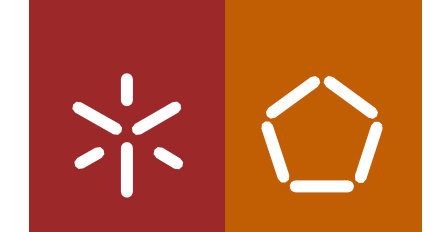




Afonso Augusto de Paula Lobo

Sistemas para a Inteligência do Negócio
Adaptativos em Saúde- Gestão de doentes

Universidade do Minho
Escola de Engenharia





Universidade do Minho
Escola de Engenharia

Afonso Augusto de Paula Lobo
A83946

Sistemas para a Inteligência do Negócio Adaptativos em Saúde- Gestão de doentes

Dissertação de Mestrado
Mestrado Integrado em Engenharia e Gestão de Sistemas de
Informação

Trabalho efetuado sob a orientação do
Professor Doutor Manuel Filipe Santos
Doutor Hugo Daniel Abreu Peixoto

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

Este é um trabalho académico que pode ser utilizado por terceiros desde que respeitadas as regras e boas práticas internacionalmente aceites, no que concerne aos direitos de autor e direitos conexos.

Assim, o presente trabalho pode ser utilizado nos termos previstos na licença abaixo indicada.

Caso o utilizador necessite de permissão para poder fazer um uso do trabalho em condições não previstas no licenciamento indicado, deverá contactar o autor, através do RepositóriUM da Universidade do Minho.

Licença concedida aos utilizadores deste trabalho



Atribuição-NãoComercial-SemDerivações
CC BY-NC-ND

<https://creativecommons.org/licenses/by-nc-nd/4.0/>

DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

AGRADECIMENTOS

Os meus agradecimentos vão para todos aqueles que de alguma forma contribuíram para que conseguisse concluir esta etapa com sucesso.

Gostava de agradecer de uma forma especial ao Professor Doutor Manuel Filipe Vieira Torres dos Santos que no papel orientador desta dissertação sempre se mostrou disponível a ajudar com todas as questões burocráticas, logística e desenvolvimento. Queria agradecer também o seu voto de confiança que me permitiu realizar a dissertação no tema que eu pretendia. Deixo também um agradecimento especial ao Doutor Hugo Daniel Abreu Peixoto que no papel de coorientador sempre se mostrou disponível para resolução das questões técnicas e de investigação.

Agradeço de forma sentida a toda a minha família e amigos que sempre me apoiaram e motivaram na realização deste projeto, foram sempre um suporte em todos os momentos do meu percurso académico.

RESUMO

Sistemas para a Inteligência do Negócio Adaptativos em Saúde- Gestão de doentes

A má planificação dos recursos nas entidades hospitalares consiste num fator comprometedor da qualidade do serviço prestado e da saúde financeira dos mesmos. Tal acontece devido a uma falta de planos que implementem uma boa gestão e previsão do fluxo de doente tendo uma otimização da alocação dos recursos condizente, permitindo melhorar a eficiência de serviço e uma minimização nos seus custos.

Nesta dissertação o projeto consiste na análise dos dados relativos a pacientes, suas admissões e recursos alocados na prestação de serviço do Centro Hospitalar do Tâmega e Sousa. Com uma análise posterior, são definidos e implementados dois tipos de modelos, primeiramente um conjunto de modelos de *Machine Learning* que nos permitam prever o fluxo temporal de admissão dos pacientes, seguido da implementação de modelos de otimização lógica para a gestão de recursos, mais concretamente a alocação ótima das camas que responda eficientemente às necessidades da unidade hospitalar tendo em conta o fluxo de doentes previsto.

Após a implementação bem-sucedida destes modelos, é realizada uma integração dos mesmos numa plataforma *ABI (Adaptive Business Intelligence)* permitindo a gestão e otimização em tempo real.

Palavras chave: Fluxo de doentes; Gestão de Recursos; Machine Learning; Otimização; Previsão.

ABSTRACT

Adaptive Business Intelligence in Healthcare – Patients Management

Poor planning of hospital resources is a compromising factor for the quality of the service provided and their financial health. This happens due to a lack of plans that implement good management and forecast of patient flow, having a consistent optimization of the allocation of resources, allowing to improve the efficiency of the service and minimization of its costs.

In this dissertation, the project consists of the analysis of data related to patients, their admissions, and resources allocated in the service provision of the Tâmega e Sousa Hospital Center. With further analysis, two types of models are defined and implemented, firstly a set of Machine Learning models that allow us to predict the temporal flow of patient admissions, followed by the implementation of logical optimization models for resource management, more specifically the optimal allocation of beds that efficiently meet the needs of the hospital unit considering the expected flow of patients.

After the successful implementation of these models, they are integrated into an ABI platform (Adaptive Business Intelligence) allowing this management and optimization in real-time.

Keywords: Forecast; Machine Learning; Optimization; Patient Flow; Resource Management

ÍNDICE

DECLARAÇÃO DE INTEGRIDADE.....	ii
Agradecimentos	iii
Resumo	iv
Abstract	v
Lista de abreviaturas/Siglas.....	viii
Lista de figuras	ix
Lista de tabelas.....	x
1. Introdução.....	1
1.1 Enquadramento	1
1.2 Motivações.....	2
1.2.1 Motivação Técnica.....	2
1.2.2 Motivação Pessoal.....	2
1.3 Objetivos e Resultados Esperados	3
1.4 Organização do Documento.....	4
2. Metodologias, Arquiteturas e Ferramentas	5
2.1 Metodologia de Investigação - DSR	5
2.2 Metodologia de Desenvolvimento – CRISP DM	7
2.3 Ligação e Cruzamento das metodologias.....	9
2.4 Arquitetura ABI	10
2.5 Ferramentas	13
3. Conceitos base e Estado da arte	14
3.1 Estratégia de Pesquisa Bibliográfica	14
3.2 <i>Data Mining</i>	14

3.2.1	Algoritmos e Técnicas de <i>Data Mining</i>	15
3.2.2	Técnicas e algoritmos de <i>data mining</i> aplicados na gestão de pacientes	16
3.2.3	Ferramentas <i>Data Mining</i> aplicadas na indústria da Saúde	16
3.3	Inteligência Artificial	17
3.4	Machine Learning	18
3.4.1	Aprendizagem Supervisionada	18
3.4.2	Aprendizagem não supervisionada	19
3.4.3	Aprendizagem por reforço	20
3.4.4	Modelos Preditivos de Séries Temporais	21
3.5	Diagrama de Relações	22
3.6	Otimização.....	22
3.6.1	Modelos de Otimização – Hill Climbing.....	23
3.6.2	Modelos de Otimização – Simulated Annealing.....	23
3.6.3	Modelos de Otimização – Genetic Algorithm	24
3.6.4	Comparação dos Modelos.....	24
3.6.5	Modelo de Otimização de Recursos numa Instituição Hospitalar.....	25
4.	Trabalho realizado e resultados obtidos	27
4.1	Compreensão do Negócio.....	27
4.2	Compreensão dos Dados.....	29
4.3	Preparação dos Dados.....	32
4.4	Modelação	37
4.5	Modelo de Previsão do Fluxo de doentes	37
4.6	Modelos de Otimização.....	40
5.	Conclusão	57
6.	Referências Bibliográficas	59

LISTA DE ABREVIATURAS/SIGLAS

IA	Inteligência Artificial
MLOA	Machine Learning Optimization Algorithms
RS	Random Search
SA	Simulated Annealing
HC	Hill Climbing
GA	Genetic Algorithm
AG	Algoritmo Genético
ABI	Adaptative Business Intelligence
CRISPDM	Cross-Industry Standart Process for Data Mining
DSRM	Design Science Research Methodology
PCA	Principal Component Analysis
ARIMA	Auto Regressive Integrated Moving Average
CHTS	Centro Hospitalar do Tâmega e Sousa
SVM	Support Vector Machines
MLP	Multilayer Perceptron
VRDP	Variable Resolution Dynamic Programming
LSTM	Long short-term memory
STD	Desvio Padrão
ML	Machine Learning

LISTA DE FIGURAS

Figura 1 - DSRM Process Model - [1]	6
Figura 2 - Fases do CRISP-DM [3]	7
Figura 3 - Domínios de aplicação do CRISP-DM [4]	9
Figura 4 - Módulo Preditivo [5]	11
Figura 5 - Módulo de Ótimização [5]	11
Figura 6 - Módulo de Adaptação [5]	11
Figura 7 - Arquitetura de um Sistema AB I[5]	12
Figura 8 - Ferramentas Utilizadas	13
Figura 9 - K-means Clustering [18]	20
Figura 10 - Diagrama de conceitos	22
Figura 11 - Gráfico representativo dos valores Null e NaN presentes no dataset.	31
Figura 12 - Exemplo de DataSet Resultante	34
Figura 13 - Exemplo de DataSet Resultante	35
Figura 14 - Exemplo de DataSet Resultante	36
Figura 15 - Fluxo de doentes de Medicina Interna de 2019 a 2022	38
Figura 16 - Divisão dados treino e dados teste	39
Figura 17 - Resultados da previsão	40
Figura 18 -. Score obtido pelo GA durante as 100 gerações	50
Figura 19 - Execução do Random Search	51
Figura 20 - Execução do Hill Climbing	51
Figura 21 - Execução do Algoritmo Genético	52
Figura 22 - Execução do Simulated Annealing	52

LISTA DE TABELAS

Tabela 1 - Cruzamento das Metodologias.....	9
Tabela 2 - Métricas	48
Tabela 3 - Parâmetros de input.....	49
Tabela 4 - Resultados Obtidos.....	50

1. INTRODUÇÃO

1.1 Enquadramento

A gestão do internamento compreende um conjunto de processos que visam, através de um eficiente planeamento, maximizar a disponibilidade de camas e condições para o doente tendo em conta a racionalização de custos. Este planeamento torna-se extremamente complexo quando os recursos são limitados e o conjunto de variáveis cujo valor é necessário conhecer com antecedência é considerável. A incorreta adequação das camas às necessidades é crítica pois acarreta um desajuste de meios e recursos humanos. Este desajuste tem custos significativos que é imperioso evitar. Quando o desajuste corresponde a uma sobrelotação implica a degradação da qualidade de serviço, o aumento de risco de doenças contagiosas e o escalonamento de recursos humanos em regime de urgência. Por outro lado, quando a disponibilidade de camas é superior às necessidades implica o custo desnecessário com recursos humanos. A planificação ajustada do número de camas disponível é essencial se os valores de que depende essa planificação forem conhecidos com antecedência e acuidade. Se for possível estimar com acuidade o tempo de internamento de cada doente dadas as suas condições clínicas, torna-se possível ajudar os profissionais a fazer um planeamento mais assertivo.

