMRA e o respetivo médico responsável pela prática clínica, a fim de obter um parecer clínico dos resultados obtidos e a respetiva entrega do formulário de inserção de consultas de avaliação clínica.

Referente ao formulário digital entregue, foi indicado que o mesmo se encontrava perfeitamente adequado às necessidades do centro e apenas pediram a inclusão de mais uma verificação antes da inserção, sendo esta a validação das datas (*Figura 27*).

Data de Nascimento < Data do AVC < Data de aplicação da Toxina

Figura 27 - Validação das datas (Formulário digital)

Foi igualmente possível reunir com o diretor informático do CMRA, para compreender como seria possível realizar a entrega e implementação do formulário nos servidores do Centro, pelo que o mesmo referiu que como o formulário não requeria nenhum tipo de aquisição de software, nem nenhuma preparação complexa de uma máquina virtual, seria facilmente possível incorporar o formulário nos servidores e disponibilizar aos responsáveis clínicos o respetivo link de acesso.

Compreendeu-se, contudo, que o diretor informático do CMRA gostaria de poder ter um método de conseguir compreender quando os registos foram efetuados, isto é, a que data e hora o formulário tinha sido submetido, a fim de conseguir colmatar possíveis registos duplicados por falhas na rede ou erros humanos. Assim sendo, consideramos as necessidades do mesmo e implementamos um campo adicional oculto que regista a data, hora e fuso horário do respetivo sistema, de modo a ser possível realizar uma rastreabilidade das inserções realizadas no dia-a-dia.

Concernente ao estudo dos dados, foi possível averiguar junto do responsável clínico que os resultados obtidos poderiam ter em influência de dados não balanceados. Deste modo, por motivos de pouca aplicação da toxina botulínica em determinados músculos, o modelo compreende que como são poucas as aplicações nos músculos. Estas representam uma importância significativa uma vez que automaticamente caso o paciente não receba uma dose nesse músculo fica excluído de um grupo restrito de aplicações. Foi indicado que certas variáveis, como Supraspinatus, Infraspinus, Supinator, Fibularis longus e Fibularis brevis, são alguns casos cuja aplicação é bastante reduzida. No caso da Supraspinatus contamos com 29 aplicações, contudo o modelo considerou ser uma variável de extrema preponderância uma vez que auxilia bastante na explicação da variância dos dados e na caracterização de um registo.

Ressalva-se que, na continuação deste trabalho, tendo em vista a continuidade do mesmo para trabalhos futuros, deverá ser considerado o balanceamento dos dados, com o intuito de perceber quais as variáveis que melhor explicam a variância dos dados, não tendo variáveis mais preponderantes devido à sua raridade.

Um agradecimento especial ao CMRA, em particular à diretora de unidade, ao responsável clínico pela prática clínica inovadora e ao diretor de informática, por toda a disponibilidade nos pareceres, opiniões e sugestões.

5.6 Abrangência

Relativamente à abrangência das cadeiras lecionadas na licenciatura, é possível indicar de forma sucinta que todas elas desenvolvem uma importância na realização do TFC. Porém, neste caso, é necessário realçar que certas cadeiras lecionadas como Programação Web e Data Mining tiveram um grande peso na concretização deste projeto.

Mais concretamente, na primeira fase, foi aplicado HTML e CSS sendo estes dois temas fulcrais da programação web, lecionada pelos docentes Lúcio Studer e Rui Santos. Posteriormente, na segunda fase, foi aplicada PCA e correlação de variáveis, bem como outras técnicas de análise exploratória de dados e *data wrangling*, técnicas estas que foram lecionadas na cadeira de *Data Mining* este ano curricular, pela docente e coordenadora deste TFC, a professora Iolanda Velho.

Complementarmente, tal como indicado anteriormente, todas as cadeiras trouxeram uma componente importante para a realização deste TFC, sendo de ressaltar que todos/as os/as docentes comportaram um papel importante e os seus ensinamentos acabam por transparecer na solução e redação do TFC.

6 Método e Planeamento

6.1 Planeamento inicial

O cronograma da ordem de trabalhos é desenhado utilizando ferramentas de metodologia Agile. Neste caso, a ferramenta de planeamento de projeto conhecida como Diagrama de Gantt (*Gantt Chart*) transpõe as expectativas iniciais, bem como, as previsões temporais da duração da realização do projeto, atendendo a fatores externos como, pausas letivas, épocas de exames, trabalhos de cadeiras paralelas, entre demais.

Deste modo, é primeiramente necessário realçar as principais tarefas (*Figura 28*).

Na realização da **1ª fase** (formulário):

- Pesquisa Bibliográfica
- Criação de um formulário digital
- Recolha de dados
- Preparação do ambiente de desenvolvimento

Na realização da **2ª fase** (análise de dados):

- Caracterização dos dados
- Preparação e Limpeza dos dados
- Análise exploratória
- Aplicação de algoritmos de ML
- Conclusão / Sugestões de melhoria de procedimentos

Work Breakdown Structure

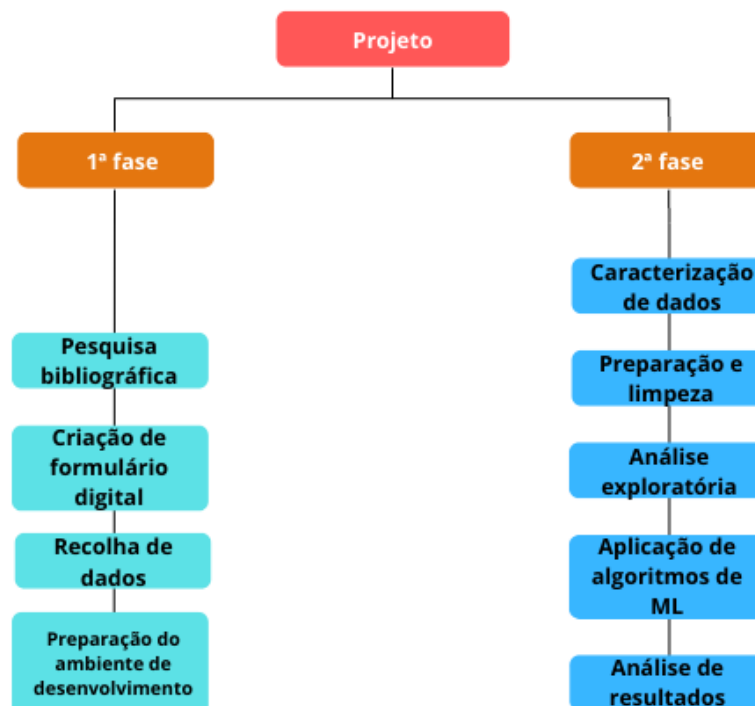


Figura 28 - Work Breakdown Structure

Além dos pontos principais de cada fase, é pertinente salientar que semanalmente, foi realizada uma reunião com as Orientadoras, Iolanda Velho e Maria Almeida Silva, a fim de transmitir o ponto de situação do trabalho final de curso e realizar qualquer esclarecimento de dúvidas.

