



UNIVERSIDADE
LUSÓFONA

Previsão de Tempos de Espera

Trabalho Final de Curso

Entrega Final

Ana Toscano, a22207260, Ciência de Dados

Nome do Orientador: Iolanda Velho

Nome do Coorientador: Sofia Naique

Entidade Externa: Centro de Medicina de Reabilitação de Alcoitão

Departamento de Engenharia Informática e Sistemas de Informação

Universidade Lusófona, Centro Universitário de Lisboa

27 de junho de 2025

www.ulusofona.pt

Direitos de cópia

Previsão de Tempos de Espera, Copyright de Ana Margarida Teixeira Toscano, Universidade Lusófona.

A Escola de Comunicação, Arquitectura, Artes e Tecnologias da Informação (ECATI) e a Universidade Lusófona (UL) 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 objectivos educacionais ou de investigação, não comerciais, desde que seja dado crédito ao autor e editor.

Este documento foi gerado com o processador (pdf/Xe/Lua)LaTeX e o modelo ULThesis (v1.0.0) [1].

Agradecimentos

Ao longo do desenvolvimento deste trabalho, tive sempre a sorte de contar com o apoio necessário nos momentos certos, nunca tendo esse apoio sido negado. Nesse sentido, gostaria de expressar o meu mais sincero agradecimento à Engenheira Mariana Matos, cuja constante disponibilidade, prontidão e atenção em todas as fases do projeto foram essenciais. A sua colaboração foi um contributo muito importante, facilitando de forma significativa a superação dos diversos desafios encontrados.

Agradeço igualmente ao CMRA, pela disponibilização dos dados utilizados neste estudo, os quais se revelaram absolutamente indispensáveis para a concretização do trabalho. Sem esse acesso, não teria sido possível desenvolver as análises e modelos que sustentam este projeto.

Dirijo ainda um agradecimento especial às professoras Iolanda Velho e Sofia Naique, pela dedicação, orientação e apoio contínuo demonstrados ao longo de todo o percurso. A sua disponibilidade constante para esclarecer dúvidas, bem como o acompanhamento, foram indispensáveis para a evolução e realização deste trabalho.

Resumo

Este trabalho é desenvolvido no âmbito da unidade curricular Trabalho Final de Curso (TFC) da Licenciatura de Ciência de Dados.

A Engenheira Mariana Matos, Diretora da Unidade de Gestão de Doentes do Centro de Medicina de Reabilitação de Alcoitão (CMRA), Santa Casa da Misericórdia de Lisboa, entrou em contato com a Professora Iolanda Velho, propondo a cooperação científica na base deste TFC.

O CMRA dedica-se à reabilitação de pessoas portadoras de deficiência de predomínio físico ou multideficiência congénita, de todas as idades. No ano de 2020 recebeu 665 utentes tratados em internamento, além de ter realizado 72 859 atos terapêuticos em ambulatório. Dada a amplitude do atendimento realizado pelo CMRA e a sua procura pela excelência em reabilitação, a instituição aposta no desenvolvimento tecnológico dos seus processos.

Através da utilização de técnicas de Machine Learning para apoiar a gestão hospitalar, este projeto tem como objetivos:

- realizar a análise de dados, cedidos pelo CMRA
- desenvolver um modelo preditivo, que permita antecipar o tempo de espera de doentes que chegam ao CMRA, para internamento, ou ambulatório
- no âmbito deste modelo preditivo, caracterizar o referido tempo de espera, em função da sua patologia, e área clínica do CMRA a que se refere.

Uma análise interessante, baseada nos mesmo dados, seria a caracterização por clusters (i.e., identificação de grupos de doentes por modelos de I.A. adequados) dos pacientes do CMRA. Esta caracterização pode abrir novas perspetivas na gestão hospitalar, constituindo um objetivo secundário neste trabalho, a ser abordado em função do tempo.

Palavras-chave: CMRA, Inteligência Artificial, Análise Exploratória de Dados, Modelos Preditivos.

Abstract

This work is developed within the scope of the curricular unit Final Course Work (TFC) of the Bachelor's Degree in Data Science.

Engineer Mariana Matos, Director of the Patient Management Unit at the Alcoitão Rehabilitation Medicine Center (CMRA), Santa Casa da Misericórdia de Lisboa, reached out to Professor Iolanda Velho, proposing the scientific collaboration underlying this TFC.

The CMRA is dedicated to the rehabilitation of individuals with predominantly physical disabilities or congenital multiple disabilities, serving all age groups. In 2020, it received 665 inpatients and conducted 72,859 outpatient therapeutic procedures. Given the broad scope of services provided by the CMRA and its pursuit of excellence in rehabilitation, the institution invests in the technological development of its processes.

Through the use of Machine Learning techniques to support hospital management, this project aims to:

- analyze data provided by the CMRA;
- develop a predictive model that allows for anticipating the waiting time for patients arriving at the CMRA, whether for inpatient care or not;
- within the scope of this predictive model, characterize the waiting time based on the patient's pathology and the clinical area of the CMRA involved.

An interesting analysis, based on the same data, would be to characterize CMRA patients by clusters (i.e. identifying groups of patients using appropriate A.I. models). This characterization could open up new perspectives in hospital management, and is a secondary objective of this work, to be addressed over time.

Keywords: CMRA, Artificial Intelligence, Exploratory Data Analysis, Predictive Models.

Conteúdo

Agradecimentos	2
Resumo	3
Abstract	4
Conteúdo	5
Lista de Figuras	8
Lista de Tabelas	15
1 - Introdução	16
2 - Pertinência e Viabilidade	18
2.1 Viabilidade	18
2.2 Pertinência	19
3 - Conceitos Fundamentais	20
3.1 Conceitos Teóricos	20
3.1.1 Machine Learning	20
3.1.2 Machine Learning na saúde	20
3.1.3 Tipos de Machine Learning	21
3.1.4 Preparação dos Dados	21
3.1.5 Métricas de Desempenho	21
3.1.6 Modelos a utilizar	23
3.1.7 Modelos para previsão de tempos de espera	23
3.1.8 Modelo para caracterização dos pacientes	28
4 - Estado da Arte	29

4.1	Artigos Analisados	29
5	- Ferramentas a Utilizar	34
6	- Solução Proposta	35
6.1	Dados a Utilizar	36
6.1.1	Recolha e Descrição dos dados	37
6.2	Pré-Processamento dos Dados	39
6.2.1	Análise Inicial dos dados	40
6.3	Análise Exploratória de Dados	41
6.3.1	Dados 1 ^{as} consultas	42
6.3.2	Dados 1 ^{as} consultas e sessões	46
6.3.3	Dados 1 ^{as} consultas e internamentos	49
6.3.4	Dados 1 ^{as} consultas e info de internamentos	53
6.3.5	Dados sessões e internamentos	58
6.3.6	Dados sessões e info internamentos	60
6.3.7	Dados internamento e info internamentos	64
6.3.8	Junção das 4 tabelas - 1 ^{as} consultas + internamento + sessões + info internamentos	67
6.3.9	Análise da variável Tempos de espera	72
6.4	Modelos e Algoritmos Escolhidos	76
6.5	Abrangência	76
7	- Método e Planeamento	78
7.1	Desafios	79
8	- Resultados e Discussão	80
8.1	Comparação de Modelos e Abordagens	80
8.2	Resultados das Análises	81
8.2.1	Análise dos resultados do melhor modelo de previsão de tempos de espera	81
8.2.2	Análise dos resultados do modelo de classificação de pacientes	82
8.3	Limitações da Análise	83

10 - Conclusão	85
10.1 Conclusão	85
10.2 Trabalhos Futuros	85
Bibliografia	86
Anexo	88
1.1 União das tabelas 1 ^a s consulta e Sessões	88
1.2 União das tabelas 1 ^a s consultas e info de internamentos	94
1.3 União das tabelas internamento e info internamentos	98
1.4 União das tabelas sessões e info internamentos	100
1.5 União das tabelas 1 ^a s consulta e internamentos	102
1.6 União das tabelas internamento e sessões	105
1.7 União das tabelas 1 ^a s consultas, internamento e sessões	107

Lista de Figuras

1	Centro de Medicina de Reabilitação de Alcoitão	16
2	Regressão Linear Simples	24
3	Regressão Linear Múltipla	24
4	Exemplo de uma árvore de decisão	25
5	Demonstração do funcionamento do algoritmo Random Forest	26
6	Exemplo de uma aplicação de SVM	27
7	Aplicação do algoritmo KNN	27
8	Resultado da aplicação do algoritmo K-Means	28
9	Fases propostas para a resolução do problema	36
10	Variáveis das quatro tabelas utilizadas	37
11	Transformação de valores em falta na tabela info internamentos	41
12	Distribuição do género dos utentes que atenderam a uma 1ª consulta	42
13	Distribuição dos utentes que compareceram a uma 1ª consulta, por concelho	43
14	Distribuição dos utentes que compareceram a uma 1ª consulta, por país de origem	43
15	Distribuição por serviço dos utentes que compareceram a uma 1ª consulta	44
16	Distribuição das entidades financeiras reponsáveis pelos utentes que compareceram a uma 1ª consulta	44
17	Distribuição das causas que deram origem aos utentes a comparecerem a uma 1ª consulta	45
18	Distribuição por tipologia dos utentes que compareceram a uma 1ª consulta	45
19	Distribuição por patologia dos utentes que compareceram a uma 1ª consulta	46
20	Distribuição dos profissionais responsáveis pelo atendimento dos utentes que compareceram a uma 1ª consulta	46
21	Distribuição das sessões por utentes que posteriormente à 1ª consulta realizaram sessões	48
22	Distribuição dos meios complementares terapêuticos dos utentes que posteriormente à 1ª consulta realizaram sessões	48

23	Distribuição dos profissionais responsáveis pelas sessões dos utentes que posteriormente à 1ª consulta realizaram sessões	49
24	Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por serviços (na 1ª consulta)	50
25	Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por serviços (no internamento)	51
26	Distribuição das entidades financeiras responsáveis pelas 1ª consulta dos utentes que posteriormente à 1ª consulta ficaram em internamento	51
27	Distribuição das entidades financeiras responsáveis pelos internamentos dos utentes que posteriormente à 1ª consulta ficaram em internamento	52
28	Distribuição dos profissionais responsáveis pelo atendimento dos utentes que posteriormente à 1ª consulta ficaram em internamento	52
29	Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento	53
30	Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por serviços (1ª consulta)	55
31	Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por serviços (info internamento)	55
32	Distribuição das entidades financeiras responsáveis pela 1ª consultas dos utentes que posteriormente à 1ª consulta ficaram em internamento	56
33	Distribuição das entidades financeiras responsáveis pelos internamentos dos utentes que posteriormente à 1ª consulta ficaram em internamento	56
34	Distribuição dos actos realizados pelos utentes que posteriormente à 1ª consulta ficaram em internamento	57
35	Distribuição dos diagnósticos obtidos pelos utentes que posteriormente à 1ª consulta ficaram em internamento	57
36	Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por patologia	58
37	Distribuição das idades dos utentes que posteriormente ao internamento realizaram sessões	59
38	Distribuição dos serviços em que são acompanhados os utentes que posteriormente ao internamento realizaram sessões	59
39	Distribuição dos utentes que posteriormente ao internamento realizaram sessões, por quantidade de sessões realizadas	60
40	Distribuição dos meios complementares terapêuticos dos utentes que posteriormente ao internamento realizaram sessões	60
41	Distribuição das sessões por utentes que realizaram sessões	62

42	Distribuição dos meios complementares terapêuticos realizados pelos utentes	62
43	Distribuição dos profissionais responsáveis pelas sessões realizadas pelos utentes	63
44	Distribuição dos profissionais responsáveis pelos utentes em internamento	63
45	Distribuição sobre o destino pós alta dos utentes	63
46	Distribuição da classificação da independência funcional no momento da admissão dos utentes	64
47	Distribuição da classificação da independência funcional no momento da alta dos utentes	64
48	Distribuição do destino pós alta dos utentes em internamento	66
49	Distribuição da classificação da independência funcional no momento da admissão dos utentes em internamento	66
50	Distribuição da classificação da independência funcional no momento da alta dos utentes em internamento	67
51	Evolução dos números de internamentos	68
52	Distribuição das durações dos internamentos	68
53	Variação das idades médias por mês	69
54	Distribuição do género por mês	69
55	Evolução do número de sessões realizadas por mês	70
56	Variação do número de diagnósticos realizados por mês	70
57	Variação no destino pós alta (lar + transf para outra unid hospitalar)	71
58	Variação no destino pós alta (domicilio + domicilio com apoio)	71
59	Variação no destino pós alta (transf para unid rede cuid continuados integr + outros)	72
60	Criação da variável Tempos de espera	72
61	Tempos de espera por patologia	73
62	Scatter plot tempos de espera por idade	74
63	Distribuição dos tempos de espera por idade	74
64	Distribuição dos tempos de espera por entidade financeira responsável . .	75
65	Distribuição dos tempos de espera por género	75
66	Evolução dos tempos de espera	76
67	Diagrama Gantt - Entrega Final	79

68	Aplicação do K-Means	83
69	Distribuição dos utentes que posteriormente à 1ª consulta realizaram sessões, por género	88
70	Distribuição dos utentes que posteriormente à 1ª consulta realizaram sessões, por concelho	89
71	Distribuição dos utentes que posteriormente à 1ª consulta realizaram sessões, por serviço	89
72	Distribuição dos utentes que posteriormente à 1ª consulta realizaram sessões, por nacionalidade	90
73	Distribuição dos utentes que posteriormente à 1ª consulta realizaram sessões, por causa	90
74	Distribuição dos utentes que posteriormente à 1ª consulta realizaram sessões, por tipologia	91
75	Distribuição dos utentes que posteriormente à 1ª consulta realizaram sessões, por patologia	91
76	Distribuição dos utentes que posteriormente à 1ª consulta realizaram sessões, por profissional responsável	92
77	Distribuição dos utentes que posteriormente à 1ª consulta realizaram sessões, por acto	92
78	Distribuição dos utentes que posteriormente à 1ª consulta realizaram sessões, por sub-serviços	93
79	Distribuição dos utentes que posteriormente à 1ª consulta realizaram sessões, por entidade financeira responsável	93
80	Distribuição dos utentes que posteriormente à 1ª consulta realizaram sessões, por diagnóstico	94
81	Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por género	94
82	Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por concelho	95
83	Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por nacionalidade	95
84	Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por sub-serviço	96
85	Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por causa	96
86	Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por tipologia	97

87	Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por profissional responsável	97
88	Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por idades	98
89	Distribuição dos utentes em internamento por género	98
90	Distribuição dos utentes em internamento por entidade financeira responsável	99
91	Distribuição dos utentes em internamento por serviço	99
92	Distribuição dos utentes em internamento por diagnóstico	99
93	Distribuição dos utentes em internamento por patologia	100
94	Distribuição dos utentes em internamento por profissional responsável . . .	100
95	Distribuição dos utentes em internamento e sessões, por diagnóstico . . .	101
96	Distribuição dos utentes em internamento e sessões, por patologia	101
97	Distribuição dos utentes em internamento e sessões, por entidade financeira responsável	101
98	Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por género (1ª consulta)	102
99	Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por género (internamento)	102
100	Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por nacionalidade	102
101	Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por concelho (1ª consulta)	103
102	Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por concelho (internamento)	103
103	Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por acto	103
104	Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por sub-serviço	104
105	Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por diagnóstico	104
106	Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por patologia	104
107	Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por causa	105

108	Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por tipologia	105
109	Distribuição dos utentes em internamento e sessões, por género	106
110	Distribuição dos utentes em internamento e sessões, por concelho	106
111	Distribuição dos utentes em internamento e sessões, por entidade financeira responsável	107
112	Distribuição dos utentes em internamento e sessões, por profissional responsável	107
113	Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por género (1ª consulta)	108
114	Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por género (internamento)	108
115	Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por idades	109
116	Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por nacionalidade	109
117	Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por concelho (1ª consulta)	109
118	Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por concelho (internamento)	109
119	Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por entidade responsável por consultas	110
120	Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, entidade responsável por internamentos	110
121	Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por serviço (1ª consulta)	110
122	Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por serviço (internamento)	110
123	Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por sub-serviço	111
124	Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por diagnóstico	111
125	Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por patologia	112
126	Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por causa	112

127	Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por tipologia	113
128	Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por acto	113
129	Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por sessão	113
130	Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por meios complementares terapêutico	114
131	Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por profissional responsável pelas 1ª consultas . .	114
132	Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por profissional responsável pelas sessões	114

Lista de Tabelas

1	Tabela de Comparação dos Estudos sobre Previsão de Tempos de Espera e Classificação de Pacientes	32
2	Valores em falta	40
3	Comparação entre os modelos para a previsão dos tempos de espera . . .	80

1 - Introdução

O Centro de Medicina e Reabilitação de Alcoitão (CMRA), Santa Casa da Misericórdia de Lisboa, fundado na década de 1950, é considerado o centro de Reabilitação mais importante do país. Durante 40 anos foi o único centro de reabilitação de Portugal tendo, desde o início, começado a apostar na educação de profissionais na área da reabilitação. Atualmente, continua a fornecer serviços exclusivos apostando em tecnologias inovadoras e na melhoria de condições de acolhimento dos utentes [2].

O CMRA dedica-se à reabilitação pós-aguda de pacientes de todas as idades, com incapacidades físicas, motoras ou múltiplas, tanto congénitas quanto adquiridas, atendendo pessoas de todo o país. O seu principal objetivo é contribuir para o cumprimento de princípios de ética e da bioética na prestação de cuidados de saúde e na realização de investigação clínica, especialmente no campo das ciências da saúde [3]

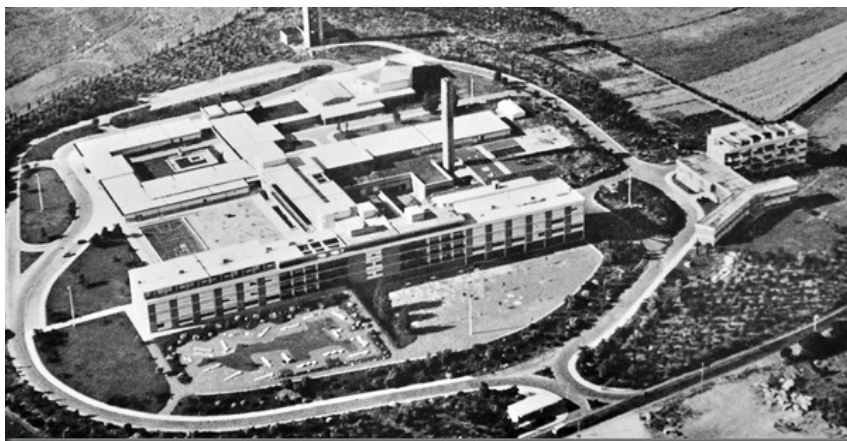


Figura 1: Centro de Medicina de Reabilitação de Alcoitão

O centro está organizado em três serviços de acordo com o grupo etário e regime de prestação de cuidados:

- **Serviço de Reabilitação de Adultos, com duas valências** – tendo a primeira 66 camas que se destinam ao internamento de utentes com lesões vertebro-medulares e outras patologias neurológicas e a segunda 68 camas destinadas ao internamento de pessoas que sofreram um acidente vascular cerebral ou traumatismo crânio-encefálico, ou que tenham sido amputadas [4].
- **Serviço de Reabilitação Pediátrica e de Desenvolvimento** – com 16 camas, destina-se ao internamento de crianças e jovens com patologias neurológicas, osteoarticulares, medulares, entre outras. Este serviço mantém as crianças na escola, através de um professor que dá as aulas nas instalações do centro [4]

Em função dos seus recursos, o centro enfrenta desafios críticos para responder à crescente procura por parte dos pacientes, sendo o tempo de espera para internamento

um dos principais problemas. Com uma capacidade total de apenas 150 camas (134 para adultos e 16 para pediatria), o CMRA enfrenta desafios relacionados com os tempos de espera prolongados que eventualmente poderão ter impacto na eficiência do atendimento. Além disso, a diversidade e complexidade das patologias atendidas aumenta a pressão sobre os recursos disponíveis e exige uma gestão eficaz para otimizar o atendimento [4].

Assim, o objetivo principal deste Trabalho Final de Curso é aplicar modelos de Machine Learning baseados nos dados sobre a gestão dos pacientes, fornecidos pelo centro. O objetivo principal é criar uma ferramenta que permita prever os tempos de espera para internamento e para tratamento e caracterizar os pacientes de acordo com a sua patologia. Neste sentido, com este TFC, pretende-se obter uma solução que permita apoiar a decisão, no que diz respeito aos recursos do CMRA, tanto a nível humano como de infraestruturas. Pretende-se, assim, melhorar a capacidade de gestão e a resposta à procura por parte dos pacientes. Consequentemente, espera-se obter uma solução que torne mais eficiente a distribuição de recursos, melhorando a capacidade de resposta do centro e adaptando-o à procura.

O propósito é oferecer ao centro uma visão global, mais clara e mais precisa sobre os padrões e tendências relacionados ao atendimento e à ocupação das camas de internamento. Ao promover uma análise detalhada dos dados, pretende-se que seja possível tomar decisões de forma mais informada e estratégica, o que pode ajudar a reduzir os tempos de espera.

Além disso, espera-se que a solução a propor por este TFC, aumente a autonomia do centro na gestão das suas atividades, possibilitando ajustes dinâmicos e respostas mais rápidas às necessidades dos pacientes. Ao disponibilizar informações que favoreçam uma administração proativa, o centro estará mais preparado para adaptar-se a mudanças na procura e para responder com facilidade a emergências, assegurando um atendimento de maior qualidade e eficiência.

2 - Pertinência e Viabilidade

Nos últimos anos, a aplicação Machine Learning (ML) em diferentes contextos clínicos, tem-se tornado cada vez mais valioso na otimização de recursos e de gestão de atendimento. No CMRA, onde a crescente procura e os limites existentes na sua capacidade de internamento são pontos críticos, o desenvolvimento de um modelo preditivo que estime os tempos de espera e que caracterize pacientes é extremamente importante, pois oferece soluções eficazes para melhorar a qualidade e eficiência dos cuidados oferecidos.

2.1 Viabilidade

Este TFC responde diretamente a uma necessidade real identificada no CMRA: otimizar a capacidade de gestão de internamentos e tratamentos e a caracterização de pacientes de modo a melhorar o atendimento.

Este TFC tem fundamento:

- na cooperação entre o CMRA e o DEISI que resultou na conclusão de diversos TFC nos últimos dois anos letivos;
- num protocolo de cooperação científica entre o Centro de Medicina de Reabilitação de Alcoitão e a COFAC - Cooperativa de Formação e Animação Cultural, CRL. que será firmado em breve;
- em dados fornecidos pela Diretora da Unidade de Gestão de Utentes do CMRA, Engenheira Mariana Matos, assim como a sua contínua colaboração com a equipa deste TFC, ao longo deste ano lectivo;
- no reforço da adequação do modelo a desenvolver por este TFC, aos desafios específicos do centro em resultado do diálogo com a instituição, de modo a garantir que a solução a obter responda às prioridades do CMRA.

Além do potencial suporte na previsão dos tempos de espera e na caracterização de pacientes, o modelo a desenvolver por este TFC representa também uma alternativa financeiramente viável, proporcionando resultados precisos sem custos. A implementação de um modelo preditivo poderá eliminar a necessidade de soluções mais complexas e onerosas, como a implementação de sistemas adicionais ou o aumento das equipas, permitindo que o centro faça uma gestão mais otimizada dos seus recursos.

Adicionalmente, é importante observar que a flexibilidade do modelo a obter, embora desenvolvido com foco nas necessidades do CMRA, permitirá uma fácil adaptação a diferentes conjuntos de dados.

2.2 Pertinência

Este trabalho irá resultar na implementação de um modelo que permitirá trazer melhorias diretas ao planeamento de internamento de pacientes do CMRA e à caracterização dos pacientes, com potencial de aumentar a eficiência e a qualidade dos serviços oferecidos.

- **A previsão de tempos de espera de doentes para internamento e tratamento** vai permitir ao CMRA otimizar o atendimento de doentes, permitindo uma distribuição mais eficiente dos recursos e, consequentemente, reduzindo o tempo de espera para os pacientes.
- **A caracterização de pacientes**, permite a otimização da distribuição de recursos e permitirá um atendimento mais organizado.
- **Com melhor planeamento**, o centro poderá atender a um maior número de pacientes sem comprometer a qualidade dos cuidados prestados.

Assim sendo, a implementação do modelo preditivo neste TFC irá abordar uma necessidade do CMRA, contribuindo para a sustentabilidade e a melhoria contínua dos serviços de saúde prestados.

Outro aspeto importante resulta da relevância do CMRA no contexto da investigação científica na área da reabilitação. Esta relevância do CMRA resulta do número de doentes que atende, comparativamente com outras unidades de saúde na Europa. O modelo preditivo a aplicar por este TFC permitirá, mais tarde, que se desenvolva investigação no que toca ao seu desenvolvimento, consequências, melhorias a implementar, possível implementação a outros processos clínicos ou unidades hospitalares.

3 - Conceitos Fundamentais

O nosso objetivo, neste ponto, é apresentar fundamentos teóricos necessários para uma melhor compreensão do desenvolvimento deste trabalho, incluindo a discussão sobre o conceito de Machine Learning e a exploração de técnicas e modelos relevantes que poderão vir a ser utilizados ao longo do projeto.

3.1 Conceitos Teóricos

Nesta secção queremos apresentar, e analisar, as técnicas que irão ser aplicadas nos capítulos seguintes, com principal destaque para os princípios de Machine Learning e a importância da preparação dos dados. Estes conceitos são essenciais para garantir a eficácia do modelo preditivo que queremos desenvolver, assim como, compreender as métricas que poderão avaliar o seu desempenho.

3.1.1 Machine Learning

Machine Learning é uma área da Inteligência Artificial (IA) que desenvolve algoritmos capazes de aprender, tomar decisões e fazer previsões ou classificações baseadas em dados [5].

A eficácia destes modelos depende dos dados a utilizar, quanto mais relevantes, diversos e de qualidade estes forem, melhor será o desempenho do modelo. Para garantir a eficácia, é necessário um grande volume de dados, permitindo ao modelo detetar padrões e relações que podem ser difíceis de perceber manualmente.

3.1.2 Machine Learning na saúde

Na área da saúde, o principal objetivo do Machine Learning é possibilitar o processamento eficiente de grandes quantidades de dados, a identificação de padrões e a obtenção de informações essenciais para a tomada de decisões por parte dos profissionais que trabalham na área.

Em contextos hospitalares, por exemplo, a gestão de tempo e recursos é essencial, fazendo com que as técnicas sejam determinantes, para otimizar a distribuição de profissionais, prever a necessidade de internamento e classificar os pacientes de acordo com as suas patologias.

Ao longo do desenvolvimento do projeto, este ponto será revisto e, para que seja mais perceptível, na secção 4.1 estão apresentados artigos que ajudam a entender melhor o papel da aplicação de ML na área da saúde.

3.1.3 Tipos de Machine Learning

O Machine Learning está dividido em dois tipos principais de aprendizagem. Estes são essenciais para entender como os modelos são desenvolvidos e treinados:

- **Aprendizagem Supervisionada** – Neste tipo de aprendizagem, o modelo é treinado a partir de dados rotulados, ou seja, as variáveis independentes (características) e as dependentes (rótulos), são previamente conhecidas. O objetivo é criar um modelo capaz de fazer previsões precisas, com base em novos dados. Esta aprendizagem, é especialmente útil, em problemas de classificação e regressão [6].
- **Aprendizagem Não Supervisionada**– Trata-se de um tipo de aprendizagem usada quando os dados não estão rotulados. O objetivo é identificar padrões ocultos nos dados, sem a necessidade de rótulos específicos [6].

3.1.4 Preparação dos Dados

Antes dos modelos de Machine Learning serem utilizados, é necessário preparar os dados. A sua qualidade tem um papel essencial no desempenho dos modelos preditivos. Assim sendo, dados inconsistentes, incompletos ou redundantes, podem comprometer o desempenho de qualquer algoritmo.

Esta preparação, é um passo muito importante no desenvolvimento de modelos. Os principais processos incluem:

- **Limpeza dos dados** – remoção de valores ausentes, duplicados ou incorretos, de modo a garantir a consistência.
- **Transformação de dados** – normalização ou padronização para que as variáveis estejam todas na mesma escala.
- **Feature engineering** – seleção e criação de variáveis com maior impacto no modelo.
- **Divisão de Dados** – separação em conjuntos de treino e teste, de forma a evitar o overfitting (quando o modelo decora os dados de treino e não funciona bem para novos dados).

3.1.5 Métricas de Desempenho

Para que seja possível avaliar o desempenho do modelo é necessário utilizar certas métricas para avaliar a eficácia do mesmo.

As principais métricas de previsão que poderão vir a ser utilizadas ao longo do desenvolvimento do trabalho são:

- **Coeficiente de determinação (R^2)** - Medida estatística que mostra o grau de proximidade entre os dados e os valores ajustados pelo modelo de regressão. Varia

entre 0 e 1, sendo que, quanto mais próximo de 1 o valor estiver, melhor é o ajuste e menor o erro [7].

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (1)$$

- **Erro Quadrático Médio (MSE)** - Representa a média das diferenças quadráticas entre os valores reais e os valores previstos pelo modelo. Quanto maior for o resultado, pior é o desempenho do modelo, já que este indica que as previsões estão longe dos valores reais. [7]

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad (2)$$

- **Raiz do Erro Quadrático Médio (RMSE)** - É a raiz quadrada do erro quadrático médio e mede o desvio padrão das diferenças entre os valores reais e os valores previstos. Quanto menor for o resultado, melhor é o desempenho do modelo [7].

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (3)$$

- **Erro Médio Absoluto (MAE)** - Esta métrica calcula a média das diferenças absolutas entre os valores reais e os valores previstos [8].

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (4)$$

- **Erro Médio (ME)** - É uma métrica que calcula apenas a média das diferenças entre os valores reais e os valores previstos pelo modelo, dando assim uma visão mais geral do seu desempenho. Quanto mais próximo de 0 o seu resultado estiver, melhor é o desempenho do modelo [9].

$$ME = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i) \quad (5)$$

- **Erro Percentual Absoluto Médio (MAPE)** - Mede a precisão do modelo ao calcular a média dos erros percentuais absolutos. Este calcula a soma dos erros de previsão absolutos, expressos em percentagem, e divide pelo número total de valores reais. Quanto mais próximo o resultado estiver de 0 melhor é o desempenho do modelo [7].

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100 \quad (6)$$

- **Soma de Erro Absoluto (SAE)** - É uma medida que calcula a soma das diferenças entre os valores reais e os valores previstos pelo modelo, ajudando a compreender os erros acumulados num conjunto de dados [10].

$$\sum_{i=1}^n |y_i - \hat{y}_i| \quad (7)$$

A utilização destas métricas, oferece uma visão mais concreta do desempenho do modelo, permitindo ajustar e comparar diferentes abordagens e modelos durante o desenvolvimento do trabalho.

As métricas descritas serão consideradas durante a avaliação dos modelos a implementar neste TFC. No entanto, a escolha de métricas a utilizar, dependerá igualmente dos dados e do modelo a obter.

3.1.6 Modelos a utilizar

De modo a facilitar a compreensão dos modelos que poderão ser utilizados ao longo do projeto, seguem-se observações sobre os mesmos. Estes modelos, representam diversas maneiras de abordar os problemas propostos no âmbito do trabalho: previsão de tempos de espera e caracterização dos pacientes.

3.1.7 Modelos para previsão de tempos de espera

Regressão Linear

A regressão linear, é um modelo estatístico utilizado para analisar a existência de uma relação linear entre variáveis e prever valores. Este modelo baseia-se na relação entre uma variável dependente (variável a prever) e uma ou mais variáveis independentes (utilizadas para a previsão). O seu objetivo principal, é identificar as relações entre as variáveis, ajustando uma linha reta que representa os dados [11].

A equação da regressão linear simples é:

$$y = \beta_0 + \beta_1 \cdot x + \epsilon \quad (8)$$

Onde:

- x é a variável independente (utilizada para prever y);
- y é a variável dependente (variável a prever);
- β_0 é a ordenada na origem (valor de y quando $x=0$);
- β_1 é o coeficiente que mede o impacto da variável x sobre a y ;
- ϵ é o erro (diferença entre valores reais e previstos).

Tipos de regressão linear:

- **Regressão Linear Simples** – apenas uma variável independente é considerada para a previsão. (Figura 2)
- **Regressão Linear Múltipla** – são consideradas várias variáveis independentes para a previsão. (Figura 3)

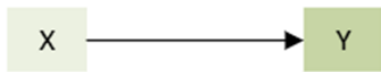


Figura 2: Regressão Linear Simples

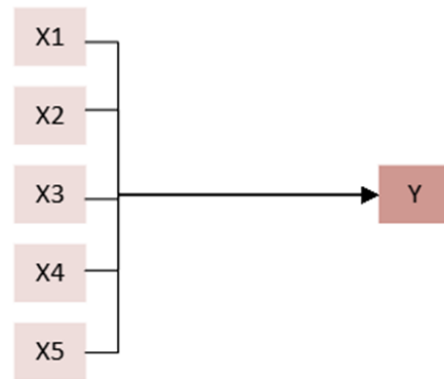


Figura 3: Regressão Linear Múltipla

Árvores de Decisão

Uma árvore de decisão é representada por uma estrutura semelhante a um fluxograma, onde cada ramificação interna representa uma característica e cada extremidade representa o valor previsto, permitindo visualizar de forma simples como o modelo chega a determinada conclusão.

Esta é uma abordagem preditiva e é utilizada para problemas de classificação e regressão. As árvores de decisão são construídas através de uma abordagem algorítmica que identifica maneiras sobre como dividir um conjunto de dados dependendo das condições.

Existem diferentes tipos de árvores de decisão, dependendo do que queremos obter:

- **Árvores de classificação** - onde a variável alvo pode assumir um conjunto discreto de valores.
- **Árvores de regressão** - onde a variável alvo pode assumir valores contínuos.

Estas árvores utilizam algoritmos que identificam o melhor atributo para dividir os dados em cada etapa, otimizando a precisão das previsões[12].

De seguida, na figura 4, está apresentado um exemplo de uma árvore de decisão:

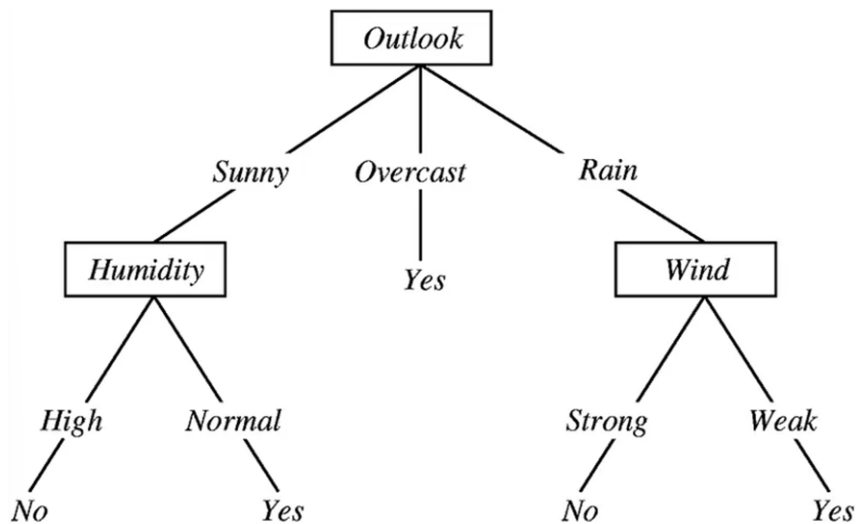


Figura 4: Exemplo de uma árvore de decisão

Random Forest

Random Forest é um algoritmo utilizado para classificação e regressão, que funciona através da criação de várias árvores de decisão durante o treino do modelo. Cada árvore é construída com um subconjunto do conjunto inicial de dados, que é selecionado aleatoriamente.

No caso da classificação, a previsão final é selecionada pela classe que mais aparece nas árvores, e no caso da regressão, pela média das previsões.

É um modelo que ajuda a reduzir o overfitting e melhora a precisão preditiva. Ao contrário das árvores de decisão, que avaliam todas as possíveis divisões de características, este seleciona aleatoriamente um subconjunto de características para cada divisão, resultando num melhor desempenho. A aleatoriedade na seleção do subconjunto de características é essencial para reduzir a correlação entre as árvores e garantir um modelo mais fiável [13].

Na imagem a seguir, figura 5, pode-se observar o funcionamento do Random Forest.

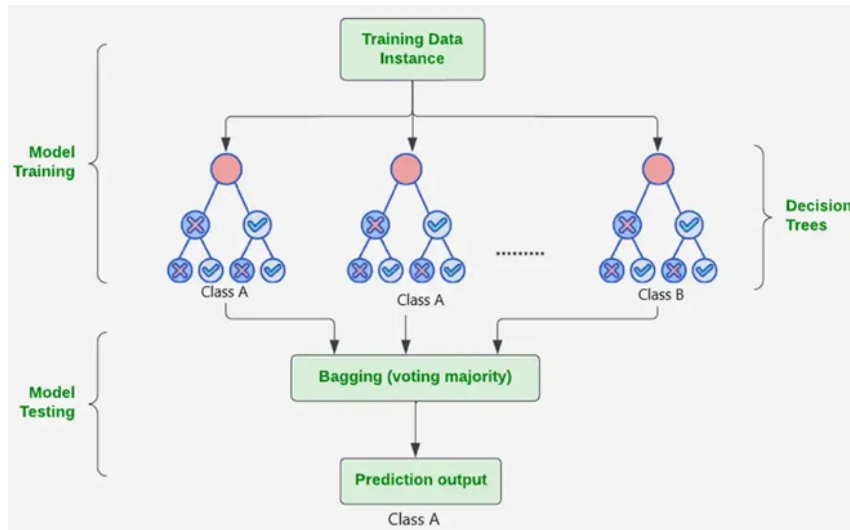


Figura 5: Demonstração do funcionamento do algoritmo Random Forest

Tal como mostra a imagem, o treino do modelo envolve a criação de várias árvores de decisão com dados seleccionados aleatoriamente. Cada árvore faz uma previsão e, durante o teste, as previsões são combinadas através do processo de bagging, onde a classe mais frequente é escolhida para previsão total.