1.2 Motivações

1.2.1 Motivação Técnica

A sobrelotação das unidades Hospitalares é cada vez mais uma realidade, devido ao envelhecimento global da população e à degradação do estilo de vida levado nos tempos recentes. Este fenómeno leva a que as respetivas unidades e serviços de saúde não sejam capazes de suprir as necessidades apresentadas e verifiquem uma perda de serviço e performance, tendo como consequências finais o aumento de custos operacionais e o insucesso nas questões medicinais adjacentes. É por isso imperativo pensar em formas de minimizar este impacto e tentar otimizar ao máximo todos os processos e recursos necessários de modo a aumentar a eficiência e eficácia dos serviços de saúde. Com recursos à Inteligência Artificial e técnicas de *Data Mining* é possível analisar os dados Hospitalares e perceber de que forma podem ser utilizados modelos de *Machine Learning* com a finalidade de prever o fluxo de doentes em determinados períodos e otimizar os recursos disponíveis com base nos resultados previstos, permitindo assim com que os serviços de saúde estejam melhores preparados para responder às necessidades apresentadas.

1.2.2 Motivação Pessoal

A área de *Computer, Data Science* e toda a sua envolvente sempre causou em mim um despertar de interesse que aliado à resolução de problemas matemáticos, logísticos e ainda representando um valor tão importante como o sucesso das unidades de saúde e melhoria geral da qualidade de vida faz com que me sinta completamente motivado para desenvolver e participar neste projeto de dissertação.

1.3 Objetivos e Resultados Esperados

Tendo como questão de investigação formulada para este projeto – “Até que ponto é possível incrementar a eficiência da gestão de camas no internamento hospitalar com recurso a técnicas de inteligência artificial?”, os objetivos e resultados esperados na realização deste projeto de dissertação decorrem no sentido de promover o respetivo aumento da eficiência no planeamento e gestão de camas. Para tal torna-se necessário a investigação e definição de modelos preditivos de *Machine Learning* que permitam prever o fluxo dos doentes, a definição do conjunto de medidas de eficiência para avaliar cenários de disponibilidade de camas e afetação de recursos e a definição de modelos de otimização da disponibilidade de camas e afetação dos recursos humanos tendo como base os modelos preditivos definidos, os recursos disponíveis e a programação de cirurgias.

De modo atingir os objetivos esperados e responder ao problema proposto é necessário, numa primeira fase, a implementação dos modelos preditivos, com o devido tratamento e dedução de dados relativos ao fluxo de doentes no hospital. Com os resultados obtidos nestes modelos, decorre a implementação de um modelo de otimização que tem como base os mesmo resultados, os recursos disponíveis e as cirurgias programadas.

Após a definição e implementação destes modelos prossegue a implementação dos mesmos num sistema ABI que permita, em tempo real, a otimização da gestão do fluxo de doentes e do planeamento das camas disponíveis.

Em suma o propósito deste trabalho é o de promover a implementação de um sistema de ABI que permita, em tempo real, otimizar a gestão e planeamento de camas e do fluxo de doentes alguns objetivos foram definidos:

- A definição de um conjunto de modelos preditivos sobre o fluxo dos doentes;
- A definição de um conjunto de medidas de eficiência para avaliar cenários de disponibilidade de camas e afetação de recursos;
- A definição de modelos de otimização da disponibilidade de camas e afetação de recursos humanos com base nos modelos preditivos desenvolvidos nos pontos anterior, os recursos disponíveis e na programação de cirurgias.

1.4 Organização do Documento

Este documento apresenta-se dividido em quatro grandes capítulos de modo a abordar cada tópico de forma separada e detalhada facilitando a percepção do mesmo e apresentado todos os conteúdos relevantes para a Dissertação.

O capítulo 1 foca na introdução ao tema e ao problema da dissertação, identificando o problema e motivação do mesmo, os objetivos propostos para sua resolução, resultados esperados e quais as metodologias utilizadas para desenvolvimento deste projeto, uma relativa ao processo de investigação e outra relativa aos processos e técnicas de *Machine Learning* e *Data Mining*.

No capítulo 2 é introduzido e desenvolvido o Estado da Arte, onde são explicados todos os conceitos necessários a serem percebidos para a execução da dissertação. Começando com uma introdução ao conceito de *Data Mining*, quais as suas formas, aplicações e o seu contributo no projeto, seguindo-se para a componente de *Machine Learning* e abordando as principais técnicas e modelos constituintes, quer de previsão quer de otimização, comparando os vários existentes não de forma generalizadas, mas também aplicada ao contexto do problema.

No terceiro capítulo é apresentado o planeamento do trabalho a ser desenvolvido no âmbito da dissertação, assim como o prazo de entrega dos artefactos a serem produzidos.

O quarto e último capítulo é apenas destinado às referências bibliográficas utilizadas na realização do documento.

2. METODOLOGIAS, ARQUITETURAS E FERRAMENTAS

Na secção seguinte são abordadas as metodologias a ser utilizadas para investigação e desenvolvimento do projeto, a arquitetura constituinte e as ferramentas a utilizar.

2.1 Metodologia de Investigação - DSR

A metodologia de investigação recomendada e utilizada no desenvolvimento deste projeto foi a *Design Science Research Methodology for Information Systems* realizada por *Peffers, Tuunanen, Rothenberger e Chatterjee* em 2007. Esta foi desenvolvida com o intuito de criar uma “framework” para resolução e pesquisa de problemas correntemente investigados. São distinguidas seis principais iterações:

- *Problem identification and motivation*: Fase inicial de identificação do problema a ser investigado e respetiva solução. No projeto de dissertação em questão consiste na integração a uma plataforma de ABI (*Adaptive Business Intelligence*) de modelos preditivos relativos ao fluxo e admissão dos doentes num hospital aliado à criação de modelos de otimização de camas e restantes recursos baseados nas previsões efetuadas. Permitindo assim estimar o tempo de chegada e internamento dos pacientes, ajustando as necessidades do hospital ao mesmo. Aumentado a eficiência da execução nos processos e minimizando os custos.
- *Define the objectives for a solution*: Aqui são definidos os objetivos a que a solução traçada terá de responder. Neste caso concreto passará pelos modelos de previsão do fluxo de doentes aliado aos modelos de otimização de recursos. Para tal é necessário a condução de um estudo investigacional acerca dos vários modelos e técnicas, aplicadas ao contexto, de ML existentes, quer a nível de previsão, quer a nível de otimização.
- *Design and development*: Nesta fase é esperado o planeamento de um artefacto representativo do projeto em questão seja sob forma de relatório, *paper*, modelo ou um pedaço tecnologia sob forma de software / hardware. Relativamente ao projeto da

dissertação serão então correspondidos os modelos de *Machine Learning* de previsão de doentes e otimização de camas e recursos devidamente implementados.

- *Demonstration* : Demonstração devida, do artefacto desenvolvido que dará as respostas ao problema proposto.
- *Evaluation*: Fase de avaliação da solução final desenvolvida, a partir das devidas métricas aplicadas ao contexto.
- *Communication*: Iteração final do DP (*Design Process*) que consiste na comunicação do problema, da sua devida resolução, da importância da mesma e em que medida responde aos objetivos propostos na fase inicial [1].

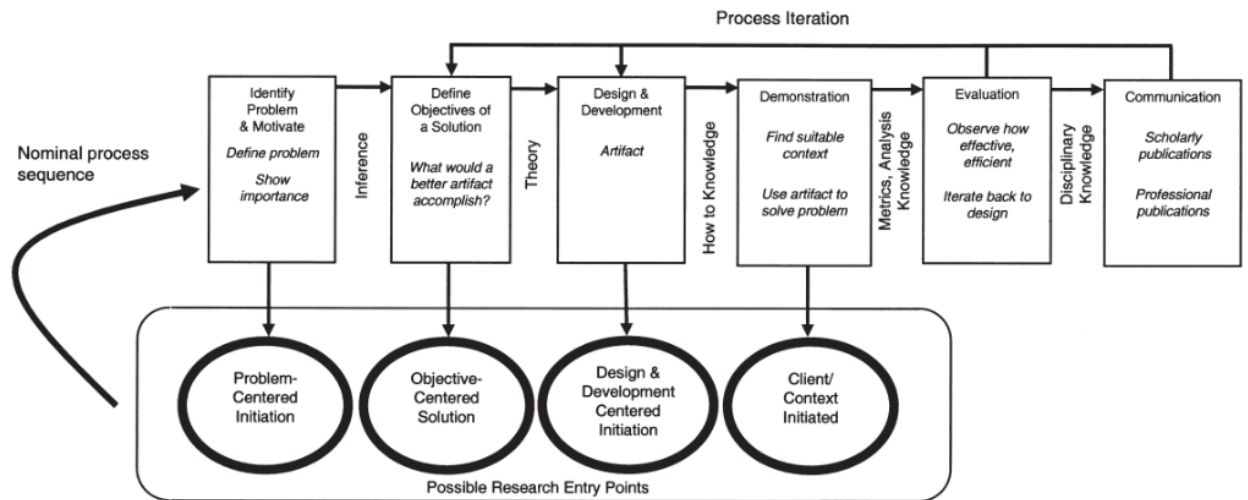


Figura 1 - DSRM Process Model - [1]

2.2 Metodologia de Desenvolvimento – CRISP DM

De modo a promover o sucesso na realização e cumprimento desta dissertação e seus respectivos objetivos propostos, torna-se necessário adotar um conjunto de metodologias previamente conhecidas que sirvam como guia e direção para todas as decisões e implementações a tomar. Será por este motivo escolhida a metodologia *CRISP-DM* (*Cross-Industry Standard Process for Data Mining*), metodologia publicada em 1999 com o âmbito de auxiliar e indicar todos os processos e passos necessários para aplicação de projetos *data mining*.

Neste modelo metodológico estão presentes quatro diferentes níveis de abstração, as fases, tarefas genéricas, tarefas específicas e instâncias dos processos [2] [3]. No nível de topo encontram-se as fases, cada fase é constituída por tarefas genéricas e abrangentes a todos os possíveis ambientes e implementações de um projeto *data mining*. Com o objetivo de descrever como devem ser efetuadas ações nas tarefas genéricas aplicadas a situações específicas foram constituídas as tarefas específicas. Por último as instâncias dos processos representam um conjunto de ações e decisões tomadas resultantes do desenvolvimento de um projeto *data mining* [3].

No primeiro e mais relevante nível de abstração são distinguidas seis diferentes fases ou iterações: *Business-Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, *Deployment*, contendo cada uma tarefas específicas correspondentes:

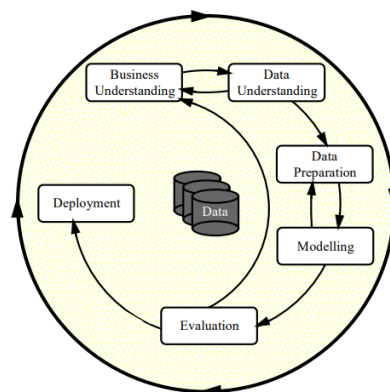


Figura 2 - Fases do CRISP-DM [3]

Buisness Understanding, os objetivos e requisitos são definidos, percepção do tipo de problema de *data mining* em questão e definição dos critérios de sucesso.

Data Understanding, nesta fase o foco centra-se na coleta de dados, no seu devido tratamento, exploração e descrição para determinar os atributos analisar e estudar a relação entre si.