No diagrama de *Gantt* (*Figura 29*) encontra-se o diagrama final da execução das tarefas durante o desenvolvimento do TFC, é também possível verificar que se encontram *Milestones* com as entregas a realizar de cada relatório (*Tabela 10 – Anexos*).

6.2 Análise crítica ao planeamento

Averigua-se que o sucesso do desenvolvimento deste estudo contou com um planeamento enquadrado com a disponibilidade e adversidades conhecidas, bem como margem de erro devidamente planeada em caso de necessidade de mudança no plano originalmente definido. Ao longo de todo o trabalho realizado, o feedback recebido por parte da orientadora e coorientadora, foi que todo o trabalho apresenta excelência e qualidade e que todos os prazos foram cumpridos dentro do cronograma estabelecido.

O sucesso para qualquer projeto advém de uma boa comunicação entre todas as partes interessadas, e uma gestão de tempo eficiente por parte dos desenvolvedores. Todo este sucesso para cumprimento do planeamento e sucesso advém de bastante trabalho árduo em esclarecer dúvidas aquando existentes e trabalho extracurricular.

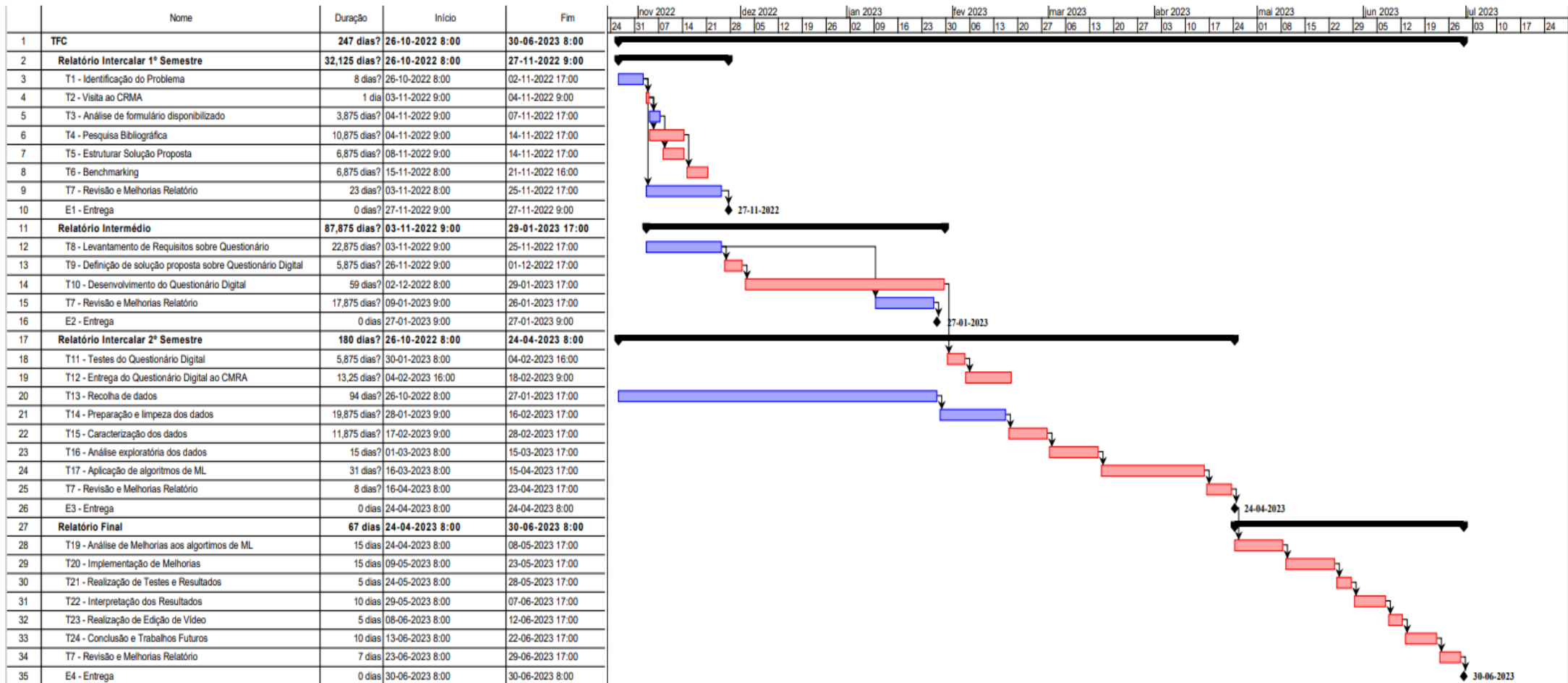


Figura 29 - Diagrama de Gantt Semanal (Realizado através do software Project Libre)