Support Vector Machine (SVM)

O SVM (Support Vector Machine) é um algoritmo particularmente eficaz quando usado em datasets mais pequenos, pois este pode ser demorado no processamento de datasets grandes.

O seu principal objetivo é encontrar o melhor hiperplano que separa as classes de dados num espaço N-dimensional, maximizando a distância entre elas. Esse hiperplano é o limite de decisão que determina a classe de um novo ponto.

Quando os dados não podem ser separados de forma linear, utiliza-se o kernel (função matemática que calcula a semelhança entre pontos, facilitando a separação entre classes) para transformar os dados num espaço de maior dimensão. O modelo pode ser utilizado tanto em problemas de classificação como de regressão. No entanto, tende a ser mais eficaz em problemas de classificação [14].

Na imagem a seguir, figura 6, pode-se observar uma aplicação de SVM. Os pontos azuis e verdes representam as classes de dados e o hiperplano (linha vermelha) é o que separa as classes. Através da imagem torna-se mais fácil entender como este modelo trabalha para encontrar o melhor limite.

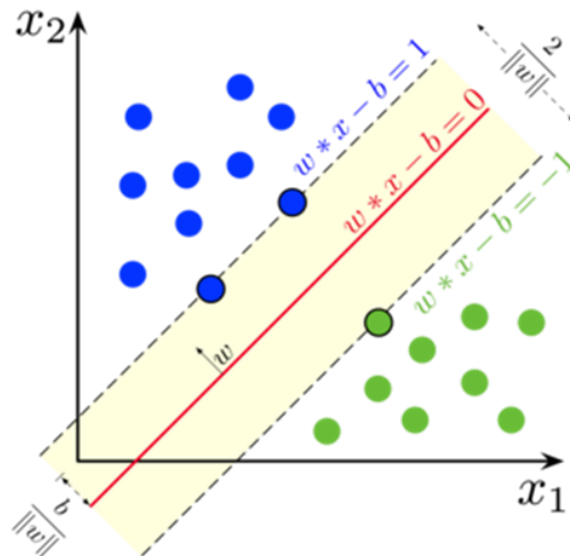


Figura 6: Exemplo de uma aplicação de SVM

K- Nearest Neighbours (KNN)

O KNN é um algoritmo muito usado para problemas de classificação e de previsão, baseando-se no princípio que, os pontos semelhantes estão próximos uns dos outros. Na classificação, o rótulo de um ponto é atribuído com base no rótulo mais frequente entre os k vizinhos mais próximos, enquanto na regressão, utiliza-se a média dos pontos vizinhos como previsão. Para medir as distâncias e identificar os pontos mais próximos, a métrica mais utilizada é a distância euclidiana (distância mais curta entre dois pontos).

Como o modelo armazena apenas os dados de treino, exige muita memória e pode não ser eficiente em grandes conjuntos de dados. Assim sendo, a escolha do k é muito importante, valores pequenos podem levar a overfitting e valores grandes aumentam o bias. O melhor ajuste é, geralmente, alcançado através da validação cruzada[15].

Na imagem abaixo, figura 7, é possível observar o processo do algoritmo:

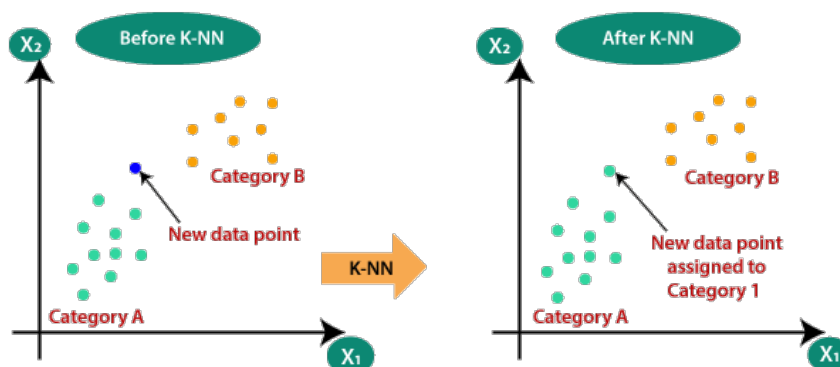


Figura 7: Aplicação do algoritmo KNN

O ponto a identificar (representado a azul) é classificado com base nos seus vizinhos mais próximos. Antes da aplicação KNN o ponto não tem classe, e após a aplicação, é

classificado, de acordo com os seus vizinhos mais próximos, na categoria A (representado a verde).

3.1.8 Modelo para caracterização dos pacientes

K-Means

De modo a garantir uma aplicação mais eficiente e eficaz do algoritmo K-Means, a Análise de Componentes Principais (PCA — Principal Component Analysis) mostra-se importante para recorrer inicialmente. O PCA é um método estatístico que preserva a maior parte da informação mais relevante. Este transforma conjuntos de dados de grande dimensão em representações de menor dimensão. Esta transformação dá-se através da identificação de novos eixos, os componentes principais, ao longo dos quais os dados mais variam [16].

O K-Means é um algoritmo destinado a resolver problemas de clustering, uma técnica que tem como objetivo agrupar pontos de dados semelhantes, tendo em conta as suas características. O clustering divide um conjunto de dados em clusters distintos, onde cada cluster consiste em pontos de dados mais semelhantes entre si.

Este é um algoritmo de agrupamento, utilizado para classificar dados não rotulados em grupos (clusters) tendo em conta a sua similaridade. O processo envolve a definição do valor ideal de K pontos centrais (centróides) através de um processo iterativo e, de seguida, atribui cada ponto ao seu centróide mais próximos criando assim os clusters [17].

Na figura abaixo, figura 9, é possível observar o resultado a obter através do processo:

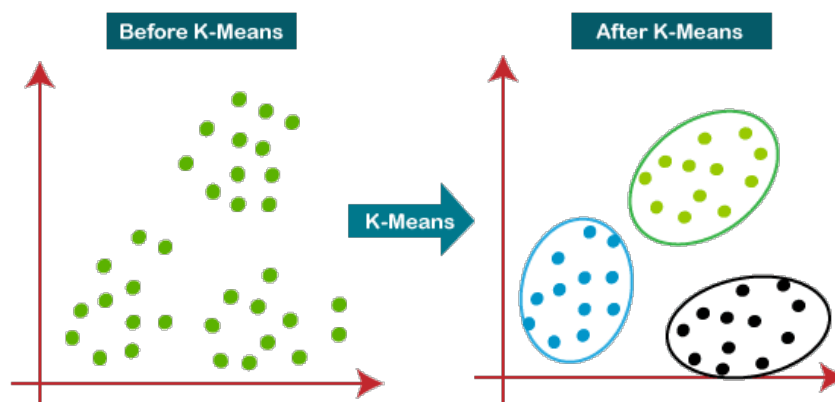


Figura 8: Resultado da aplicação do algoritmo K-Means

Este algoritmo será utilizado na caracterização dos pacientes, segmentando-os tendo em conta características semelhantes. O objetivo é agrupar os pacientes de maneira que cada grupo represente um conjunto distinto de pacientes, facilitando a gestão de recursos e permitindo ao CMRA tomar decisões mais informadas e adequadas para cada tipo de característica.

4 - Estado da Arte

Este capítulo tem como objetivo analisar as vantagens e desafios da implementação de modelos de Machine Learning (ML) na previsão de tempos de espera e na caracterização de pacientes no CMRA, com base numa pesquisa sobre modelos já desenvolvidos para objetivos semelhantes. A partir da pesquisa realizada, foram identificados diversos modelos e abordagens que utilizam técnicas de ML, variando em complexidade e eficácia. Estes modelos refletem diferentes estratégias para otimizar a distribuição de recursos e aumentar a eficiência nos serviços de saúde.

Nos últimos anos, a aplicação de algoritmos de ML na saúde, tem gerado resultados promissores, principalmente devido ao aumento da disponibilidade de dados clínicos e à capacidade dos modelos de lidar com a complexidade e variabilidade dos dados. Esta evolução permitiu o desenvolvimento de diversos métodos capazes de otimizar a gestão de recursos hospitalares e melhorar o atendimento aos pacientes.

Diversos estudos analisam diferentes processos de Machine Learning como, por exemplo, Regressão Linear e K-Nearest Neighbors (KNN), Gradient Boosting Decision Trees (GBDT) e Redes Neurais (NN). Cada uma destas abordagens, apresenta vantagens específicas dependendo do tipo de dados, número de variáveis envolvidas e nível de precisão exigida. A seguir, será realizada uma revisão detalhada dos estudos mais relevantes encontrados sobre o tema, com o objetivo de fornecer uma visão aprofundada do estado de arte na previsão de tempos de espera e na classificação de pacientes no contexto hospitalar.

4.1 Artigos Analisados

Esta é uma lista contendo os artigos que analisei e que exploram a aplicação de Machine Learning (ML) em contextos semelhantes ao do CMRA, com foco na previsão de tempos de espera e classificação de pacientes.

Artigo 1 – Anat Ratnovskya e outros [18]

"Statistical learning methodologies and admission prediction in an emergency department", investiga aplicação de Machine Learning (ML) para prever a admissão hospitalar de pacientes no setor de emergências. O artigo, propõe um modelo preditivo baseado em dados iniciais de triagem. Para identificar os principais indicadores de desempenho (KPIs) que influenciam a previsão, são realizadas análises exploratórias (AED) e confirmatórias (ACD) dos dados. Os algoritmos aplicados incluem a Regressão Logística, a Árvore de Decisão e as Redes Neurais Artificiais (RNA). A métrica AUC (Área Sob a Curva) da curva ROC (Receiver Operating Characteristic), que avalia a taxa de verdadeiros e falsos positivos, é usada para medir a eficácia dos modelos, tendo o modelo de RNA, demonstrado o melhor desempenho preditivo.

Artigo 2 - Jeffin Joseph e outros [19]

"Machine Learning for Prediction of Wait Times in Outpatient Clinics", explora o uso

dos algoritmos Random Forest e Extreme Gradient Boosting (XGBoost) para prever tempos de espera e o tempo de processamento em clínicas em regime ambulatorio. As métricas utilizadas para avaliar o desempenho dos modelos incluem precisão, acurácia e a curva ROC. Para lidar com o desequilíbrio de classes, foi aplicada a técnica de super amostragem de vizinhança sintética (SMOTE), resultando numa melhoria significativa dos resultados. O XGBoost, especialmente quando combinado com SMOTE, obteve uma precisão de 0,86 e uma especificidade muito alta, destacando-se como uma ferramenta eficaz na gestão de filas e na melhoria da eficiência dos serviços de saúde.

Artigo 3 - Anton Pak e outros [20]

No artigo "Predicting waiting time to treatment for emergency department patients", são propostos algoritmos de ML para prever, com mais precisão, os tempos de espera em serviços de urgência. O estudo compara diferentes modelos, como LASSO (Least Absolute Shrinkage and Selection Operator), Random Forest e Quantile Regression, que superaram métodos tradicionais, como a média móvel, ao capturar, de forma mais eficaz, a variabilidade nos tempos de espera. Os resultados indicam que modelos como LASSO e Random Forest, são mais eficazes em cenários de alta procura, oferecendo previsões mais precisas e uteis.

Artigo 4 - A/Prof Katie Walker e outros [21]

O artigo "Emergency medicine patient wait time multivariable prediction models", apresenta a validação de modelos de ML para prever os tempos de espera em serviços de urgência, através dos dados de 12 hospitais na Austrália. A pesquisa compara modelos como Random Forest e Regressão Linear com modelos tradicionais, como a media móvel. Os modelos desenvolvidos para cada hospital, mostraram maior precisão, principalmente, em situações excepcionais como a pandemia de COVID-19. O estudo destaca a importância da validação externa dos modelos antes da sua implementação noutros hospitais, devido à limitação da transferência de modelos específicos entre locais.

Artigo 5 – Amjed Al-Mousa e outros [22]

O estudo "A machine learning-based approach for wait-time estimation in healthcare facilities with multi-stage queues", propõe uma abordagem baseada em ML, para estimar tempos de espera em hospitais, tendo em consideração, a complexidade das filas multifacetadas. A pesquisa utiliza algoritmos como Random Forest, Ridge Regression, LASSO, Elastic Net e Redes Neurais Artificiais (RNA). O modelo Random Forest obteve o menor valor de erro quadrático médio (RMSE), superando métodos tradicionais. O estudo destaca a capacidade dos modelos de ML de melhorar significativamente a precisão das previsões, principalmente, em cenários com múltiplas etapas de filas.

Artigo 6 - Parthasarathi Pattnayak e outros [23]

Este estudo, "Deep Learning based Patient Queue Time Forecasting in the Emergency Room", propõe um modelo baseado em deep learning (DL) para prever, com maior precisão, os tempos de espera em serviços de emergência, focando em pacientes de baixa gravidade. Motivado pela alta taxa de erro de modelos anteriores, o artigo utiliza técnicas de DL e teoria das filas de espera, para estimar os tempos de espera com base em dados de Registos Eletrônicos de Saúde (EHR). O modelo melhora a utilização de variáveis preditoras como hora e dia de chegada, e tempo de permanência na fila, por meio de otimizações como Adagrad, Adam, RMSprop e SGD. O modelo de DL mostrou-se mais eficaz que métodos convencionais, apresentando um erro menor e contribuindo

para a priorização dos pacientes.

Artigo 7 – Xiaoqing Li e outros [24]

No estudo "Prediction of outpatient waiting time", são explorados diferentes modelos de ML, incluindo Regressão Linear, Random Forest, Gradiente Boosting Decision Tree (GBDT) e K-Nearest Neighbors (KNN), para prever os tempos de espera num hospital pediátrico chinês. O Random Forest, obteve um desempenho superior no Departamento de Medicina Interna I, com um erro médio absoluto (MAE) 47,6% inferior ao da Regressão Linear. Já o GBDT, apresentou os melhores resultados em outros departamentos. O estudo enfatiza ainda a importância da engenharia de características e da validação cruzada de modo a melhorar a precisão dos modelos.

Artigo 8 - Suebsarn Ruksakulpiwat e outros [25]

"Machine learning-based patient classification system for adults with stroke: A systematic review", explora o uso de técnicas de ML na classificação de pacientes adultos com AVC. São discutidas abordagens supervisionadas, não supervisionadas e por reforço, com destaque para o uso de DL, que obteve os melhores resultados na previsão de tempo de detecção e de internamento de pacientes com AVC.

Artigo 9 - Shatha Melhem e outros [26]

O estudo "Patient care classification using machine learning techniques", explora a classificação de pacientes (internados ou em ambulatório), utilizando técnicas de ML com os dados EHR de um hospital indonésio. Foram implementados modelos como SVM, Árvore de Decisão, Random Forest e KNN, destacando o Random Forest como o mais eficaz, com uma precisão de 77%. O estudo mostra que a classificação de pacientes, pode melhorar a eficiência dos serviços de saúde, reduzindo tempos de espera e erros nas decisões.

De modo a que seja mais fácil a comparação dos artigos segue-se uma tabela comparativa (Tabela 2):

Título do Artigo	Ano	Objetivo	Modelos e Desempenho	Dados	Utiliza- dos
"Statistical learning methodologies and admission prediction in an emergency department"	2021	Prever admissão hospitalar em emergência	Regressão Logística, Árvore de Decisão, RNA (melhor desempenho)	Dados de triagem inicial	
"Machine Learning for Prediction of Wait Times in Outpatient Clinic"	2022	Prever tempo de espera e tempo de processamento em clínicas ambulatoriais	Random Forest, XGBoost (um melhor desempenho com SMOTE)	Dados de consultas ambulatoriais	
"Predicting waiting time to treatment for emergency department patients"	2021	Prever tempo de espera em serviços de emergência	Quantile Regression, LASSO (melhores previsões)	Dados de um serviço de emergência	
"Emergency medicine patient wait time multivariable prediction models: a multicentre derivation and validation study"	2021	Prever tempos de espera em múltiplos serviços de emergência	Random Forest, Regressão Linear (melhor desempenho)	Dados de 12 serviços de emergência na Austrália	
"A machine learning-based approach for wait-time estimation in healthcare facilities with multi-stage queues"	2023	Prever tempos de espera em estabelecimentos de saúde com filas multifacetadas	Random Forest (melhor desempenho com menor RMSE)	Dados de transações de um grande hospital	
"Deep Learning based Patient Queue Time Forecasting in the Emergency Room"	2023	Prever com precisão o tempo de espera dos pacientes em salas de emergência	Deep Learning (superou as técnicas convencionais)	Dados de registros eletrônicos de saúde (EHR)	
"Prediction of outpatient waiting time: using machine learning in a tertiary children's hospital"	2023	Prever o tempo de espera num hospital pediátrico	GBDT (melhor desempenho), Random Forest, KNN	Dados do sistema de informação hospitalar (HIS)	
"Machine learning-based patient classification system for adults with stroke: A systematic review"	2023	Explorar o uso de ML na classificação de doentes com AVC	Deep Learning (melhores resultados na detecção precoce e internamento de pacientes com AVC, no entanto não houve consenso sobre o algoritmo mais eficaz)	Dados clínicos de pacientes com AVC (detalhes não especificados)	
"Patient care classification using machine learning techniques"	2021	Classificar pacientes (internados ou ambulatoriais) para melhorar a gestão hospitalar	Random Forest (alcançou 77% de precisão), SVM, Árvore de Decisão, KNN	Dados EHR (resultados dos testes laboratoriais dos doentes) de um hospital indonésio	

Tabela 1: Tabela de Comparação dos Estudos sobre Previsão de Tempos de Espera e Classificação de Pacientes

Este capítulo analisou a aplicação de modelos de Machine Learning (ML) na previsão de tempos de espera e na classificação de pacientes, com base em estudos relevantes. A pesquisa revelou que algoritmos como Random Forest, Redes Neurais Artificiais (RNA), Regressão Linear, XGBoost, Deep Learning e GBDT, apresentaram resultados positivos, dependendo do tipo de dados e do contexto em que foram aplicados.

Modelos como o Random Forest, que se destacou em diversos estudos pela sua eficácia na previsão de tempos de espera e classificação de pacientes, mostram ser uma boa escolha no desenvolvimento do modelo proposto.

Resumindo, os modelos de ML analisados, oferecem soluções favoráveis na otimização da gestão de recursos hospitalares e na melhoria da eficiência no atendimento aos pacientes.

5 - Ferramentas a Utilizar

Para o desenvolvimento deste trabalho foi usada a linguagem Python pois, para além de ser a mais praticada ao longo da licenciatura, é também a mais utilizada em trabalhos direcionados a machine learning.

Quanto à escolha de ambiente de desenvolvimento Jupyter Notebook, esta justifica-se pela sua capacidade de ser utilizado sem acesso à internet e pela sua eficiência que permite a junção de texto com código, garantindo a organização do trabalho. Além disso, possibilita, ainda, a execução de cada célula de código individualmente, tornando o processo mais dinâmico ao permitir testar e ajustar o código de forma isolada, sem a necessidade de o executar todo de uma vez.

Para o processamento, manipulação, visualização dos dados e implementação do modelo, serão utilizadas as bibliotecas:

- **Numpy** – Utiliza arrays multidimensionais eficientes, realiza operações matemáticas sem necessidade da criação de loops detalhados, disponibiliza variadas funções algébricas, estatísticas, entre outras, e serve ainda como base para diversas bibliotecas de Machine Learning, como, por exemplo, o Pandas [27].
- **Pandas** – Permite a manipulação e organização de grandes volumes de dados de forma mais eficiente, possibilita a leitura de ficheiros de diversos formatos (csv, excel, json, entre outros) e fornece ainda ferramentas que permitem lidar com diferentes tipos de dados, garantindo a sua consistência [28].
- **Matplotlib** – Facilita a criação de diversos tipos de gráficos, oferecendo ferramentas para personalizar e ajustar elementos gráficos consoante o que é pretendido obter [29].
- **Scikit-learn** – Oferece uma grande coleção de algoritmos (como regressão, classificação, entre outros) e, além de incluir várias métricas que possibilitam a avaliação do desempenho do modelo, disponibiliza, também, ferramentas para a normalização e seleção de características, otimizando o processo [30].
- **Seaborn** – É uma biblioteca para a criação de gráficos estatísticos em python. Cria gráficos em conjunto com o matplotlib, tornando tudo mais simples e informativo. É usada principalmente quando se pretende observar padrões ou relações nos dados [31].
- **Plotly Express** – É um módulo da biblioteca Plotly que contém funções que conseguem criar gráficos de forma mais rápida, eficiente e simples. É útil, principalmente, para a criação de gráficos interativos, sem ser necessária a especificação de grandes detalhes [32].

Estas ferramentas e bibliotecas não otimizam apenas o desenvolvimento e a análise de dados, como também garantem a consistência dos resultados e a clareza no processo de documentação.

6 - Solução Proposta

A solução proposta para a realização deste projeto, consiste na aplicação de um modelo preditivo, utilizando técnicas de Machine Learning. O objetivo é prever os tempos de espera e caracterizar os pacientes de modo a comparar as métricas de cada um dos modelos estudados e escolher a melhor solução.

Apesar da importância da aplicação de Machine Learning, o sucesso do modelo não dependerá apenas disso. O processo de preparação dos dados será o primeiro passo a ser dado, após a obtenção dos dados.

O próximo passo será a realização da análise exploratória dos dados. Nesta fase, a distribuição e o comportamento das variáveis serão estudados através de visualizações gráficas e análises estatísticas, permitindo a identificação de padrões, a detecção de outliers e uma melhor compreensão do comportamento das variáveis. Esta análise é essencial para a detecção de relações entre variáveis e para um melhor entendimento do conjunto de dados. Através da análise de correlação entre as variáveis, será possível selecionar as mais importantes para o modelo.

Após a análise exploratória dos dados, será realizado o processo de feature engineering. Este consiste na transformação e seleção de variáveis para criar um modelo preditivo, com o objetivo de descobrir padrões ocultos nos dados e melhorar o desempenho do modelo. Nesta fase, serão aplicadas técnicas, como a normalização ou padronização de dados, entre outras, para que seja possível ajustar as escalas e destacar as variáveis mais relevantes para o modelo.

Com os dados já preparados, o próximo passo será a construção e o treino do modelo. Nesta fase, serão escolhidos e treinados os algoritmos mais adequados à resolução do problema, e os dados serão divididos, aleatoriamente, em conjuntos de treino e teste, de modo a garantir uma avaliação eficaz do modelo, evitando problemas como o overfitting e underfitting. A divisão dos dados será feita da seguinte forma:

- **80% dos dados** serão destinados ao conjunto de treino, permitindo ao modelo aprender os padrões presentes nas variáveis.
- Os restantes **20% dos dados** serão destinados ao conjunto de teste, utilizado para avaliar a capacidade do modelo perante novos dados.

De seguida, serão aplicados alguns dos modelos referidos nos conceitos teóricos, e será feita uma comparação dos seus desempenhos de modo a escolher o mais adequado. A avaliação do modelo escolhido será realizada, através da análise do seu desempenho utilizando métricas, como o coeficiente de determinação, o erro quadrático médio, entre outras, de modo a assegurar a capacidade de previsão do modelo.

Por fim, com o modelo validado, a sua implementação poderá ocorrer, assegurando que este está preparado para ser utilizado de forma eficaz.

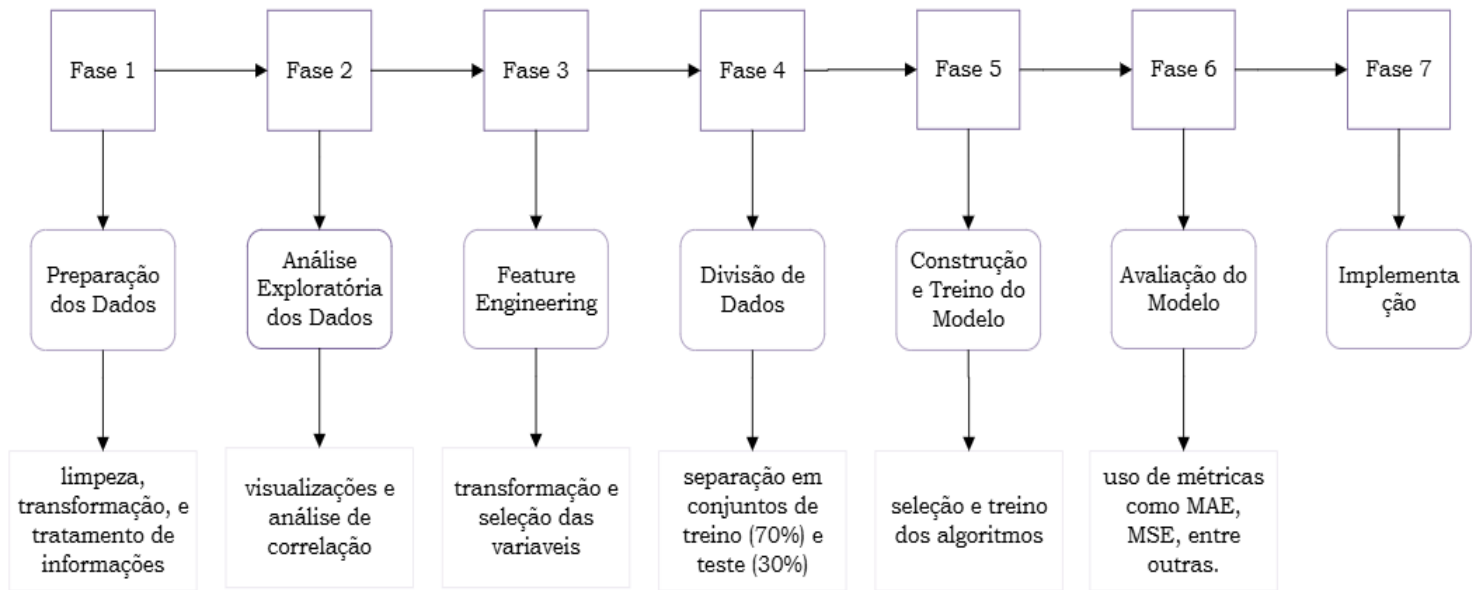


Figura 9: Fases propostas para a resolução do problema

6.1 Dados a Utilizar

A preparação deste TFC iniciou-se este ano letivo (setembro de 2024), com o convite da Engenheira Mariana Matos à Professora Iolanda Velho, para esta colaboração. Na altura antecipávamos que o CMRA partilharia, de uma forma célere e simples, o conjunto de dados na base do desenvolvimento do modelo preditivo. No entanto, colocou-se entretanto a necessidade de obter a aprovação da comissão de Ética do CMRA, que só se reuniu no dia 4 de dezembro de 2024, o que acabou por prejudicar o tempo de realização do trabalho.

Nos dados que nos foram fornecidos temos 4 diferentes tabelas, cada uma sobre:

- 1^{as} consultas (Figura 10, (a))
- Info Internamento (Figura 10, (b))
- Sessões (Figura 10, (c))
- Internamento (Figura 10, (d))

(a) 1ª consulta		(b) Info internamentos		(c) Sessões		(d) Internamentos	
Variável	Tipo	Variável	Tipo	Variável	Tipo	Variável	Tipo
N_PC	int64	N_PROC	float64	Dia	object	N_PC	int64
SER	object	Sexo	object	Sessões	int64	Sexo	object
Género	object	Data Nascimento	object	MCT's	int64	ConcResidencia	object
DNasc	object	Data Início	object	ID_PROF	int64	DataInternamento	object
Nacionalidade	object	Data Fim	object	N_PC2	int64	Servico	object
Concelho	object	Data Diag	object			Idade	int64
CodEFR	int64	Cod_Diag	float64			Entidade	object
EFR	object	Diagnostico	object				
CodActo	int64	ID_PROF	object				
Designação Acto	object	Serviço	object				
Data Consulta	object	EFR_Nome	object				
CodServ	int64	EFR_Cod	float64				
Serviço2	object	Pat_Nome	object				
CodDiag	float64	DestinoPosAlta	object				
Diagnóstico	object	MIFAdm	float64				
Tipo Consulta	object	MIFalta	float64				
Cod Pat	float64						
Patologia	object						
CAUSA	object						
TIPOLOGIA	object						
ID_PROF	object						

Figura 10: Variáveis das quatro tabelas utilizadas

Ao utilizar variáveis detalhadas do historial de saúde dos utentes, o modelo poderá identificar padrões e relações entre estas. Dessa forma, poderá ser possível antecipar não apenas a necessidade imediata de internamento, como também estimar a possível evolução dos casos, prevendo, por exemplo, a duração esperada de recuperação e o tipo de cuidados necessários após a alta.

Este tipo de análise preditiva permite aos profissionais de saúde uma visão mais abrangente sobre cada caso. Através da análise do percurso dos pacientes, o sistema poderá prever tempos de espera, por exemplo, indicando quanto tempo um paciente terá de esperar para ser atendido ou internado, o que se traduz em decisões mais fundamentadas e uma gestão hospitalar mais eficaz.

O projeto tem então como objetivo explorar essas variáveis de modo que seja possível prever os tempos de espera e caracterizar os pacientes. Esta abordagem permitirá, principalmente, otimizar recursos e melhorar a qualidade dos cuidados prestados.

6.1.1 Recolha e Descrição dos dados

Nesta secção, vão ser introduzidas as variáveis que nos foram disponibilizadas e ao longo das seguintes, exploraremos os dados mais detalhadamente, com o objetivo de encontrar padrões e tendências, de modo a que seja possível, mais à frente, implementar os modelos que ajudem a chegar ao objetivo final.

Começando pela análise das tabelas separadas, a tabela relativa às 1ªs consultas contem 6675 linhas e 21 colunas, a tabela de internamentos tem 2185 linhas e 7 colunas, a tabela relativa às sessões contém 159600 linhas e 5 colunas e, por último, a tabela que nos fornece as informações relativas aos internamentos tem 3478 linhas e 16 colunas. As colunas de cada tabela representam as variáveis descritas que se seguem.

Durante a análise exploratória, estes valores acabarão por ser modificados conforme for sendo necessário.

De modo a ter uma maior contextualização dos dados segue-se uma explicação de todas as variáveis:

1^{as} consultas

- **N_PC** - Número de identificação do paciente (tipo: int64)
- **SER** - Serviço ao qual o paciente está associado (tipo: object)
- **Género** - Género do paciente (Masculino / Feminino) (tipo: object)
- **DNasc** - Data de nascimento do paciente (tipo: object)
- **Nacionalidade** - Nacionalidade do paciente (tipo: object)
- **Concelho** - Concelho de residência do paciente (tipo: object)
- **CodEFR** - Código da entidade financeira responsável (tipo: int64)
- **EFR** - Nome da entidade financeira responsável (tipo: object)
- **CodActo** - Código do acto médico realizado (tipo: int64)
- **Designação Acto** - Descrição do acto médico realizado (tipo: object)
- **Data Consulta** - Data da realização da consulta (tipo: object)
- **CodServ** - Código do serviço médico (tipo: int64)
- **Serviço2** - Nome do serviço médico associado (tipo: object)
- **CodDiag** - Código do diagnóstico (tipo: float64)
- **Diagnóstico** - Descrição do diagnóstico do paciente (tipo: object)
- **Tipo Consulta** - Tipo de consulta realizada (Primeira) (tipo: object)
- **Cod Pat** - Código da patologia do paciente (tipo: float64)
- **Patologia** - Nome da patologia do paciente (tipo: object)
- **CAUSA** - Motivo que levou o paciente à consulta (tipo: object)
- **TIPOLOGIA** - Categoria ou classificação da consulta (tipo: object)
- **ID_PROF** - Identificação do profissional de saúde responsável (tipo: float64)

DA Internamentos

- **N_PC** - Número de identificação do paciente (tipo: int64)
- **Sexo** - Género do paciente (Masculino / Feminino) (tipo: object)
- **ConcResidencia** - Concelho de residência do paciente (tipo: object)
- **DataInternamento** - Data em que o paciente foi internado (tipo: object)
- **Serviço** - Serviço ao qual o paciente está associado (tipo: object)
- **Idade** - Idade do paciente (tipo: int64)
- **Entidade** - Nome da entidade financeira responsável (tipo: object)

Sessões

- **Dia** - Data respetiva a cada sessão (tipo: int64)
- **Sessões** - Sessões terapêuticas a que o paciente foi (tipo: int64)
- **MCT's** - Número de meios complementares terapêuticos (cada sessão terapêutica é composta por um ou vários MCT's) (tipo: int64)
- **ID_PROF** - Identificação do profissional de saúde responsável (tipo: int64)
- **N_PC2** - Número de identificação do paciente (tipo: int64)

Info Internamento

- **N_PROC** - Número de identificação do paciente (tipo: float64)
- **Sexo** - Género do paciente (Masculino / Feminino) (tipo: object)
- **Data Nascimento** - Data de nascimento do paciente (tipo: object)
- **Data Inicio** - Data em que o paciente foi internado (tipo: object)
- **Data Fim** - Data em que o paciente teve alta (tipo: object)
- **Data Diag** - Data em que o diagnostico foi realizado (tipo: object)
- **Cod_Diag** - Código do diagnostico (tipo: float64)
- **Diagnostico** - Diagnóstico do paciente (tipo: object)
- **ID_PROF** - Id do profissional responsável (tipo: object)
- **Serviço** - Serviço em que o paciente é atendido (tipo: object)
- **EFR_Nome** - Nome da entidade financeira responsável (tipo: object)
- **EFR_Cod** - Código da entidade financeira responsável (float64)
- **Pat_Nome** - Nome da patologia (tipo: object)
- **DestinoPosAlta** - Destino para qual vai o paciente apos ser dada a alta (tipo: object)
- **MIFAdm** - Classificação da independência funcional no momento da admissão (float64)
- **MIFalta** - Classificação da independência funcional no momento da alta (float64)

6.2 Pré-Processamento dos Dados

O pré-processamento de dados é um conjunto de ações que envolvem a preparação, organização e estruturação dos dados, convertendo dados brutos em dados limpos para análise.

É uma etapa muito importante para que seja possível garantir que os dados estão sempre num formato apropriado. Esta permite que a análise e modelação dos dados seja feita com uma maior precisão e fiabilidade.

Durante o pré-processamento, foi também realizada uma análise à qualidade dos dados, nomeadamente no que diz respeito à presença de valores em falta e de valores duplicados. No caso dos valores em falta foi aplicada a estratégia de remoção ou substituição, consoante os casos. Foram ainda identificados valores duplicados que foram então removidos de modo a evitar enviesamentos e erros na análise.

6.2.1 Análise Inicial dos dados

Para a preparação dos dados, o primeiro passo foi identificar os valores duplicados e os valores em falta existentes.

Na tabela das **1^{as} consultas**, foram encontrados 3 valores duplicados e 9 colunas com valores em falta: *TIPOLOGIA*, *CAUSA*, *Patologia*, *Cod Pat*, *Concelho*, *Diagnóstico*, *CodDiag* e *DNasc*. Na tabela que se segue, tabela 2, é possível observar a quantidade de missing values existentes.

TIPOLOGIA	0.168165
CAUSA	0.168165
Patologia	0.142986
Cod Pat	0.142986
Concelho	0.005695
Diagnóstico	0.002398
CodDiag	0.002398
DNasc	0.000150

Tabela 2: Valores em falta

Estes problemas foram resolvidos através da eliminação dos valores duplicados e da remoção da maior parte das linhas com valores em falta, com exceção das linhas pertencentes à coluna *Concelho*, onde os valores em falta foram substituídos pela moda da coluna.

Na tabela relativa aos **internamentos** foram encontrados 13 valores duplicados (que acabaram por ser apagados) e não foram encontrados valores em falta.

Na tabela relativa às **sessões**, não foram encontrados valores duplicados nem valores em falta.

Por último, na tabela **info internamentos**, não foram encontrados valores duplicados e, foram encontrados vários valores em falta, no entanto como nem todos estavam identificados da mesma maneira foi necessária a análise dos valores únicos de cada coluna e depois a transformação dos valores em falta para NA, tal como esta representado na figura 11. Por último, para o tratamento dos valores em falta existentes em todas as colunas, foi feito:

- Drop das linhas com valores em falta nas colunas: *Cod_Diag*, *MIFalta*, *MIFAdm*, *N_PROG*, *Sexo*, *Data Nascimento*, *Data Inicio*, *Data Fim*, *Data Diag*, *ID_PROF*,

Serviço, EFR_Nome, EFR_Cod.

- Substituição dos valores em falta pela moda da respetiva coluna, para as colunas: *DestinoPosAlta, Pat_Nome, Diagnostico.*

```
dados_info_internamento.replace('#N/D', pd.NA, inplace=True)

dados_info_internamento.replace('NaT', pd.NA, inplace=True)
```

Figura 11: Transformação de valores em falta na tabela info internamentos

De seguida, foi essencial renomear algumas colunas de modo a permitir juntar as tabelas tendo em conta as variáveis comuns. As colunas com variáveis renomeadas foram as seguintes:

- **Dados de Internamento** – coluna *Serviço* para *SER*; coluna *ConcResidencia* para *Concelho*; coluna *Entidade* para *EFR*; coluna *Sexo* para *Género*.
- **Dados de Sessões** – coluna *N_PC2* para *N_PC*
- **Dados Info Internamento** – coluna *EFR_Nome* para *EFR*; coluna *EFR_Cod* para *CodEFR*; coluna *Sexo* para *Género*; coluna *N_PROG* para *N_PC*; coluna *Data Nascimento* para *DNasc*

Foi ainda necessária a transformação de algumas variáveis para que fosse mais fácil tanto a análise exploratória de dados como a implementação do modelo preditivo.