Data Preparation, responsável pelas tarefas de *data cleaning*, lidando com a possível má qualidade de dados e construção de novos atributos derivados a partir dos existentes, tendo em conta o modelo e objetivos de *data mining*.

Modeling, centra-se aqui a escolha da técnica de modelação a ser utilizada, dividindo os dados em dados teste e dados de treino ou de modelo. Para a escolha acertada dos modelos é necessário ter em conta o tipo de dados disponíveis, o problema de negócio e qual será o modelo que melhor desempenhará na resolução do problema, avaliando-o segundo os critérios de avaliação de modelos.

Evaluation, decorre a avaliação dos resultados obtidos, interpretados à luz dos objetivos propostos. São definidos um conjunto de ações a ser tomadas e é realizada uma revisão de todo o processo.

Deployment, é a instância concreta do processo realizado, pode estar em forma de software ou um relatório final. Consiste na implementação, monitorização e manutenção do produto final [2].

A implementação e aplicação de projeto *data mining* seguindo a metodologia *CRISP-DM* é naturalmente aplicada nas diversas áreas e contextos em que se torna necessário, entre as quais se encontram, em segundo e primeiro posto na frequência de aplicação, a indústria da Educação e Investigação e a indústria da Saúde, podendo comprovar através da figura 3, resultante de um estudo realizado com o objetivo de perceber os principais domínios de aplicação do modelo *CRISP-DM*, pelo *Christoph Schröer et al. / Procedia Computer Science* [2].

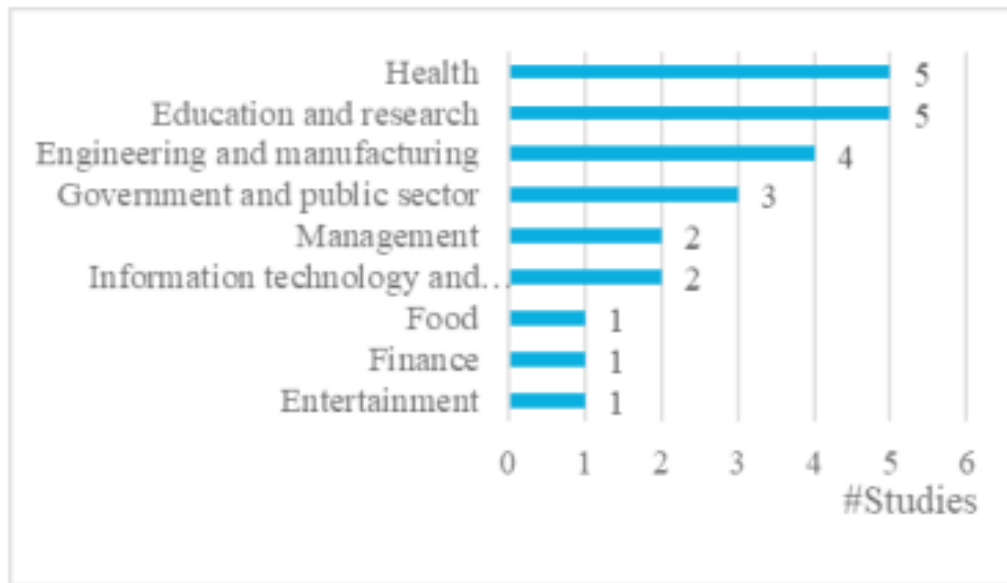


Figura 3 - Domínios de aplicação do CRISP-DM [4]

2.3 Ligação e Cruzamento das metodologias

De maneira a facilitar o desenvolvimento do projeto de dissertação seguindo ambas as metodologias sugeridas, foi desenvolvida uma tabela demonstrativa de ligação entre as iterações do *Design Science Research Methodology* e as fases do *CRISP-DM*.

Tabela 1 - Cruzamento das Metodologias

CRISP-DM DSRM	Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Problem identification and motivation	X					

Define the objectives for a solution	X	X				
Design and development			X	X		
Demonstration						X
Evalutation					X	
Comunication						X

2.4 Arquitetura ABI

Uma vez que este projeto tem como resultado a implementação dos modelos a criar no sistema ABI, a arquitetura seguida no mesmo será a arquitetura de um sistema ABI respetivo.

O Sistema ABI é um sistema representado pela sua capacidade de constante adaptação aos dados e ocorrências em tempo real. Este tipo de sistemas é caracterizado por três grandes módulos:

- Previsão, constituído pelos conjuntos de modelos preditivos;
- Otimização, conjunto de algoritmos aplicados aos dados previstos de maneira a obter as melhores soluções para os problemas em questão;

- Adaptação, módulo responsável pela aprendizagem constante do sistema, pegando no output dos modelos, corrigindo erros e aprendendo à medida que as execuções são realizadas.

Nas Figuras 4 a 6 abaixo podemos ver representados de forma ilustrativa o funcionamento e relação de cada um dos módulos:

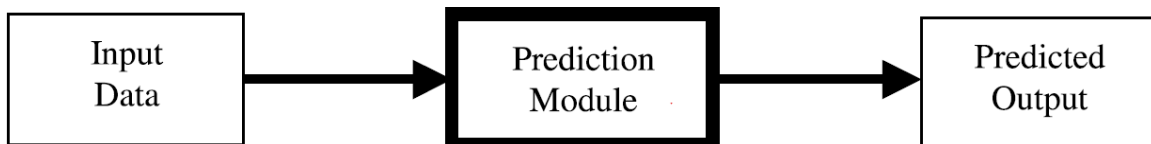


Figura 4 - Módulo Preditivo [5]

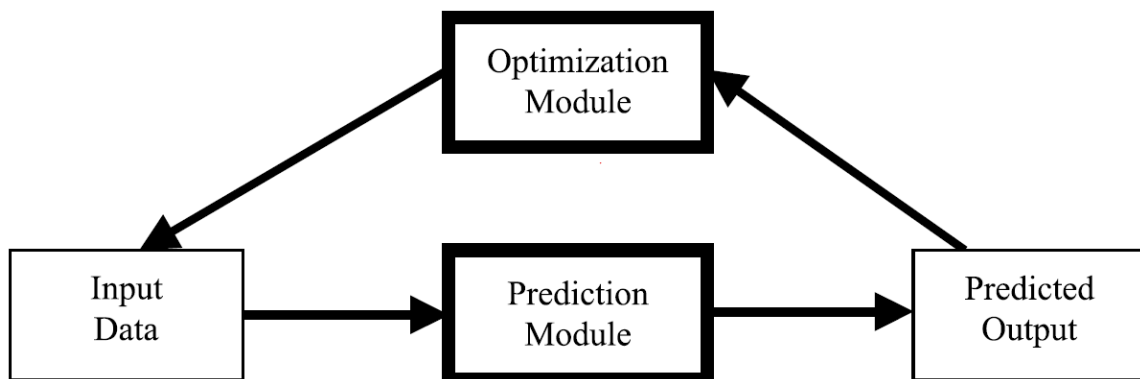


Figura 5 - Módulo de Otimização [5]

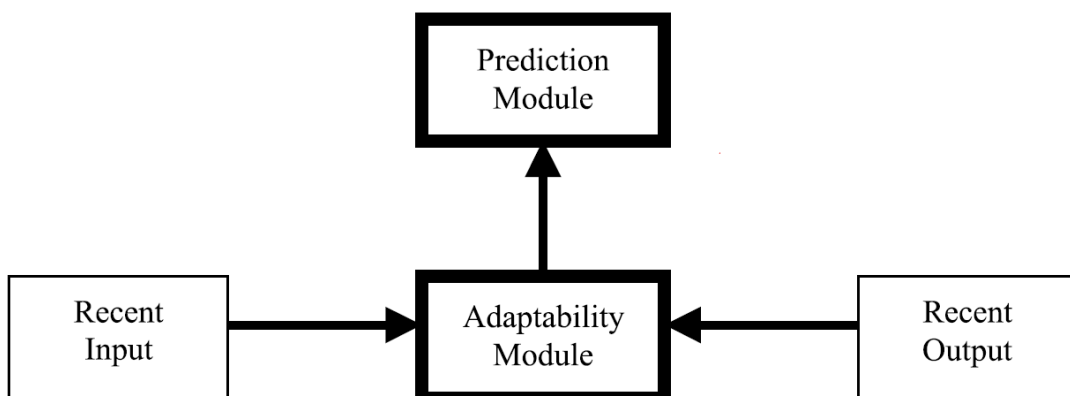


Figura 6 - Módulo de Adaptação [5]

Com estes três módulos este tipo de sistemas fornece-nos várias vantagens, assim como:

- Modelos de previsão baseados em resultados de *Data Mining*;
- Algoritmos de otimização completamente integrados com os modelos preditivos;
- Módulo de adaptação que permite a evolução e aprendizagem constante e atual do sistema garantindo a sua capacidade de contextualização em diferentes cenários;
- Transformação de dados em informação [6].

A arquitetura característica destes sistemas segue então uma linha de funcionamento onde conjuntos de dados em grandes quantidades são previamente preparados e analisado através das técnicas de *Data Mining* criando informação útil capaz de constituir um input para o módulo preditivo. Neste conjunto de dados previstos são realizadas as devidas otimizações com o intuito de chegar à melhor solução possível. Este output final é analisado pelo módulo de adaptação que corrige erros possíveis transforma outra vez este conjunto de dados num input de previsão, garantindo assim a evolução e adaptação constante dos dados e do respetivo sistema. Podemos ver representada esta arquitetura na figura 7, abaixo [7].

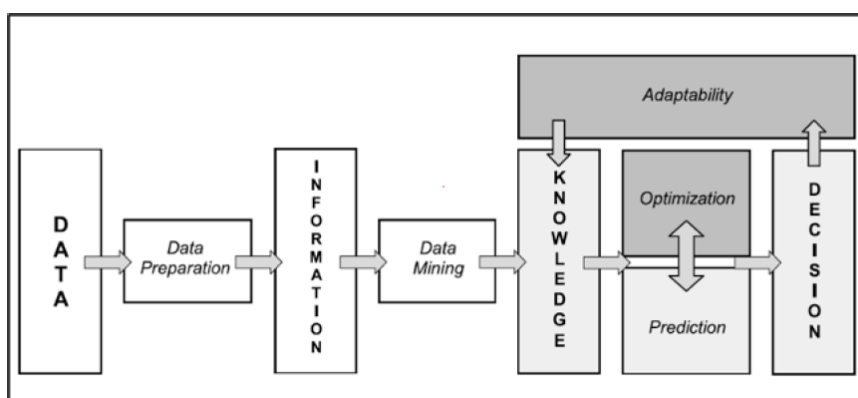


Figura 7 - Arquitetura de um Sistema AB [5]

A implementação destes sistemas complexos na indústria da saúde representa um grande passo na evolução e modernização dos mesmos. Com os mesmos é possível garantir uma melhor

performance e tomada de decisão em tempo real. Os benefícios apresentados vão desde a detecção e prevenção de doenças, a otimização de soluções clínicas até à resolução de questões logísticas de alocação e gestão de camas e recursos, garantindo assim o aumento da eficiência por parte das unidades Hospitalares que ficam capazes de responder às suas necessidades e prestar um melhor serviço aos pacientes[7].