Bibliografia

- [1] American Heart Association, “Spasticity | American Stroke Association,” [Online]. Available: <https://www.stroke.org/en/about-stroke/effects-of-stroke/physical-effects-of-stroke/physical-impact/spasticity>. [Acedido em 09 Novembro 2022].
- [2] American Heart Association, “Let's Talk About Spasticity After Stroke,” [Online]. Available: https://www.stroke.org/-/media/stroke-files/lets-talk-about-stroke/life-after-stroke/lta_s_spasticity_english_0419.pdf. [Acedido em 09 Novembro 2022].
- [3] World Stroke Organization, “Global Stroke Fact Sheet 2022,” [Online]. Available: https://www.world-stroke.org/assets/downloads/WSO_Global_Stroke_Fact_Sheet.pdf. [Acedido em 11 Novembro 2022].
- [4] World Stroke Organization, “Learn about stroke,” [Online]. Available: <https://www.world-stroke.org/world-stroke-day-campaign/why-stroke-matters/learn-about-stroke>. [Acedido em 11 Novembro 2022].
- [5] Centro de Reabilitação de Alcoitão, “História - Centro de Reabilitação de Alcoitão,” [Online]. Available: <http://cmra.pt/centro/historia>. [Acedido em 30 Outubro 2022].
- [6] Centro de Reabilitação de Alcoitão, “Organização - Centro de Reabilitação de Alcoitão,” [Online]. Available: <http://cmra.pt/centro/organizacao>. [Acedido em 30 Outubro 2022].
- [7] CHU de São João, “Espasticidade,” [Online]. Available: <https://portal-chsj.min-saude.pt/pages/998>. [Acedido em 04 Novembro 2022].
- [8] A. Khan, A. Hazart, O. Galarraga, S. Garcia-Salicetti e V. Vigneron, “Treatment Outcome Prediction Using Multi-Task Learning: Application to Botulinum Toxin in Gait Rehabilitation,” *Sensors*, vol. 22, n.º 21, p. 8452, 03 Novembro 2022.
- [9] Y. Zhang e Y. Ma, “Application of supervised machine learning algorithms in the classification of sagittal gait patterns of cerebral palsy children with spastic diplegia,” *Computers in biology and medicine*, vol. 106, p. 33–39, Março 2019.
- [10] A. Suppa, G. S. Francesco Ascì, L. Marsili, D. Casali, Z. Zarezadeh, G. Ruoppolo, A. Berardelli e G. Costantini, “Voice analysis in adductor spasmodic dysphonia: Objective diagnosis and response to botulinum toxin,” *Parkinsonism & Related Disorders*, vol. 73, pp. 23-30, Abril 2020.
- [11] M. Raluy-Callado, A. Cox, S. MacLachlan, A. M. Bakheit, A. P. Moore, J. Dinét e S. Gabriel, “A retrospective study to assess resource utilization and costs in patients with post-stroke

spasticity in the United Kingdom,” *Current Medical Research and Opinion*, vol. 34, nº 7, pp. 1317-1324, 29 Março 2018.

- [12] D. D. V. Richa Jain, “Data Mining Algorithms in Healthcare: An,” em *Fifth International Conference on I-SMAC*, Punjab, India, 2021.
- [13] S. C. Pandey, “Data Mining Techniques for Medical Data: A Review,” em *International conference on Signal Processing, Communication, Power and Embedded System*, Paralakhemundi, Odisha, India., 2016.
- [14] F.-C. Hsu, S. B. Kritchevsky, Y. Liu, A. Kanaya, A. B. Newman, S. E. Perry, M. Visser, M. Pahor, T. B. Harris e B. J. Nicklas, “Association Between Inflammatory Components and Physical Function in the Health, Aging, and Body Composition Study: A Principal Component Analysis Approach,” 19 Fevereiro 2009.
- [15] R. J. Martis, U. R. Acharya, K.M.Mandana, A.K.Raya e C. Chakraborty, “Application of principal component analysis to ECG signals for automated diagnosis of cardiac health,” 5 Maio 2012.
- [16] O. Tadevosyan-leyfer, M. Dowd, R. Mankoski, B. Winklosky, S. Putnam, L. McGrath, H. Tager-flusberg e S. E. Folstein, “Journal of the American Academy of Child & Adolescent Psychiatry,” *A Principal Components Analysis of the Autism Diagnostic Interview-Revised*, pp. 864-872, Julho 2003.
- [17] C. Rasetti-Escargueil, E. Lemichez e M. Popoff, “Variability of Botulinum Toxins: Challenges and Opportunities for the Future,” *Toxins*, vol. 10, nº 9, p. 374, Setembro 2018.
- [18] B. Jabbari, “History of Botulinum Toxin Treatment in Movement Disorders,” *Tremor and Other Hyperkinetic Movements*, Novembro 2016.
- [19] Z. Jaadi, “A Step-by-Step Explanation of Principal Component Analysis (PCA),” 26 Setembro 2022. [Online]. Available: <https://builtin.com/data-science/step-step-explanation-principal-component-analysis>. [Acedido em 01 Janeiro 2023].
- [20] R. M, “MLmuse: Correlation and Collinearity — How they can make or break a model,” 15 Julho 2019. [Online]. Available: <https://blog.clairvoyantsoft.com/correlation-and-collinearity-how-they-can-make-or-break-a-model-9135fbe6936a>. [Acedido em 31 Dezembro 2022].
- [21] M. Gudesblatt, M. Zaffaroni, V. Stevenson, F. Béthoux, C. Tornatore, A.-M. Thomas, R. Plunkett, S. Sadiq, A. Erwin e S. Koelbel, “Intrathecal baclofen in multiple sclerosis: Too little, too late?,” *Multiple sclerosis*, vol. 17, pp. 623-629, 2011.
- [22] Z. Liu, C. Heffernan e J. Tana, “Caregiver burden: A concept analysis,” *International journal of nursing sciences*, vol. 7, nº 4, pp. 438-445, 25 Julho 2020.