Foram transformadas 8 variáveis para datetime e 5 outras variáveis para o tipo string.

6.3 Análise Exploratória de Dados

A análise exploratória de dados é a etapa que nos ajuda a conhecer e compreender melhor o conjunto de dados a utilizar. Nesta etapa o objetivo é observar padrões, tendências, possíveis relações entre variáveis e outros comportamentos relevantes. Ajudando assim na tomada de decisões para as etapas seguintes.

Para facilitar a interpretação dos dados recolhidos, foram desenvolvidas visualizações e tabelas adaptadas a cada conjunto analisado. Recorreu-se a gráficos de barras para representar a distribuição das categorias em cada variável, possibilitando uma leitura clara e objetiva. Procedeu-se ainda à junção de tabelas com o objetivo de organizar e consolidar a informação mais importante. Agruparam-se também certos dados, em variáveis com demasiadas categorias ou com categorias de peso muito pequeno, o que permitiu reduzir o ruído visual e destacar as categorias mais representativas. Por último, recorreu-se ainda ao uso de cores diferenciadas para evidenciar visualmente as categorias de maior peso, reforçando a sua importância na análise.

6.3.1 Dados 1^{as} consultas

Para uma melhor compreensão dos dados é necessária a criação de visualizações que os representem de forma clara.

A distribuição por *género*, visível no gráfico de barras que se segue, indica que 60.3% dos pacientes que participaram em primeiras consultas são do sexo masculino, comparativamente a 39.7% do sexo feminino. (Figura 12)

Existe uma percentagem maior de homens do que de mulheres a ir a 1^{as} consultas

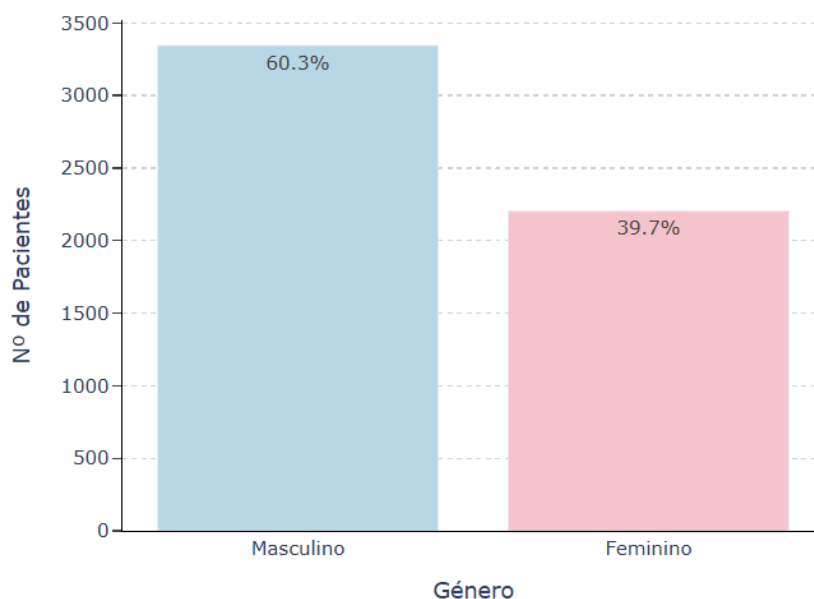


Figura 12: Distribuição do género dos utentes que atenderam a uma 1^a consulta

Seguindo para a variável *Concelho*, verifica-se que, entre os 143 concelhos, o concelho de Cascais é aquele de onde provêm mais pacientes, com 1457 pacientes. Por outro lado, alguns concelhos, como Anadia, Estremoz, Calheta contam com apenas 1 paciente cada. Na figura 13, com o top 10 de concelhos com mais pacientes, destaca-se o concelho de Cascais e de Sintra com 26.3% e 16.7% de pacientes, respetivamente.

Grande parte dos pacientes pertencem ao concelho de Cascais e de Sintra

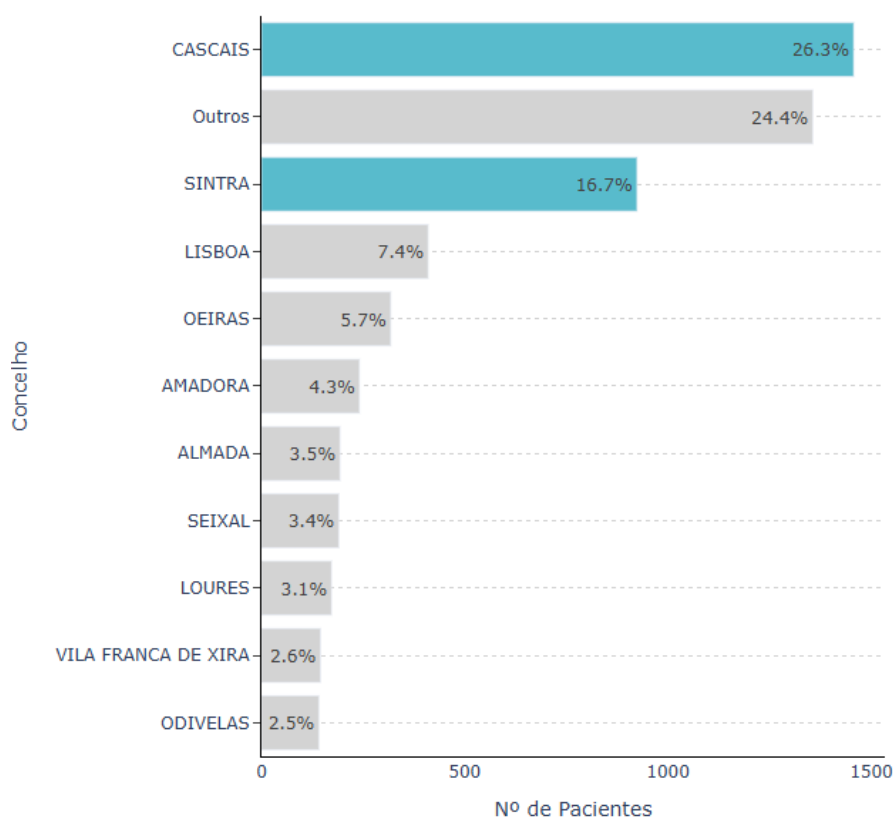


Figura 13: Distribuição dos utentes que compareceram a uma 1ª consulta, por concelho

A variável *nacionalidade* revela que cerca de 96.5% dos pacientes têm nacionalidade portuguesa, enquanto os restantes se distribuem por vários outros países, não havendo destaque substancial em nenhum dos outros. No gráfico representado na figura 14, observa-se que apenas 3.5% dos pacientes são de diferentes nacionalidades, representando uma minoria.

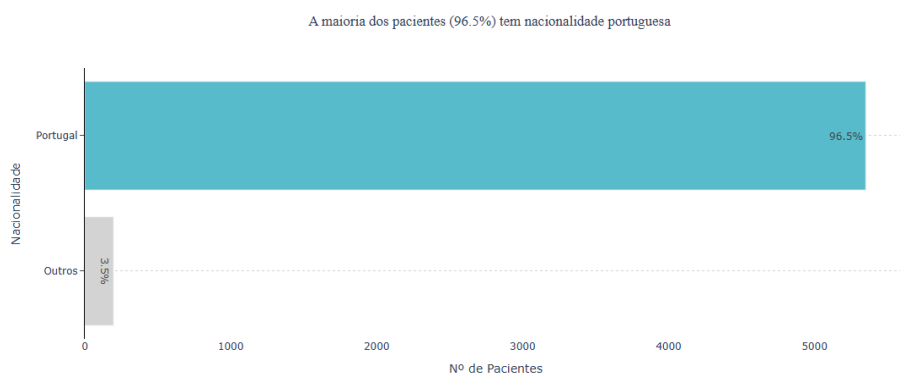


Figura 14: Distribuição dos utentes que compareceram a uma 1ª consulta, por país de origem

Na variável *SER* é possível analisar que o serviço de adultos concentra um número maior de pacientes quando em comparação com o serviço pediátrico. Especificamente, os serviços para adultos representam 75.9% do total de pacientes, sendo os restantes

pacientes acompanhados no serviço pediátrico. No entanto, dentro do serviço de adultos, aquele com mais pacientes é o direccionado para amputações, acidentes vasculares cerebrais ou traumatismos crânio-encefálicos, conforme se pode observar na figura 15.

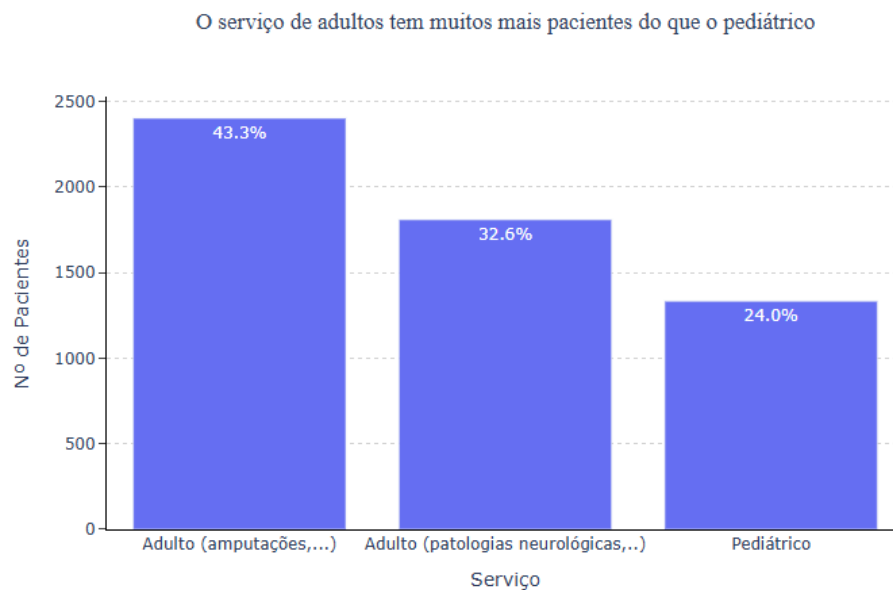


Figura 15: Distribuição por serviço dos utentes que compareceram a uma 1ª consulta

Em relação às *entidades financeiras responsáveis* destaca-se a entidade “ARSLVT-ACORDO”, com 2615 pacientes, o que corresponde a 47.1% do total. No gráfico que se segue observam-se as percentagens referentes às 4 entidades mais comuns entre os pacientes. As restantes entidades estão agrupadas na categoria “outros”, que representa 9.4% de pacientes. (Figura 16)

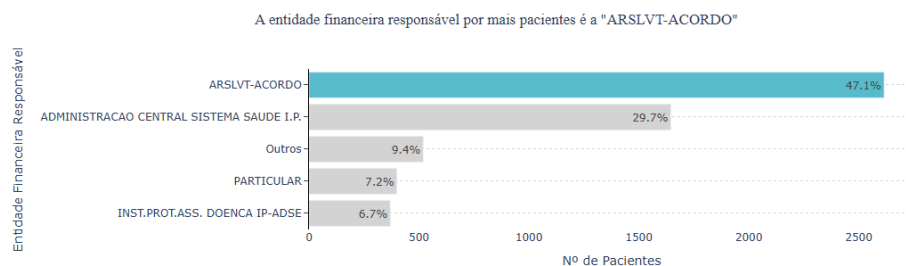


Figura 16: Distribuição das entidades financeiras responsáveis pelos utentes que compareceram a uma 1ª consulta

Tal como se pode observar na figura 17, a maioria das primeiras consultas marcadas, mais especificamente 88.6%, está associada à existência de doença por parte do paciente. Ainda que em menor número, também se registam consultas motivadas por acidentes de viação, acidentes de trabalho e outras causas. Assim sendo, a *Causa* principal para a marcação de uma primeira consulta é maioritariamente devido a problemas de saúde.

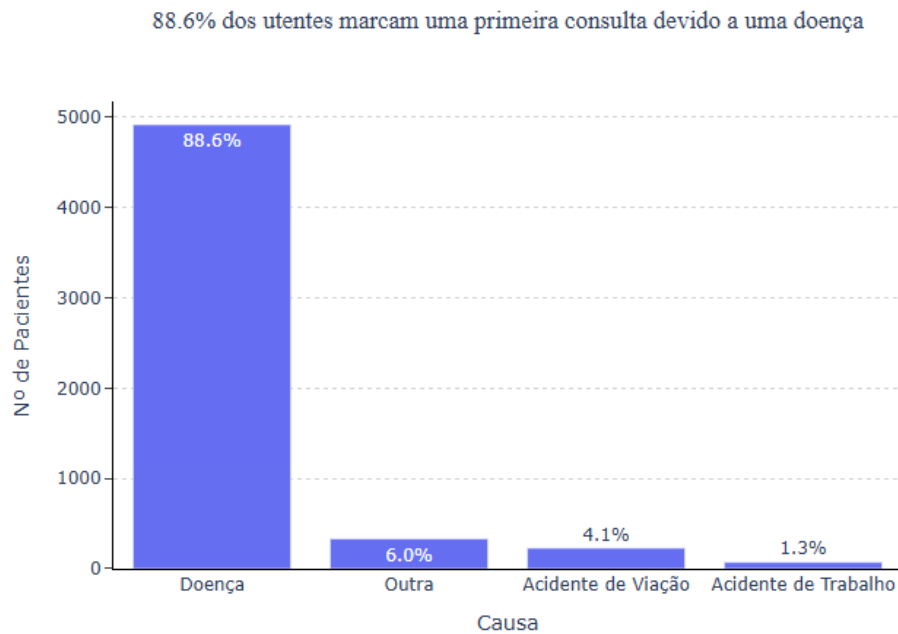


Figura 17: Distribuição das causas que deram origem aos utentes a comparecerem a uma 1ª consulta

Quanto às *tipologias*, apesar de serem apenas 10 existe uma grande discrepância na quantidade de pacientes que a elas pertencem. No gráfico da figura 18, é possível observar que a reabilitação neurol-pat encefálica tem quase 40% da quantidade de pacientes total.

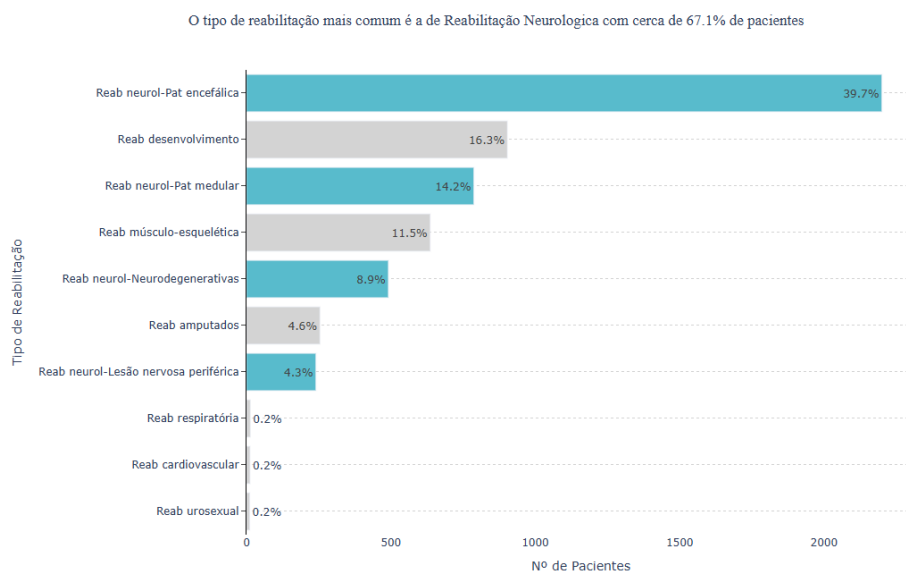


Figura 18: Distribuição por tipologia dos utentes que compareceram a uma 1ª consulta

O gráfico da figura 19 permite observar a distribuição das 10 patologias mais comuns. A *patologia* mais comum é “lesão encefálica não traumática (AVC)” representando 19.8% do total dos pacientes. As diferenças na distribuição das patologias são bem visíveis no gráfico, sendo a categoria “outros” uma agregação das restantes 31 patologias com percentagens mais baixas.

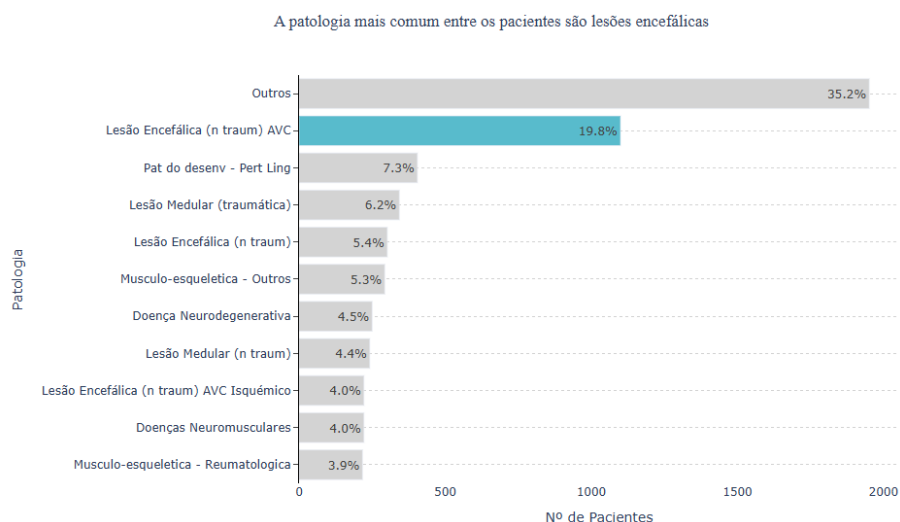


Figura 19: Distribuição por patologia dos utentes que compareceram a uma 1ª consulta

Estas consultas são dadas por diversos profissionais, no gráfico da figura 20 é possível observar os profissionais que mais consultas dão e as diferenças na quantidade de pacientes que cada profissional atende. Dos 26 profissionais, através da análise do gráfico, é possível identificar os 3 que mais pacientes atendem, sendo estes os profissionais com IDs 27, 3 e 17, respetivamente.

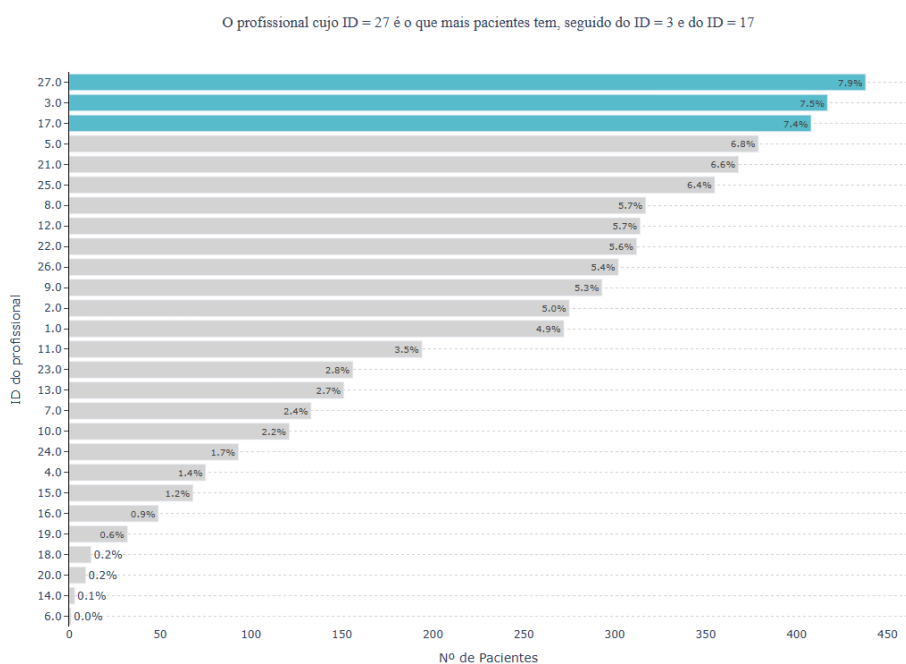


Figura 20: Distribuição dos profissionais responsáveis pelo atendimento dos utentes que compareceram a uma 1ª consulta

6.3.2 Dados 1ªs consultas e sessões

A tabela *1ªs consultas* foi unida à tabela *Sessões* através do número de identificação do paciente. A tabela "Sessões" acrescenta:

- o dia em que a sessão foi realizada (ou eventualmente duas sessões);
- os meios complementares terapêuticos realizados (no máximo de 11);
- o profissional de saúde envolvido.

As variáveis cujo nome inclui "_1ªconsulta" ou "_sessões" indicam que essas variáveis não são comuns às duas tabelas, sendo exclusivas de uma delas.

Esta nova tabela contém 57138 linhas e 26 colunas, não contendo nem valores em falta nem valores duplicados.

A união dos dados em 1ª consulta com sessões permite analisar para cada paciente que tenha comparecido a uma 1ª consulta (em internamento ou não) o plano de sessões realizado, i.e. o dia, o número de sessões, e o número de meios complementares terapêuticos realizados. Concretamente a tabela 1ª consulta tem 6675 linhas, e a tabela 1ª consulta unida com sessões tem 57138 linhas. A análise desta última tabela considera:

- Género
- Concelho
- Nacionalidade
- Serviço (Adulto(amputações...), Adulto(patologias neurológicas...), Pediatria)
- Número de sessões (uma ou duas)
- Número de meios complementares terapêuticos (entre 1 e 11)
- Entidade financeira responsável
- Causa que leva a uma 1ª consulta
- Tipologia (tipos de reabilitação)
- Patologia
- ID do profissional responsável pela 1ª consulta
- ID do profissional responsável pela sessão
- Actos (consultas de avaliação para admissão acompanhamento (Adultos), consultas de avaliação para admissão acompanhamento (Pediátrico),...)
- Serviços2 (cada serviço contém outros serviços, a designação Serviço2 representa esta subdivisão)
- Diagnóstico

A união entre estas tabelas faculta um número muito maior de dados que caracterizam sessões. Ainda assim existe uma repetição de variáveis (exemplo: género, concelho, diagnóstico,...) presentes em 1ª consulta. Por isso vamos apresentar a análise de algumas destas variáveis em anexo.

É possível concluir, na figura 21, que entre todos os pacientes, apenas 0.2% atenderam a 2 sessões.

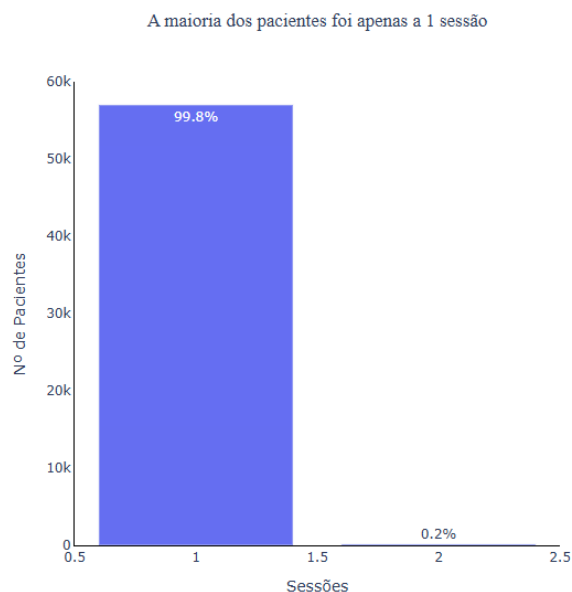


Figura 21: Distribuição das sessões por utentes que posteriormente à 1ª consulta realizaram sessões

Cada sessão pode ser composta por vários meios complementares terapêuticos. No gráfico que se segue, figura 22, é possível analisar que 37.8% dos pacientes atendem a sessões compostas por apenas 1 meio complementar terapêutico.

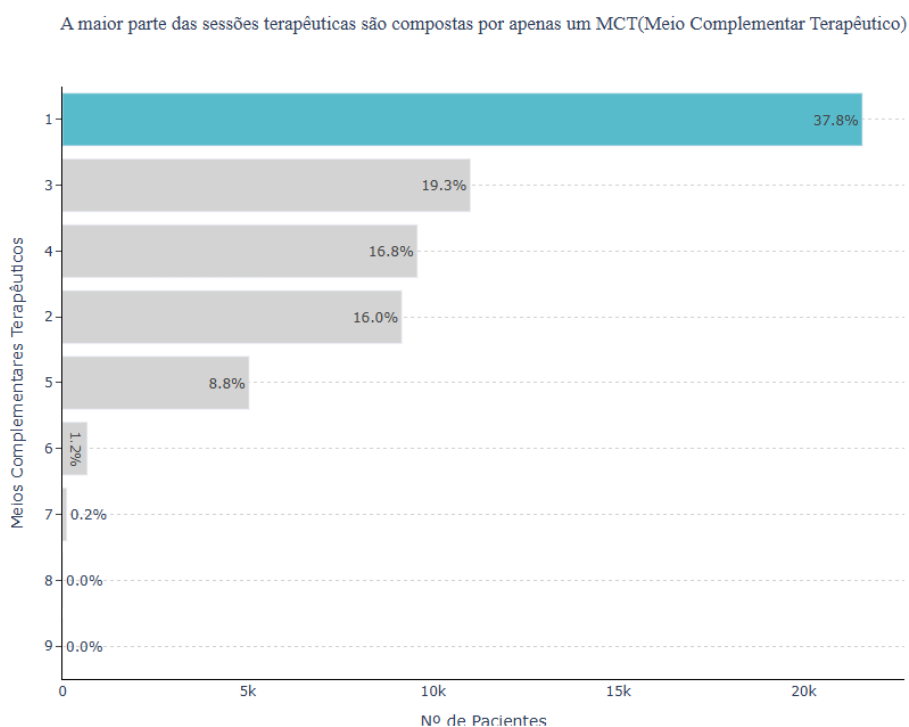


Figura 22: Distribuição dos meios complementares terapêuticos dos utentes que posteriormente à 1ª consulta realizaram sessões

Quanto aos profissionais responsáveis, podemos ver que são menos os profissionais responsáveis por 1ªs consultas, apenas 24, quando em comparação com os profissionais responsáveis pelas sessões, 111. Nas 1ªs consultas (em anexo) observamos que os

3 profissionais responsáveis por mais pacientes são os com os IDs = 5.0, 27.0 e 3.0 enquanto os responsáveis por mais pacientes nas sessões (Figura 23) são os com IDs = 39, 135, 151.

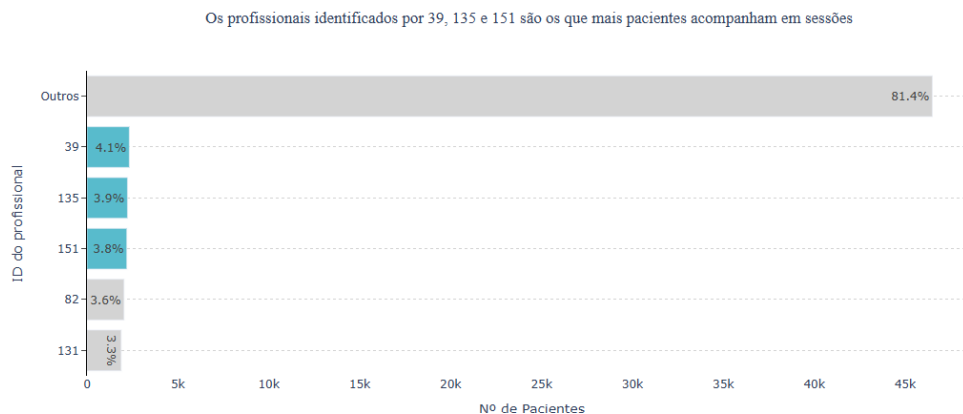


Figura 23: Distribuição dos profissionais responsáveis pelas sessões dos utentes que posteriormente à 1ª consulta realizaram sessões

6.3.3 Dados 1ªs consultas e internamentos

Considerando agora a tabela criada através da junção dos dados relativos às *1ªs consultas* e a *internamentos*:

- Esta nova tabela contém 819 linhas e 28 colunas.
- Não tem dados duplicados nem valores em falta.

A união dos dados em *1ª consulta* com *internamentos* permite analisar para cada paciente que tenha comparecido a uma 1ª consulta e ficado em internamento o seguinte:

- Serviços (relativos a 1ª consulta e internamento);
- Género;
- Concelho;
- Nacionalidade;
- Idade;
- Entidade financeira responsável;
- Actos (consultas de avaliação para admissão acompanhamento (Adultos), consultas de avaliação para admissão acompanhamento (Pediátrico),...);
- Serviços2 (cada serviço contém outros serviços, a designação Serviço2 representa esta subdivisão);
- Diagnóstico;
- Patologia;

- Causa;
- Tipologia (tipos de reabilitação);
- ID do profissional responsável pela 1ª consulta e internamento;
- Tipo Consulta.

A união entre estas duas tabelas facultava um número menor de dados que caracterizam internamentos. Ainda assim existe uma repetição de variáveis (exemplo: género, concelho, diagnóstico,...) presentes em 1ª consulta. Por isso vamos apresentar a análise de algumas destas variáveis em anexo.

De modo a realizar uma análise mais aprofundada dos dados desta nova tabela foram desenvolvidas visualizações adicionais.

As figuras 24, 25 permitem observar que:

- A união entre doentes adultos (amputações,...) e adultos (patologias neurológicas,...) verifica uma diminuição de 0.6% entre 1ª consulta e internamento.
- Os doentes pediátricos registam um aumento de 0.6% entre 1ª consulta e internamento.

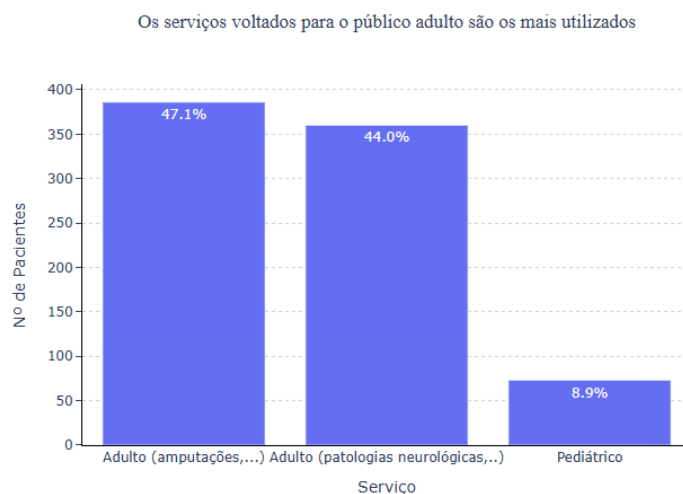


Figura 24: Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por serviços (na 1ª consulta)

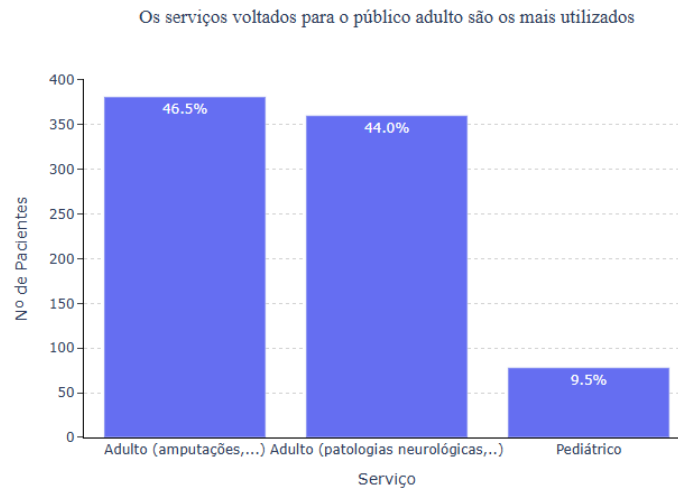


Figura 25: Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por serviços (no internamento)

Já as entidades financeiras responsáveis são distintas para as variáveis: *EFR_1ªconsulta* e *EFR_internamento*.

A entidade financeira responsável mais utilizada pelos pacientes para compartilhar as 1ªs consultas é a “ARSLVT-ACORDO” com 61.9%, seguida da entidade “ADMINISTRACAO CENTRAL SISTEMA SAUDE I.P.” com 30.5%, tendo todas as restantes entidades um valor igual ou inferior a 2% de pacientes. (Figura 26)

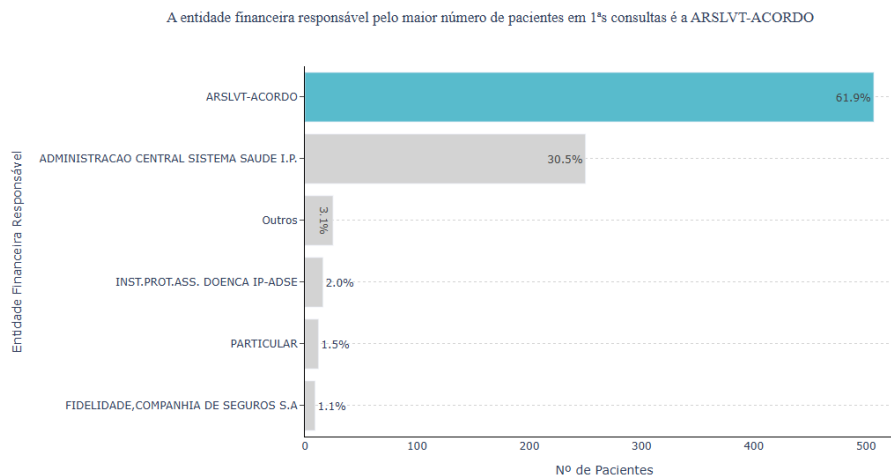


Figura 26: Distribuição das entidades financeiras responsáveis pelas 1ª consulta dos utentes que posteriormente à 1ª consulta ficaram em internamento

Já para a comparticipação de internamentos, a entidade mais utilizada é a “ADMINISTRACAO CENTRAL SISTEMA SAUDE I.P.” com 92.7% do total de pacientes, sendo que as restantes entidades contêm todas menos de 2.5% dos pacientes. (Figura 27)

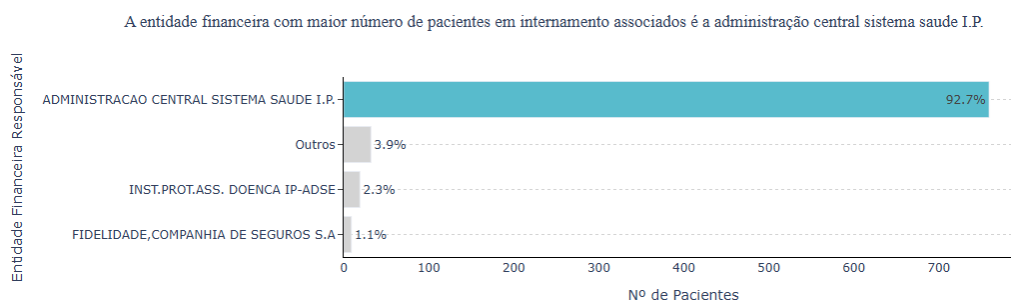


Figura 27: Distribuição das entidades financeiras responsáveis pelos internamentos dos utentes que posteriormente à 1ª consulta ficaram em internamento

A figura 28 descreve a identificação do profissional de saúde responsável pela 1ª consulta e internamento de um paciente. Uma vez que o número de 1ªs consultas é inferior ao número de internamentos, o número de pacientes por profissional é também menor. Concluimos igualmente que que o profissional com o ID = 5.0 é o responsável por um maior número de pacientes.

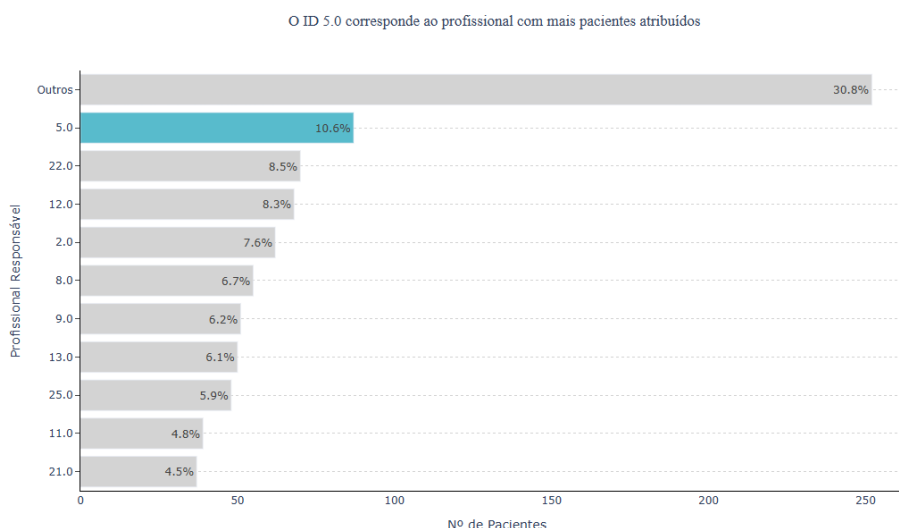


Figura 28: Distribuição dos profissionais responsáveis pelo atendimento dos utentes que posteriormente à 1ª consulta ficaram em internamento

Por último, relativamente a esta tabela, foi possível ainda, através de um histograma, entender melhor a distribuição de idades e perceber o porquê do serviço direcionado a adultos ter mais pacientes. Aqui observamos que a faixa etária entre os 55 e os 59 é a mais comum entre os pacientes. (Figura 29)

A distribuição etária dos pacientes mostra uma maior concentração entre os 55 e os 59 anos

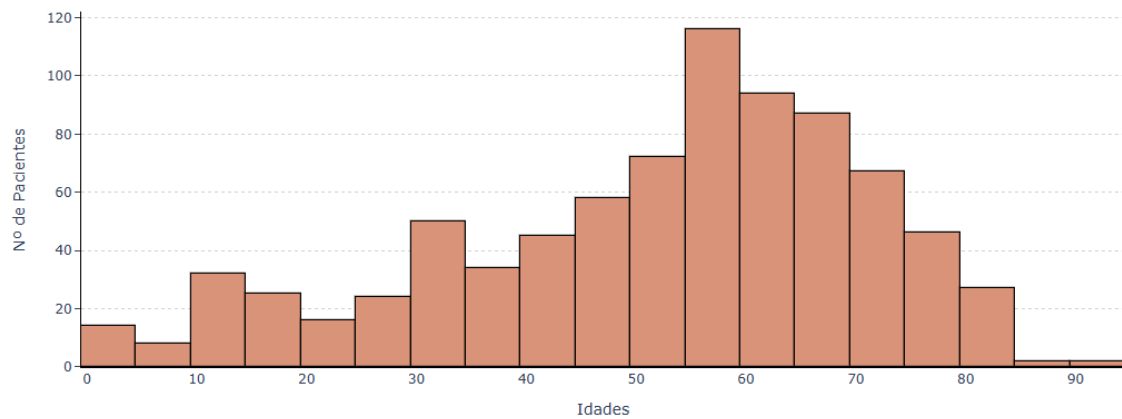


Figura 29: Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento

6.3.4 Dados 1ªs consultas e info de internamentos

Através da junção das tabelas relativas aos dados de *1ªs consultas* e aos dados da *informação dos internamentos* obtivemos uma nova tabela:

- Com 819 linhas e 28 colunas.
- Sem valores duplicados ou em falta.