2.5 Ferramentas

As ferramentas utilizadas para o desenvolvimento deste projeto foram reduzidas uma vez que não necessita de grandes recursos. A nível de base de dados foram apenas ficheiros Excel com os dados provenientes dos fluxos de doentes no Hospital em questão, o Centro Hospitalar do Tâmega e Sousa (CHTS), que posteriormente foram transformados em CSV's e tratados como *datasets*.

A linguagem de programação utilizada foi o *Python*, uma vez que oferece uma grande quantidade de bibliotecas e packages para desenvolvimento de algoritmos de *Machine Learning* assim como para realizar tratamento e análise de dados com as respetivas técnicas de *Data Mining*.

O ambiente de programação escolhido foi o *Google Colab*, pelo facto de não sobrecarregar as máquinas pessoais com a compilação e execução do código, apesar dos seus recursos computacionais serem limitados foi possível desenvolver o projeto sem grandes constrangimentos.



Figura 8 - Ferramentas Utilizadas

3. CONCEITOS BASE E ESTADO DA ARTE

Na secção abaixo estão abrangidos e explicados os conceitos necessários para compreender o trabalho desenvolvido, assim como a aplicação das tecnologias implementadas em situações testadas anteriormente.

3.1 Estratégia de Pesquisa Bibliográfica

Para realização da secção em questão foi definida uma estratégia de maneira a obter fontes credíveis e relevantes. Estas fontes constituem maioritariamente livros, artigos científicos, teses, dissertações e artigos de conferências.

Os motores de busca utilizados foram primariamente o *Google Scholar*, *Scopus*, *Science Direct* e *Web of Science*.

3.2 Data Mining

Data Mining ou *Knowledge Discovery in Databases (KDD)* é o nome a que se designa o processo de extração de valor e conhecimento a partir de um grande conjunto de dados compostos com *n* atributos, descobrindo e identificando diferentes tipos de padrões e análises que sirvam o propósito em questão. “*KDD* é um processo não trivial, interativo e iterativo, para a identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados” [8].

Os padrões identificados no processo de *Data Mining* podem ser de dois tipos, padrões descritivos ou padrões preditivos. Padrões descritivos são aqueles que descrevem o grande conjunto de dados e nos dão a informação acerca dos mesmos e seus atributos. Os padrões preditivos são padrões construídos com vista a solução de um determinado problema onde o foco é a previsão de um ou mais atributos tendo em conta os valores dos outros atributos [8].

3.2.1 Algoritmos e Técnicas de *Data Mining*

De modo a serem obtidos e interpretados os padrões escondidos nos dados é necessário a aplicação de diferentes algoritmos e técnicas às quais ocorreremos consoante o objetivo em questão. Na identificação de padrões no processo de *Data Mining* distinguem-se técnicas como a associação, classificação, *clustering*, previsão e padrões sequenciais.

Associação dos dados, uma das técnicas mais utilizadas, consiste na descoberta de padrões a partir da relação entre os atributos dos dados.

Modelo de Classificação dos dados para identificação de padrões, é a técnica mais conhecida e baseada em algoritmos de *machine learning* onde cada atributo pertencente a um *DataSet* é classificado e inserido num conjunto de classes/padrões pré-definidos de acordo com as suas características.

A técnica de *Clustering* é similar à de técnica de classificação, porém por sua vez os padrões são construídos a partir da organização dos dados em grupos com atributos semelhantes, "*Clustering is "the process of organizing objects into groups whose members are similar in some way"* [9]. Um Cluster por sua vez é por definição, uma coleção de dados cujos atributos são similares entre si e distintos aos atributos da coleção de dados constituintes dos *Clusters* restantes.

A Previsão tem como base a descoberta a relação entre variáveis independentes e entre variáveis dependentes e independentes. Por definição, em *Data Mining*, entende-se por variável independente são atributos já conhecidos e variáveis de resposta que pretendemos prever.

Podem também ser utilizado *Clustetring* para ajudar à previsão de certos atributos como é o caso do algoritmo *K-Means*.

A análise de padrões sequenciais é uma técnica de *data mining* utilizada na análise de negócio onde o foco persiste em descobrir padrões idênticos nas transações de dados num determinado período de negócio (in & 2012, 2012) [10].

3.2.2 Técnicas e algoritmos de *data mining* aplicados na gestão de pacientes

O uso de *data mining* nos Sistemas de saúde, mais concretamente na gestão de pacientes tem vindo a ser uma vertente cada vez mais explorada devido aos benefícios inerentes à aplicação das suas técnicas, que permitem não só um tratamento mais eficiente dos pacientes atuando na área de previsão de patologias, assim como na gestão e alocação de pacientes numa infraestrutura hospitalar [11] [10].

De modo a efetuar um processo de *data mining* adequado e uma correta identificação de padrões que seja por sua vez útil à indústria dos Sistemas de Saúde torna-se necessário explorar quais as principais técnicas que mais se adequam na extração da grande quantidade de dados característicos dos mesmos. Segundo o “*Journal of Kind Stad University – Computer and Information Sciences*” destacam-se as técnicas de: classificação, mais concretamente os algoritmos de *SVM(Support Vector Machines)* e *Artificial Neural Networks*, atuando na deteção e prevenção de certas patologias com base nos históricos dos pacientes, associação onde permite prever determinadas patologias com base na sua relação com os restantes atributos, *clustering*, com o algoritmo *K-means*, criando e identificando padrões dos dados dos pacientes, efetuando análises descritivas e preditivas sobre os diferentes tipos de pacientes e a suas respetivas admissões a determinada infraestrutura e por fim a deteção de *outliers* de modo a identificar atributos fora do comum ou erros nos dados adquiridos [11].

3.2.3 Ferramentas *Data Mining* aplicadas na indústria da Saúde

Para realizar o processo de *Data Mining* e pôr em prática todos as suas técnicas e algoritmos correspondentes é necessário a utilização de ferramentas que nos permitam lidar com a grande quantidade de dados característicos e retirar a informação contida nos padrões escondidos que os mesmos apresentam. De entre as muitas ferramentas existentes destacam-se cinco de grande reconhecimento:

- *KNIME*, que constitui uma plataforma com base de java, com interface gráfica de integração, análise e tratamento dos dados;

- *R*, uma linguagem de programação destinada à computação estatística, elaboração de algoritmos de *Machine Learning* e execução de tarefas de *data mining*, é facilmente integrada com linguagens como *Python*, *C* e *C++*;
- *Rapid Miner*, uma ferramenta de análise, com base Java destinada a *data mining*, *text mining*, *predictive* e *business analytics*, esta ferramenta suporta tarefas como, integração, limpeza, tratamento, integração, validação, exploração e modelação dos dados, permite também a criação de aplicações web e lidar com a complexidade dos processos de análises da grande quantidade de dados com o acesso à tecnologia *Hadoop*;
- *Scikit-Learn*, uma biblioteca de *machine learning* em *python* que contém todos os algoritmos e técnicas de *data mining*, contendo a capacidade de minerar todo o tipo de dados assegurando uma boa performance;
- *Spark*, uma ferramenta desenhada para lidar com dados de larga escala executando modelos de *data mining* através de vários nodos quando os dados são demasiado extensos e complexos para um computador apenas [11].

3.3 Inteligência Artificial

A Inteligência Artificial (IA) é um campo de pesquisa em rápido desenvolvimento que visa criar máquinas capazes de realizar tarefas que normalmente requerem inteligência humana. Isso inclui tarefas como reconhecimento de fala, aprendizado, resolução de problemas e compreensão de linguagem natural. Existem aplicações em diversas áreas, desde jogos, finanças, transporte e saúde. Na indústria da saúde, é utilizada com o intuito de melhorar a precisão do diagnóstico, desenvolver tratamentos personalizados e otimizar a alocação de recursos.

Dentro da IA, existem várias subáreas, estas incluem, *Machine Learning*, processamento de linguagem natural e visão computacional. O *Machine Learning* é uma técnica que permite que as máquinas aprendam com os dados e melhorem seu desempenho ao longo do tempo. O

processamento de linguagem natural é usado para permitir que as máquinas entendam e gerem linguagem humana. A visão computacional é usada para permitir que as máquinas interpretem imagens e vídeos.

Nos dias de hoje IA é um termo cada vez mais familiar e menos desconhecido, é o presente e o futuro, com diversos benefícios práticos à vida humana, destacando o seu contributo na indústria da saúde.

3.4 Machine Learning

Machine Learning, por definição é, um conjunto evolutivo de algoritmos computacionais cujo objetivo é a simulação da inteligência humana na capacidade de aprendizagem automática a partir do ambiente inserido e dados fornecidos [12]. Constitui o processo de previsão realizada através destes algoritmos, utilizando dados passados na forma de *datasets*, onde a qualidade e o tamanho dos mesmos constituem dois fatores cruciais para o sucesso dos mesmos [13]. Este é aplicado nas diversas áreas compreendidas desde a classificação de texto, reconhecimento facial, previsão de vendas até ao diagnóstico de doenças e a restante aplicação na indústria da saúde.

No capítulo de aprendizagem automática é possível distinguir três grandes tipos de aprendizagem: aprendizagem supervisionada, aprendizagem não supervisionada e aprendizagem por reforço.

3.4.1 Aprendizagem Supervisionada

Aprendizagem supervisionada consiste no tipo de aprendizagem onde um computador recebe uma série de dados, relativos a experiências passadas, na forma de *dataset*, como input, é destacado um determinado atributo conhecido como *target* para prever os seus valores futuros. É através da aplicação de algoritmos característicos que se realiza a previsão do mesmo para todos os pontos desconhecidos com base nos restantes atributos presentes no *dataset* (Mohri et al., 2018) [15].

Os algoritmos e problemas mais associados a este tipo de aprendizagem são os de classificação, onde o objetivo se centra na aprendizagem de um sistema de classificação presente no input fornecido e os dados representam valores categóricos. Quando estes problemas se focam na previsão de valores numéricos são associados aos mesmos, algoritmos de regressão [13].

De entre os vários algoritmos existentes aplicados nas áreas de Classificação e Regressão os mais conhecidos e aplicados são algoritmos como *Linear Classifiers (Logical Regression, Support Vector Machines e Naive Bayes Classifier)*, *Neural Networks* e *Decision tree* como o *Random Forest* [13]. Num estudo realizado pelo departamento de *computer science* da Cornell University, onde foram comparados todos estes principais algoritmos característicos a este tipo de aprendizagem e problemas, testados através de 11 diferentes fontes de dados e avaliados através das métricas de performance e avaliação de modelos de *machine learning*, foram distinguidos com indicadores de melhor performance no *overall* as *Calibrated Boosted Trees*. *Random Forest* foi o algoritmo com a segunda melhor performance seguida de algoritmos como, *Uncalibrated Bagged Trees*, *Calibrated SVM's* e *Uncalibrated Neural Networks*. Os modelos ou algoritmos que se destacaram pela negativa na sua performance em problemas de Classificação foram modelos como, *Naive Bayes*, *Logistic Regression*, *Decision Trees* e *Boosted Stumps* (Caruana & Niculescu-Mizil, 2006).