- [23] P. Vasconcellos, “Como selecionar as melhores features para seu modelo de Machine Learning,” [Online]. Available: <https://paulovasconcellos.com.br/como-selecionar-as-melhores-features-para-seu-modelo-de-machine-learning-2e9df83d062a>. [Acedido em 16 Novembro 2022].
- [24] J. L. Marks, “What Is Spasticity? Symptoms, Causes, Diagnosis, and Treatment,” [Online]. Available: <https://www.everydayhealth.com/spasticity/guide/>. [Acedido em 11 Novembro 2022].
- [25] Johns Hopkins Medicine, “Spasticity,” [Online]. Available: <https://www.hopkinsmedicine.org/health/conditions-and-diseases/spasticity>. [Acedido em 11 Novembro 2022].
- [26] IPSEN Brasil, “Blog / Impacto da espasticidade em pacientes pós-AVC,” [Online]. Available: <https://caminhosposavc.com.br/impacto-da-espasticidade-em-pacientes-pos-avc/>. [Acedido em 11 Novembro 2022].
- [27] J. Brownlee, “How to Choose a Feature Selection Method For Machine Learning,” 27 Novembro 2019. [Online]. Available: <https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/>. [Acedido em 16 Novembro 2022].
- [28] Centro de Reabilitação de Alcoitão, “O Tratamento da Espasticidade com Toxina Botulínica,” [Online]. Available: <http://cmra.pt/o-tratamento-da-espasticidade-com-toxina-botulinica/>. [Acedido em 16 Novembro 2022].
- [29] British Society of Rehabilitation Medicine, “Botulinum Toxin cover,” Janeiro 2009. [Online]. Available: <https://www.bsrm.org.uk/downloads/spasticity-in-adults-management-botulinum-toxin.pdf>. [Acedido em 17 Novembro 2022].
- [30] E. Domingues, “ML — Feature Selection,” 11 Julho 2020. [Online]. Available: <https://pt.linkedin.com/pulse/ml-feature-selection-eduardo-domingues>. [Acedido em 19 Novembro 2022].
- [31] B. Dykes, *How to Drive Change with Data, Narrative and Visuals*, 1ª ed., Wiley, 2019.
- [32] D. R. N. d. Oliveira, I. M. B. Paiva e R. F. Anomal, “O uso da toxina botulínica no tratamento da espasticidade após acidente vascular encefálico,” *Revista Pesquisa em Fisioterapia*, vol. 7, nº 2, pp. 289-297, 2017.
- [33] H. Zeng, J. Chen, Y. Guo e S. Tan, “Prevalence and Risk Factors for Spasticity After Stroke: A Systematic Review and Meta-Analysis,” *Frontiers in neurology*, vol. 11, 2021.
- [34] P. S. S. a. V. K. Jeffrey S Shilt, “Optimal management for people with severe spasticity,” *Degenerative neurological and neuromuscular disease*, vol. 2, pp. 133-140, 2012.

- [35] M. J. Leathley, J. M. Gregson, A. P. Moore, T. L. Smith, A. K. Sharma e C. L. Watkins, "Predicting spasticity after stroke in those surviving to 12 months," *Clinical Rehabilitation*, vol. 18, nº 4, pp. 438-443, 2004.

Anexos

Tabela 5 - Exemplos de ferramentas de avaliação de pacientes no mercado

Escala de Ashworth Modificada [21]	Originalmente desenvolvida para pacientes de esclerose múltipla, em 1987 foi modificada adicionando a avaliação "+1" que demonstrou ser uma boa representação da espasticidade nos cotovelos. Atualmente é a mais utilizada em estudos da aplicação da toxina nos pacientes, contudo a sua aplicação nos restantes membros é bastante questionável.
Escala Visual Analógica	Permite ao paciente atribuir um valor numérico, por exemplo, à dor que sente a realizar uma determinada tarefa.
Caregiver Burden Scale [22]	Criada em 1980, foi a primeira escala desenvolvida para aferir o efeito dos cuidados dispostos pelos cuidadores. Atualmente existem versões "lite" como as ZBI-22, ZBI-12 e ZBI-7, que deu origem a outras novas escalas como CSI, CBI, entre outras.
Escala de frequência de espasmos	Uma escala de 0-4 que permite perceber a frequência dos espasmos do paciente.
Tempos de marcha	Tempo, nº de passos e velocidade que o paciente demora a percorrer 10 metros.
Tempos de reação	Qual a reação associada ao fim de quanto tempo, passos e velocidade.

Tabela 6 - Tabela de requisitos

COD	Descrição
RQ1	Permitir a indicação da data do primeiro AVC
RQ2	Utilização de <i>sliders</i> aquando for indicador de escala das dores
RQ3	Não implementar a avaliação reação associada (Obsoleta)
RQ4	Não implementar a avaliação <i>Caregiver</i> (Obsoleta)
RQ5	Na seleção da toxina, apenas se pode selecionar 1 por consulta.
RQ6	Na indicação da localização, pode-se realizar escolha múltipla.
RQ7	Alterar o todos os nomes “Ultrasom” para “ECO”
RQ8	Não implementar a avaliação descrição da marcha.
RQ9	Introduzir uma nova escala de avaliação que atendesse a amplitude articular passiva e ativa do paciente. A medida de avaliação encontra-se em percentagem.
RQ10	A data de inserção da ficha deve ser realizada pelo sistema.
RQ11	Introduzir a avaliação da força muscular após a realização da avaliação na escala de Ashworth. A mesma tem de comportar um <i>score</i> de 0 a 5 e fazendo recurso a <i>sliders</i> .
RQ12	No fim do formulário o mesmo deve ter uma componente de observações.
RQ13	Na análise dos músculos superiores, adicionar: <ul style="list-style-type: none"> • Grande redondo • Interósseos dorsais • Lumbricoides
RQ14	Na análise aos músculos inferiores, adicionar: <ul style="list-style-type: none"> • Adutor magnum • Adutor longo • Adutor breve • Gracilis
RQ15	Na análise aos músculos inferiores, deve-se também remover adutor da coxa.
RQ16	Alterar o nome “Flexor profundo dos dedos” para “Flexor longo dos dedos”
RQ17	Alterar o nome “Curto flexor dos dedos” para “Flexor curto dos dedos”
RQ18	Alterar o nome “Gas Score” para “Gas T-Score”
RQ19	Incorporar as informações do paciente: <ul style="list-style-type: none"> • Nº Processo

	<ul style="list-style-type: none">• Nome• Sexo Biológico• Diagnóstico
RQ20	É obrigatória a introdução dos seguintes parâmetros: <ul style="list-style-type: none">• Avaliação em escalada de Ashworth• Avaliação na escala de dor• Avaliação da FAC• Informações sobre o paciente• Informação sobre qual a toxina aplicada e qual a quantidade• Indicação do método guia da amplitude passiva e amplitude ativa• Avaliação da força muscular

Tabela 7 – Requisitos complementares e pedidos de alterações/modificações

COD	Descrição
RQC1	Alteração de campos fechados para campos abertos.
RQC2	Implementação de somatórios aquando da inserção da quantidade de toxina aplicada em cada músculo do paciente.
RQC3	Remoção da imagem representativa da escala de dor.
RQC4	Implementação do cálculo da cadência (passos por minuto) do paciente.
RQC5	Implementação do cálculo da velocidade por segundo.
RQC6	Introdução de campos apenas presentes no ficheiro Excel, não implementados no formulário em papel anteriormente.
RQC7	Alteração das cores e <i>layout</i> .