A junção dos dados em *1ª consulta* com *info internamentos* permite analisar para cada paciente que tenha comparecido a uma 1ª consulta e ficado em internamento as informações mais detalhadas relativas a esse internamento, através das variáveis:

- SER_1ªconsulta (Serviço relativo à primeira consulta)
- Género_1ªconsulta
- DNasc
- Nacionalidade
- Concelho_1ªconsulta
- CodEFR
- EFR_1ªconsulta (Entidade Financeira Responsável)
- CodActo
- Designação Acto
- Data Consulta
- CodServ
- Serviço2

- CodDiag
- Diagnóstico
- Tipo Consulta
- Cod Pat
- Patologia
- Causa
- Tipologia
- ID_Prof
- Mes-Ano
- Género_info_internamento
- Concelho_info_internamento
- DataInternamento
- SER_info_internamento
- Idade
- EFR_info_internamento

A união entre estas duas tabelas faculta um número menor de dados que caracterizam internamentos. Ainda assim existe uma repetição de variáveis (exemplo: género, concelho, diagnóstico,...) presentes em 1ª consulta. Por isso vamos apresentar a análise de algumas destas variáveis em anexo.

Nas seguintes visualizações pode-se obter uma análise mais aprofundada relativa a esta nova tabela (dados 1ª consulta + info internamento).

Os serviços, distribuídos por 3 grupos diferentes, dão-nos a informação que a grande maioria dos pacientes estão no serviço de adultos direcionado a amputações e outras patologias. Através da análise das duas variáveis: “SER_1ªconsulta” e “SER_info_internamento”, observamos ainda um aumento no serviço pediátrico na parte das informações sobre os internamentos o que pode querer dizer que alguns pacientes podem-se ter enganado a marcar a 1ª consulta indo assim, por engano, ao serviço de adultos. (Figuras 30 e 31)

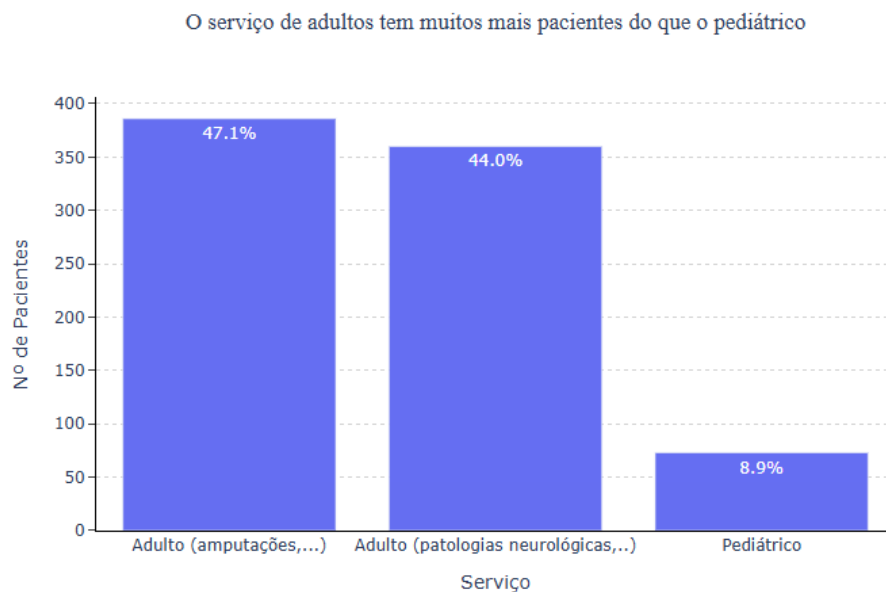


Figura 30: Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por serviços (1ª consulta)

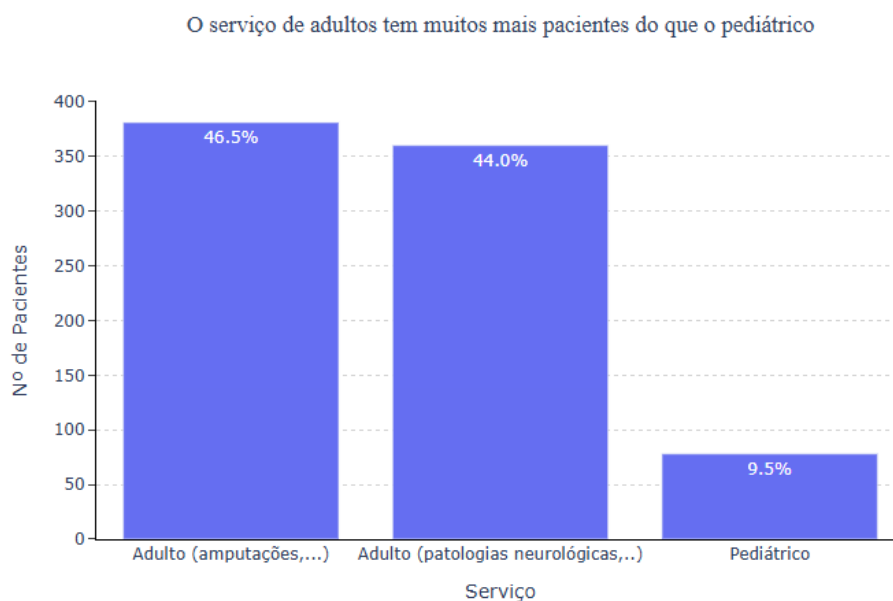


Figura 31: Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por serviços (info internamento)

Alguns pacientes utilizam certos seguros de modo a compartilhar as consultas/internamento. Para as primeiras consultas são muitos os pacientes que utilizam a entidade “ARSLVT-ACORDO” enquanto que, para os internamentos, a grande maioria dos pacientes utilizam a entidade “ADMINISTRACAO CENTRAL SISTEMA SAUDE I.P.”. (Figuras 32 e 33)

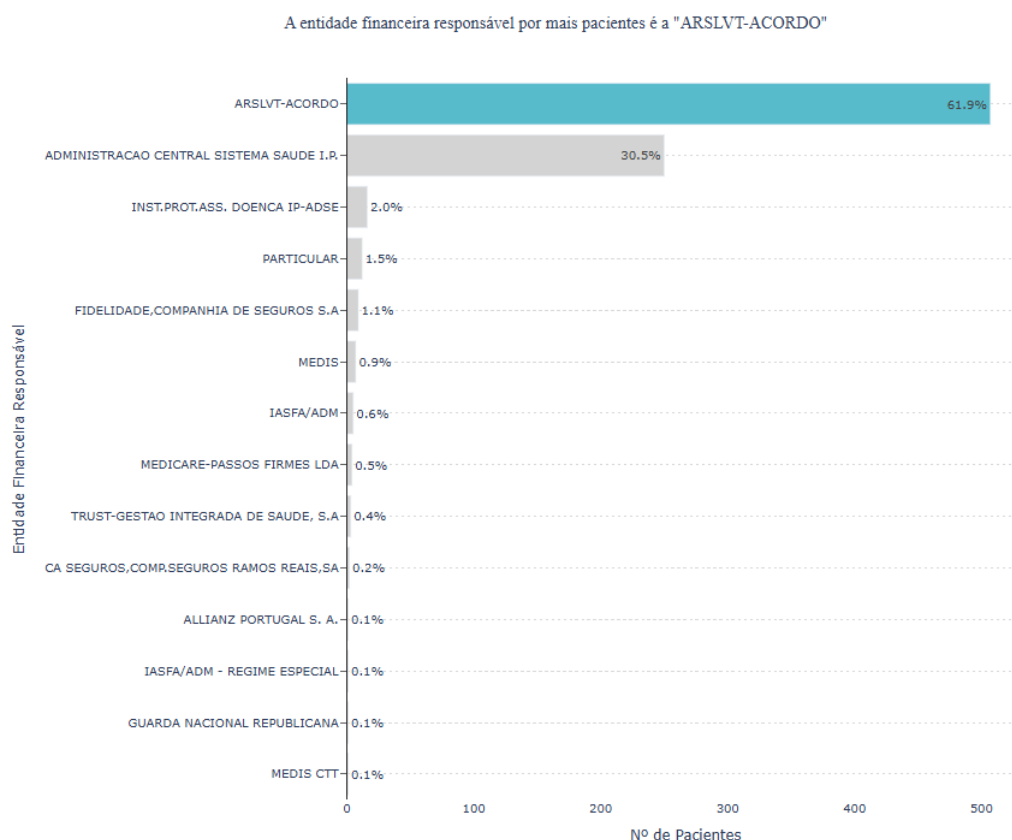


Figura 32: Distribuição das entidades financeiras responsáveis pela 1ª consultas dos utentes que posteriormente à 1ª consulta ficaram em internamento

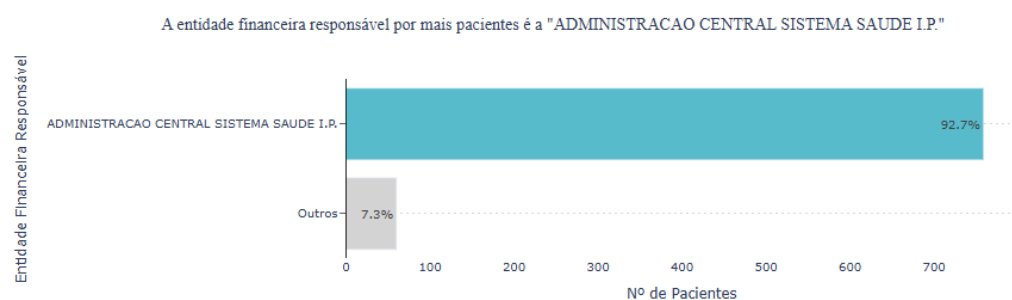


Figura 33: Distribuição das entidades financeiras responsáveis pelos internamentos dos utentes que posteriormente à 1ª consulta ficaram em internamento

As designações de actos mais comuns são as consultas de avaliação para admissão de adultos contendo cerca de 79.9% dos pacientes como se observa na figura 34, podemos também observar que as consultas de medicina interna são as menos comuns.

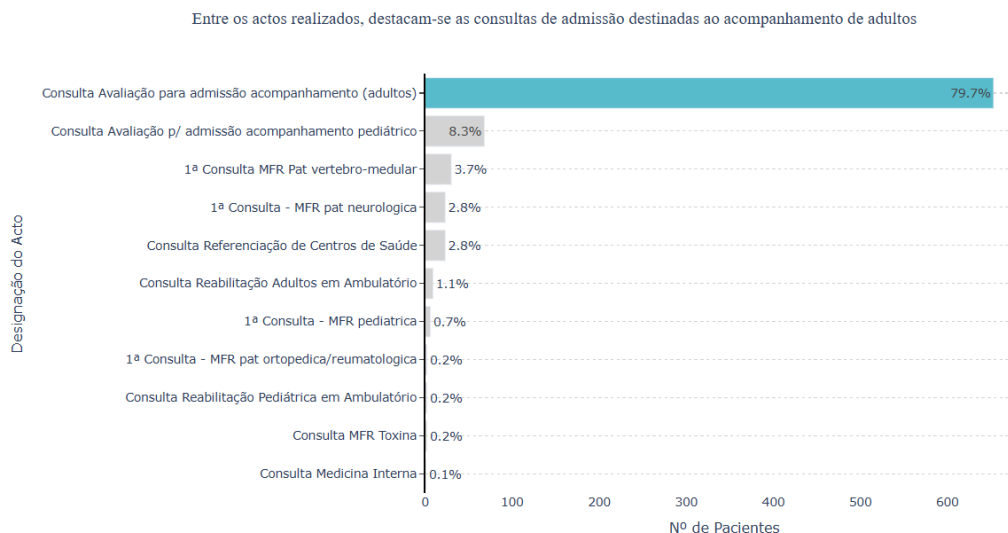


Figura 34: Distribuição dos actos realizados pelos utentes que posteriormente à 1ª consulta ficaram em internamento

O diagnóstico mais comum, tal como é possível observar na figura 35, é a seq de doença cerebrovascular com cerca de 37.2% de pacientes, os restantes 62.8% de pacientes estão distribuídos pelas restantes categorias.

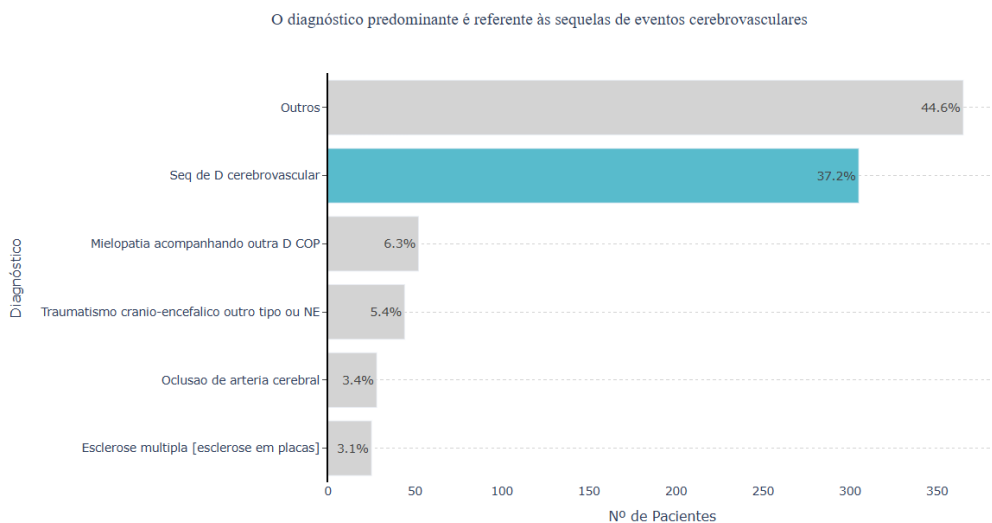


Figura 35: Distribuição dos diagnósticos obtidos pelos utentes que posteriormente à 1ª consulta ficaram em internamento

Já as patologias apesar de serem várias a mais comum é a “lesão encefálica (n traum) AVC” com 34.4% dos pacientes, estando os restantes distribuídos pelas 32 restantes categorias. (Figura 36)

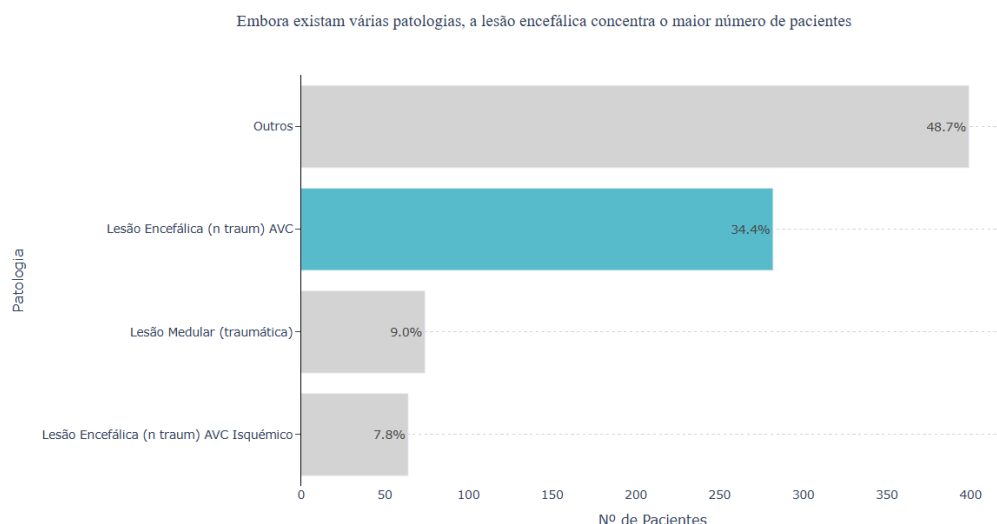


Figura 36: Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por patologia

6.3.5 Dados sessões e internamentos

De modo a analisar as sessões realizadas para utentes em internamento, uniram-se as tabelas relativas a *sessões* e *internamento*. A tabela resultante contém 22438 linhas e 11 colunas, sem dados duplicados nem valores em falta.

A junção dos dados em internamentos e sessões permite analisar para cada paciente internado as sessões realizadas, em particular:

- Serviços (relativos ao internamento);
- Género;
- Concelho;
- Idade;
- Entidade financeira responsável;
- Número de sessões (uma ou duas)
- Número de meios complementares terapêuticos (entre 1 e 11)
- ID do profissional responsável pelas sessões;

A união entre as duas tabelas faculta um número menor de dados do que a junção entre "1ª consulta" e "sessões" (respetivamente, 22438 vs. 57138). Existem variáveis em comum (exemplo: género, concelho,...), de modo que vamos apresentar a análise de algumas destas variáveis em anexo.

Seguem-se algumas visualizações relativas à nova tabela (dados internamentos + sessões).

Os pacientes encontram-se, maioritariamente, entre as faixas etárias dos 55 aos 59 anos, dos 60 aos 64 anos e dos 70 aos 74 anos. (Figura 37)

As faixas etárias mais predominantes situam-se entre os 55-59, 60-64 e 70-74

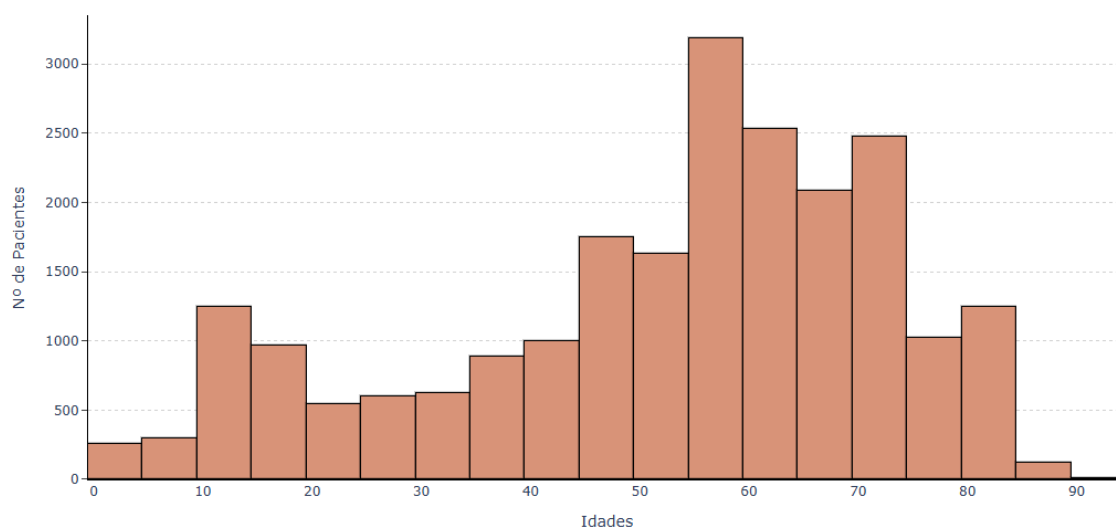


Figura 37: Distribuição das idades dos utentes que posteriormente ao internamento realizaram sessões

Dos 3 serviços existentes é possível concluir, através da observação da figura 38, que o serviço de adultos tem mais pacientes que o serviço pediátrico, com uma diferença de 87.3% entre os serviços.

O serviço de adulto é o que contém a grande maioria dos utentes

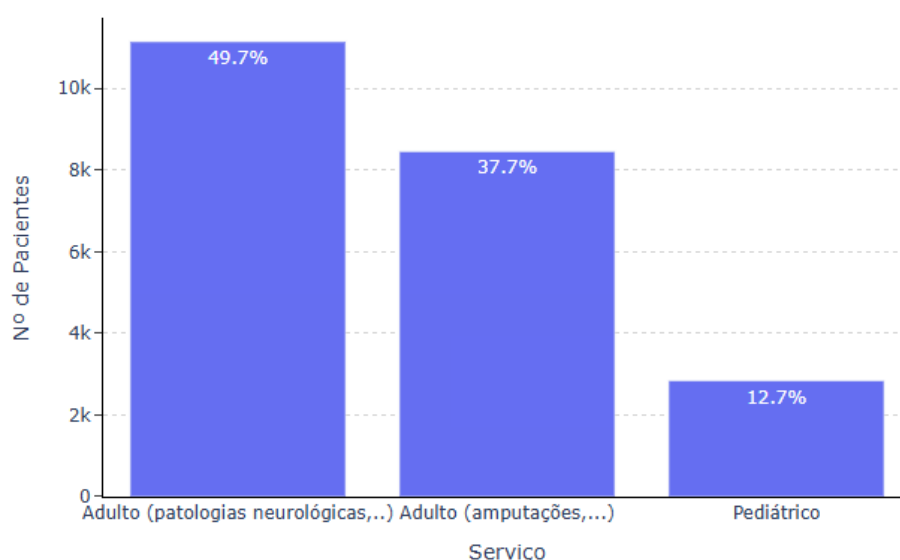


Figura 38: Distribuição dos serviços em que são acompanhados os utentes que posteriormente ao internamento realizaram sessões

A grande maioria dos utentes comparecem a apenas uma sessão (Figura 39). Essas sessões podem ser compostas por vários meios complementares terapêuticos. Assim sendo, 37.3% dos pacientes comparecem em sessões com apenas um meio complementar terapêutico (Figura 40).

Apenas 0.1% dos pacientes atendem a 2 sessões

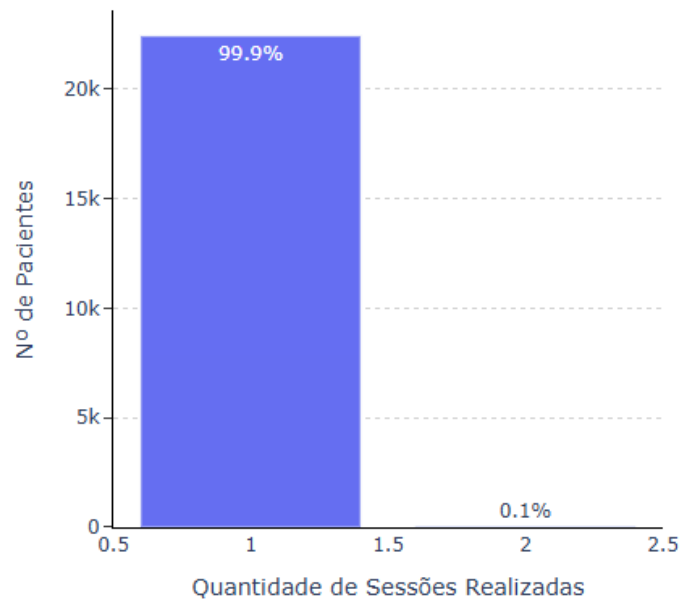


Figura 39: Distribuição dos utentes que posteriormente ao internamento realizaram sessões, por quantidade de sessões realizadas

37.3% dos pacientes atendem a sessões compostas por apenas 1 método complementar terapêutica

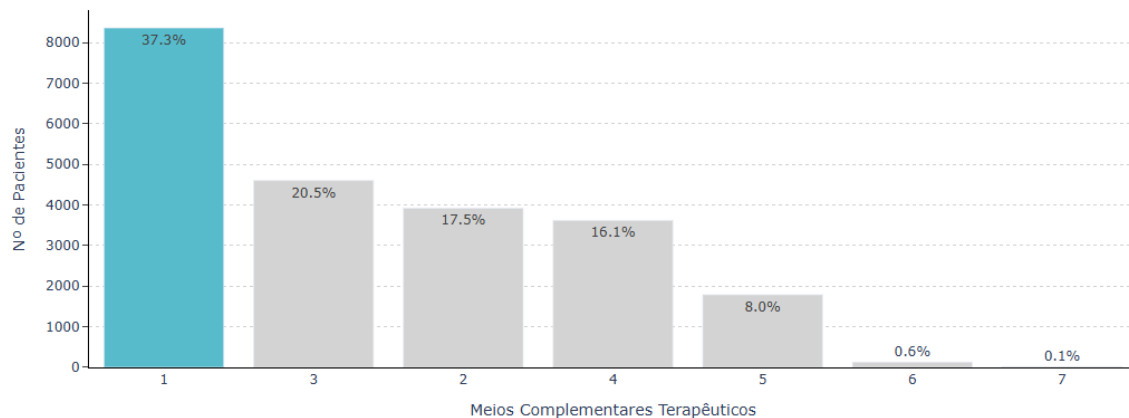


Figura 40: Distribuição dos meios complementares terapêuticos dos utentes que posteriormente ao internamento realizaram sessões

6.3.6 Dados sessões e info internamentos

Através da junção das tabelas *Sessões* e *info internamentos* obtivemos uma nova tabela com 22592 linhas e 20 colunas, sem valores em falta nem valores duplicados e com as seguintes variáveis:

- Dia
- Sessões

- MCT's
- ID_Prof_sesoes
- Género
- DNasc
- Data Inicio
- Data Fim
- Data Diag
- Cod_Diag
- Diagnostico
- ID_Prof_info_internamentos
- Serviço
- EFR
- CodEFR
- Pat_Nome
- DestinoPosAlta
- MIFAdm
- MIFalta

As seguintes visualizações têm como objetivo dar a conhecer algumas informações sobre esta nova tabela.

Começando pelas sessões observa-se, no gráfico da figura 41 que a grande maioria dos pacientes, 99.9%, apenas realizaram 1 sessão. E, dos pacientes que atendem a sessões, 38.9% dos pacientes atendem a sessões compostas por apenas um método complementar terapêutico. Apenas 0.1% dos pacientes atendem a sessões compostas por 7 meios complementares terapêuticos. (Figura 42)

A maioria dos pacientes atende a apenas 1 sessão

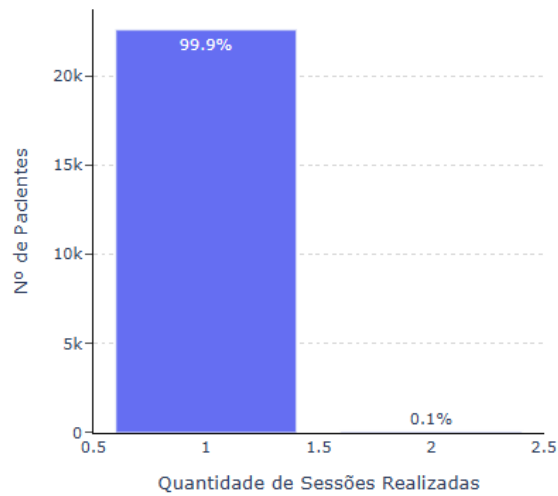


Figura 41: Distribuição das sessões por utentes que realizaram sessões

38.9% dos pacientes atendem a sessões compostas por apenas 1 método complementar terapêutico

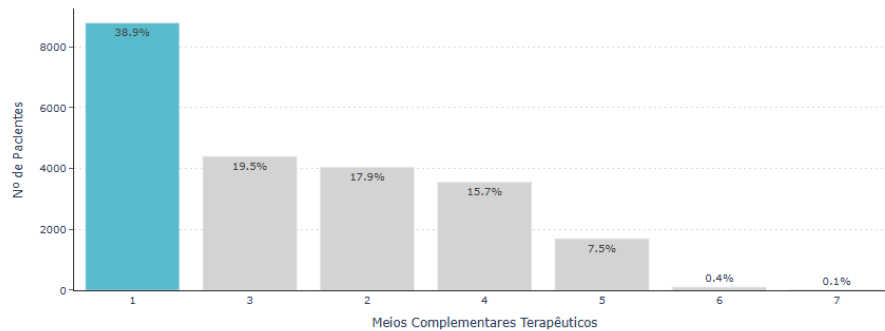


Figura 42: Distribuição dos meios complementares terapêuticos realizados pelos utentes

Essas sessões são realizadas com o apoio de vários profissionais, no entanto alguns sobressaem-se na quantidade de pacientes por quem são responsáveis, tal como é o caso do profissional com ID = 82. (Figura 43)

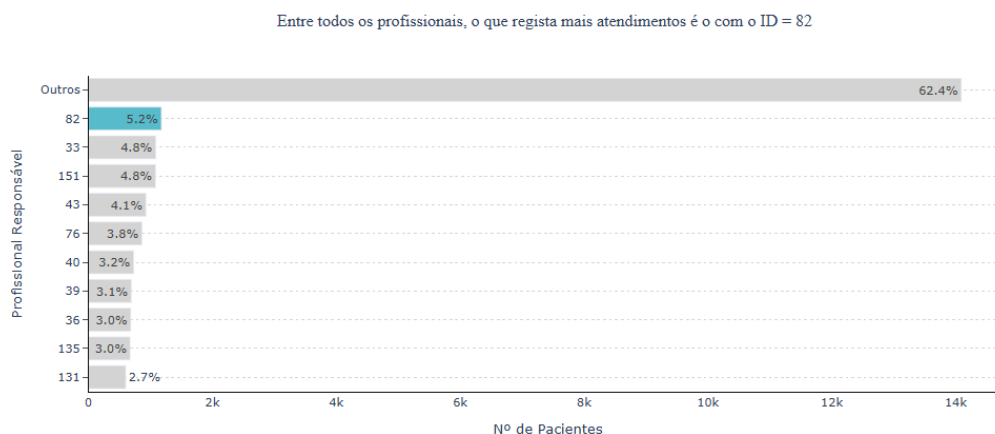


Figura 43: Distribuição dos profissionais responsáveis pelas sessões realizadas pelos utentes

Já nos profissionais de internamentos o profissional que se destaca é o com ID = 23, tal como se pode observar na figura 44.

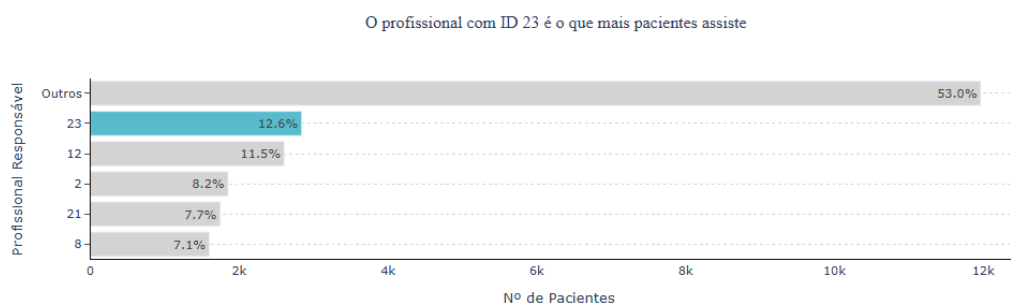


Figura 44: Distribuição dos profissionais responsáveis pelos utentes em internamento

Como destino pós alta, no gráfico da figura 45, 73.8% dos pacientes seguem para o domicílio.

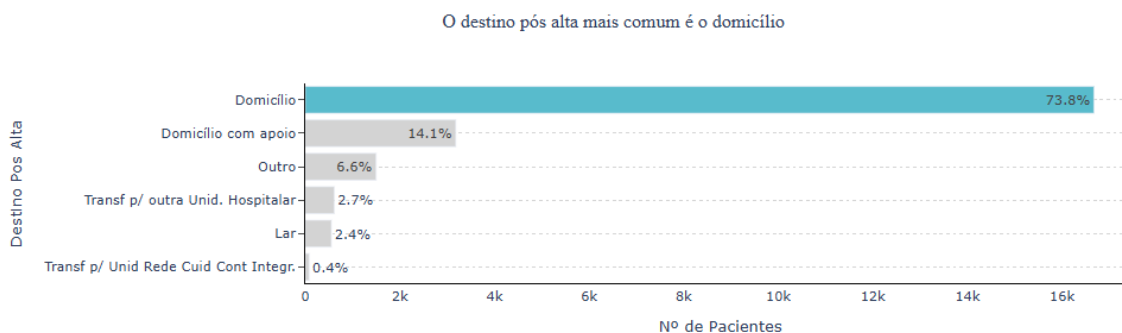


Figura 45: Distribuição sobre o destino pós alta dos utentes

Através das variáveis "MIFalta" e "MIFAdm" é possível observar que a classificação mais comum da independência funcional no momento da admissão é de 113, tal como se observa na figura 46 e que a classificação mais comum da independência funcional no momento da alta é de 114, tal como se observa na figura 47

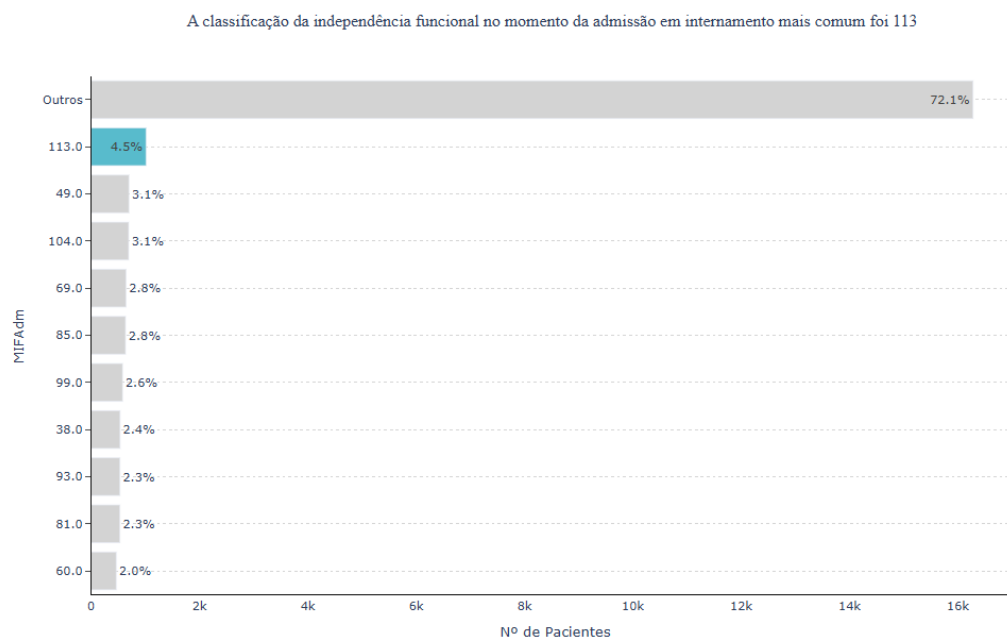


Figura 46: Distribuição da classificação da independência funcional no momento da admissão dos utentes

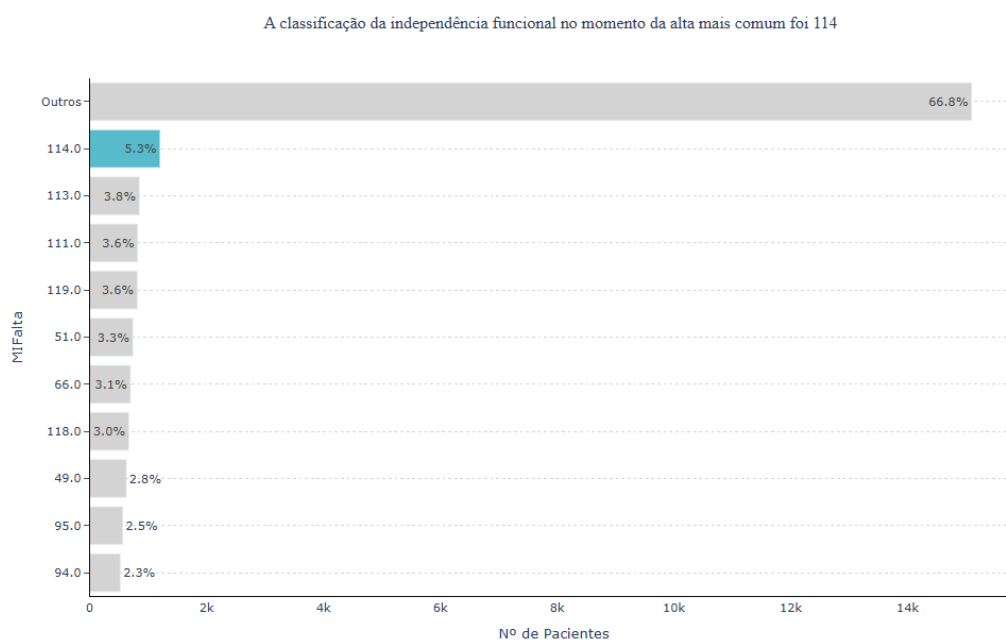


Figura 47: Distribuição da classificação da independência funcional no momento da alta dos utentes

6.3.7 Dados internamento e info internamentos

Através da junção das tabelas relativas aos dados de *internamento* e de *info internamentos* foi obtida uma nova tabela com 2196 linhas e 22 colunas, sem dados nulos nem valores em falta.