3.4.2 Aprendizagem não supervisionada

Aprendizagem não supervisionada diferencia-se da supervisionada na medida que o software não necessita de um atributo target no seu input para realizar o processo de aprendizagem automática [17]. Constitui o modelo de aprendizagem mais comum presente no cérebro humano.

Os principais algoritmos constituintes deste tipo de aprendizagem centram-se no *Principal Component Analysis* (PCA), técnica utilizada para reduzir o número de dimensões de um conjunto de dados, agrupando variáveis semelhantes numa só e na identificação de padrões em algoritmos como *K-Means* realizando *Clustering* num conjunto de dados sem um atributo target, sendo por sua vez um dos algoritmos mais simples e utilizados de aprendizagem não supervisionada [18].

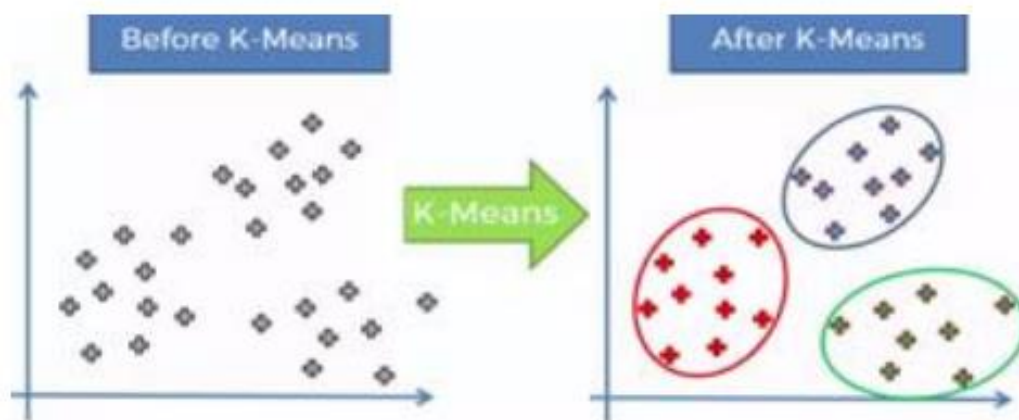


Figura 9 - K-means Clustering [18]

Este modelo de aprendizagem permite realizar análises preditivas e descritivas, contendo várias aplicações no mundo real na detecção de *outliers* nos dados, detecção de fraudes e reconhecimento facial [19].

3.4.3 Aprendizagem por reforço

Neste tipo de aprendizagem as decisões efetuadas pelo software de modo a acumular e aprender a informação que lhe é atribuída é realizada através do reforço positivo, isto é, o software pode seguir vários caminhos de modo a chegar a um objetivo de previsão, uns constituem a solução, outros não, cada vez que o software segue um “caminho” ou decide acertadamente é lhe atribuída uma recompensa, realizando assim a aprendizagem em questão. É definido assim um modelo onde a aprendizagem do agente com o meio ambiente é efetuada através da tentativa e erro em que são descobertas quais ações geram maior recompensa através da experiência, podendo tal influenciar não só o resultado e a solução final como a próxima ação a ser tomada.

Os algoritmos mais utilizados neste tipo de aprendizagem são, *Decision Trees*, *Logic Based Methods* e *Variable Resolution Dynamic Programming* (VRDP). [20].

3.4.4 Modelos Preditivos de Séries Temporais

A previsão de séries temporais nada mais é que a previsão de dados com base em conjuntos de registos que acontecem ao longo do tempo. Esta previsão por sua vez constitui uma ferramenta útil para prever o fluxo de pacientes nas unidades hospitalares. Ao analisar dados históricos sobre admissões, altas e transferências de pacientes, permite aos hospitais antecipar melhor a demanda futura e alocar recursos de acordo.

Vários estudos e artigos científicos discutem o uso da previsão de séries temporais para o fluxo de pacientes em hospitais.

Um artigo de Shruti Kaushik intitulado “*AI in Healthcare: Time-Series Forecasting Using Statistical, Neural, and Ensemble Architectures*” compara o desempenho de diferentes técnicas estatísticas, neurais e de agrupamento na previsão das despesas médias semanais dos pacientes com certos medicamentos para dor. Foi descoberto no mesmo que um modelo de agrupamento combinando previsões do modelo *autoregressive integrated moving average* (ARIMA), do *multilayer perceptron* (MLP) e do *long short-term memory* (LSTM) superou os modelos individuais [21].

Outro estudo realizado por Fouzi Harrou com o nome de “*Monitoring patient flow in a hospital emergency department: ARMA-based nonparametric GLRT scheme*” propõe um método para prever fluxos anormalmente altos de pacientes em departamentos de emergência usando dados de séries temporais. Foi desenvolvido um modelo ARMA para descrever o fluxo nominal de pacientes no departamento de emergência e um teste de razão de verossimilhança generalizada para detetar estas anomalias [22].

Concluindo, a previsão de séries temporais apresenta ser uma ferramenta valiosa para prever o fluxo de pacientes em hospitais. Analisando dados históricos e através métodos estatísticos ou de modelos de ML apropriados, as unidades hospitalares conseguem antecipar melhor a o fluxo de pacientes futuro e alocar recursos em conformidade.

3.5 Diagrama de Relações

De maneira compreender melhor os conceitos acima abordados nas suas aplicações práticas, segue um diagrama representativo das ligações e relações que cada conceito tem com os restantes.

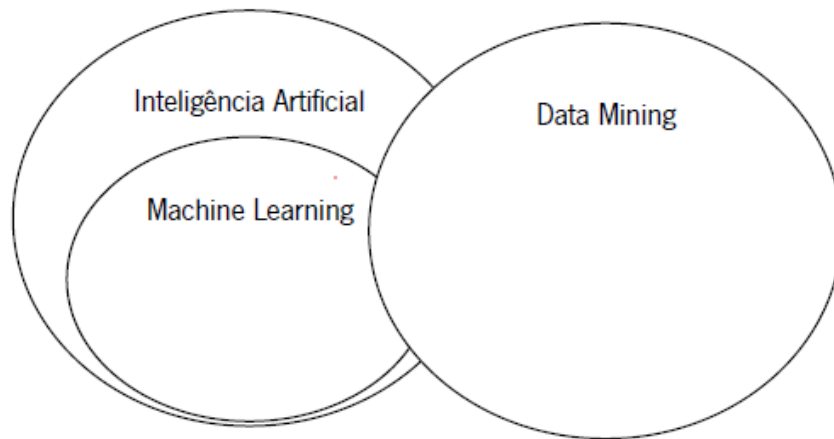


Figura 10 - Diagrama de conceitos

Sendo a Inteligência Artificial uma tecnologia que permite a criação de sistemas capazes de simular a inteligência Humana, o ML apresenta-se como uma subsecção concentrada no conjunto de técnicas e modelos capazes de produzir aprendizagem de máquinas. Este modelo por sua vez tem por bases grandes conjuntos de dados, dados esses que são previamente tratados e analisados através de técnicas de *Data Mining*.

3.6 Otimização

Algoritmos de Otimização são um tipo de algoritmos cuja função é, a partir de um conjunto soluções possíveis de um problema, obter a solução ótima aplicada ao seu contexto, minimizando custo ou maximizando lucro. Estes algoritmos são conhecidos também como algoritmos heurísticos, destacam-se entre os mesmos: *Hill Climbing*, *Simulated Annealing* e o *Genetic Algorithm*.

A aplicação deste tipo de técnicas torna-se necessário no desenvolvimento da dissertação devido à sua utilidade na resolução de problemas de mapeamento, ajustando-se assim à gestão de camas e recursos presentes nos hospitais, tendo em conta o conjunto solução dos modelos preditivos da gestão de clientes.

3.6.1 Modelos de Otimização – Hill Climbing

O *Hill Climbing* é um modelo de otimização cujo objetivo é encontrar um local mínimo ou máximo dependendo do objetivo. O modelo começa com uma solução inicial gerada aleatoriamente, ou não, são aplicadas alterações locais a elementos da solução consultando os seus vizinhos. O custo de cada um é avaliado através da aplicação da *fitness_function*, caso se observe uma melhora no mesmo, o processo volta ao início partindo da solução que representa o vizinho em questão, não se observando melhora em nenhum dos vizinhos locais da solução, esta mantém-se e é repetido também o processo. Este acaba quando as alterações efetuadas à solução inicial deixam de representar uma melhora positiva no custo da mesma, não existindo nenhum vizinho melhor que o atual sendo este um mínimo ou máximo ótimo ao problema [23].

3.6.2 Modelos de Otimização – Simulated Annealing

Simulated Annealing é um método análogo à termodinâmica, mais concretamente com o arrefecimento de metais e líquidos, onde a temperatura é reduzida gradualmente atingindo o ponto de equilíbrio térmico. Neste existe um parâmetro T , que define a temperatura inicial, um parâmetro denominado como *cooling*, responsável pela taxa de arrefecimento e o *step*, responsável pela alteração dos elementos da solução consultando assim os seus vizinhos. Na inicialização do processo é gerada uma solução inicial em que é escolhido um dos seus elementos ao acaso para ser alterado conforme o parâmetro *step* definido, ao efetuar a alteração esta solução passa a corresponder a um vizinho da inicial. Através da *fitness function* é calculado o custo da nova solução. A grande diferença neste método é que nem sempre o vizinho com melhor custo é escolhido uma vez que pode existir a probabilidade de estarmos a escolher um caminho que à partida tende a ser melhor, mas pode levar a atingir um mínimo local e não mínimo ótimo, esta escolha é baseada numa probabilidade calculada no algoritmo. Este processo segue todas as iterações até que a temperatura atinga um ponto previamente definido, a cada iteração é efetuado um arrefecimento da temperatura [24].

3.6.3 Modelos de Otimização – Genetic Algorithm

O Algoritmo Genético é um algoritmo de procura que tem como base a teoria da evolução de Charles Darwin, que atua sobre um conjunto de possíveis soluções, funcionando através da genética e seleção “natural”. Todas as soluções possíveis são chamadas de indivíduos de uma população ou genes de uma solução.

Para inicialização deste algoritmo é gerado um conjunto de n soluções denominado como população inicial, através da mesma decorrem transformações dos seus indivíduos com recurso à mutação dos seus elementos ou ao seu *cross-over* (junção de duas soluções, gerando uma só) realizando assim a criação de novas soluções e novas populações. Tendo em conta o número de elitismo, número dos n melhores indivíduos, denominados como “pais”, a serem escolhidos para gerar a nova população, são então escolhidos os que verificam melhor resultado na *fitness_function* representado assim uma melhor solução. O algoritmo de pesquisa acaba aquando é atingido o número de populações / gerações pretendidas na inicialização do mesmo [25].