Tabela 8 – Variáveis removidas devido à elevada multicolinearidade

Variável	Motivo
ID	Indiferente para o estudo uma vez que o ID apenas representa o número atribuído ao paciente.
% Dmax Dysport	Variante da variável <i>Dysport</i>
% Dmax Xeomin	Variante da variável <i>Xeomin</i>
Xeomin dose	Variante da variável <i>Xeomin</i>
% Dmax Botox	Variante da variável <i>Botox</i>
Botox dose	Variante da variável <i>Botox</i>
Stroke-First BoNTA interval	Cálculo diferencial entre <i>First BoNTA administration – ever</i> e <i>Stroke Date</i>
Stroke-BoNTA interval	Cálculo diferencial entre <i>BoNTA treat. Date</i> e <i>Stroke Date</i>
BoNTA treat. Date	Variante da <i>Sheet Date</i>
First BoNTA administration - ever	Apresenta a data da primeira administração que pode ser percebida realizando uma pesquisa pelo ID do paciente.

Tabela 9 – Loadings mais impactantes por PC (PC2- PC20)

Most impactful variables for PC2:		Most impactful variables for PC7:		Most impactful variables for PC12:	
Lower limbs	0.772141	D1-Sintomas/Défices	0.444051	Supinator	0.434291
LL Dose	0.764351	Fibularis Longus	0.420966	Flexor pollicis longus	0.389913
Gastrocnemius medialis	0.631735	Principal goal Subcategory	0.411575	Fibularis Longus	0.357160
Gastrocnemius lateralis	0.609075	Fibularis brevis	0.397104	Fibularis brevis	0.321894
UL Dose	0.598479	Adductor pollicis	0.358861	Digitorum extensor	0.310478
Soleus	0.536159	Age	0.333024	Flexor digitorum profundus	0.264025
Name: PC2, dtype: float64		Name: PC7, dtype: float64		Name: PC12, dtype: float64	
Most impactful variables for PC3:		Most impactful variables for PC8:		Most impactful variables for PC13:	
Impairment	0.758051	Fibularis brevis	0.543115	Others.1	0.347934
Treated side	0.726646	Fibularis Longus	0.543054	Age	0.347393
Diagnosis	0.716111	D1-Sintomas/Défices	0.370885	Gender	0.318732
Flexor hallucis longus	0.432471	Principal goal Subcategory	0.351525	Vastus lateralis	0.307011
Pectoralis major	0.299295	Systemic medication	0.310965	Rhomboideus	0.296639
Flexor digitorum brevis	0.297461	Ultrasound	0.309452	Age at Stroke	0.284168
Name: PC3, dtype: float64		Name: PC8, dtype: float64		Name: PC13, dtype: float64	
Most impactful variables for PC4:		Most impactful variables for PC9:		Most impactful variables for PC14:	
Age at Stroke	0.618868	Flexor digitorum brevis	0.446078	Neurostimulation	0.367673
Age	0.578326	Others	0.421819	Ultrasound	0.347709
Treated side	0.484793	Upper limbs	0.372034	Principal goal Subcategory	0.315471
Impairment	0.473803	Upper+Lower limbs	0.356928	D1-Sintomas/Défices	0.303093
Diagnosis	0.416073	Dysport	0.349229	Extensor hallucis longus	0.297927
Neurostimulation	0.405653	Principal goal Subcategory	0.312625	Interossei dorsales	0.275278
Name: PC4, dtype: float64		Name: PC9, dtype: float64		Name: PC14, dtype: float64	
Most impactful variables for PC5:		Most impactful variables for PC10:		Most impactful variables for PC15:	
Semimembranosus	0.586461	Deltoideus	0.459440	Primary goal outcome score	0.388791
Semitendinosus	0.536391	Supraspinatus	0.420184	Interossei dorsales	0.302288
Stroke date	0.439971	Flexor carpi ulnaris	0.397340	Digitorum extensor	0.293760
Ultrasound	0.417167	Flexor carpi radialis	0.365823	Vastus internus	0.287443
Sheet date	0.396987	Trapezius	0.345203	Biceps brachii	0.278976
Gender	0.366707	Sheet date	0.322184	Brachioradialis	0.259664
Name: PC5, dtype: float64		Name: PC10, dtype: float64		Name: PC15, dtype: float64	
Most impactful variables for PC6:		Most impactful variables for PC11:		Most impactful variables for PC16:	
Fibularis Longus	0.522932	Flexor hallucis longus	0.401299	Carpi extensor	0.370699
Fibularis brevis	0.495824	Vastus internus	0.340605	Digitorum extensor	0.297955
Sheet date	0.453930	Rectus anterior	0.321739	Vastus lateralis	0.296514
Brachialis	0.391091	Vastus lateralis	0.292513	Pronator quadratus	0.277511
Flexor carpi radialis	0.339063	Others.1	0.282076	Vastus internus	0.262110
Ultrasound	0.315027	Upper+Lower limbs	0.274879	EMG	0.247662
Name: PC6, dtype: float64		Name: PC11, dtype: float64		Name: PC16, dtype: float64	
		Most impactful variables for PC17:			
		Trapezius	0.337154		
		Digitorum extensor	0.328349		
		Supraspinatus	0.259854		
		Flexor carpi ulnaris	0.258905		
		Supinator	0.236966		
		Rectus anterior	0.235580		
		Name: PC17, dtype: float64			
		Most impactful variables for PC18:			
		Psoas iliacus	0.414813		
		Teres major	0.347984		
		Biceps cruris	0.326553		
		Supinator	0.296057		
		Digitorum extensor	0.280409		
		Primary goal outcome score	0.269785		
		Name: PC18, dtype: float64			
		Most impactful variables for PC19:			
		Intercorrências	0.441772		
		Occupational therapy	0.366455		
		Palpation	0.305399		
		Discharge	0.285480		
		Physiotherapy	0.284694		
		Others.1	0.237226		
		Name: PC19, dtype: float64			
		Most impactful variables for PC20:			
		Latissimus dorsi	0.377943		
		Digitorum extensor	0.261544		
		Supinator	0.237091		
		Flexor digitorum brevis	0.223674		
		Interossei dorsales	0.214379		
		Discharge	0.213228		
		Name: PC20, dtype: float64			