A união dos dados em internamentos com informação de internamentos permite ana-

lisar para cada paciente que tenha ficado internado as informações relativas a esse acontecimento através das variáveis:

- Género_internamento
- Concelho
- DataInternamento
- SER
- Idade
- EFR_internamento
- Género_info_internamentos
- DNasc
- Data Inicio
- Data Fim
- Data Diag
- Cod_Diag
- Diagnostico
- ID_Prof
- Serviço
- EFR_info_internamentos
- CodEFR
- Pat_Nome
- DestinoPosAlta
- MIFAdm
- MIFalta

A união entre estas duas tabelas faculta um número menor de dados que caracterizam internamentos. Ainda assim existe uma repetição de variáveis (exemplo: género, concelho, diagnóstico,...) presentes em 1ª consulta. Por isso vamos apresentar a análise de algumas destas variáveis em anexo.

Para ter uma melhor compreensão sobre a tabela seguem-se algumas visualizações.

O destino pós alta mais comum é o domicílio com cerca de 69.4% de pacientes. (Figura 48)

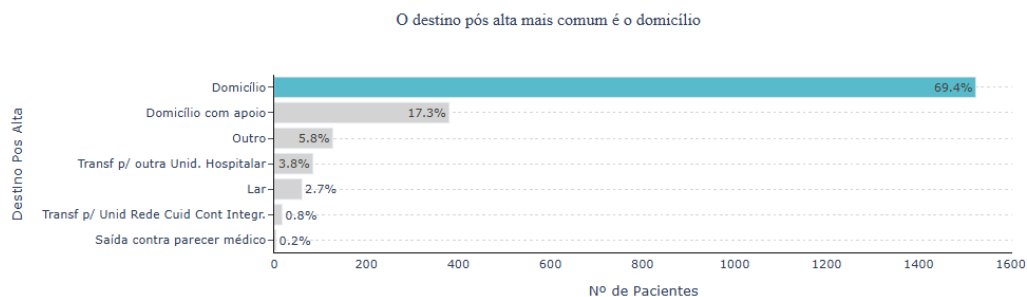


Figura 48: Distribuição do destino pós alta dos utentes em internamento

Por último, através das informações que as colunas “MIFalta” e “MIFAdm” nos fornecem, podemos perceber que a classificação de independência funcional no momento da admissão é em 1.9% dos pacientes 107, estando os restantes pacientes distribuídos entre todas as outras classificações, figura 49. Já na classificação da independência funcional no momento da alta é de 111, em 3.1% dos pacientes (figura 50).

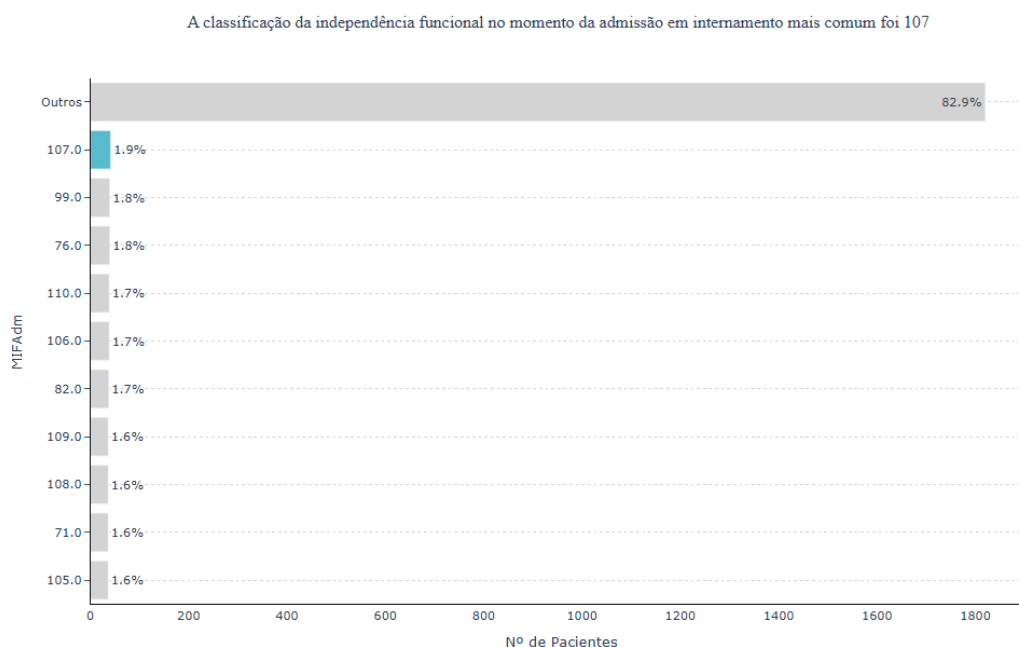


Figura 49: Distribuição da classificação da independência funcional no momento da admissão dos utentes em internamento

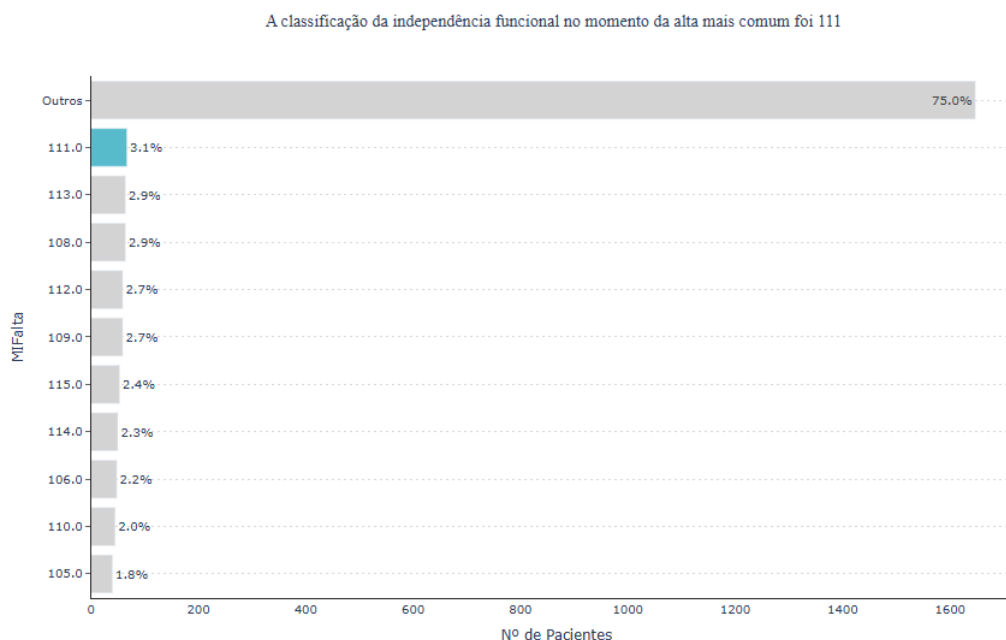


Figura 50: Distribuição da classificação da independência funcional no momento da alta dos utentes em internamento

6.3.8 Junção das 4 tabelas - 1^as consultas + internamento + sessões + info internamentos

Por último, de modo analisar efetivamente os dados todos juntos e a chegar ao objetivo pretendido pelo trabalho, foi necessário juntar as 4 tabelas.

Esta nova tabela contém 44386 linhas e 47 colunas, sem dados duplicados nem valores em falta. Foi ainda necessário transformar alguns dados em datetime para que fosse possível fazer alguns dos gráficos de linhas.

De todas as visualizações realizadas muitas já foram várias vezes faladas ao longo do projeto, como o género, concelhos, serviços entre muitos outros e, por esse motivo, nesta secção só estarão as visualizações mais importantes, as restantes estarão no anexo.

Começando pela distribuição dos internamentos ao longo dos anos, no gráfico representado na figura 51 é possível observar que desde janeiro 2022 a janeiro de 2025 existiram muitas oscilações, não sendo possível tirar conclusões sobre as alturas do ano que são mais comuns, no entanto observa-se um pico substancial em maio 2023 com cerca de 4000 internamentos apenas nesse mês.

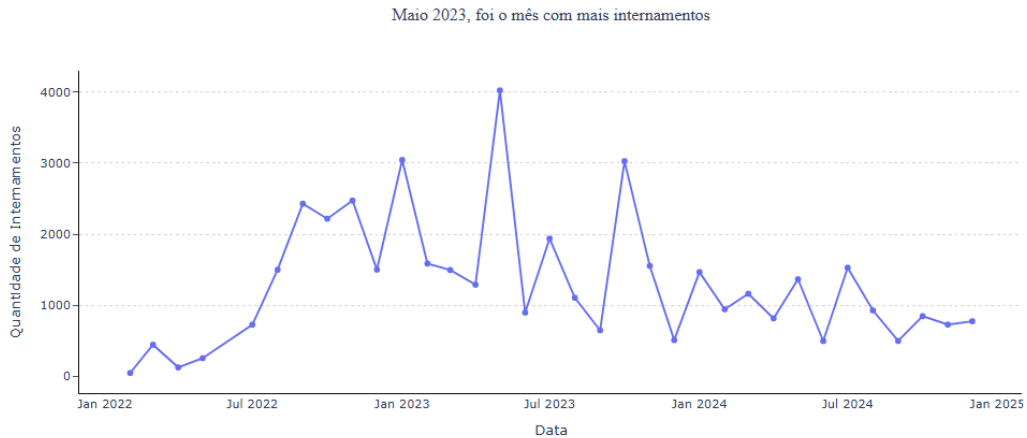


Figura 51: Evolução dos números de internamentos

Quanto a duração de cada internamento, esta distribui-se entre os 0 e os 250 dias. Na figura 52 é possível observar que a grande maioria dos internamentos tem uma duração entre os 50 e 100 dias, sendo que internamentos com mais de 150 dias são muito raros.

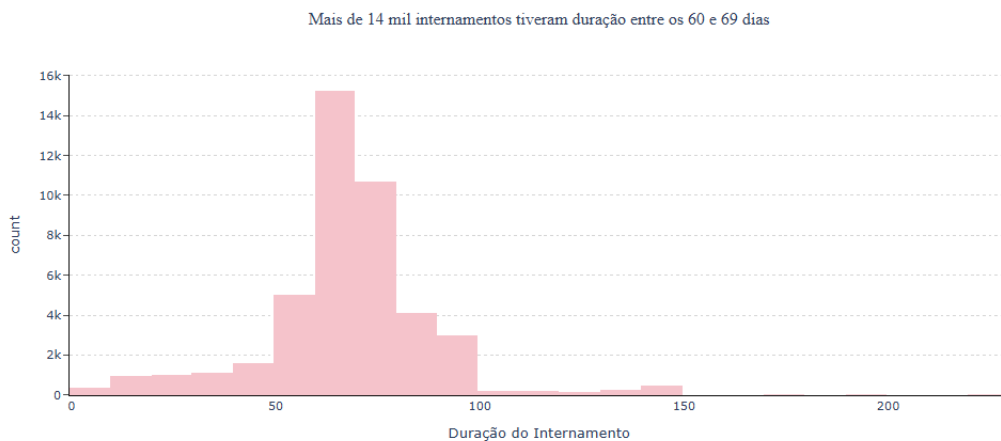


Figura 52: Distribuição das durações dos internamentos

Em termos de idade podemos observar, na figura 53, que se verifica que ao longo dos anos a idade media mais é acima dos 40 anos.

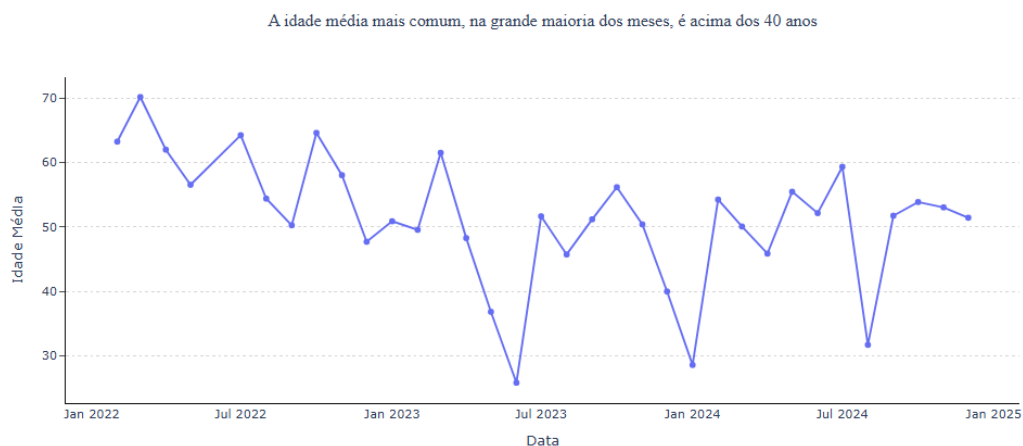


Figura 53: Variação das idades médias por mês

Através da distribuição dos géneros ao longo dos anos observa-se algo que se tem vindo a verificar ao longo da AED feita, isto é: mesmo ao longo dos meses o género mais comum é na grande maioria dos meses o masculino, havendo apenas 9 meses, desde janeiro 2022 a janeiro 2025, em que o género feminino predomina. (Figura 54)

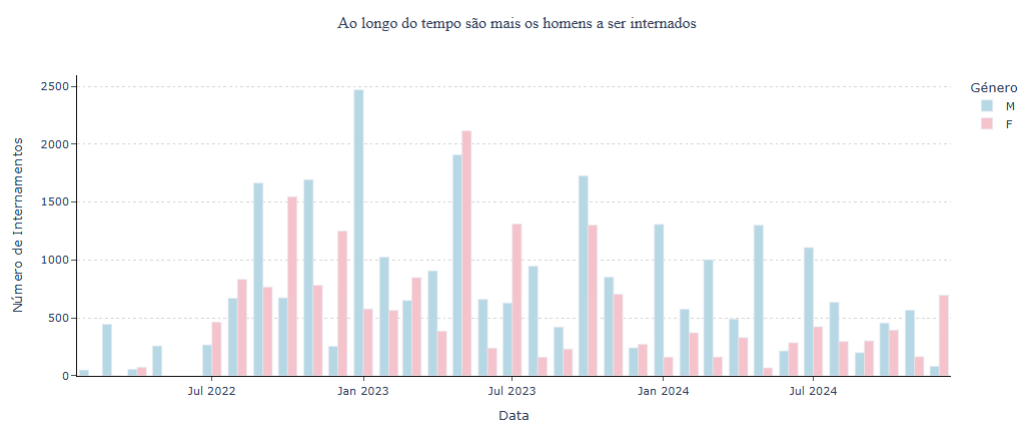


Figura 54: Distribuição do género por mês

Quanto às sessões é possível verificar, na figura 55, que maio de 2023 foi o mês com mais sessões realizadas e fevereiro de 2022 o com menos sessões realizadas.

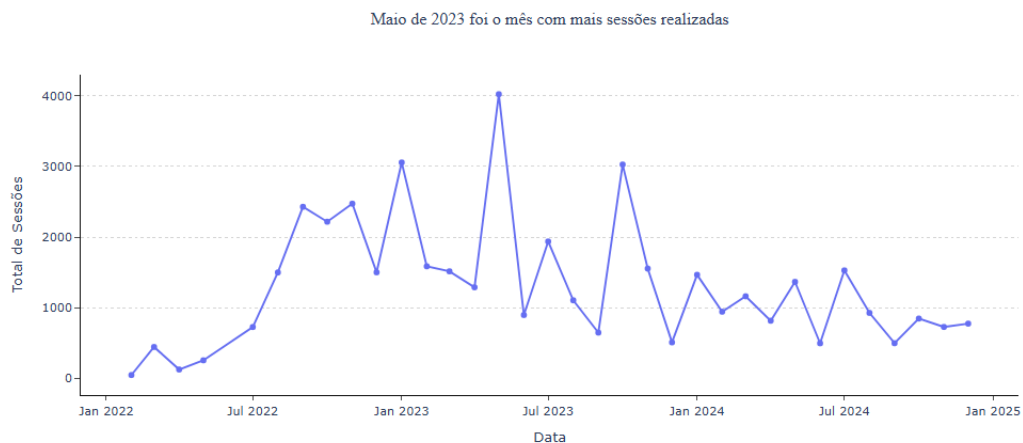


Figura 55: Evolução do número de sessões realizadas por mês

Já nos diagnósticos, observa-se que em outubro de 2022 foram realizados mais diagnósticos do que no resto dos meses entre janeiro de 2022 e janeiro de 2025. (Figura 56)

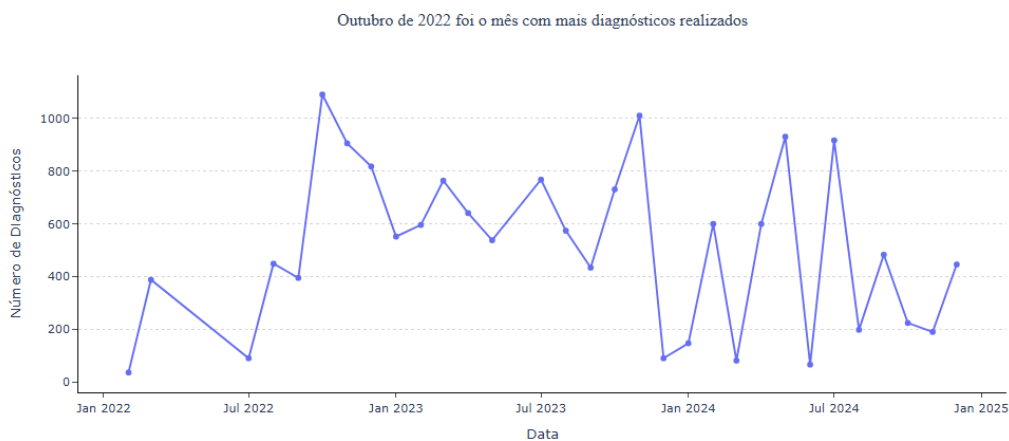


Figura 56: Variação do número de diagnósticos realizados por mês

Após os internamentos os pacientes recebem alta e, nos gráficos das figuras 57, 58 e 59, é possível observar a distribuição de cada um dos destinos entre janeiro de 2022 e janeiro de 2025.

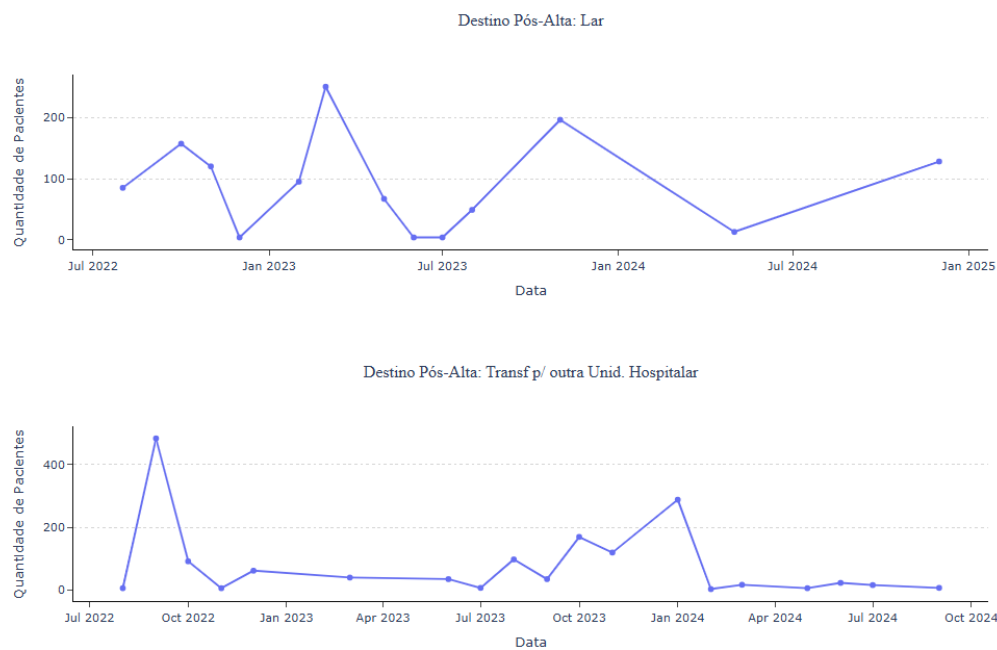


Figura 57: Variação no destino pós alta (lar + transf para outra unid hospitalar)

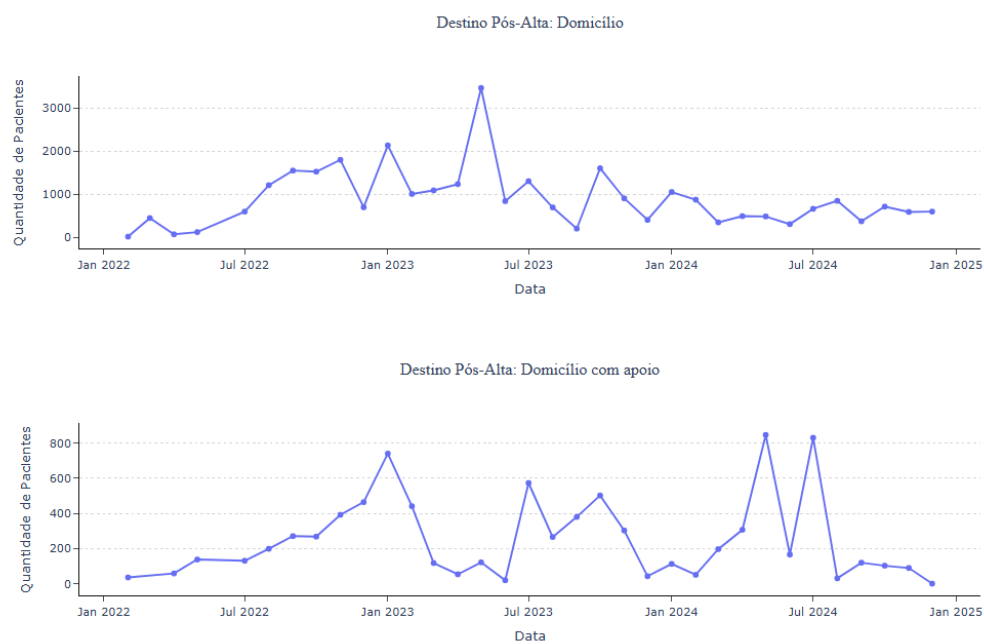


Figura 58: Variação no destino pós alta (domicilio + domicilio com apoio)

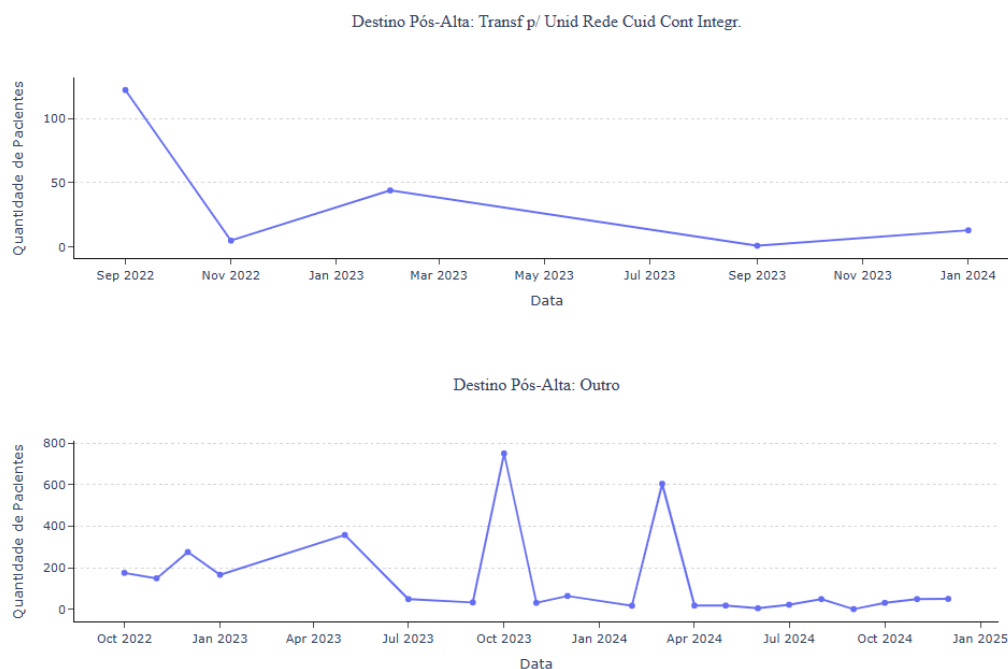


Figura 59: Variação no destino pós alta (transf para unid rede cuid continuados integr + outros)

6.3.9 Análise da variável Tempos de espera

De modo a obter o objetivo esperado pela resolução deste trabalho, foi necessária a criação de uma nova variável à qual demos o nome de *Tempos de espera*. Esta variável foi criada através da subtração da variável *DataInternamento* pela variável *Data Consulta*, figura 60, dando-nos assim informação sobre o tempo de espera resultante desde a primeira consulta até ao internamento do paciente.

```
dados['Tempos de Espera'] = (dados['DataInternamento'] - dados['Data Consulta']).dt.days
```

Figura 60: Criação da variável Tempos de espera

Após a análise dos valores resultantes nesta nova variável foi necessário fazer uma separação dos dados devido ao facto de existirem valores negativos, numa coluna em que deveriam ser todos os valores positivos. Assim sendo foi realizado o tratamento de dados de modo a apenas utilizar os valores positivos para os modelos pretendidos. Com os dados corrigidos agora é possível analisar melhor esta nova variável.

Através da função `.describe()` foi possível obter:

- **Total de observações (count):** Foram analisados 33623 registos.
- **Média (mean):** Valor médio de tempo de espera de 121.58 dias.
- **Desvio padrão (std):** O desvio padrão é de 155.99 dias o que dá a entender que ha uma dispersão à volta da média
- **Valor mínimo (min):** O valor mínimo observado foi de 0 dias.

- **1º quartil (25%):** O que nos indica que 25% dos dados têm valores iguais ou inferiores a 27 dias.
- **Mediana (50%):** O que nos diz que o valor central é 52 dias.
- **3º quartil (75%):** O que nos indica que 75% dos dados têm valores iguais ou inferiores a 145 dias.
- **Valor máximo (max):** O valor máximo observado foi de 940 dias.

Através desta análise é possível observar que a diferença entre a média e a mediana sugere uma distribuição assimétrica, existindo valores muito elevados que puxam a média para cima, o desvio padrão mostra-nos uma grande variação dos dados e o facto de haver o valor mínimo 0 pode querer dizer que há pacientes que são imediatamente internados após a primeira consulta.

Nos gráficos que se seguem podemos observar a distribuição dos tempos de espera:

- Distribuição dos tempos de espera pelas patologias existentes - indica-nos que certas patologias devido a terem mais pacientes acabam por ter, em média, tempos de espera mais elevados, como por exemplo a lesão encefálica (n traum) Anóxia. (Figura 61)

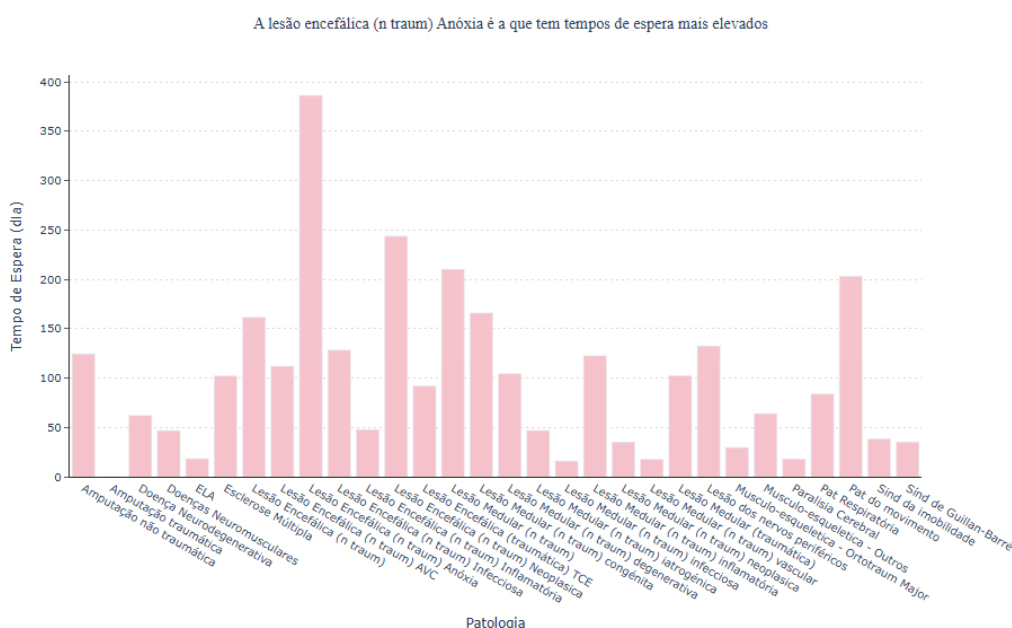


Figura 61: Tempos de espera por patologia

- Distribuição da média do tempo de espera por idades em:
 1. Gráfico de dispersão (Scatter Plot) - de modo a observar a distribuição de cada paciente, que nos mostra que um paciente com 57 anos esperou 940 dias para ser internado após a sua primeira consulta. (Figura 62)

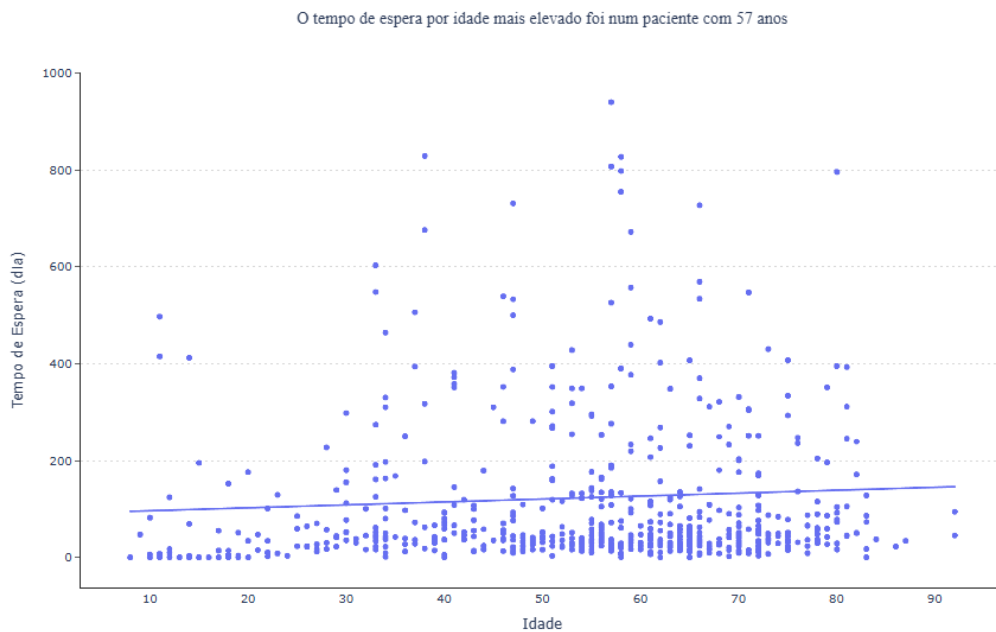


Figura 62: Scatter plot tempos de espera por idade

2. Gráfico de barras - de modo a ser possível observar a distribuição de cada idade e, tal como é possível observar na figura 63, os pacientes com 57 são os que maior média de tempo de espera têm.

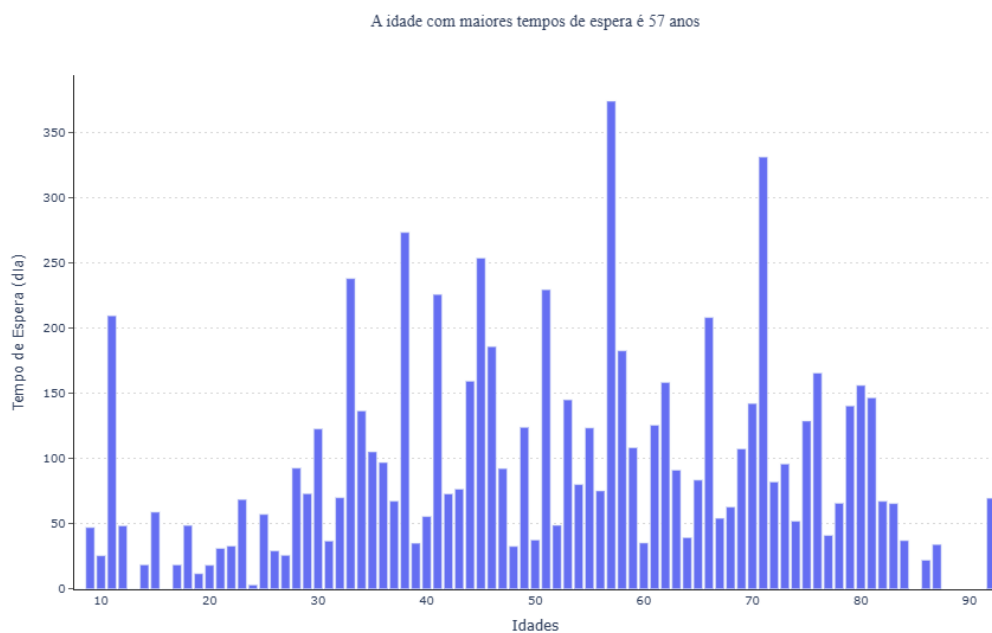


Figura 63: Distribuição dos tempos de espera por idade

- Neste gráfico, figura 64, é possível observar a distribuição das médias dos tempos de espera por entidade financeira responsável, sendo a ADSE a que maior tempo de espera têm.

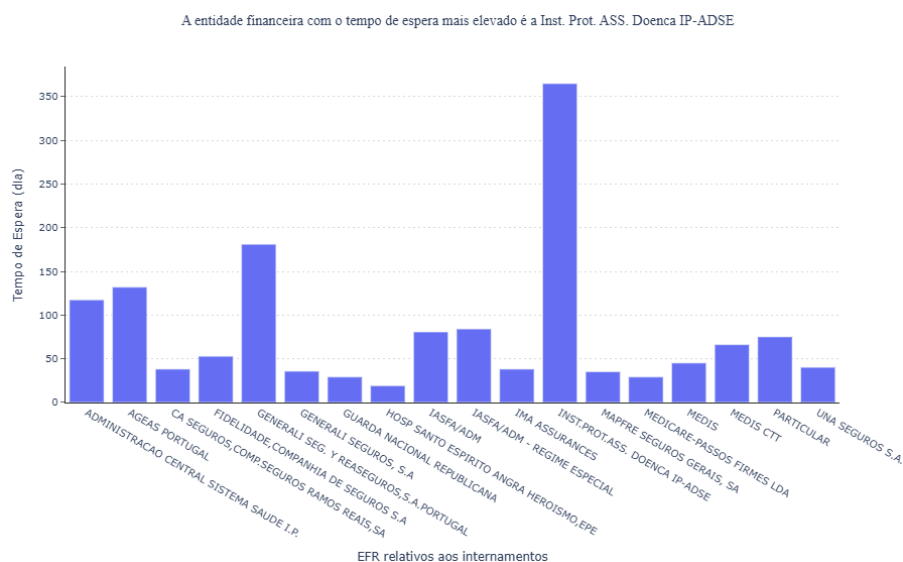


Figura 64: Distribuição dos tempos de espera por entidade financeira responsável

- Em termos de distribuição de género é possível observar que o género masculino em comparação com o género feminino tem uma média de tempo de espera mais elevado. (Figura 65)

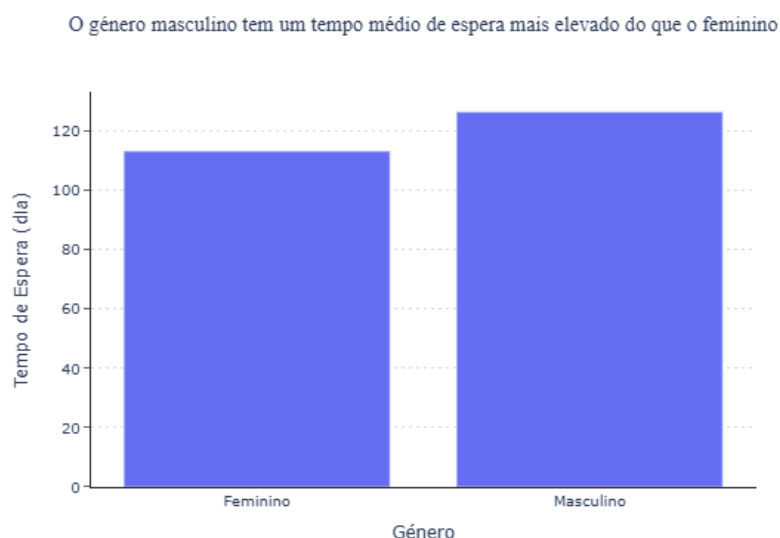


Figura 65: Distribuição dos tempos de espera por género

- Por último, neste último gráfico (figura 66) é possível observar a distribuição da média dos tempos de espera ao longo dos anos, de janeiro de 2022 a dezembro de 2024, sendo possível observar um pico em maio de 2022 que estará possivelmente relacionado com o COVID-19 pois, a partir de julho de 2022 notou-se uma mudança drástica com a média de tempos de espera daí para a frente sempre abaixo de 200 dias.

Ao longo dos anos, a altura em que se registou um tempo de espera particularmente elevado foi em Maio de 2022

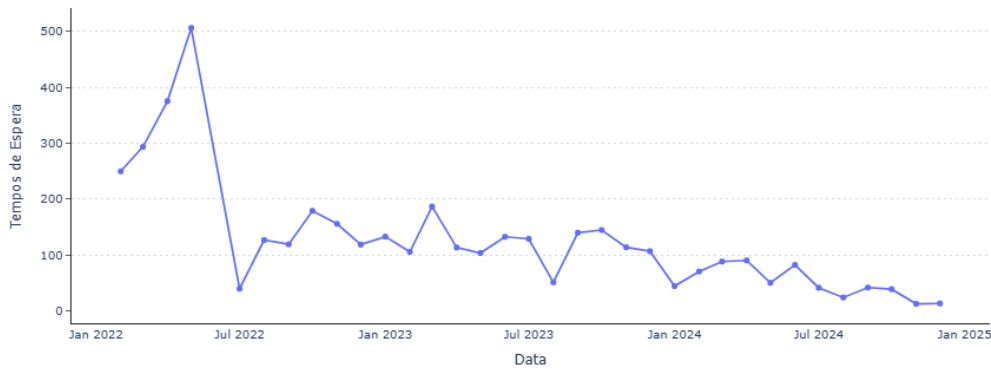


Figura 66: Evolução dos tempos de espera

6.4 Modelos e Algoritmos Escolhidos

Após a análise exploratória dos dados e a junção dos dados das quatro tabelas numa única, os modelos, explicados anteriormente nas secções 3.1.7 e 3.1.8, foram aplicados nos nossos dados.