A classe dos algoritmos genéticos tem aplicações nos diversos problemas de *machine learning*, atuando na pesquisa de uma solução ótima através de um conjunto solução gerado pelos métodos de ML em questão. Atuam sobre espaços solução discretos onde métodos de gradiente não tem aplicação, por norma, com melhor performance do que os métodos referidos anteriormente (*Hill-Climbing*, *Simulated-Annealing*). As principais utilizações centram-se na definição de soluções para problemas de mapeamento ou gestão de recursos e problemas de planeamento com restrição horária.

3.6.4 Comparação dos Modelos

De maneira a facilitar a realização do projeto da dissertação é necessário obter uma perceção acerca das três hipóteses de algoritmos de otimização e qual se adequa melhor ao contexto atendendo à sua performance.

Já vários estudos foram realizados com o objetivo de perceber qual o método/algoritmo que permite chegar a uma “melhor” solução ótima. Para resolver o problema de *timetabling* ou

definição horária para ocorrência de eventos com certos recursos disponíveis, o Departamento de Inteligência Artificial da Universidade de Edinburgo realizou um estudo que tinha como objetivo a otimização horária dos eventos, aulas ou exames, nas instituições escolares. Para tal foi definido um problema onde se pretendia a definição de um calendário de exames para uma instituição de ensino em que nenhum exame poderia acontecer ao mesmo tempo, os alunos tinham de ter pelo menos duas horas de intervalo entre os exames no mesmo dia e a capacidade da instituição é limitada dependendo do horário, quantos menos horas extra foram utilizadas na definição deste calendário melhor é considerada a solução. Após realização do estudo e comparação das performances foi concluído que os métodos *Hill Climbing* e *Simulated Annealing* ofereciam uma melhor qualidade da solução, porém o Algoritmo Genético permite obter uma variedade de soluções distintas para o problema ainda que não tão ótimas [26].

De modo a comparar estes métodos, não só em problemas de *timetabling* mas também nos de mapeamento, gestão ou alocação de recursos, como a distribuição de camas num hospital, foi atendido um estudo realizado pela Universidade de Grenoble onde o objetivo prendia pela resolução de um problema de mapeamento ou alocação de processos de comunicação nos processadores de uma máquina de memória distribuída comparando a performance dos três métodos ou algoritmos tendo em conta a qualidade da solução ótima encontrada e o tempo de computação necessário. Foi verificado após comparação das métricas de avaliação de performance que a qualidade das soluções oferecida pelo *Hill Climbing* era inferior ao *Simulated Annealing* obtidas, porém com menor tempo computacional, os resultados obtidos pelo *Simulated Annealing* eram comparáveis ao Algoritmo Genético apresentando o último menos tempo de computação. Foi também estudada a hipótese da implementação híbrida do Algoritmo Genético com o *Hill Climbing* o que por sua vez oferecia a melhor solução em termos de qualidade, mas o custo computacional era demasiado elevado, sendo considerada por isso o Algoritmo Genético a melhor opção para resolução deste tipo de problemas. A mesma Universidade em questão utiliza o Algoritmo Genético como método preferencial para otimização em problemas de mapeamento, com grande sucesso, nas áreas de medicina, robótica e redes neuronais [23].

3.6.5 Modelo de Otimização de Recursos numa Instituição Hospitalar

A má gestão dos recursos nas instituições hospitalares constitui um problema presente na indústria da saúde, comprometendo não só a qualidade de serviço e representando um aumento nos custos inerentes da manutenção do mesmo. Torna-se por isso uma necessidade real, realizar uma otimização de recursos humanos e materiais destas instituições. Neste intuito o Departamento de Engenharias Industrial e Gestão Sistemas de Informação da Universidade de Tehran realizaram um estudo acerca da eficiência de aplicação do Algoritmo Genético para otimização dos recursos humanos no Tous Hospital em Tehran. Os dados utilizados foram colecionados num período de 36 meses contendo informação acerca do fluxo de clientes, em quantidade e espaçamento temporal. O objetivo passava pelo ajuste dos horários laborais dos médicos e nove unidades especializadas hospitalares tendo em conta os dados recolhidos, de modo a encontrar qual o mínimo ótimo de pessoal staff que permitisse um aumento da eficiência dos processos e a minimização dos custos. Para tal foi executado um algoritmo genético em que a *fitness function* tinha como parâmetros base os salários do pessoal necessário e número de clientes atendidos por cada um em três turnos de trabalho. Após 500 execuções / gerações foi encontrada, com sucesso, uma solução ótima de alocação do staff para o atendimento máximo de clientes, maximizando a eficiência e minimizando o custo, de 69 funcionários necessários comparativamente aos 108 utilizados frequentemente no hospital.

No mesmo estudo é referida a limitação acerca da obtenção de dados das unidades hospitalares e da polivalência da aplicação deste método, com sucesso, em outros hospitais [27].

4. TRABALHO REALIZADO E RESULTADOS OBTIDOS

Nesta secção são apresentados os trabalhos realizados e os resultados obtidos. A secção está organizada em torno das fases do processo CRISP-DM.

4.1 Compreensão do Negócio

Primeiramente, antes de qualquer desenvolvimento ou receção e tratamento de dados foi importante perceber aquilo que eram os problemas presentes e respetivos os objetivos propostos para realização do projeto e como cumprir os mesmos.

O Hospital de Marco Canaveses tinha várias questões por ver resolvidas, entre as quais, primeiramente se centra, ter uma noção de quantos pacientes ou doentes são esperados nos próximos períodos, sejam estes dias, semanas ou meses. De modo a poder planificar a gestão dos seus recursos de uma forma adequada e garantindo que responde às necessidades apresentadas foi então proposto criar um modelo de previsão diária do fluxo de doentes, com base nos dados passados de admissões, altas e internamento dos mesmos.

Posteriormente à realização do modelo preditivo e conseguindo ter então uma previsão, dentro dos possíveis, acerca do número de doentes esperado em cada especialidade no seguinte dia, segue-se o próximo problema, a má gestão/distribuição de camas pelas especialidades fazendo com que fosse necessário alocar doentes em especialidades que não as suas contribuindo para o mau funcionamento do Hospital. Foi para isso, elaborado um modelo de otimização de camas, visando, de acordo com o fluxo de doentes e pensando nas camas como um recurso movível entre os Hospital e não fixo de cada especialidade, a melhor distribuição de camas possível pelas especialidades, visando minimizar a alocação errada de doentes.

Importante referir no contexto do problema em questão que para uma melhor perceção do mesmo foram criados três importantes conceitos:

- **Doente Físico**, o doente que se encontra internado no espaço físico da especialidade e é da responsabilidade da mesma especialidade, exemplificando, doente da

responsabilidade da especialidade de Medicina Interna internado no espaço físico da mesma.

- **Doente Responsável**, o doente que se encontra internado numa especialidade fora da responsável pelo mesmo, devido à falta de camas nessa especialidade, exemplificando, doente da responsabilidade de Medicina Interna internado no serviço de pneumologia.
- **Doente Previsto**, doente cuja responsabilidade não é da especialidade onde se encontra internado devido à sublocação da mesma, exemplificando, doente de pneumologia internado no serviço de Medicina Interna, representando assim um 'Doente Responsável' de pneumologia e um 'Doente Previsto' na Medicina Interna.

Conseguindo assim obter uma melhor perceção do estado de cada especialidade e saber que o objetivo do modelo de otimização se centrava em minimizar os 'Doentes Responsáveis' e 'Previstos' em cada especialidade, dentro das limitações do Hospital.

Paralelamente ao problema da gestão de camas surge também o problema de uma má gestão e planificação horária dos enfermeiros no Hospital, fazendo isto com que haja uma constante necessidade de realização de horas extra e uma sobrecarga horária nos mesmos. De modo a perceber se era possível obter uma otimização significativa da planificação horária do pessoal, sob o número de recursos apresentado, foi elaborado um modelo de otimização cujo objetivo centrava em minimizar a realização de horas extra, garantindo ao mesmo tempo que todas as necessidades do Hospital eram supridas e percebendo se havia uma falta de recursos ou efetivamente má gestão dos mesmos.

Em suma esta dissertação pode ser dividida em três partes:

- A elaboração de um modelo preditivo para o fluxo de doentes no Hospital.
- Desenvolvimento de um modelo de otimização de distribuição de camas, de modo a minimizar a alocação errada dos doentes pelas especialidades.
- Implementação de um modelo de otimização horária dos enfermeiros com intuito de perceber se é possível sob o contexto atual do hospital minimizar a realização de horas extra do pessoal.

4.2 Compreensão dos Dados

Para realização dos modelos, foram fornecidos dados relativos, às camas do hospital, à descrição de cada especialidade e ao fluxo de doentes num espaço temporal de quatro anos, 2019 a 2022 respetivamente.

Os ficheiros em formato Excel, correspondente às camas do Hospital, tinham como finalidade informar, quantas camas existiam no mesmo, em que especialidade se encontravam, a sua respetiva unidade e sala, o tipo de cama e se encontrava ativa ou ocupada. Esta informação estava presente nos seguintes atributos:

- NUM_REGISTO_CAMA, número de identificação da cama;
- COD_ESPECIALIDADE, código da especialidade onde a cama se encontrava;
- COD_UNIDADE_INTER, código da unidade de internamento onde a cama se encontrava;
- COD_SALA, código da sala onde a cama estava localizada;
- NUM_CAMA,
- TIPO_CAMA, indicava o tipo de cama em questão;
- OCUPADA, indicava se a cama estava ocupada;
- ATIVA, indicava se a cama se encontrava ativa;

Este *dataset* continha então 8 atributos e 551 registos, podendo compreender assim, para o âmbito da gestão de camas e otimização de recursos que o Hospital continham 551 camas disponíveis para distribuir pelas várias especialidades.

De modo a obter uma descrição de cada especialidade e complementar a análise, posterior, do fluxo de doentes, foi fornecido um *dataset*, com dois atributos:

- COD_ESPECIALIDADE, código da especialidade em questão;
- DES_ESPECIALIDADE, nome e descrição da especialidade;

Acerca da principal estrutura de dados correspondente ao fluxo de doentes, cada estrutura de era dividida anualmente e tinha o formato Excel, com quinze atributos e diferentes números de registos. Para uma análise facilitada foram agrupados em um só *dataset* o que resultou num ficheiro CSV com quinze atributos e 708482 registos. Cada registo era então composto por

quinze diferentes várias que visavam guardar informação acerca do tráfego de doentes no Hospital ao longo do tempo, entre as quais eram constituintes:

- Episódio, número único de identificação de cada paciente;
- Especialidade, código da especialidade onde o paciente se encontra internado;
- Unidade, número da unidade hospital onde o paciente se encontra internado;
- Especialidade_Prev, código da especialidade responsável pelo paciente;
- Unidade_Prev, número da unidade hospitalar responsável pelo paciente;
- DataMov, data atual do registo, corresponde a um dia de internamento de um paciente;
- DataMovs, atributo idêntico ao DataMov;
- ADM_Servico, data da admissão do paciente no serviço da especialidade;
- HADM_Servico, hora da admissão do paciente no serviço da especialidade;
- ALTA_Servico, data da alta do paciente no serviço da especialidade;
- HALTA_Servico, hora da alta do paciente no serviço da especialidade;
- ADM_Hospital, data da admissão do paciente no hospital;
- HADM_Hospital, hora da admissão do paciente no hospital;
- ALTA_Hospital, data da alta do paciente no hospital;
- HALTA_Hospital, hora da alta do paciente no hospital;

Depois de uma compreensão efetuada dos atributos deste *dataset*, foi necessária uma análise pormenorizada acerca da qualidade que os mesmo conferiam, isto é, verificar a quantidade de valores Null's, Nan's e outros que não se enquadrem no campo do atributo em questão. Esta verificação foi realizada após uma simples operação em *python*, onde o output está representado na imagem abaixo.