Tabela 10 - Tabela Milestones e Tarefas

COD	Tarefa
T1	Primeira reunião com as orientadoras do TFC e introdução do problema que o CMRA carece.
T2	Primeira visita ao CMRA.
T3	Receção do formulário em formato papel e análise do mesmo.
T4	Realização de pesquisa bibliográfica a fim de perceber as práticas clínicas realizadas em outros centros de reabilitação. Análise de práticas de ML.
T5	Análise e Estruturação da proposta de solução.
T6	Comparação das práticas do centro com as restantes práticas nos outros centros. Comparação da proposta de solução com outras ofertas no mercado.
T7	Revisão e melhorias no relatório.
T8	Levantamento de requisitos relativamente ao Questionário Digital.
T9	Definição de solução proposta sobre Questionário Digital.
T10	Desenvolvimento do Questionário Digital com o uso de HTML, PHP e CSV.
T11	Realização de testes no Questionário Digital, a fim de assegurar qualidade e fiabilidade.
T12	Entrega do Questionário Digital junto do CMRA e feedback do centro.
T13	Recolha dos dados para estudo sobre variáveis do questionário.
T14	Preparação e Limpeza dos dados.
T15	Caracterização dos dados.
T16	Análise exploratória dos dados.
T17	Aplicação de algoritmos de ML.
T19	Análise de possíveis melhorias para os algoritmos de ML.
T20	Implementação de Melhorias (Afinação de modelos).
T21	Realização de Testes e Resultados.
T22	Interpretação de Resultados.

T23	Realização de Edição de Vídeo.
T24	Conclusão e Trabalhos Futuros.
	<i>Milestones</i>
E1	Entrega do relatório intercalar do 1º Semestre (27/10/2022)
E2	Entrega do relatório intermédio (27/01/2023)
E3	Entrega do relatório intercalar do 2º Semestre (24/04/2023)
E4	Entrega do relatório final (30/06/2023)

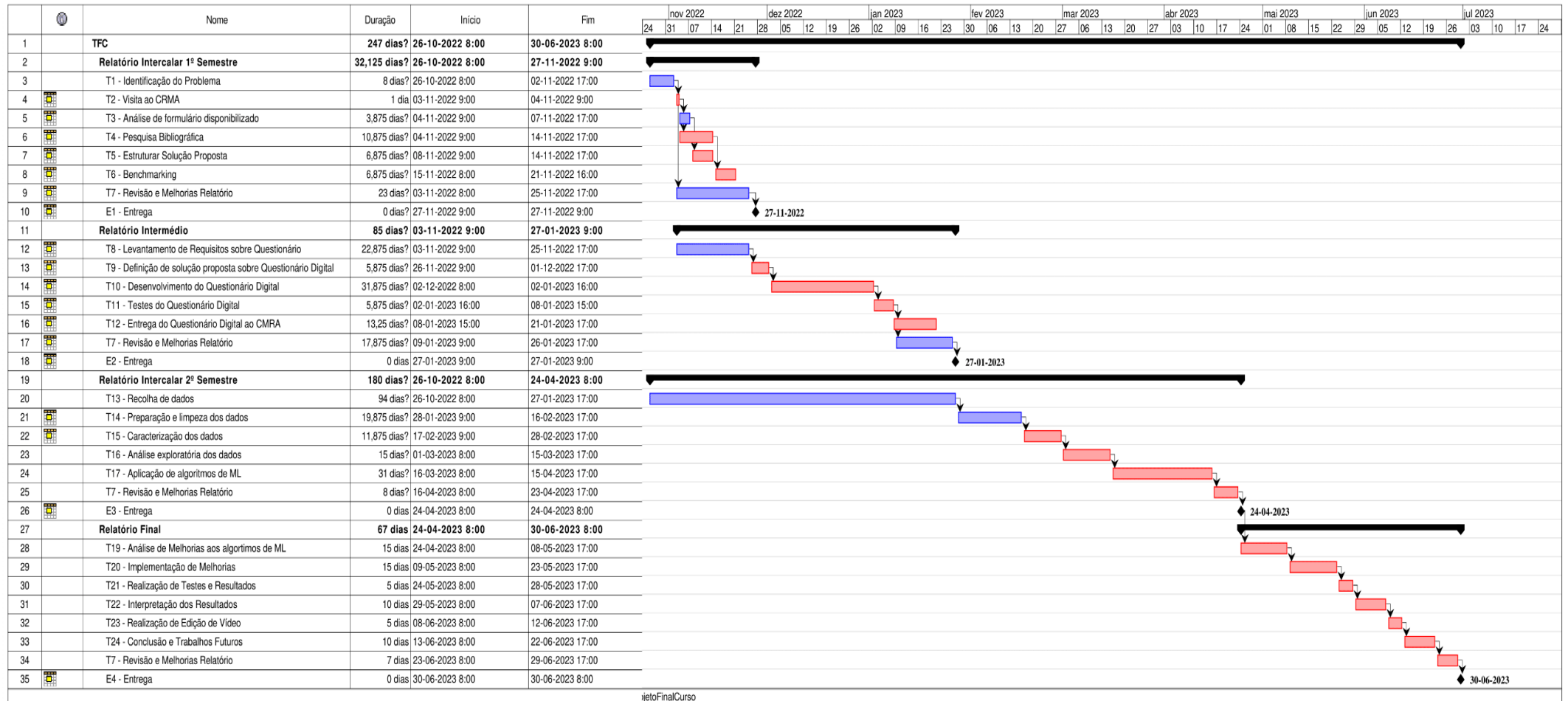


Figura 30 - Diagrama de Gantt anterior semanal (Realizado através do software Project Libre)