A seleção dos modelos teve por base a sua simplicidade e eficácia em contextos semelhantes, conforme foi observado na pesquisa bibliográfica apresentada no Capítulo 3.1.8 – Estado da Arte. Optou-se assim por modelos clássicos de aprendizagem supervisionada, nomeadamente a Regressão Linear, Árvore de Decisão, Random Forest, Support Vector Machine (SVM) e K-Nearest Neighbors (KNN), dada a sua capacidade de obter bons resultados.

Adicionalmente, foi aplicado o algoritmo K-Means para a caracterização dos pacientes, com o objetivo de identificar possíveis grupos com padrões semelhantes dentro do conjunto de dados.

Para garantir uma avaliação fiável, o conjunto de dados foi dividido em subconjuntos de treino e teste, com objetivo de garantir uma avaliação fiável do desempenho dos modelos. Esta divisão foi feita através da utilização do método *train_test_split*, permitindo que o modelo possa ser treinado com 80% dos dados e testado com os restantes 20%. Esta abordagem é o que garante ao modelo ter capacidade de generalizar com dados novos.

6.5 Abrangência

Neste projeto, diversas Unidades Curriculares serão essenciais, refletindo a importância das aprendizagens adquiridas, ao longo da licenciatura, para a realização do TFC. Destacam-se as Unidades Curriculares de Análise Exploratória de Dados, Aprendizagem Automatizada I e II, e Linguagens de Programação, fundamentais para o tratamento e estruturação dos dados, desenvolvimento de modelos preditivos e a implementação de algoritmos eficazes.

Na fase inicial, conhecimentos de Probabilidades e Estatística serão cruciais para a

análise estatística, garantindo a precisão dos resultados obtidos. Além disso, na fase seguinte, serão aplicadas técnicas de modelação e otimização de modelos, com o suporte de métodos aprendidos em Inteligência Artificial, que permitirão explorar novas abordagens para a resolução do problema.

Desta forma, pode-se afirmar que todas as Unidade Curriculares, mesmo as não mencionadas, trarão uma componente importante para a realização do projeto e para a qualidade da solução a ser desenvolvida.

7 - Método e Planeamento

Durante o desenvolvimento do projeto, ao longo do ano letivo, foi seguido um plano de trabalho, que serviu como referência para a realização do projeto.

Este plano está dividido em três fases principais, com os seguintes pontos de referência:

- Entrega do Relatório Intercalar do 1º semestre (01/12/2024)
- Entrega do Relatório Intercalar do 2º semestre (27/04/2025)
- Entrega do Relatório Final (27/06/2025)

Estas fases têm objetivos específicos a seguir, estes correspondem a:

- **1ª fase – Entrega do Relatório Intercalar do 1º semestre**
 - Identificação do problema
 - Leitura e pesquisa bibliográfica
 - Desenvolvimento da proposta de solução
 - Definição de conceitos fundamentais
- **2ª fase – Entrega do Relatório Intercalar do 2º semestre**
 - Recolha e preparação dos dados
 - Análise exploratória dos dados
- **3ª fase – Entrega do relatório Final**
 - Continuação e finalização da análise exploratória dos dados
 - Análise e implementação dos modelos a utilizar
 - Registo dos resultados obtidos
 - Análise dos resultados obtidos
 - Conclusão do relatório

Na figura 67, apresenta-se um diagrama de Gantt, que mostra o tempo previsto para o desenvolvimento de cada etapa. Este cronograma será atualizado ao longo do desenvolvimento do projeto, conforme necessário.

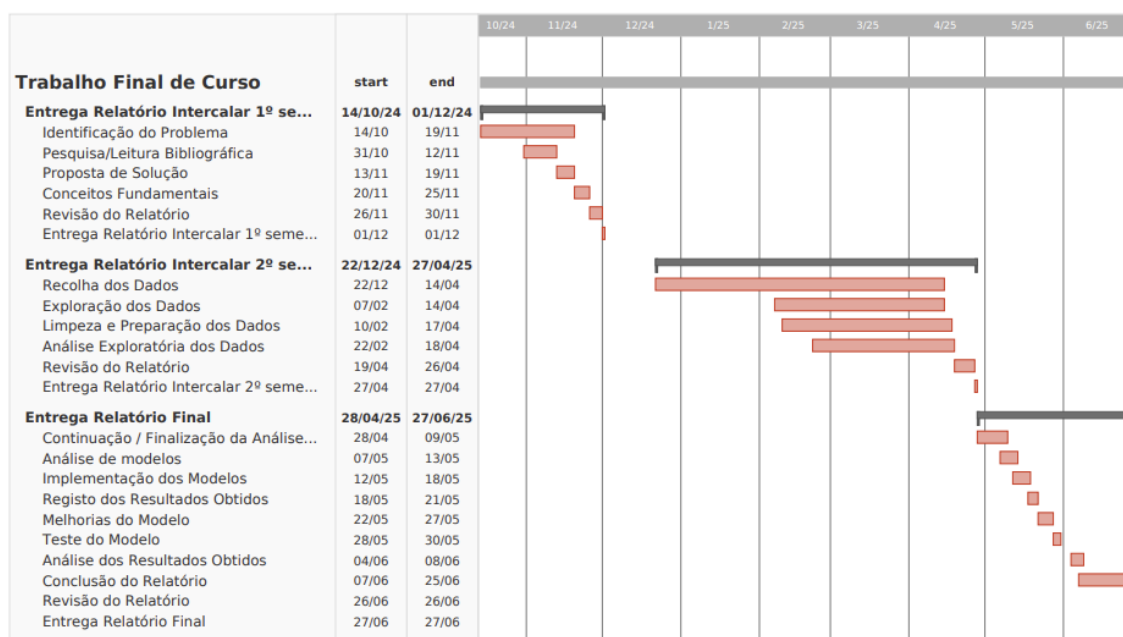


Figura 67: Diagrama Gantt - Entrega Final

7.1 Desafios

Devido ao tempo limitado desde a disponibilização completa dos dados por parte do CMRA, grande parte do trabalho concentrou-se na análise exploratória. Por esse motivo, o tempo revelou-se insuficiente para desenvolver certas análises adicionais que poderiam ter enriquecido o projeto, como uma caracterização mais aprofundada dos pacientes ou a construção de um modelo preditivo para os valores negativos da variável "Tempos de Espera".

Importa ainda referir que o repositório no GitHub inclui todo o código referente às análises aqui apresentadas. As correspondentes bases de dados não estão incluídas no GitHub uma vez que são propriedade do CMRA, e nos comprometemos com a comissão de ética a protegê-los.

8 - Resultados e Discussão

8.1 Comparação de Modelos e Abordagens

Na tabela que se segue, tabela 3, é possível observar a comparação entre todos os modelos treinados para a previsão de tempos de espera. Os melhores valores estão destacados, sendo que correspondem todos ao mesmo modelo: Random Forest.

Métrica	Regressão Linear	Árvores de Decisão	Random Forest	KNN	SVM
RMSE	151.7887	33.6293	31.6751	48.2479	169.5516
MSE	23039.8123	1130.9280	1003.3139	2327.8631	28747.7296
MAE	107.0177	11.7357	11.7058	17.0395	94.0470
ME	0.5390	-0.4642	-0.4257	0.1414	-67.9413
MAPE	inf%	inf%	inf%	inf%	inf%
SAE	719694.0805	78922.8307	78721.7317	114590.4000	632466.1220
R2	0.06	0.95	0.96	0.90	-0.18

Tabela 3: Comparação entre os modelos para a previsão dos tempos de espera

Começando pelo modelo de base para a previsão de tempos de espera, a Regressão Linear, é possível observar que este não é um bom preditor de tempos de espera pois, para além de apresentar um RMSE de 151 dias ainda tem um valor do coeficiente de determinação de apenas 0.06, o que indica que o modelo não explica a variabilidade dos dados. Não se revelando adequado para o problema a resolver.

O modelo Support Vector Machine apresentou resultados ainda piores. Neste modelo foram utilizados os parâmetros padrão da função SVR(). Com base nos resultados que este modelo registou, o erro quadrático médio é equivalente a quase 170 dias e o R2 de -18%, demonstrando assim o seu desempenho insuficiente.

Importa referir que, ao contrário da Regressão Linear, em que o R^2 representa diretamente a proporção da variabilidade explicada, no caso do SVM — especialmente quando se trata de um modelo não linear — esta métrica é apenas uma estimativa da qualidade do ajuste, não estando diretamente relacionada com uma explicação linear da variância. Nestes casos, o R2 funciona apenas como uma medida geral da qualidade do ajuste, e valores negativos indicam que o modelo ajusta pior do que uma média dos dados.

Dado o desempenho insatisfatório obtido, seria importante investigar mais sobre as possíveis causas. Entre as hipóteses possíveis estão a má qualidade dos dados (o que não parece ser o caso), um mau ajustamento dos parâmetros do modelo ou a presença de outliers, aos quais o SVM é sensível. Uma análise mais aprofundada destas questões poderia ajudar a compreender melhor os resultados obtidos e a melhorar a qualidade do modelo.

Já o modelo K-Nearest Neighbors (implementado com parâmetros padrão), apesar de apresentar grandes melhorias quando em comparação com os dois modelos acima

mencionados, ainda não atinge os níveis pretendidos. Tendo obtido um coeficiente de determinação de 90%, o que é ótimo, no entanto, quando se olha para o erro médio nota-se que ainda continua elevado, com erros de cerca de 50 dias, o que compromete a qualidade do modelo.

Por último, os dois modelos com melhores resultados foram as Árvores de Decisão e o Random Forest. Ambos os modelos apresentaram resultados melhores quando em comparação com os acima mencionados.

O modelo de Árvores de Decisão registou métricas com bons níveis de precisão para o problema a resolver, no entanto foi o Random Forest que se destacou com um coeficiente de determinação de 96% e um erro quadrático médio inferior a 32 dias, podendo ser assim considerado o melhor modelo entre os testados.

8.2 Resultados das Análises

Através da utilização das métricas já mencionadas, na secção 3.1.5, foi possível perceber qual o melhor modelo para este problema.

Ao analisar todos os resultados obtidos foi possível identificar o modelo que apresenta uma melhor capacidade de previsão dos tempos de espera.

Esse modelo foi então o *Random Forest*.

8.2.1 Análise dos resultados do melhor modelo de previsão de tempos de espera

O modelo Random Forest foi o modelo que apresentou um melhor desempenho, com métricas que mostram previsões fiáveis para o problema em estudo.

A raiz do erro quadrático médio (RMSE) foi de aproximadamente 31.68 dias, o que indica que, em média, os desvios das previsões quando em comparação com os valores reais situam-se dentro desse valor.

O erro quadrático médio (MSE), de 1003.31 dias, confirma a boa capacidade preditiva do modelo, embora penalize de forma mais acentuada os erros maiores. Já o erro absoluto médio (MAE), de cerca de 11.71 dias, mostra que, em termos absolutos, as previsões mantêm-se próximas dos valores reais, o que é um bom sinal.

O erro médio (ME) de -0.43 mostra que, em média, o modelo tem uma ligeira tendência para subestimar os tempos de espera, o que pode indicar algum enviesamento dos dados.

A soma dos erros absolutos (SAE), apesar de apresentar um valor elevado (78721.7317), deve ser interpretada de maneira diferente das restantes métricas, pois o seu valor aumenta proporcionalmente ao número de observações, ou seja, varia consoante o tamanho do conjunto de dados.

O valor de MAPE foi infinito (inf%), o que acontece pelo facto de existirem valores reais iguais a 0 nos dados, o que torna impossível o cálculo da percentagem relativa.

Por último, o valor do coeficiente de determinação foi de 0.96, o que significa que 96% da variabilidade dos dados é explicada pelo modelo. Este é um valor bom que reforça a qualidade da previsão.

Assim sendo, os resultados obtidos pelo modelo Random Forest demonstraram uma boa capacidade preditiva.

No ponto que se segue será possível ter uma melhor percepção do que foi aqui explicado através da comparação entre todos os modelos.

8.2.2 Análise dos resultados do modelo de classificação de pacientes

Para o objetivo de caracterização dos pacientes de início foi utilizado o método PCA (Principal Component Analysis) e de seguida o algoritmo de clustering K-Means. Após a aplicação do PCA foi utilizado o método Elbow de modo a conseguir determinar o número ideal de clusters, chegando à conclusão que seriam 6.

Posto isto o K-Means foi novamente aplicado com o número de clusters obtido sobre os dados transformados. E, de modo a manter a ligação com a informação não transformada, foram de seguida recuperados os IDs (números de identificação dos pacientes) das observações possibilitando uma análise mais completa e interpretável dos grupos (clusters) formados.

Na imagem que se segue observa-se a aplicação do K-Means com os 6 clusters. (Figura 68)

```

kmeans_final = KMeans(n_clusters=6, random_state=42)
clusters = kmeans_final.fit_predict(X_pca)

ids = dados_corretos[['N_PC']]

resultado_clusters = ids.copy()
resultado_clusters['Cluster'] = clusters

dados_pca = pd.DataFrame(X_pca, columns=[f'PC{i+1}' for i in range(X_pca.shape[1])])
resultado_completo = pd.concat([ids.reset_index(drop=True), dados_pca, pd.Series(clusters, name='Cluster')], axis=1)

fig_kmeans = px.scatter(dados_pca, x='PC1', y='PC2', color=clusters.astype(str),
                        title='Grupos de Pacientes: grupo 1 e 0 são os menos espalhados',
                        labels={'color': 'Cluster'}, height=600)
fig_kmeans.update_layout(title_x=0.5, plot_bgcolor='white')
fig_kmeans.update_yaxes(showgrid=False, showline=True, linewidth=1, linecolor='black')
fig_kmeans.update_xaxes(showline=True, linewidth=1, linecolor='black')
fig_kmeans.show()

```

Grupos de Pacientes: grupo 1 e 0 são os menos espalhados

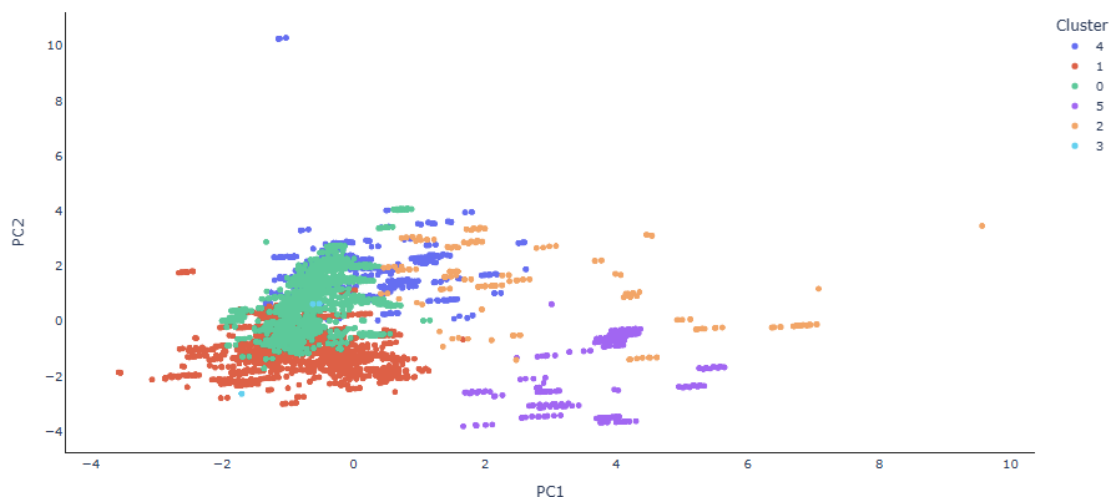


Figura 68: Aplicação do K-Means

No gráfico acima é apresentado um gráfico de dispersão dos pacientes, com os diferentes clusters identificados a cores distintas. A distribuição dos pacientes é representada no plano definido pelas duas primeiras componentes principais (PC1 e PC2), após a aplicação do algoritmo de clustering. É possível observar que os grupos 0 e 1 são os menos espalhados, o que sugere uma maior homogeneidade nas características dos pacientes que os compõem. Por outro lado, os restantes grupos apresentam uma maior dispersão, o que indica uma maior variabilidade nas características dos pacientes pertencentes a esses clusters.

Apesar destas observações visuais iniciais, não foi possível realizar uma análise mais aprofundada dos clusters, devido à limitação de tempo nesta fase do trabalho. Teria sido relevante, por exemplo, calcular medidas estatísticas como a média e a mediana das variáveis contínuas, bem como a moda das variáveis categóricas codificadas, para cada um dos seis grupos. Essa análise permitiria uma caracterização mais completa dos clusters e uma melhor interpretação das diferenças entre os grupos identificados.

8.3 Limitações da Análise

Apesar dos resultados obtidos ao nível das previsões dos tempos de espera terem sido satisfatórios, não foi possível avaliar adequadamente a qualidade da componente relativa

à classificação dos pacientes. Esta limitação deveu-se, em grande parte, à falta de tempo que permitisse validar devidamente as classificações obtidas.

O objetivo inicial passava por caracterizar os pacientes com base na patologia e, eventualmente, noutras variáveis relevantes, de modo a permitir uma segmentação mais eficiente e informada dos utentes.

Contudo, devido à limitação de tempo e à complexidade durante o tratamento e preparação dos dados necessários para este ponto, a componente relativa à classificação dos pacientes acabou por não ser desenvolvida com o nível de detalhe inicialmente previsto. Ainda assim, é importante referir que o objetivo principal do trabalho — as previsões dos tempos de espera — foi atingido, tendo produzido resultados satisfatórios.

10 - Conclusão

10.1 Conclusão

Ao longo do desenvolvimento do trabalho, foi possível realizar uma análise detalhada sobre os dados fornecidos pelo CMRA e sobre a aplicação de vários modelos de machine learning. Tal como foi referido ao longo do relatório, o principal objetivo era aplicar um modelo de previsão capaz de estimar os tempos de espera entre a data da 1ª consulta e a data de início do internamento.

Foram testados cinco modelos distintos, o que permitiu realizar uma comparação dos seus desempenhos com base em métricas adequadas. Os resultados obtidos demonstraram que modelos como o Random Forest e as Árvores de Decisão apresentam uma capacidade preditiva superior face a modelos mais simples, como a Regressão Linear. No entanto, de forma geral, os modelos apresentaram um desempenho satisfatório, cumprindo o objetivo principal do trabalho: prever com precisão os tempos de espera dos pacientes.

Relativamente ao objetivo secundário do trabalho, embora o algoritmo K-Means tenha sido aplicado, a sua exploração foi limitada devido ao tempo disponível até à entrega final.

Assim sendo, o trabalho atingiu com sucesso o objetivo principal, contribuindo para que o CMRA possa vir a ter uma melhor perceção de como gerir os seus recursos de forma mais eficiente.

10.2 Trabalhos Futuros

Em trabalhos futuros, em colaboração com o CMRA, seria importante aprofundar a classificação dos pacientes e analisar com maior detalhe os casos com tempos de espera negativos, ambos com potencial para melhorar a gestão dos serviços.

No caso da classificação dos pacientes, seria relevante aprofundar a análise através do cálculo de estatísticas descritivas, por idades ou patologias, por exemplo.

Quanto aos tempos de espera negativos, trata-se de situações em que os utentes já em internamento são encaminhados para consultas, sendo estas registadas como “1ªs consultas”. Esta questão, inicialmente sinalizada pela Eng. Mariana Matos, mereceria uma análise mais aprofundada, nomeadamente para identificar padrões por patologia ou serviço.

Assim sendo, estas propostas podem abrir caminho para melhorias substanciais na resposta dos serviços às necessidades dos utentes.

Bibliografia

- [1] João P. Matos-Carvalho. *The Lusófona \LaTeX Template User's Manual*. Lusófona University. 2024. URL: <https://github.com/jpmcarvalho/UL-Thesis>.
- [2] Santa Casa da Misericórdia de Lisboa. *Centro de Medicina de Reabilitação de Alcoitão*. 2023. URL: <https://scml.pt/media/noticias/centro-de-medicina-e-reabilitacao-de-alcoitao/> (acedido em 10/2024).
- [3] Centro de Reabilitação de Alcoitão. *Sobre nós - Centro de Reabilitação de Alcoitão*. URL: <https://cmra.scml.pt/sobre-nos/> (acedido em 10/2024).
- [4] Centro de Reabilitação de Alcoitão. *Como fazemos - Centro de Reabilitação de Alcoitão*. URL: <https://cmra.scml.pt/como-fazemos/> (acedido em 10/2024).
- [5] Rishabh Singh. *Introduction to Machine Learning*. URL: <https://medium.com/@RobuRishabh/introduction-to-machine-learning-555b0f1b62f5>.
- [6] IBM. *What is machine learning?* URL: <https://www.ibm.com/topics/machine-learning>.
- [7] Felipe Azank. *Como avaliar seu modelo de regressão*. 2020. URL: <https://medium.com/turing-talks/como-avaliar-seu-modelo-de-regress%C3%A3o-c2c8d73dab96>.
- [8] Akshita Chugh. *MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better?* 2020. URL: <https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e>.
- [9] Statistics How To. *Mean Error: Definition*. URL: <https://www.statisticshowto.com/mean-error/>.
- [10] Mohamad. *SAE - Sum of Absolute Errors*. 2016. URL: <https://support.numxl.com/hc/en-us/articles/214286506-SAE-Sum-of-Absolute-Errors>.
- [11] Saishruthi Swaminathan. *Linear Regression — Detailed View*. 2018. URL: <https://towardsdatascience.com/linear-regression-detailed-view-ea73175f6e86>.
- [12] *What is a decision tree?* URL: <https://www.ibm.com/think/topics/decision-trees>.
- [13] *What is a Random Forest?* URL: <https://www.ibm.com/think/topics/random-forest>.
- [14] Wikipedia. *Support vector machine*. URL: https://en.wikipedia.org/wiki/Support_vector_machine.
- [15] Onel Harrison. *Machine Learning Basics with the K-Nearest Neighbors Algorithm*. 2018. URL: <https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761>.
- [16] Roshmita Dey. *Understanding Principal Component Analysis (PCA)*. URL: <https://medium.com/@roshmitadey/understanding-principal-component-analysis-pca-d4bb40e12d33>.
- [17] Dishant kharkar. *Algoritmo de agrupamento K-Means*. 2023. URL: <https://medium.com/@dishantkharkar9/k-means-clustering-algorithm-ce4fbcac8fb0>.

- [18] Anat Ratnovsky et al. "Statistical learning methodologies and admission prediction in an emergency department". Em: *Australasian Emergency Care* 24.4 (2021), pp. 241–247. ISSN: 2588-994X. DOI: <https://doi.org/10.1016/j.auec.2020.11.004>. URL: <https://www.sciencedirect.com/science/article/pii/S2588994X20301147>.
- [19] Jeffin Joseph et al. "Machine Learning for Prediction of Wait Times in Outpatient Clinic". Em: *Procedia Computer Science* 215 (2022). 4th International Conference on Innovative Data Communication Technology and Application, pp. 230–239. ISSN: 1877-0509. DOI: <https://doi.org/10.1016/j.procs.2022.12.026>. URL: <https://www.sciencedirect.com/science/article/pii/S187705092202097X>.
- [20] Anton Pak, Brenda Gannon e Andrew Staib. "Predicting waiting time to treatment for emergency department patients". Em: *International Journal of Medical Informatics* 145 (2021), p. 104303. ISSN: 1386-5056. DOI: <https://doi.org/10.1016/j.ijmedinf.2020.104303>. URL: <https://www.sciencedirect.com/science/article/pii/S1386505620305219>.
- [21] Katie Walker et al. "Emergency medicine patient wait time multivariable prediction models: a multicentre derivation and validation study". Em: *medRxiv* (2021). DOI: 10.1101/2021.03.19.21253921. eprint: <https://www.medrxiv.org/content/early/2021/03/24/2021.03.19.21253921.full.pdf>. URL: <https://www.medrxiv.org/content/early/2021/03/24/2021.03.19.21253921>.
- [22] Mohammad Al-Dweik Amjed Al Moussa Hamza Al-Zubaidi. "A machine learning-based approach for wait-time estimation in healthcare facilities with multi-stage queues". Em: *IET WILEY* (2023). DOI: 10.1049/smc2.12079.
- [23] Parthasarathi Pattnayak et al. "Deep Learning Based Patient Queue Time Forecasting in the Emergency Room". Em: (2023), pp. 541–545. DOI: 10.1109/ICSSAS57918.2023.10331755.
- [24] Xiaoqing Li et al. "Prediction of outpatient waiting time: using machine learning in a tertiary children's hospital". Em: *Translational Pediatrics* 12.11 (2023). ISSN: 2224-4344. URL: <https://tp.amegroups.org/article/view/119333>.
- [25] Zhou W Benjasirisan C Phianhasin L Schiltz NK Brahmabhatt S Ruksakulpiwat S Thongking W. "Machine learning-based patient classification system for adults with stroke: A systematic review". Em: *Chronic Illn* (2023). DOI: 10.1177/17423953211067435.
- [26] Shatha Melhem, Ahmad Al-Aiad e Muhammad Al-Ayyad. "Patient care classification using machine learning techniques". Em: (mai. de 2021), pp. 57–62. DOI: 10.1109/ICICS52457.2021.9464582.
- [27] Daniel Amorim. *Python — Pandas (Essenciais)*. 2023. URL: <https://medium.com/@dev.daniel.amorim/python-pandas-24ab58577de5>.
- [28] Francesca Fuentes. *NumPy: The Fundamental Tool for Data Science in Python*. 2024. URL: <https://medium.com/@m.franfuentes/numpy-the-fundamental-tool-for-data-science-in-python-fa2b605a3bf9>.
- [29] Matplotlib. *Matplotlib: Visualization with Python*. URL: <https://matplotlib.org/>.
- [30] Hugo Habbema. *Introdução ao Scikit-Learn*. 2024. URL: <https://medium.com/@habbema/introdu%C3%A7%C3%A3o-ao-scikit-learn-f00b7201dbf7>.
- [31] Seaborn. *An introduction to seaborn*. URL: <https://seaborn.pydata.org/tutorial/introduction.html>.
- [32] Plotly. *Overview*. URL: <https://plotly.com/python/plotly-express/>.

Anexo

1.1 União das tabelas 1^as consulta e Sessões

Começando pelo *género* foi possível observar que são mais os pacientes do género masculino do que do género feminino. Com o género masculino a ter cerca de 17.8% de mais pacientes do que o género feminino. (Figura 69)

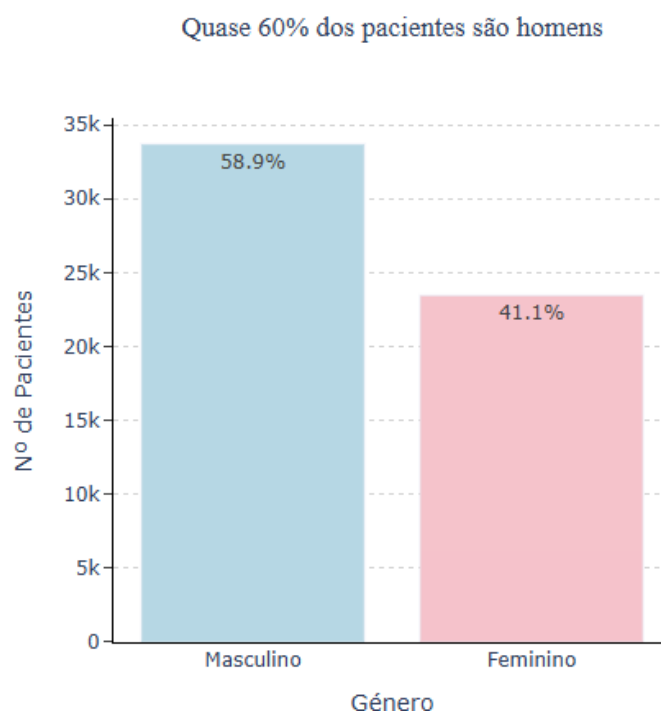


Figura 69: Distribuição dos utentes que posteriormente à 1^a consulta realizaram sessões, por género

Passando para o *concelho*, é possível observar que Cascais é o concelho ao qual mais pacientes pertencem, com 33.3% do total dos pacientes. (Figura 70)

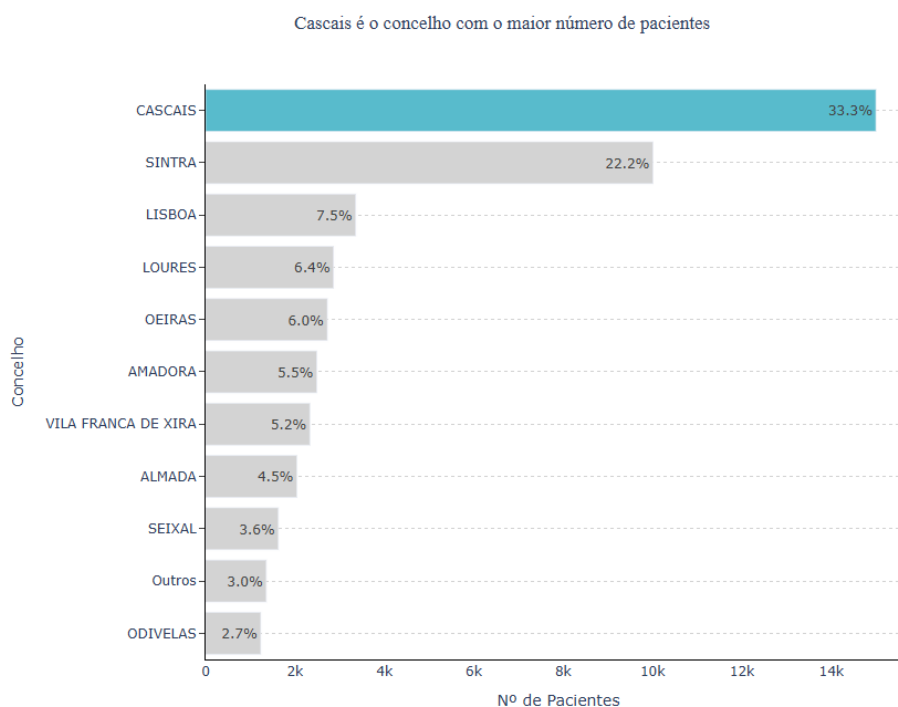


Figura 70: Distribuição dos utentes que posteriormente à 1ª consulta realizaram sessões, por concelho

Quanto ao serviço mais comum 72.7% dos pacientes são atendidos nos serviços de adultos, sendo que 41.4% pertencem ao serviço de adulto relacionado a amputações e 31.3% ao serviço de adulto que abrange patologias neurológicas. (Figura 71)

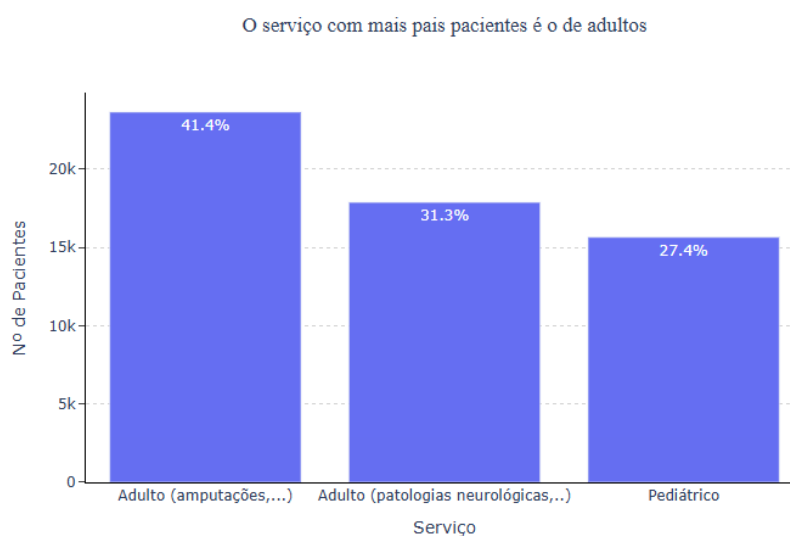


Figura 71: Distribuição dos utentes que posteriormente à 1ª consulta realizaram sessões, por serviço

Dos pacientes que atendem a consultas e sessões, temos que é possível serem de 13 diferentes nacionalidades, sendo Portugal a nacionalidade mais comum seguido da Angola com apenas 0.7% dos pacientes. A categoria "Outros" representa os restantes países com percentagens de pacientes ainda menores, tal como se observa na figura 72.

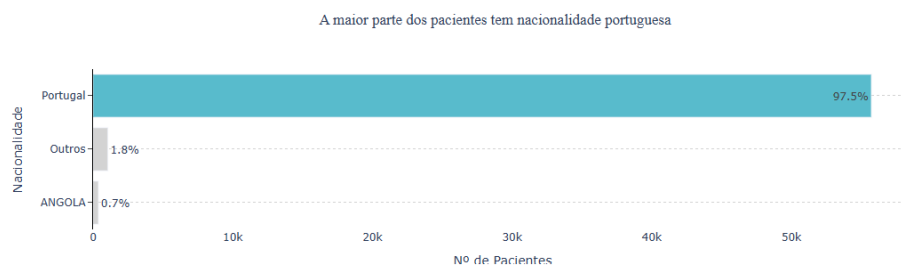


Figura 72: Distribuição dos utentes que posteriormente à 1ª consulta realizaram sessões, por nacionalidade

Como causa da maioria das entradas dos pacientes no CMRA, é possível observar, na figura 73, que os problemas de saúde aparentam ser o fator mais comum, com 89.9% dos pacientes.

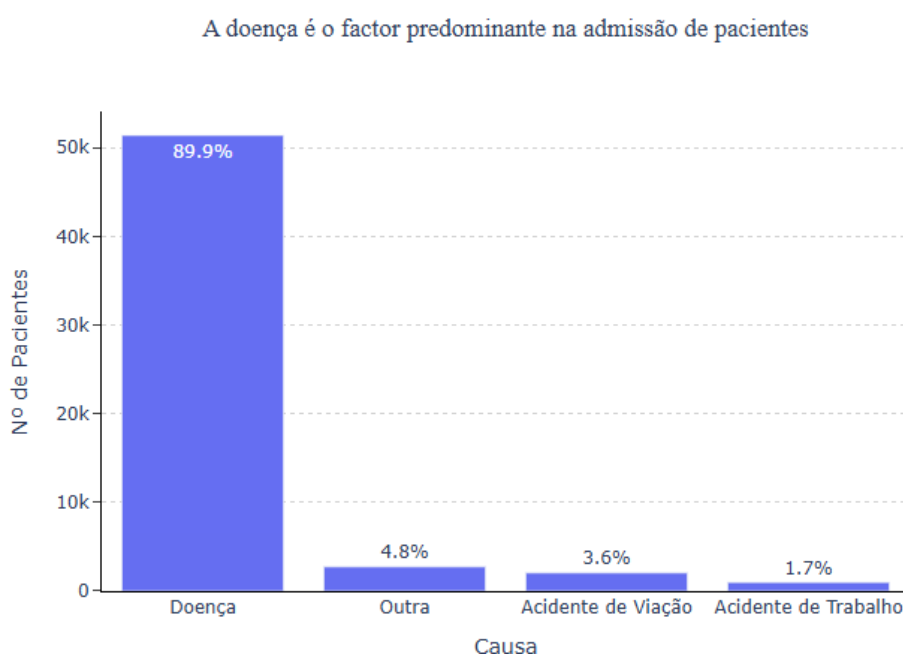


Figura 73: Distribuição dos utentes que posteriormente à 1ª consulta realizaram sessões, por causa

Perante os vários tipos de reabilitação realizados no centro é possível observar, na figura 74, uma tendência para reabilitações neurológicas com cerca de 66.4% dos pacientes totais.