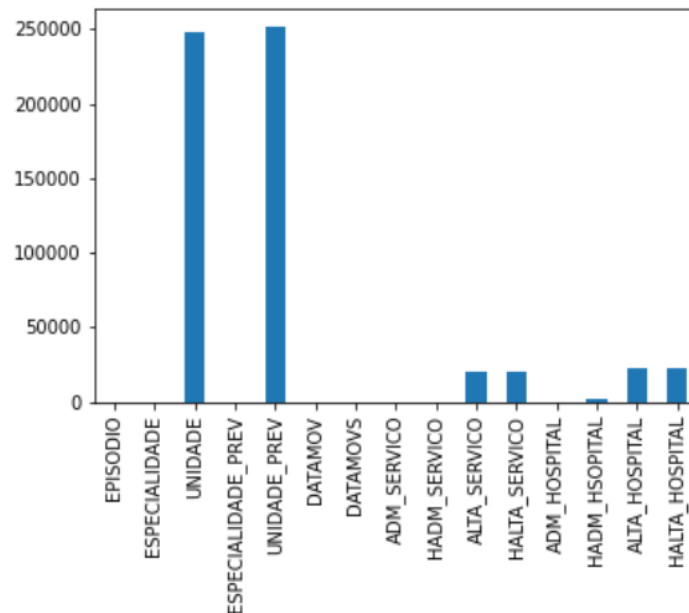


Figura 11 - Gráfico representativo dos valores Null e NaN presentes no dataset.

Na figura 11 conseguimos perceber que apenas os atributos representativos à unidade Hospitalar e Horas de Admissão ao Hospital contêm erros e faltas de registos, uma vez que as presenças dos restantes valores se devem ao facto de o doente se encontrar internado no Hospital e ainda não ter efetuado alta no dia da colheita dos dados.

Para além da análise qualitativa, foi também efetuado um estudo nos três diferentes *datasets* e correspondentes atributos com o intuito de perceber a relação entre si, quais e qual o tipo de informação possível de extrair a partir de várias operações e transformações dos mesmos. Observaram-se então, várias relações que nos permitiam extrair informação relevantes para a constituição dos modelos presentes neste projeto tais como:

- Tipos de doente em cada dia numa determinada especialidade (Físicos, Responsáveis ou Previstos), através da comparação de códigos das especialidades num determinado dia;
- Totalidade dos doentes à responsabilidade de uma certa especialidade;
- Totalidade dos doentes internados no hospital e em cada especialidade, através das Altas e Admissões;
- Quais as datas e/ou períodos com maior afluência no hospital;

Depois de registadas as informações a extrair a partir do conjunto de dados seguiu-se a sua e posterior construção destes novos campos, na fase designada *de Preparação dos Dados*.

4.3 Preparação dos Dados

Nesta etapa o objetivo era preparar e transformar os dados de modo que estes pudessem estar devidamente preparados para ser utilizados como input nos modelos a serem posteriormente implementados.

Primeiramente foram tratados todos os dados correspondentes a valores em falta nos registos, isto é, valores 'Null' ou valores 'Nan'. Nos atributos categóricos este valor foi substituído pela letra 'n' e nos atributos numéricos pelo número 0. Não foram identificados erros relativamente à constituição de registos no contexto dos seus atributos.

Seguiu-se a etapa de construção de novos atributos relevantes para o contexto do problema. Desta fase resultou a criação de nove novas colunas ou atributos, passando a enumerar:

- 'NDOENTES_FISICO' – Atributo correspondente ao número dos doentes físicos numa determinada especialidade, numa determinada data
- 'NDOENTES_RESPONSABEL' – Atributo correspondente ao número dos doentes responsáveis numa determinada especialidade, numa determinada data.
- 'NDOENTES_PREVISTO' – Atributo correspondente ao número dos doentes previstos numa determinada especialidade, numa determinada data.
- 'NDOENTES_ESPACO_FISICO' – Número real de doentes que se encontram no espaço físico da especialidade, isto é, a soma dos doentes físicos com os doentes previstos.
- 'NDOENTES_RESPONSABEL_TOTAL' – Número total de doentes cuja responsabilidade é de uma determinada especialidade num determinado dia, todos os doentes que estejam ou não internados no espaço físico de uma especialidade sendo do cuidado da mesma.
- 'NDOENTES_ASSOCIADOS_ESPECIALIDADE' – Totalidade dos doentes associados à especialidade, sejam estes físicos, responsáveis ou previstos.
- 'NUMERO_ADMISSOES' – Número de admissões num determinado dia numa determinada especialidade.

- 'NUMERO_ALTAS' – Número de altas num determinado dia numa determinada especialidade.
- 'CAMAS_OCUPADAS' – Número de camas ocupadas num determinado dia numa determinada especialidade, resultante das camas disponíveis dedicadas à especialidade menos os doentes presentes no espaço físico do serviço.
- 'CAMAS_DISPONIVEIS' – Camas disponíveis num determinado dia numa determinada especialidade, resultante da diferença entre as camas disponíveis e os doentes presentes no espaço físico da mesma.
- 'CAMAS_A_NECESSITAR' – Camas necessárias para responder às necessidades de uma determinada especialidade num determinado dia, isto é internar todos os doentes que foram alocados para o espaço físico da mesma, sendo da sua responsabilidade ou não.
- 'CAMAS_A_NECESSITAR_OTIMO' – Número de camas que seriam necessárias na especialidade num determinado dia, num cenário ideal onde o Hospital conseguiria responder a 100% das suas necessidades e cada especialidade conseguir ter camas suficientes para alocar todos os doentes da sua responsabilidade, eliminando assim a alocação errada de doentes.

Posteriormente, todos os restantes atributos que nos os extraídos a partir dos existentes foram eliminados do conjunto de dados uma vez que grande parte deles constituía informação redundante e iria servir apenas para diminuir a eficácia e eficiência dos modelos aquando a sua realização.

Com o intuito de começar também a preparar os dados para o modelo de previsão de série temporal e seguintes foram realizados dois importantes passos:

- A criação de um *dataframe* com os atributos criados em que o seu índice correspondia à data dos dados em questão, isto é cada linha correspondia a um determinado dia de registos no Hospital com a informação dos atributos construídos sobre cada especialidade.

- A divisão de este *dataframe* em conjuntos de dados mais pequenos contendo cada um a informação ao longo do tempo de uma determinada especialidade em específico, isto é dividir o *dataset* por especialidades de modo a obter uma melhor organização e perceção nos dados, ajudando isto também a que não seja necessário executar um modelo sobre todo o conjunto de

dados, reduzindo o tempo de computação e também o risco de erro, permitindo que cada especialidade tenha um conjunto de dados dedicado sendo analisado e inserido como input de forma individual em cada um dos modelos.

DATA	NDOENTES_ESPAC	NDOENTES_RESPO	NDOENTES_FISICO	NDOENTES_RESPO	NDOENTES_PREVIR	NDOENTES_ASSOC	NUMERO_ADMISSO	NUMERO_ALTAS	CAMAS_OCUPADAS	CAMAS_DISPONIVE	CAMAS_A_NECESS	CAMAS_A_NECESS
2019-01-01	18	18	18	0	0	18	1	1	17	0	1	1
2019-01-02	19	19	19	0	0	19	2	2	17	0	2	2
2019-01-03	21	21	21	0	0	21	4	4	17	0	4	4
2019-01-04	19	19	19	0	0	19	2	2	17	0	2	2
2019-01-05	17	17	17	0	0	17	0	1	17	0	0	0
2019-01-06	17	17	17	0	0	17	1	0	17	0	0	0
2019-01-07	22	23	22	1	0	23	6	4	17	0	5	6
2019-01-08	21	21	21	0	0	21	2	4	17	0	4	4
2019-01-09	18	19	18	1	0	19	2	1	17	0	1	2
2019-01-10	21	21	21	0	0	21	3	3	17	0	4	4
2019-01-11	22	22	22	0	0	22	4	4	17	0	5	5
2019-01-12	20	20	20	0	0	20	2	3	17	0	3	3
2019-01-13	19	19	19	0	0	19	2	2	17	0	2	2
2019-01-14	19	19	19	0	0	19	2	4	17	0	2	2
2019-01-15	17	17	17	0	0	17	2	0	17	0	0	0
2019-01-16	20	22	20	2	0	22	5	5	17	0	3	5
2019-01-17	19	22	19	3	0	22	5	3	17	0	2	5
2019-01-18	20	22	20	2	0	22	3	3	17	0	3	5
2019-01-19	19	22	19	3	0	22	3	2	17	0	2	5
2019-01-20	20	23	20	3	0	23	3	3	17	0	3	6
2019-01-21	20	24	20	4	0	24	4	3	17	0	3	7
2019-01-22	20	24	20	4	0	24	3	4	17	0	3	7
2019-01-23	22	25	22	3	0	25	5	6	17	0	5	8
2019-01-24	23	25	23	2	0	25	6	5	17	0	6	8
2019-01-25	20	22	20	2	0	22	2	6	17	0	3	5
2019-01-26	18	19	18	1	0	19	3	1	17	0	1	2
2019-01-27	20	21	20	1	0	21	3	4	17	0	3	4
2019-01-28	20	21	20	1	0	21	4	3	17	0	3	4
2019-01-29	18	20	18	2	0	20	2	1	17	0	1	3
2019-01-30	22	24	22	2	0	24	5	5	17	0	5	7
2019-01-31	20	21	20	1	0	21	2	3	17	0	3	4
2019-02-01	19	20	19	1	0	20	2	2	17	0	2	3
2019-02-02	18	19	18	1	0	19	1	1	17	0	1	2
2019-02-03	19	20	19	1	0	20	2	2	17	0	2	3
2019-02-04	20	21	20	1	0	21	3	3	17	0	3	4
2019-02-05	20	21	20	1	0	21	3	3	17	0	3	4
2019-02-06	20	21	20	1	0	21	3	3	17	0	3	4
2019-02-07	21	22	21	1	0	22	4	5	17	0	4	5
2019-02-08	22	22	22	0	0	22	5	5	17	0	5	5
2019-02-09	20	20	20	0	0	20	3	4	17	0	3	3
2019-02-10	17	17	17	0	0	17	1	0	17	0	0	0
2019-02-11	23	23	23	0	0	23	6	6	17	0	6	6
2019-02-12	19	19	19	0	0	19	2	8	17	0	2	2
2019-02-13	20	20	20	0	0	20	9	2	17	0	3	3
2019-02-14	20	21	20	1	0	21	3	4	17	0	3	4
2019-02-15	18	19	18	1	0	19	2	6	17	0	1	2
2019-02-16	12	13	12	1	0	13	0	1	12	5	0	0