Aplicação de Inteligência Artificial para estudo de prática clínica inovadora

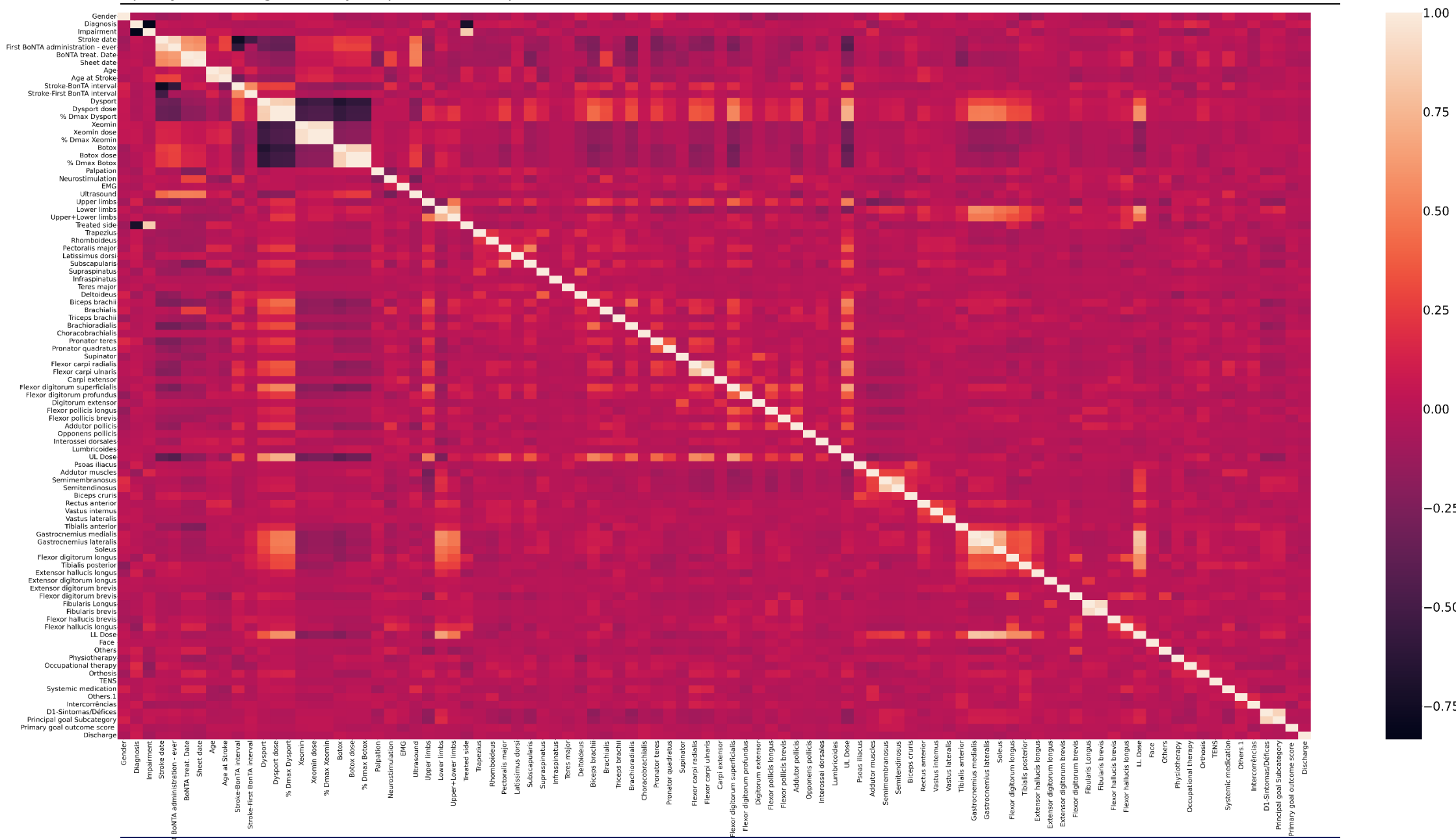


Figura 31 - Mapa de correlação entre as variáveis originais

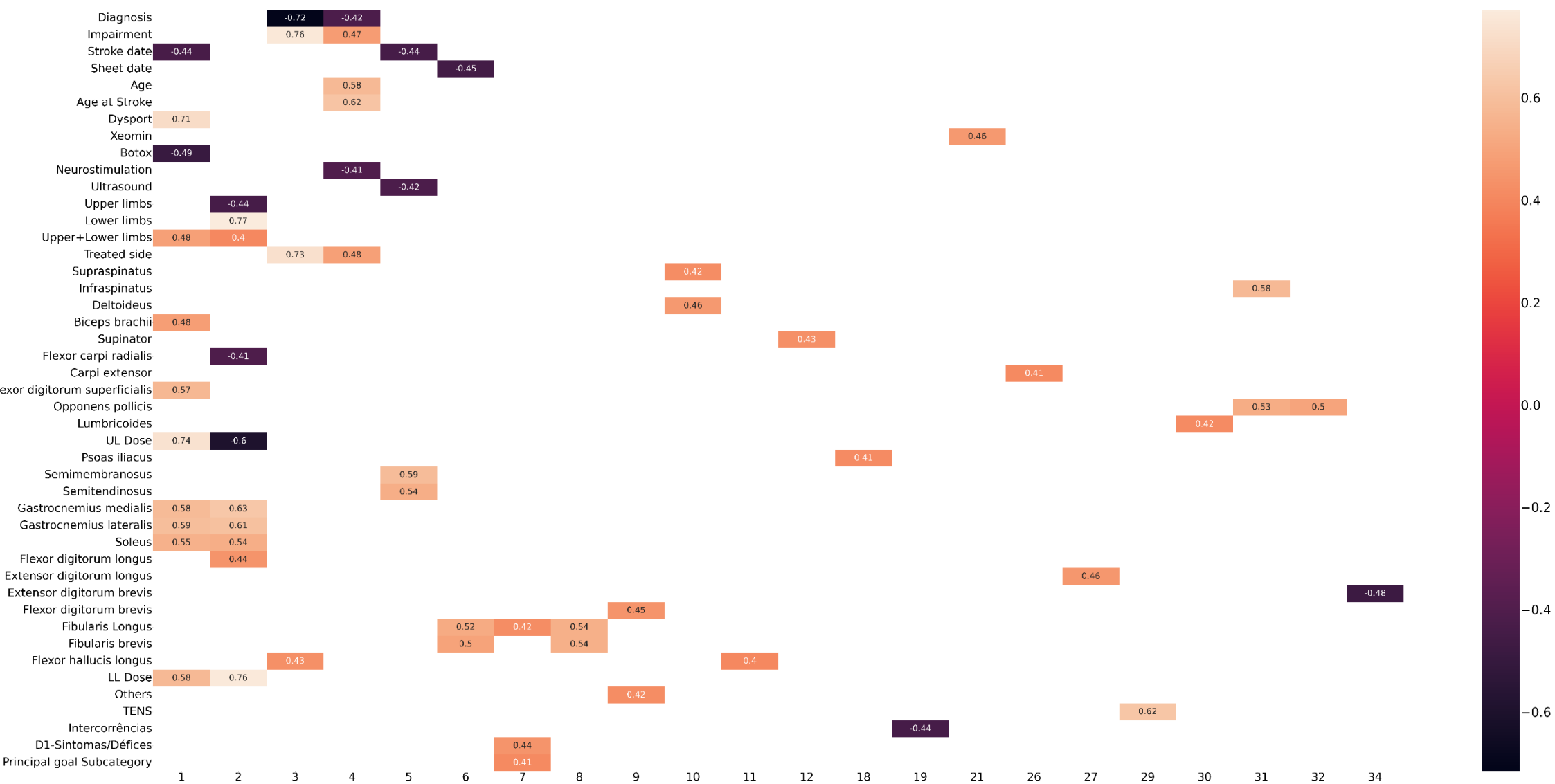


Figura 33 – Mapa de correlação das variáveis originais com as componentes principais (correlação superior a 0.4 em valor absoluto)

Tabela 11 – “Novo paciente” criado com base na média de valores do modelo PCA

Variável	Motivo
<i>Gender</i>	Male
<i>Diagnosis</i>	LH isch stroke
<i>Impairment</i>	Right hemiparesis
<i>Stroke date</i>	2011-05-12
<i>Sheet date</i>	2016-08-12
<i>Age</i>	58
<i>Age at Stroke</i>	52
<i>Dysport</i>	Yes
<i>Xeomin</i>	No
<i>Botox</i>	No
<i>Palpation</i>	Yes
<i>Neurostimulation</i>	No
<i>EMG</i>	No
<i>Ultrasound</i>	No
<i>Upper limbs</i>	Yes
<i>Lower limbs</i>	Yes
<i>Upper+Lower limbs</i>	Yes
<i>Treated side</i>	Right
<i>Trapezius</i>	0
<i>Rhomboideus</i>	0
<i>Pectoralis major</i>	30
<i>Latissimus dorsi</i>	0
<i>Subscapularis</i>	40
<i>Supraspinatus</i>	0
<i>Infraspinatus</i>	0

<i>Teres major</i>	0
<i>Deltoideus</i>	20
<i>Biceps brachii</i>	60
<i>Brachialis</i>	50
<i>Triceps brachii</i>	20
<i>Brachioradialis</i>	30
<i>Choracobrachialis</i>	0
<i>Pronator teres</i>	30
<i>Pronator quadratus</i>	0
<i>Supinator</i>	0
<i>Flexor carpi radialis</i>	40
<i>Flexor carpi ulnaris</i>	30
<i>Carpi extensor</i>	0
<i>Flexor digitorum superficialis</i>	70
<i>Flexor digitorum profundus</i>	40
<i>Digitorum extensor</i>	0
<i>Flexor pollicis longus</i>	20
<i>Flexor pollicis brevis</i>	0
<i>Adductor pollicis</i>	20
<i>Opponens pollicis</i>	0
<i>Interossei dorsales</i>	20
<i>Lumbricoides</i>	0
<i>UL Dose</i>	520
<i>Psoas iliacus</i>	0
Adductor muscles	9
Semimembranosus	25

Semitendinosus	11
Biceps cruris	0
Rectus anterior	20
Vastus internus	1
Vastus lateralis	0
Tibialis anterior	20
Gastrocnemius medialis	70
Gastrocnemius lateralis	60
Soleus	60
Flexor digitorum longus	50
Tibialis posterior	40
Extensor hallucis longus	20
<i>Extensor digitorum longus</i>	0
<i>Extensor digitorum brevis</i>	0
<i>Flexor digitorum brevis</i>	8
<i>Fibularis Longus</i>	1
<i>Fibularis brevis</i>	0
<i>Flexor hallucis brevis</i>	0
<i>Flexor hallucis longus</i>	13