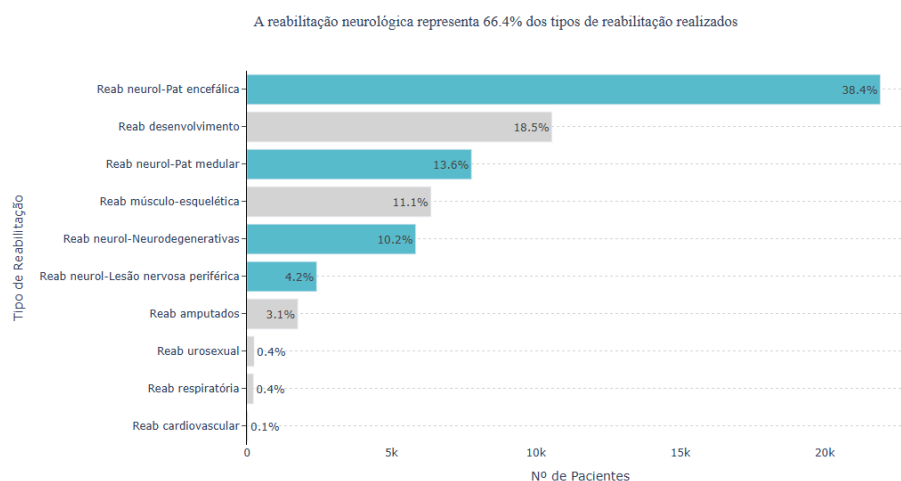


Figura 74: Distribuição dos utentes que posteriormente à 1ª consulta realizaram sessões, por tipologia

Das 40 diferentes patologias observamos também que a lesão encefálica não traumática (AVC) é a mais comum entre os pacientes. As patologias não especificadas no gráfico são as que contêm uma percentagem menor, estando todas agrupadas na categoria "Outros". (Figura 75)

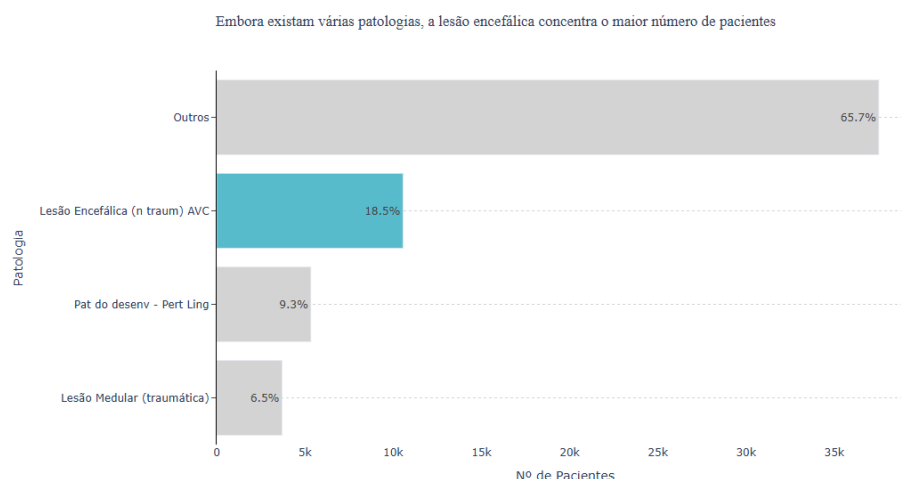


Figura 75: Distribuição dos utentes que posteriormente à 1ª consulta realizaram sessões, por patologia

Quanto aos profissionais responsáveis, podemos ver que são menos os profissionais responsáveis por 1ªs consultas, apenas 24, quando em comparação com os profissionais responsáveis pelas sessões, 111 profissionais. Nas 1ªs consultas (Figura 76) observamos que os 3 profissionais responsáveis por mais pacientes são os com IDs = 5.0, 27.0 e 3.0 enquanto os responsáveis por mais pacientes nas sessões são os com IDs = 39, 135, 151.

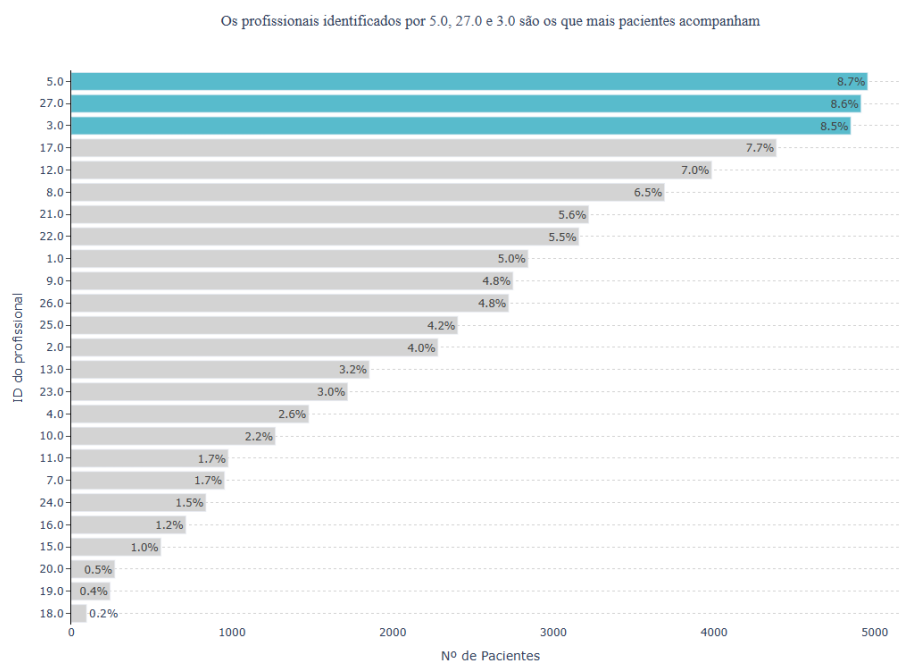


Figura 76: Distribuição dos utentes que posteriormente à 1ª consulta realizaram sessões, por profissional responsável

Entre os diversos actos realizados é possível identificar uma grande percentagem de pacientes em consultas de avaliação para admissão a acompanhamento de adultos. (Figura 77)

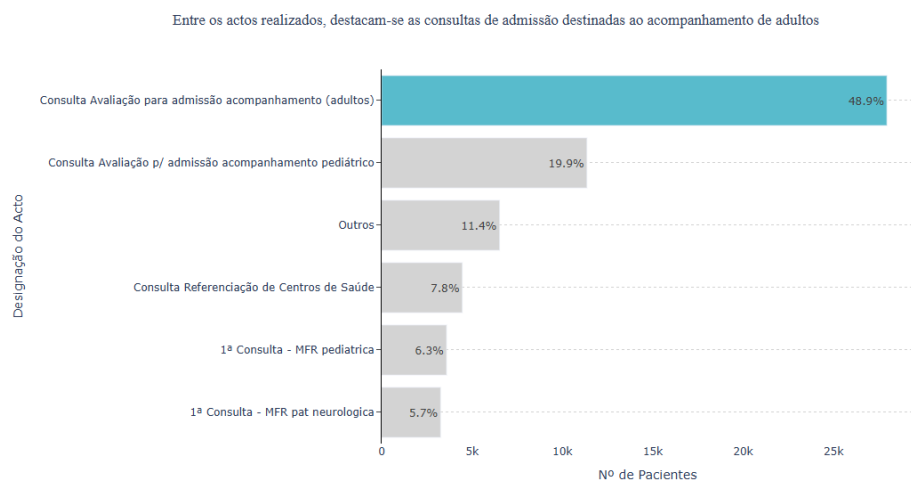


Figura 77: Distribuição dos utentes que posteriormente à 1ª consulta realizaram sessões, por acto

Dentro dos serviços existe ainda um sub-serviço “Serviço2”. Este está dividido em 7 categorias sendo, tal como no serviço normal, a categoria mais comum os serviços de adultos, a categoria menos comum neste caso é a Toxina Adultos com apenas 0.1% dos pacientes. (Figura 78)

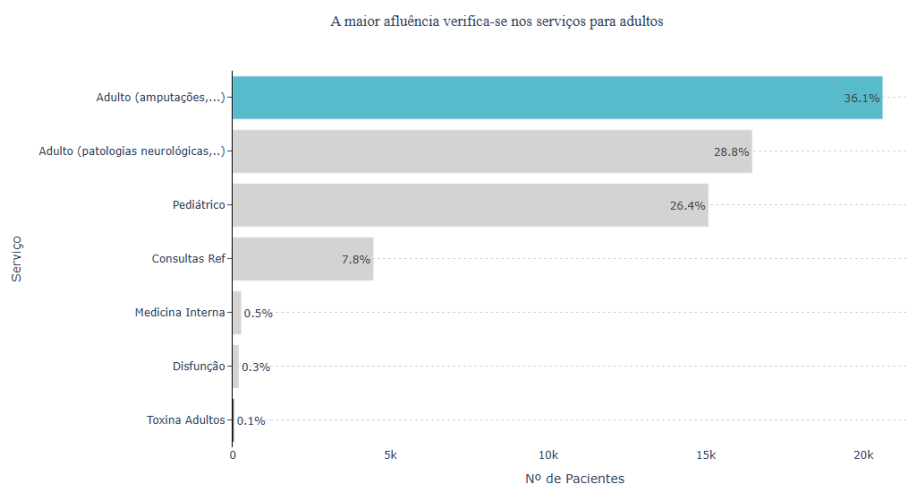


Figura 78: Distribuição dos utentes que posteriormente à 1ª consulta realizaram sessões, por sub-serviços

Estas sessões e consultas podem eventualmente ter comparticipações de diversas entidades. Neste caso, tendo em conta os pacientes em análise, são 30 as entidades financeiras utilizadas pelos diversos pacientes, sendo que a entidade “ARSLVT-ACORDO” comparticipa mais de metade dos pacientes (59.6%). (Figura 79)

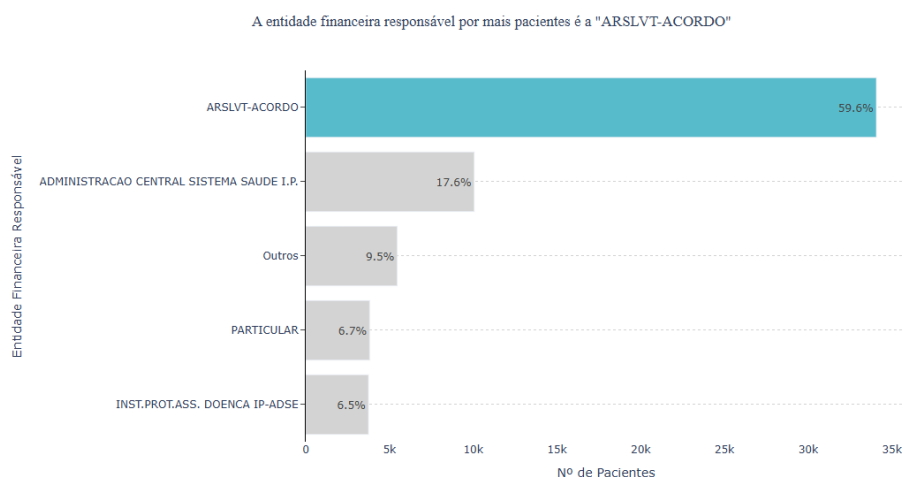


Figura 79: Distribuição dos utentes que posteriormente à 1ª consulta realizaram sessões, por entidade financeira responsável

Por último, nesta tabela temos ainda o diagnóstico que está dividido em 310 diferentes categorias, sendo a mais comum, com uma percentagem substancial, a sequência de doença cerebrovascular, com 23.6%. Esta diferença é substancial pois, todas as restantes categorias têm uma percentagem menor que 5.2%. (Figura 80)

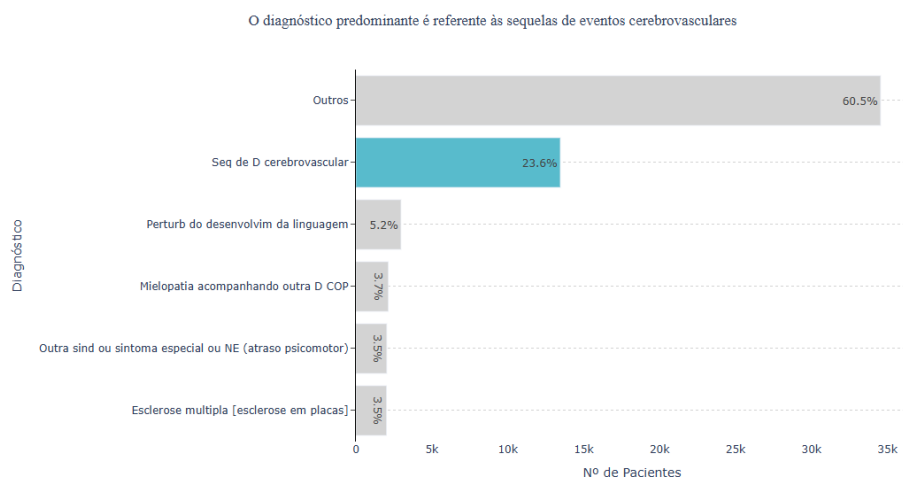


Figura 80: Distribuição dos utentes que posteriormente à 1ª consulta realizaram sessões, por diagnóstico

1.2 União das tabelas 1ªs consultas e info de internamentos

Tendo em conta ao género foi possível tirar a conclusão de que as percentagens são iguais tanto para o género relativo à variável "género_1ªconsulta" como para o relativo à variável "género_info_internamento". (Figura 81)

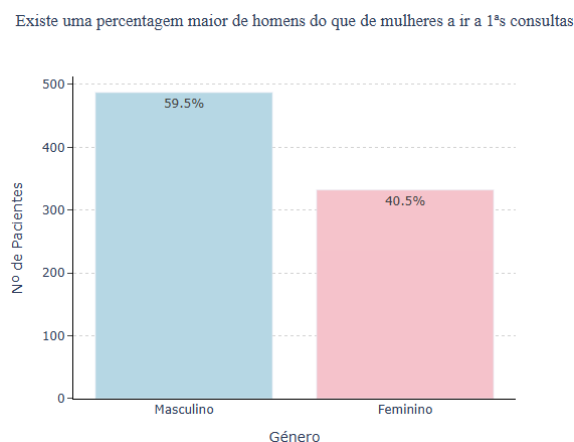


Figura 81: Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por género

Já relativamente aos concelhos, acontece exatamente a mesma coisa, chegando assim há conclusão que uma grande parte dos pacientes pertence ao concelho de Sintra. (Figura 82)

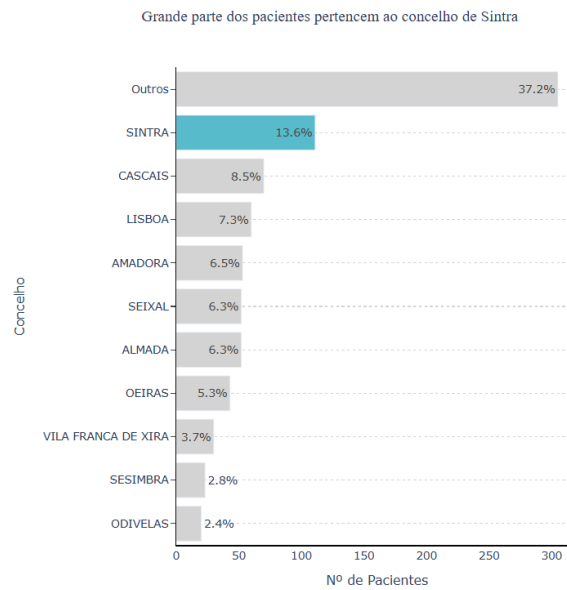


Figura 82: Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por concelho

Quanto à nacionalidade, através do gráfico que se segue observa-se uma predominância de portugueses. (Figura 83)

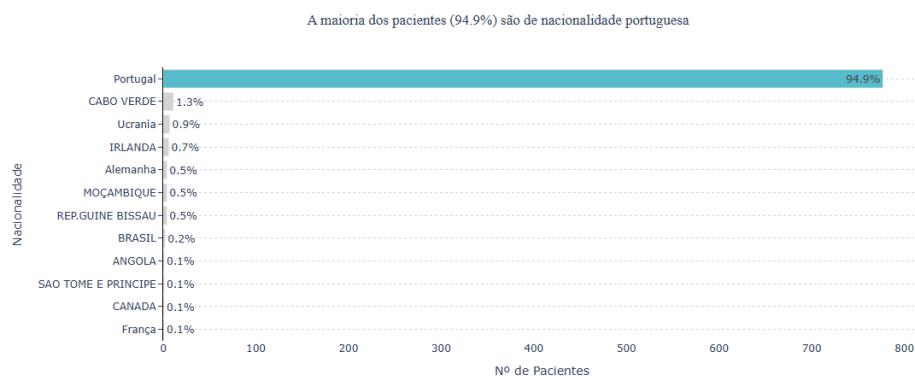


Figura 83: Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por nacionalidade

No gráfico da figura 84 podemos concluir que os serviços de adulto são os mais comuns contendo cerca de 87.5% dos pacientes.

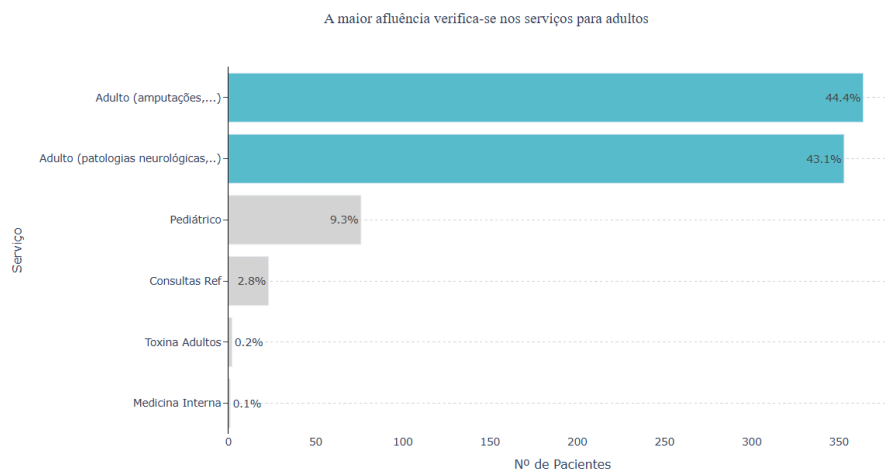


Figura 84: Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por sub-serviço

A causa mais comum de entrada de pacientes são os problemas de saúde, tal como se pode observar no gráfico da figura 85

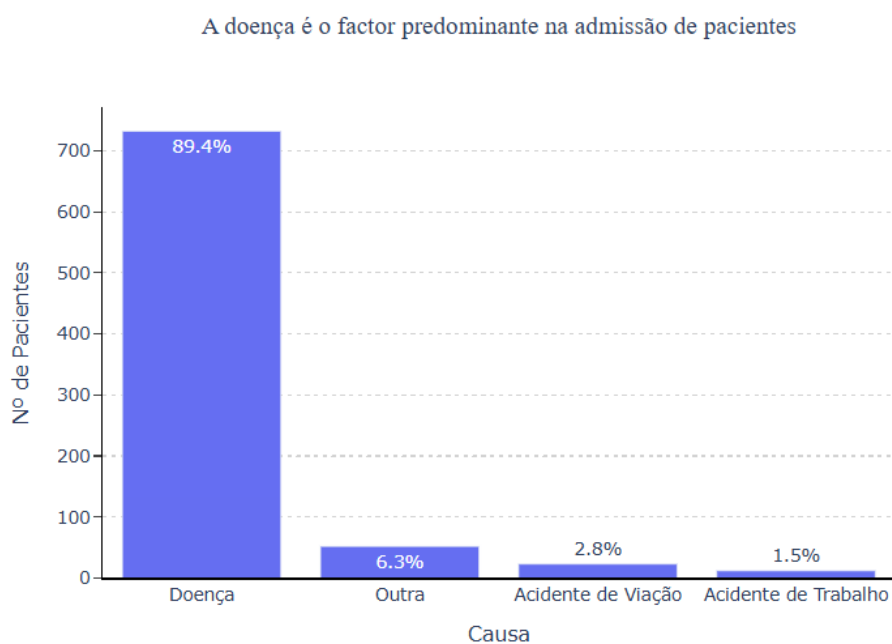


Figura 85: Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por causa

Tal como se tem vindo a observar nas restantes juncões de tabelas é possível observar uma predominância nas reabilitações neurológicas. (Figura 86)

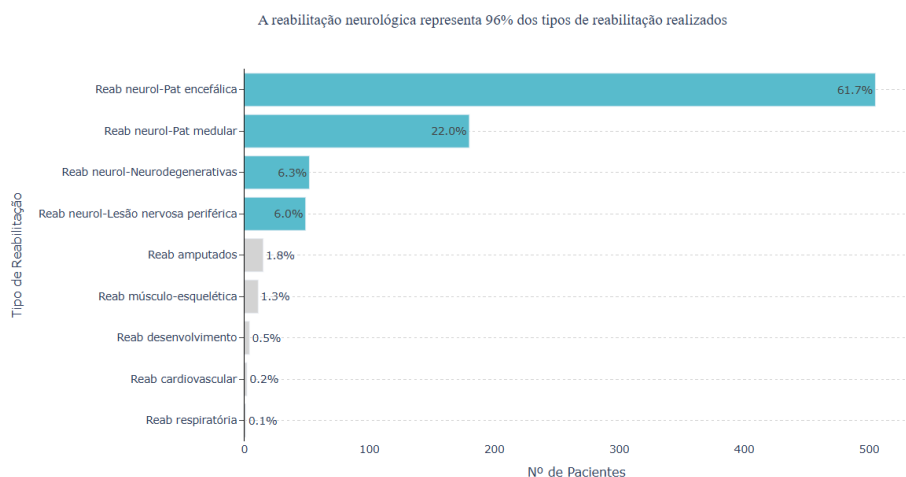


Figura 86: Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por tipologia

Quanto aos profissionais responsáveis na junção destas duas tabelas, pode-se observar que os profissionais com ID = 5 e ID = 22 são os que mais pacientes atendem, sendo os profissionais com os ID = 10 e ID = 20 os que menos pacientes atendem. (Figura 87)

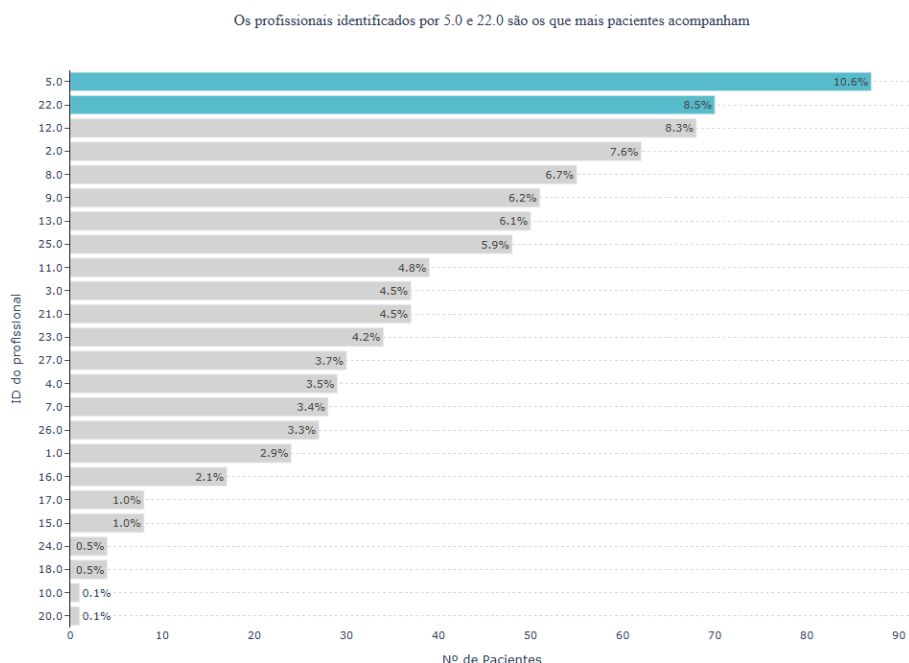


Figura 87: Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por profissional responsável

Consegue-se ainda observar, através da distribuição de idades representada no gráfico da figura 88 que há um grande grupo de pacientes com idade compreendida entre os 55 e os 59 anos.

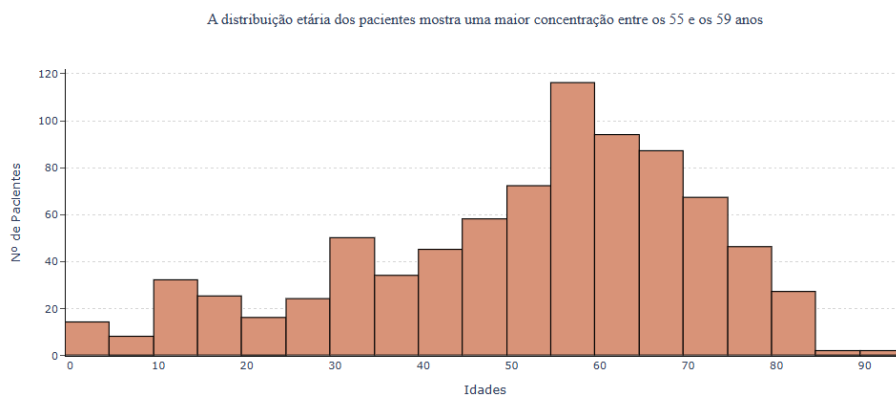


Figura 88: Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por idades

1.3 União das tabelas internamento e info internamentos

Começando pelo género, é possível observar que continua a ser o género masculino o predominante entre os pacientes. (Figura 89)

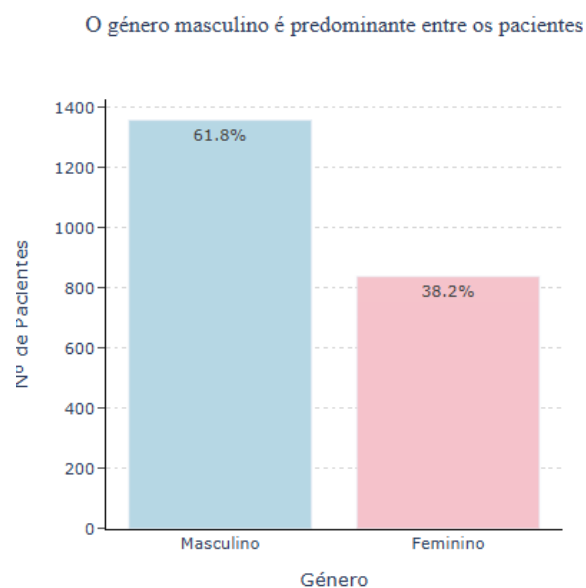


Figura 89: Distribuição dos utentes em internamento por género

De seguida temos as entidades financeiras responsáveis que são iguais tanto na variável “EFR_internamento” como na variável “EFR_info_internamentos”, com a administração centra sistema saúde I.P como a principal entidade financeira utilizada pelos pacientes. (Figura 90)

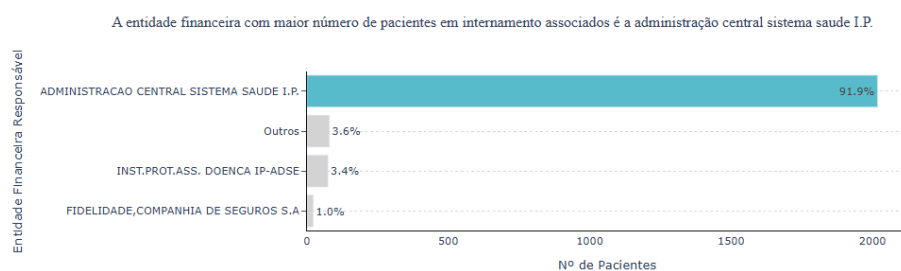


Figura 90: Distribuição dos utentes em internamento por entidade financeira responsável

A grande maioria dos pacientes são seguidos pelo serviço de adultos, sendo apenas 9.6% dos pacientes seguidos pelo serviço pediátrico, tal como se pode observar na figura 91.

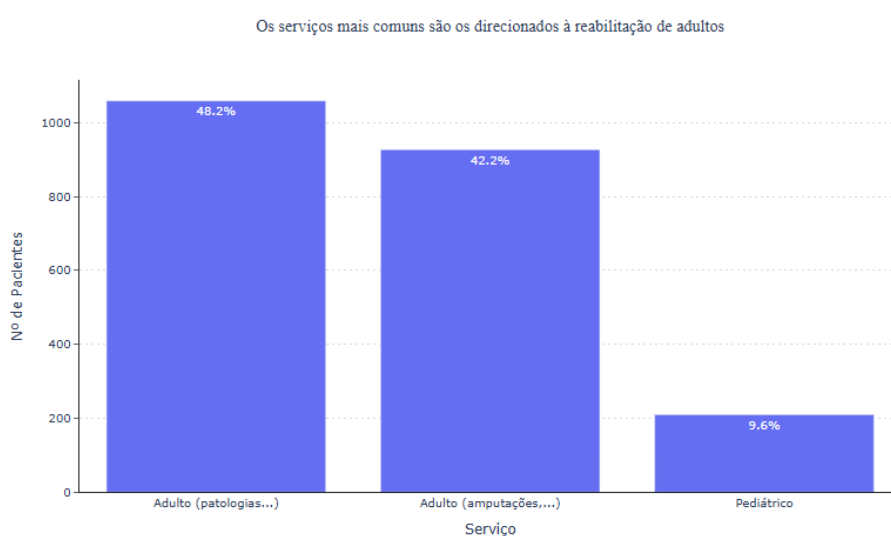


Figura 91: Distribuição dos utentes em internamento por serviço

O diagnóstico mais comum é a sequência de doença cerebrovascular, com cerca de 37.8% dos pacientes assim diagnosticados (figura 92). Sendo a patologia mais comum a lesão encefálica não traumática (AVC), figura 93

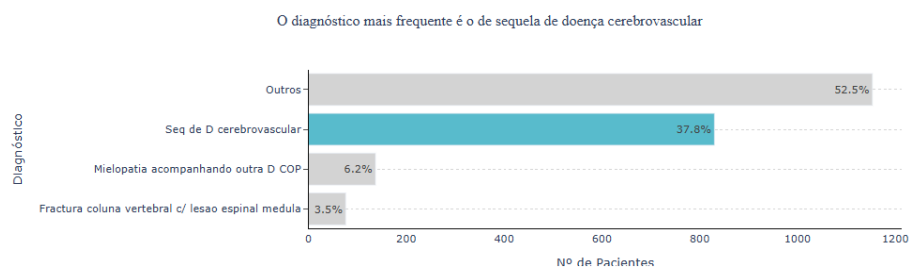


Figura 92: Distribuição dos utentes em internamento por diagnóstico

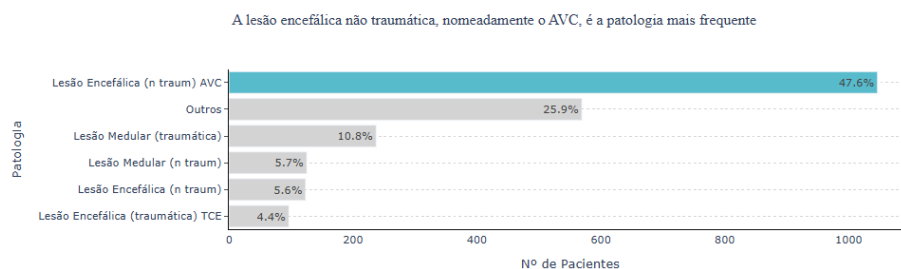


Figura 93: Distribuição dos utentes em internamento por patologia

O profissional responsável por mais pacientes é o com o ID = 23, já o profissional responsável por menos pacientes é o com ID = 15. (Figura 94)

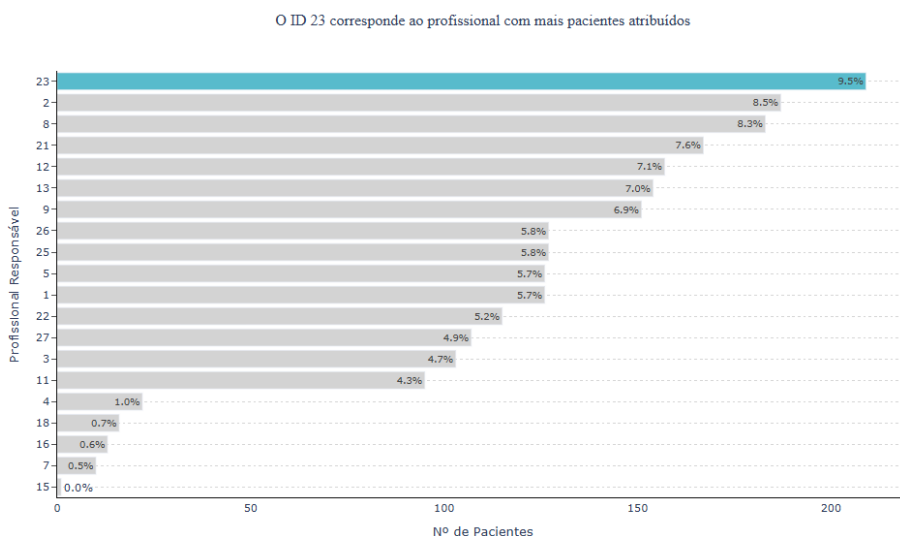


Figura 94: Distribuição dos utentes em internamento por profissional responsável

1.4 União das tabelas sessões e info internamentos

Os pacientes de género masculino são os mais comuns.

Os diagnósticos mais comuns são novamente a sequência de doença cerebrovascular, tal como se tem vindo a observar ao longo do trabalho e como se pode observar na figura 95

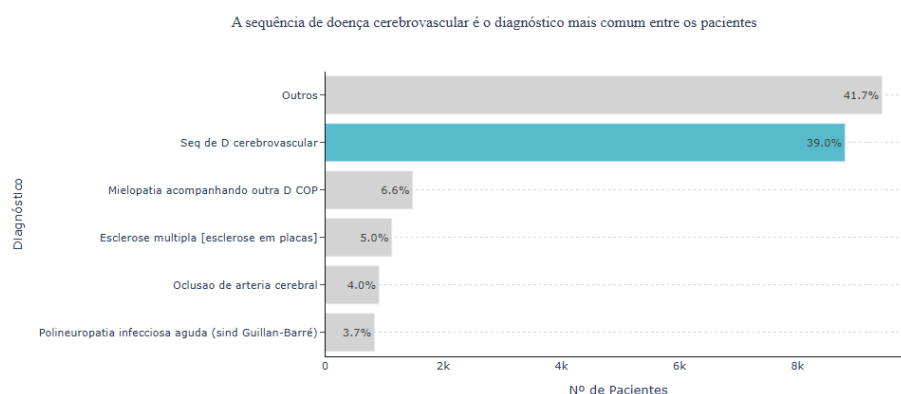


Figura 95: Distribuição dos utentes em internamento e sessões, por diagnóstico

Já as patologias, a mais comum continua a ser a lesão encefálica (n traum) AVC. (Figura 96)

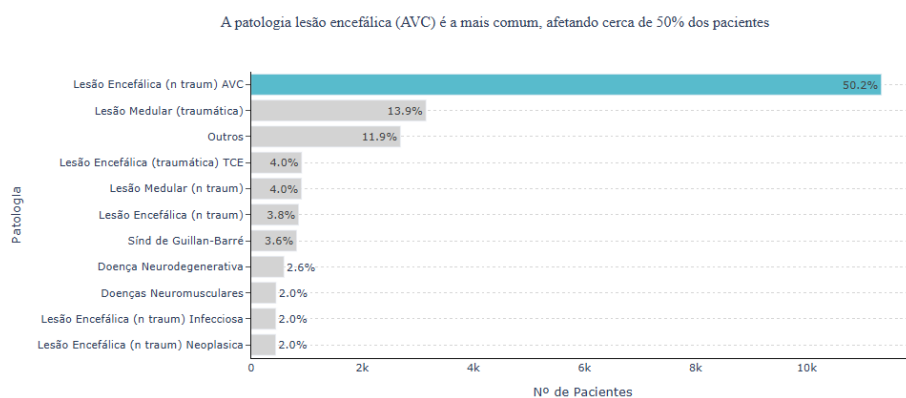


Figura 96: Distribuição dos utentes em internamento e sessões, por patologia

Com os serviços acontece a mesma questão e, sendo assim, o mais comum é o direcionado a adultos.

Como entidades financeiras responsáveis, a administração central sistema de saúde I.P é a responsável pela maioria dos pacientes, com 92.9% de pacientes. (Figura 97)

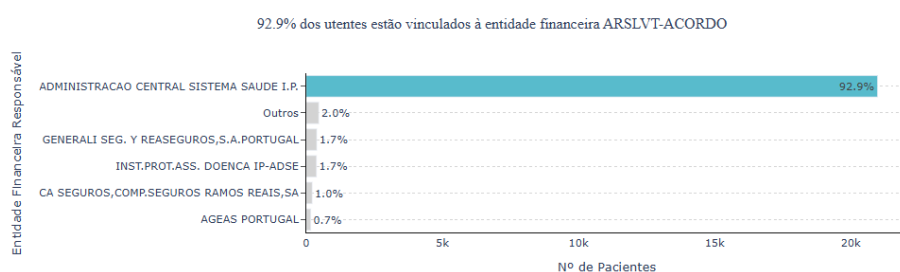


Figura 97: Distribuição dos utentes em internamento e sessões, por entidade financeira responsável

1.5 União das tabelas 1^as consulta e internamentos

Começando pelo género, através das duas variáveis *Género_1ªconsulta* e *Género_internamento* observa-se que, em ambos os casos, o resultado é o mesmo e que, 59.5% dos pacientes são do género masculino. (Figuras 98 e 99)

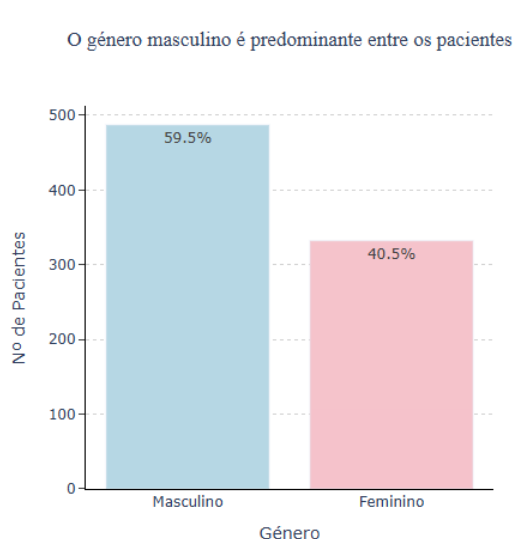


Figura 98: Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por género (1ª consulta)

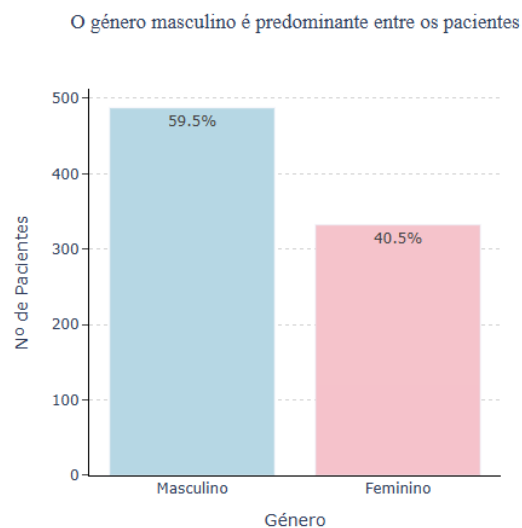


Figura 99: Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por género (internamento)

O gráfico que se segue, figura 100, mostra que a maior parte dos pacientes é português e apenas 5.1% dos pacientes são de outras nacionalidades.