<i>LL Dose</i>	408
<i>Face</i>	0
<i>Others</i>	0
<i>Physiotherapy</i>	Yes
<i>Occupational therapy</i>	No
<i>Orthosis</i>	Yes
<i>TENS</i>	No

<i>Systemic medication</i>	No
<i>Others.1</i>	No
<i>Intercorrências</i>	No
<i>D1-Sintomas/Défices</i>	D1-Sintomas/Défices
<i>Principal goal Subcategory</i>	D2- Facilitating therapy
<i>Primary goal outcome score</i>	0
<i>Discharge</i>	No

Lista de Acrónimos

- CMRA Centro de Medicina e Reabilitação de Alcoitão – 50 Anos de Excelência em Reabilitação. O CMRA pertence à Santa Casa da Misericórdia de Lisboa.
- LIG Licenciatura em Informática de Gestão
- TFC Trabalho Final de Curso
- ULHT Universidade Lusófona de Humanidades e Tecnologia
- HTML *HyperText Markup Language* – Linguagem de estruturação de websites e do seu conteúdo.
- PHP *Hypertext Preprocessor* – Linguagem de programação utilizada para desenvolver aplicações para a web e criação de websites dinâmicos, favorecendo a conexão entre servidores e cliente.
- ML *Machine Learning* – Método de análise de dados através da construção de modelos analíticos.
- FS *Feature Selection* – Processo de seleção de variáveis relevantes para o modelo de *Machine Learning*.
- PCA *Principal Component Analysis* – Técnica de redução da dimensionalidade com base na construção de componentes principais que explicam uma elevada percentagem da variabilidade dos dados. Essa técnica é comumente utilizada em análise de dados, reconhecimento de padrões e aprendizagem de máquina.
- PC *Principal Component* – Resultado obtido da aplicação de PCA, constituindo combinações lineares dos atributos originais, que capturam a maior parte da variação presente nos dados. As componentes principais são ordenadas em termos de importância, de modo que os primeiros componentes principais representam as características mais significativas dos dados.
- K-Means – Algoritmo de agrupamento utilizado para resolver problemas de classificação em modelos não supervisionados.
- KNN *K Nearest Neighbors* – Algoritmo de Machine Learning utilizado para resolver problemas de classificação.
- SVM *Support Vector Machines* – Conjunto de técnicas de modelos supervisionados usados para classificação, regressão e deteção de *outliers*.