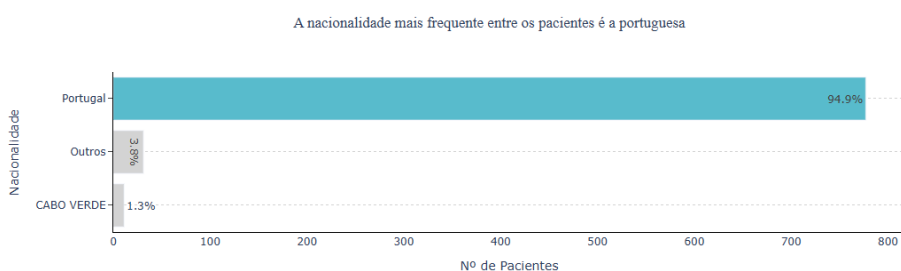


Figura 100: Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por nacionalidade

Quanto aos concelhos, é igual entre as duas variáveis: *Concelho_1ªconsulta* e *Concelho_internamento*. Sendo assim possível retirar a conclusão de que o concelho mais comum entre os pacientes é o concelho de Sintra. (Figuras 101 e 102)

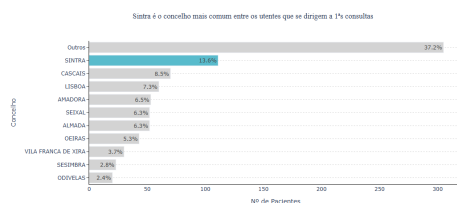


Figura 101: Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por concelho (1ª consulta)

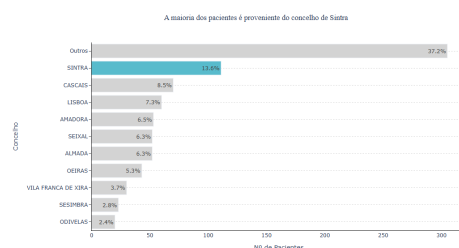


Figura 102: Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por concelho (internamento)

Através da utilização da variável *Designação Acto* é possível observar que o acto médico mais realizado pelos pacientes é a consulta de avaliação para admissão de acompanhamento em adultos com 79.7% dos pacientes a realizá-la. Como contraste é possível ainda observar que todas as restantes variáveis contêm significativamente menos pacientes. (Figura 103)

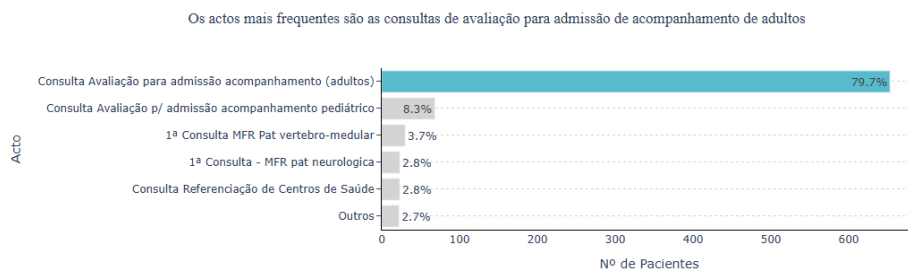


Figura 103: Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por acto

Seguindo para o sub-serviço, *serviço2*, tal como já era de esperar, é possível verificar uma percentagem significativamente maior no serviço direccionado à reabilitação de adultos com 87.5% do total dos pacientes. (Figura 104)

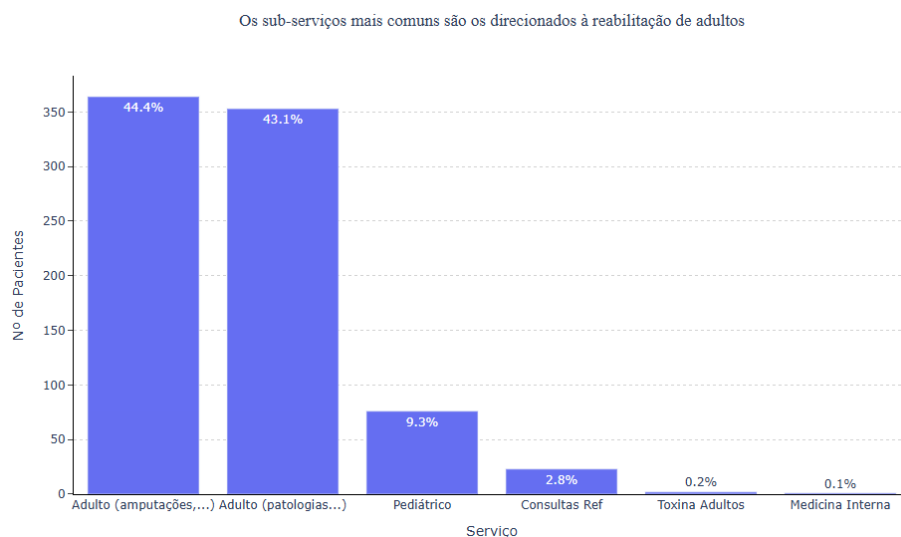


Figura 104: Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por sub-serviço

Dos vários diagnósticos possíveis, o mais comum, neste caso, continua a ser a sequência de doença cerebrovascular correspondendo a 37.2% dos pacientes. Os restantes diagnósticos encontram-se divididos, no entanto todos eles contêm menos de 6.5% dos pacientes. (Figura 105)

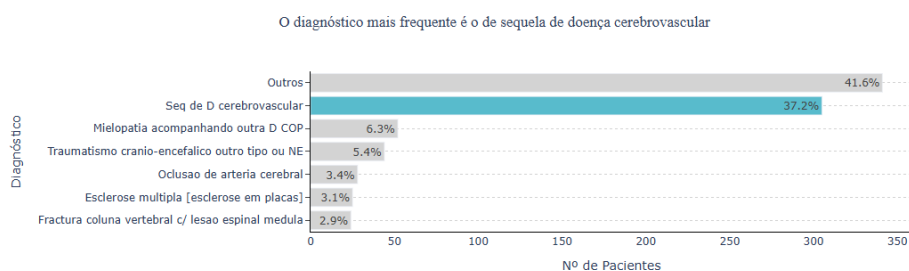


Figura 105: Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por diagnóstico

Seguindo para a patologia, tal como aconteceu no diagnóstico, a categoria que representa mais pacientes é, novamente, a lesão encefálica não traumática (AVC). Na categoria “Outros” observa-se a agrupação dos restantes dados, todos com um valor percentual inferior a 6.8% dos pacientes. (Figura 106)

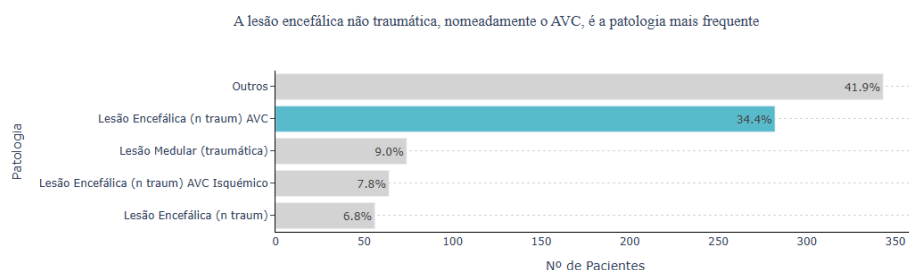


Figura 106: Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por patologia

A *causa*, tal como tem acontecido nas últimas representações gráficas, mostra que a grande maioria dos pacientes (neste caso 89.4%) entra no CMRA devido a problemas de saúde, sendo os restantes 10.6% divididos entre as restantes categorias. (Figura 107)

O principal motivo de entrada no centro por parte dos pacientes são os problemas de saúde

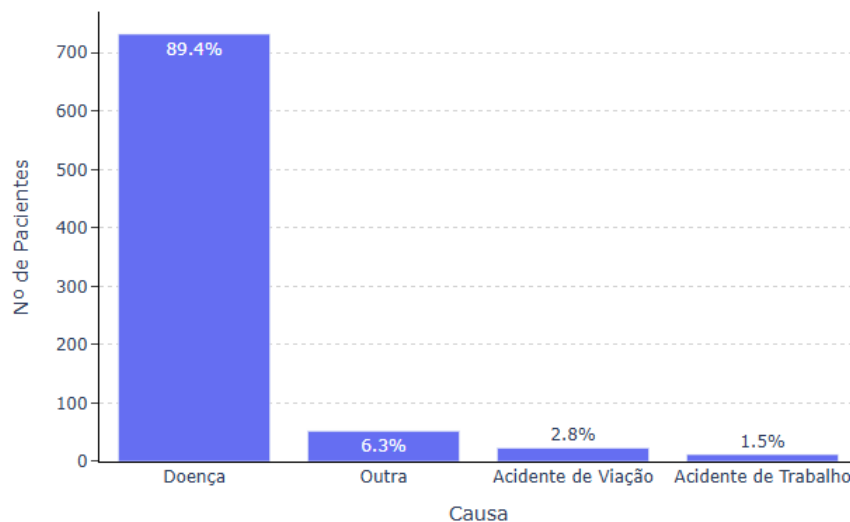


Figura 107: Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por causa

Com a observação da variável *tipologia* chega-se à conclusão de que o tipo de reabilitação mais comum é a direccionada à reabilitação neurológica, contendo mais de 90% dos pacientes. (Figura 108)

O tipo de reabilitação mais comum é a neurológica, com mais de 90% de pacientes

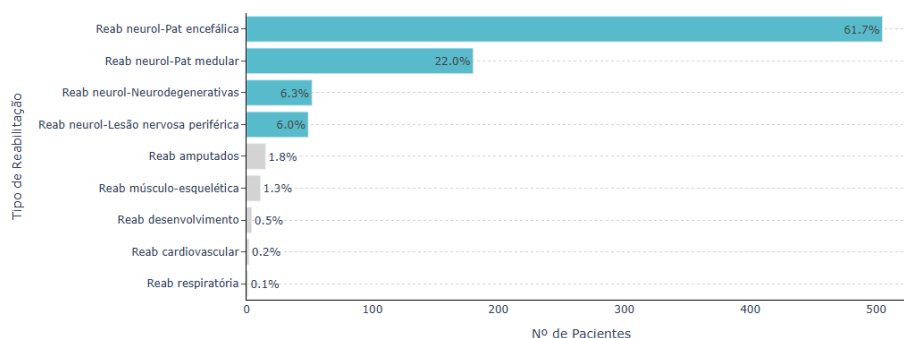


Figura 108: Distribuição dos utentes que posteriormente à 1ª consulta ficaram em internamento, por tipologia

1.6 União das tabelas internamento e sessões

Começando pelo género, observamos que o mais comum continua a ser o género masculino, com 59.5% dos pacientes. (Figura 109)

O género masculino predomina entre os pacientes

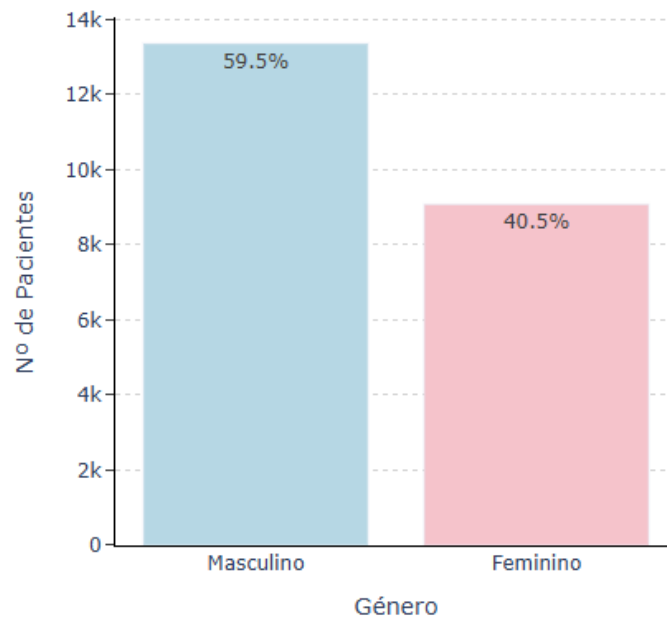


Figura 109: Distribuição dos utentes em internamento e sessões, por género

Através da variável *concelho* é possível observar que grande parte dos pacientes (16.6%) pertencem ao concelho de Sintra, estando os restantes pacientes divididos entre os restantes concelhos e contendo todos um valor percentual inferior. (Figura 110)

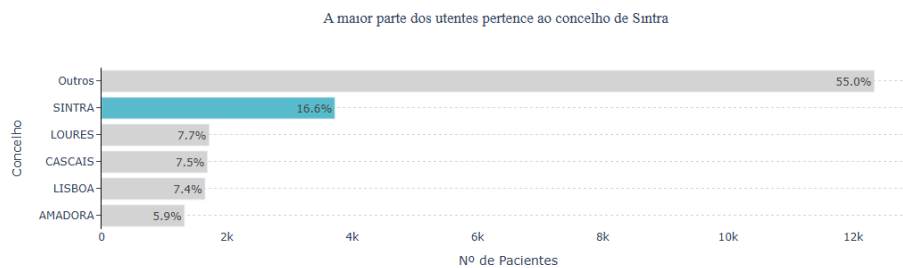


Figura 110: Distribuição dos utentes em internamento e sessões, por concelho

Quanto à entidade financeira responsável, na figura 111, é possível observar a distribuição dos pacientes pelas diversas entidades financeiras, sendo a mais comum entre a grande maioria dos pacientes, 91.7%, a “ADMINISTRACAO CENTRAL SISTEMA SAUDE I.P.”.

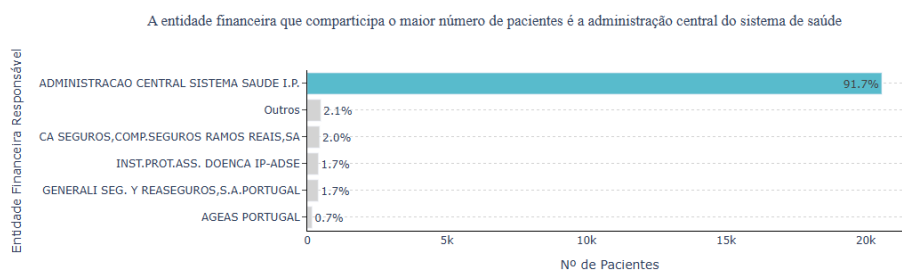


Figura 111: Distribuição dos utentes em internamento e sessões, por entidade financeira responsável

Entre os vários profissionais responsáveis, através do gráfico que se segue, figura 112, é possível observar que o profissional com o ID = 82, com 4.9% do total dos pacientes, é o responsável por mais pacientes.

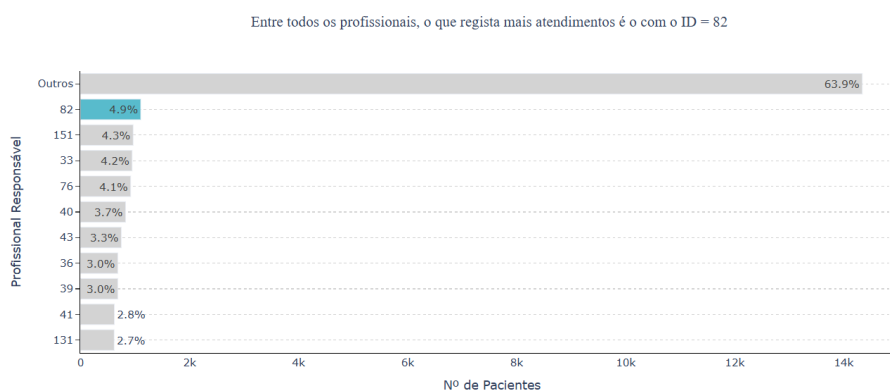


Figura 112: Distribuição dos utentes em internamento e sessões, por profissional responsável

1.7 União das tabelas 1ªs consultas, internamento e sessões

De modo a analisar efetivamente os dados todos juntos foi necessária a junção das três tabelas.

Desta junção obtivemos uma nova tabela com 32743 linhas e 32 colunas.

Esta nova tabela não continha nem valores nulos nem duplicados.

A junção dos dados em 1ª consulta, internamentos e sessões permite analisar, para cada paciente que expêenciou as três situações, o seguinte:

- Serviços (relativos a 1ª consulta e internamento);
- Género
- Concelho
- Nacionalidade
- Idade

- Entidade financeira responsável (relativos a 1ª consulta e internamento);;
- Actos (consultas de avaliação para admissão acompanhamento (Adultos), consultas de avaliação para admissão acompanhamento (Pediátrico),...);
- Serviços2 (cada serviço contém outros serviços, a designação Serviço2 representa esta subdivisão);
- Diagnóstico;
- Patologia;
- Causa;
- Tipologia (tipos de reabilitação);
- ID do profissional responsável (relativo às 1ªs consultas e às sessões);
- Número de sessões (uma ou duas)
- Número de meios complementares terapêuticos (entre 1 e 11)
- Tipo de Consulta

Esta união entre "1ª consulta", "internamento" e "sessões" centra a nossa análise de novo no paciente. Para ser possível retirar conclusões e perceber melhor os dados foram realizadas, novamente, as visualizações.

Começando pelo género, foi possível observar que o género relativo à primeira consulta e o género relativo ao internamento obtiveram os mesmos valores, como já era de esperar. Sendo o género masculino o predominante. (Figuras 113 e 114)

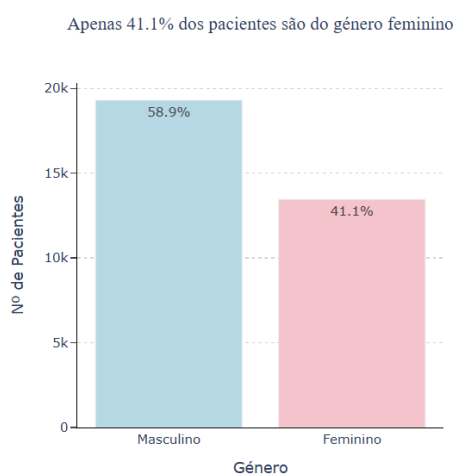


Figura 113: Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por género (1ª consulta)

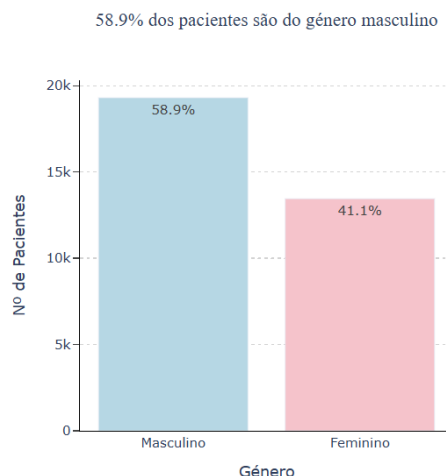


Figura 114: Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por género (internamento)

As idades mais comuns entre os pacientes são entre os 55 e os 59 anos de idade, tal como se pode observar no gráfico que se segue. (Figura 115)

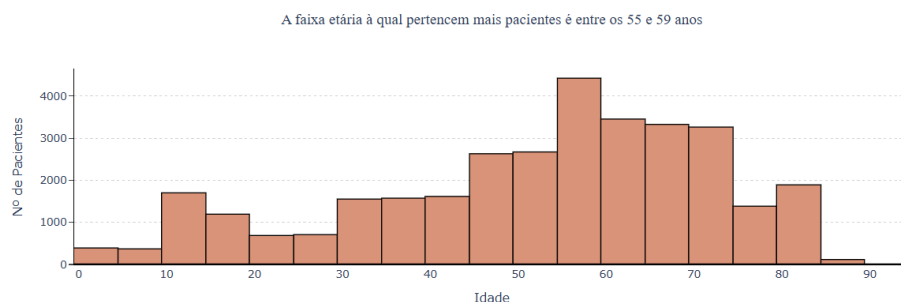


Figura 115: Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por idades

Através da variável *Nacionalidade* é possível analisar uma grande diferença entre as 12 diferentes nacionalidades, sendo que a grande maioria dos pacientes são de Portugal. Através do gráfico é possível observar que apenas 3.3% dos pacientes pertencem a outras nacionalidades. (Figura 116)

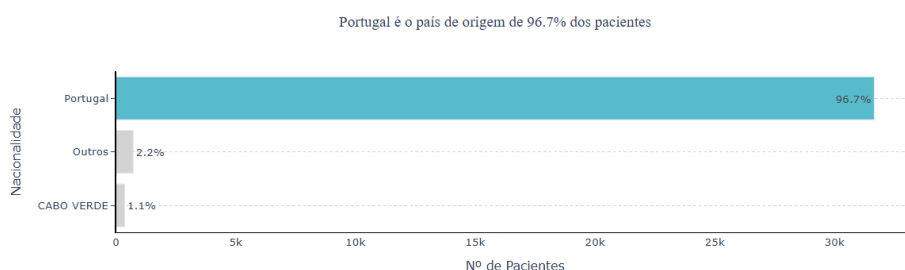


Figura 116: Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por nacionalidade

Nos gráficos relativos ao concelho acontece exatamente a mesma coisa, sendo que as variáveis *concelho_1ªconsulta* e *concelho_internamento* têm também os mesmos valores, dando gráficos iguais. Podemos então chegar à conclusão de que o concelho de Sintra é o mais comum entre os pacientes (16.1%), seguido do concelho de Vila Franca de Xira (7%). (Figuras 117 e 118)

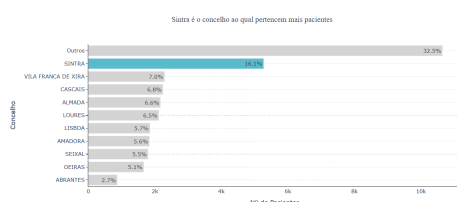


Figura 117: Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por concelho (1ª consulta)

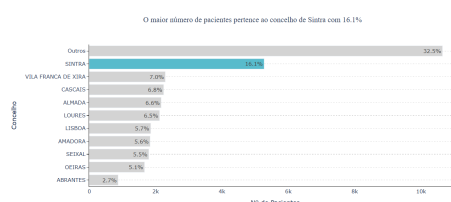


Figura 118: Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por concelho (internamento)

As entidades financeiras responsáveis divergem entre as responsáveis pelas consultas e as responsáveis pelos internamentos, na figura 119 estão representadas as entidades financeiras responsáveis na primeira consulta, sendo a entidade “ARSLVT-ACORDO” a responsável por cerca de 70% dos pacientes.

Já na figura 120, estão representadas as entidades financeiras responsáveis por internamentos, sendo a principal a “ADMINISTRACAO CENTRAL SISTEMA SAUDE I.P.” com 93.8% de pacientes.

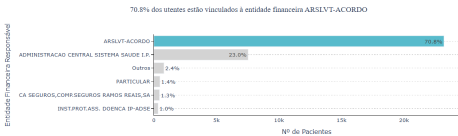


Figura 119: Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por entidade responsável por consultas

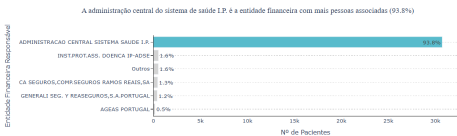


Figura 120: Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, entidade responsável por internamentos

Quanto aos serviços, existe alguma diferença entre as duas variáveis *SER_1ªconsulta* e *SER_internamento*. É possível observar, na figura 121, que o serviço de adultos é composto por 90.3% dos pacientes, sendo o serviço pediátrico composto por apenas 9.8% dos pacientes. Já na figura ??, relativa aos dados de internamento, é possível observar uma diferença, percebendo que neste caso, 88.8% dos pacientes estão internados no serviço de adultos e 11.2% no serviço pediátrico.

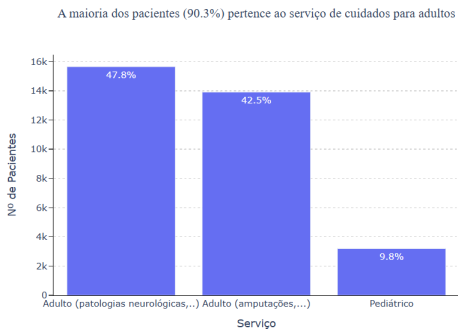


Figura 121: Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por serviço (1ª consulta)

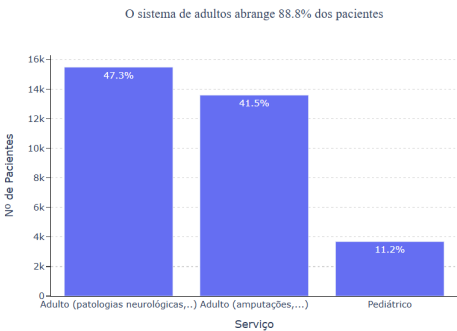


Figura 122: Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por serviço (internamento)

No sub-serviço, variável *Serviço2*, observamos, na figura 123, que o serviço relacionado à reabilitação de adultos continua a ser o com mais pacientes, cerca de 87.1%.

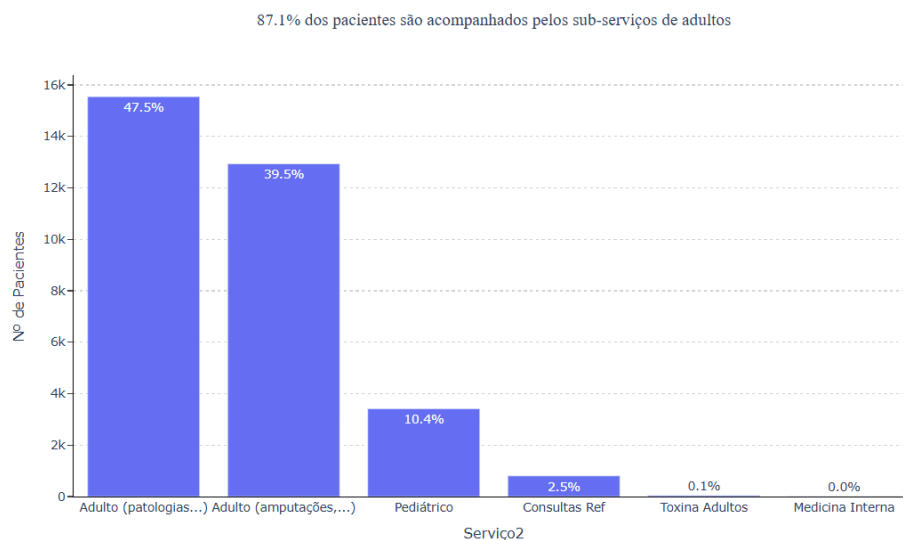


Figura 123: Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por sub-serviço

Quanto aos diagnósticos, através da utilização da variável *Diagnóstico*, é possível observar que o diagnóstico mais comum é a sequência de doença cerebrovascular, com 35.7% de pacientes. Todos os restantes diagnósticos têm uma percentagem inferior a 6%. (Figura 124)

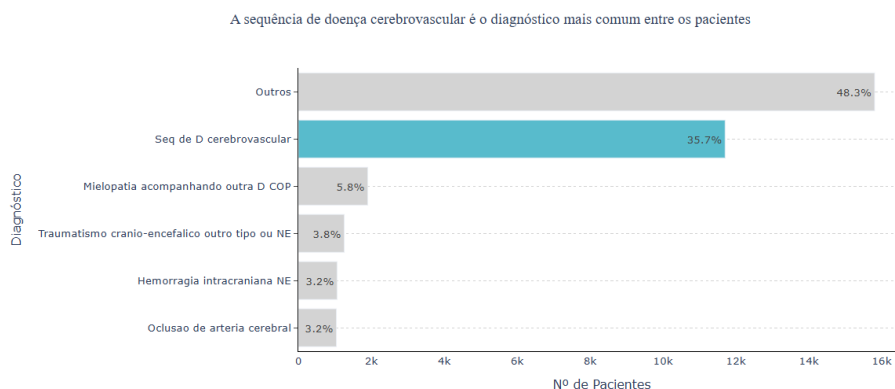


Figura 124: Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por diagnóstico

As patologias apresentadas são diversas, no entanto através do gráfico é possível identificar uma diferença significativa associada à patologia "lesão encefálica não traumática", que se destaca em relação às restantes. Enquanto as restantes não ultrapassam os 13.7% dos casos, a patologia "lesão encefálica não traumática" representa 30% dos pacientes. (Figura 125)

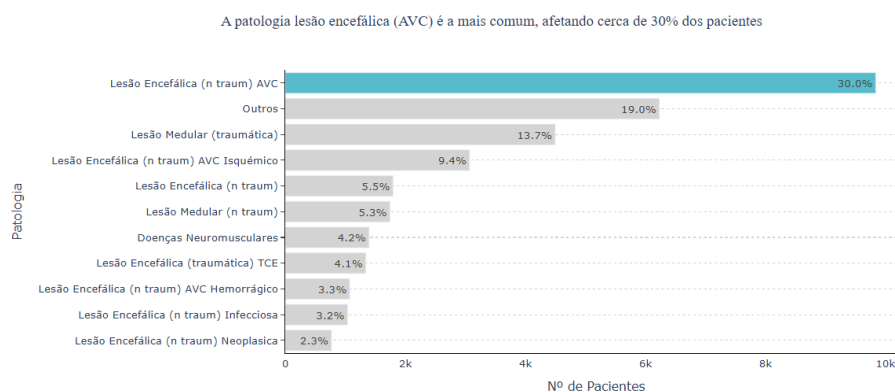


Figura 125: Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por patologia

Seguindo para a variável *Causa* observamos que 90.2% dos pacientes são seguidos devido a problemas de saúde. Os restantes pacientes estão distribuídos entre as restantes categorias. (Figura 126)

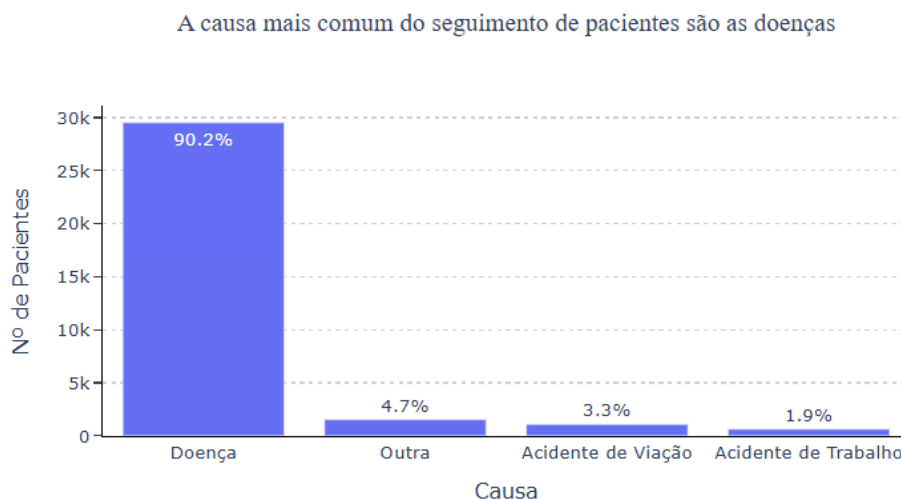


Figura 126: Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por causa

Já os tipos de reabilitação são diversos, sendo que os mais comuns têm a ver com a reabilitação neurológica. Com destaque substancial para a "reabilitação neurol-patologica encefálica" com mais de metade dos pacientes totais, 57.3%. (Figura 127)

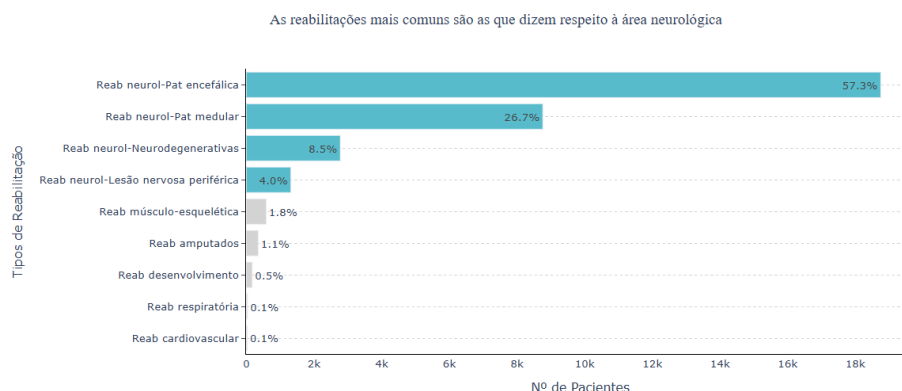


Figura 127: Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por tipologia

A designação de acto mais comum entre os pacientes foi a consulta de avaliação para admissão acompanhamento adulto, com 80.8% de pacientes, seguida da consulta de avaliação para admissão acompanhamento pediátrico, com 10% de pacientes. (Figura 128)

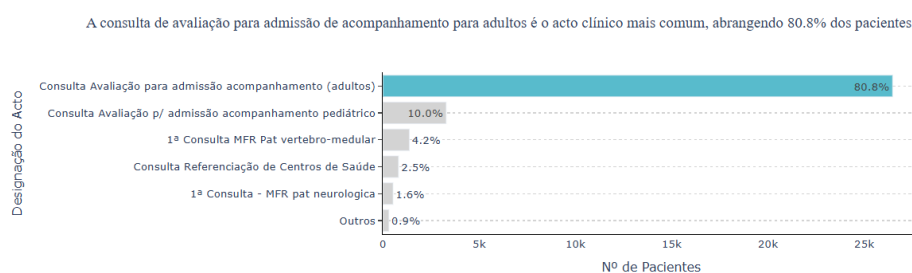


Figura 128: Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por acto

Quase todos os pacientes, neste caso 99.9% dos pacientes, atenderam a uma sessão terapêutica sendo que apenas 0.1% atenderam a mais que uma sessão. (Figura 129)

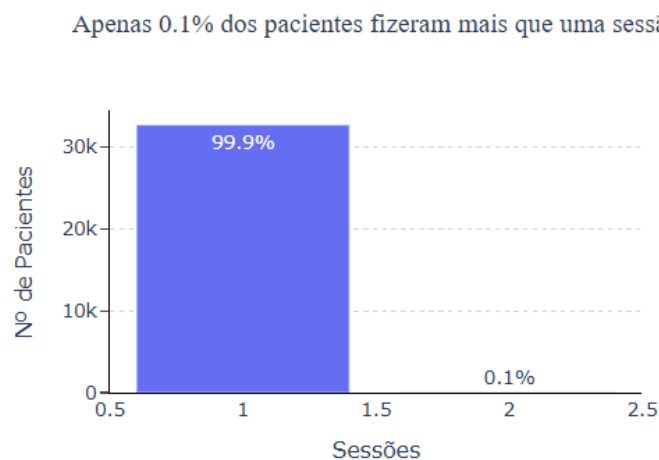


Figura 129: Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por sessão

Cada sessão é composta por um ou mais meios complementares terapêuticos e, através da utilização da variável *mcts* foi possível chegar à conclusão que a maioria das sessões tem apenas até 6 meios complementares terapêuticos, sendo raros os pacientes com sessões com 6 ou mais meios complementares terapêuticos. (Figura 130)

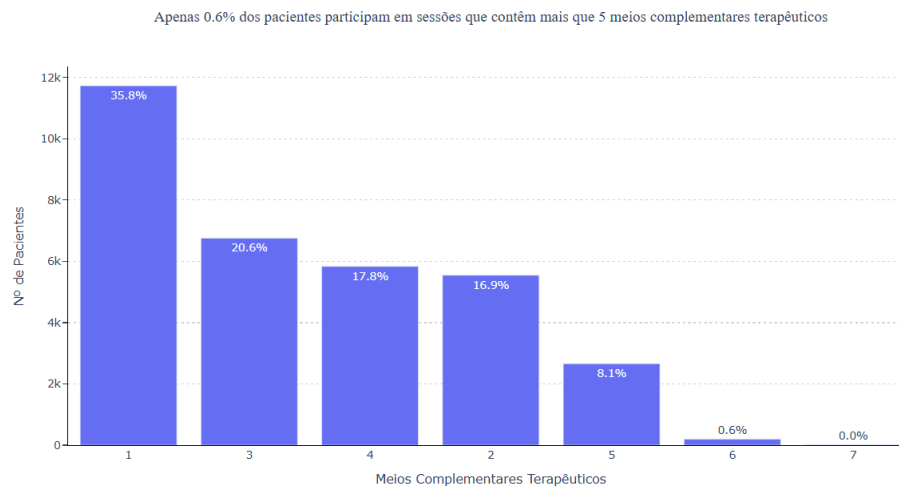


Figura 130: Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por meios complementares terapêutico

Para concluir, através das variáveis *ID_PROF* e *ID_PROF_sessoes*, verifica-se que o profissional com ID=5.0 acompanha 12.8% dos pacientes em consultas e o profissional com ID=82 é responsável por 4.8% dos pacientes em sessões, sendo estes os que assistem o maior número de pacientes nos respetivos serviços. (Figura 131 e 132)

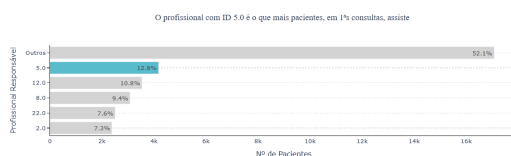


Figura 131: Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por profissional responsável pelas 1ª consultas

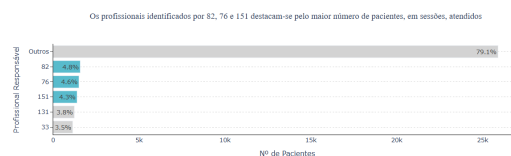


Figura 132: Distribuição dos utentes que posteriormente à 1ª consulta realizaram internamento e sessões, por profissional responsável pelas sessões