Figura 12 - Exemplo de DataSet Resultante

DATA	NDENTES_ESPAC	NDENTES_RESPC	NDENTES_FISCO	NDENTES_RESPC	NDENTES_PREVI	NDENTES_ASSOC	NUMERO_ADMISSC	NUMERO_ALTAS	CAMAS_OCUPADAS	CAMAS_DISPONIVE	CAMAS_A_NECESS	CAMAS_A_NECESS
2019-01-01	129	156	127	29	2	158	18	5	104	0	25	52
2019-01-02	137	162	133	29	4	166	11	16	104	0	33	58
2019-01-03	136	161	133	28	3	164	15	25	104	0	32	57
2019-01-04	134	160	131	29	3	163	24	24	104	0	30	56
2019-01-05	125	152	124	28	1	153	16	3	104	0	21	48
2019-01-06	128	157	127	30	1	158	8	3	104	0	24	53
2019-01-07	134	166	133	33	1	167	12	27	104	0	30	62
2019-01-08	136	164	131	33	5	169	25	25	104	0	32	60
2019-01-09	130	157	125	32	5	162	18	19	104	0	26	53
2019-01-10	130	151	125	26	5	156	13	12	104	0	26	47
2019-01-11	124	149	119	30	5	154	10	21	104	0	20	45
2019-01-12	124	147	120	27	4	151	19	4	104	0	20	43
2019-01-13	137	162	133	29	4	166	19	6	104	0	33	58
2019-01-14	144	168	138	30	6	174	12	26	104	0	40	64
2019-01-15	147	169	141	28	6	175	27	23	104	0	43	65
2019-01-16	145	172	138	34	7	179	26	22	104	0	41	68
2019-01-17	147	171	140	31	7	178	21	26	104	0	43	67
2019-01-18	145	166	138	28	7	173	21	24	104	0	41	62
2019-01-19	134	158	126	32	8	166	16	9	104	0	30	54
2019-01-20	142	176	134	42	8	184	27	5	104	0	38	72
2019-01-21	152	184	142	42	10	194	13	30	104	0	48	80
2019-01-22	138	177	130	47	8	185	23	23	104	0	34	73
2019-01-23	140	174	131	43	9	183	20	25	104	0	36	70
2019-01-24	141	170	131	39	10	180	21	12	104	0	37	66
2019-01-25	144	174	135	39	9	183	16	22	104	0	40	70
2019-01-26	136	164	127	37	9	173	12	4	104	0	32	60
2019-01-27	145	177	136	41	9	186	17	4	104	0	41	73
2019-01-28	166	199	156	43	10	209	26	19	104	0	62	95
2019-01-29	159	195	151	44	8	203	15	31	104	0	55	91
2019-01-30	152	188	144	44	8	196	24	25	104	0	48	84
2019-01-31	151	176	143	33	8	184	13	19	104	0	47	72
2019-02-01	150	175	143	32	7	182	18	28	104	0	46	71
2019-02-02	155	179	148	31	7	186	32	5	104	0	51	75
2019-02-03	163	189	157	32	6	195	15	7	104	0	59	85
2019-02-04	174	201	167	34	7	208	19	25	104	0	70	97
2019-02-05	171	201	166	35	5	206	25	28	104	0	67	97
2019-02-06	159	189	154	35	5	194	16	30	104	0	55	85
2019-02-07	149	178	144	34	5	183	19	22	104	0	45	74
2019-02-08	142	169	138	31	4	173	13	39	104	0	38	65
2019-02-09	128	148	124	24	4	152	18	1	104	0	24	44
2019-02-10	148	171	142	29	6	177	24	2	104	0	44	67
2019-02-11	165	190	158	32	7	197	21	36	104	0	61	86
2019-02-12	159	176	152	24	7	183	22	31	104	0	55	72
2019-02-13	149	163	141	22	8	171	18	22	104	0	45	59
2019-02-14	146	157	136	21	10	167	16	14	104	0	42	53
2019-02-15	151	168	142	26	9	177	25	22	104	0	47	64
2019-02-16	152	173	144	29	8	181	27	6	104	0	48	69

Figura 13 - Exemplo de DataSet Resultante

DATA	NDOENTES_ESPAC	NDOENTES_RESPC	NDOENTES_FISICO	NDOENTES_RESPC	NDOENTES_PREV	NDOENTES_ASSOC	NUMERO_ADMISSO	NUMERO_ALTAS	CAMAS_OCUPADAS	CAMAS_DISPONIVE	CAMAS_A_NECES	CAMAS_A_NECES
2019-01-01	23	23	23	0	0	23	4	2	23	15	0	0
2019-01-02	22	22	22	0	0	22	1	14	22	16	0	0
2019-01-03	15	15	15	0	0	15	7	5	15	23	0	0
2019-01-04	16	16	16	0	0	16	6	3	16	22	0	0
2019-01-05	17	17	17	0	0	17	4	6	17	21	0	0
2019-01-06	15	15	15	0	0	15	4	6	15	23	0	0
2019-01-07	13	13	13	0	0	13	4	7	13	25	0	0
2019-01-08	11	11	11	0	0	11	5	2	11	27	0	0
2019-01-09	14	14	14	0	0	14	5	2	14	24	0	0
2019-01-10	21	21	21	0	0	21	9	1	21	17	0	0
2019-01-11	32	32	32	0	0	32	12	6	32	6	0	0
2019-01-12	34	34	34	0	0	34	8	10	34	4	0	0
2019-01-13	30	30	30	0	0	30	6	5	30	8	0	0
2019-01-14	31	31	31	0	0	31	6	11	31	7	0	0
2019-01-15	29	29	29	0	0	29	9	10	29	9	0	0
2019-01-16	24	24	24	0	0	24	5	8	24	14	0	0
2019-01-17	22	22	22	0	0	22	6	7	22	16	0	0
2019-01-18	28	28	28	0	0	28	13	9	28	10	0	0
2019-01-19	26	26	26	0	0	26	7	5	26	12	0	0
2019-01-20	24	24	24	0	0	24	3	7	24	14	0	0
2019-01-21	24	24	24	0	0	24	7	10	24	14	0	0
2019-01-22	25	25	25	0	0	25	11	6	25	13	0	0
2019-01-23	21	21	21	0	0	21	2	5	21	17	0	0
2019-01-24	19	19	19	0	0	19	3	6	19	19	0	0
2019-01-25	17	17	17	0	0	17	4	7	17	21	0	0
2019-01-26	18	18	18	0	0	18	8	8	18	20	0	0
2019-01-27	16	16	16	0	0	16	6	4	16	22	0	0
2019-01-28	18	18	18	0	0	18	6	3	18	20	0	0
2019-01-29	23	23	23	0	0	23	8	5	23	15	0	0
2019-01-30	23	23	23	0	0	23	5	4	23	15	0	0
2019-01-31	24	24	24	0	0	24	5	10	24	14	0	0
2019-02-01	19	19	19	0	0	19	5	6	19	19	0	0
2019-02-02	18	18	18	0	0	18	5	5	18	20	0	0
2019-02-03	20	20	20	0	0	20	7	6	20	18	0	0
2019-02-04	21	21	21	0	0	21	7	4	21	17	0	0
2019-02-05	28	28	28	0	0	28	11	4	28	10	0	0
2019-02-06	31	31	31	0	0	31	7	5	31	7	0	0
2019-02-07	30	30	30	0	0	30	4	14	30	8	0	0
2019-02-08	20	20	20	0	0	20	4	6	20	18	0	0
2019-02-09	18	18	18	0	0	18	4	4	18	20	0	0
2019-02-10	20	20	20	0	0	20	6	4	20	18	0	0
2019-02-11	22	22	22	0	0	22	6	6	22	16	0	0
2019-02-12	24	24	24	0	0	24	8	6	24	14	0	0
2019-02-13	22	22	22	0	0	22	4	6	22	16	0	0
2019-02-14	22	22	22	0	0	22	6	8	22	16	0	0
2019-02-15	23	23	23	0	0	23	9	6	23	15	0	0
2019-02-16	23	23	23	0	0	23	6	5	23	15	0	0
2019-02-17	25	25	25	0	0	25	7	7	25	13	0	0

Figura 14 - Exemplo de DataSet Resultante

Nas figuras 12 a 14 estão representados exemplos de *datasets* resultantes depois de aplicadas todas as transformações nesta etapa de preparação dos dados. Neste caso em específico são correspondentes ao conjunto de dados relativo ao fluxo de doentes diário entre 2019 e maio de 2022 das três especialidades com maior expressão no hospital.

4.4 Modelação

Para o desenvolvimento deste estudo foram desenvolvidos, três diferentes modelos, com dois focos distintos. Primeiramente foi elaborado um modelo de previsão com intuito de prever o fluxo de doentes no CHTS, de seguida, tendo por base esta previsão, foram elaborados modelos de otimização para a alocação de camas pelas diferentes especialidades do Hospital e para alocação horária do pessoal.

4.5 Modelo de Previsão do Fluxo de doentes

De modo a realizar um modelo que nos permitisse obter uma previsão geral acerca do fluxo de doentes em cada especialidade no Hospital, foi elaborado, para caso demonstrativo, um modelo com base apenas no fluxo doentes da especialidade de Medicina Interna, sendo esta a que representa maior expressão a nível de volume de pacientes.

Uma vez que a variável a prever se centrava só e apenas no fluxo de pacientes, já representada pelo atributo “NDOENTES_RESPONSAVEL_TOTAL” foi descartada a necessidade de realizar modelos multivariados de maneira a evitar redundância e ocorrência de erros no modelo. Para tal foi escolhido o ‘ARIMA-*Auto Regressive Integrated Moving Average*’, mais concretamente o “Auto-Arima”, um modelo uni variado que apenas prevê uma variável/série temporal de cada vez. O Modelo Auto-Arima permite lidar já com dados sazonais como é o caso e realizar a diferenciação automática dos mesmos para lhes retirar a sazonalidade e quaisquer tendências presentes nos dados tornando-os estacionários. Para além destes parâmetros existem mais dois a ter conta neste modelo uni variado, o parâmetro p , número de termos autorregressivos ou o número de “observações defasadas” e determina o resultado do modelo fornecendo pontos defasados e o parâmetro q que indica a ordem do modelo de média movel. Dependendo do valor destes parâmetros depende também a previsão do modelo, no caso do Auto-Arima estes valores são também eles definidos automaticamente dentro de um range definido inicialmente sendo este entre 1 e 5 ($start_q$ & $start_p$). O critério para escolher a melhor combinação destes valores (p , d , q) é baseado no AIC, uma técnica baseada no ajuste dentro da amostra para estimar a probabilidade de um modelo prever/estimar valores futuros. A melhor combinação é aquela que apresenta o AIC menor.

Na figura 15 podemos ver representado graficamente o conjunto de dados que serviu de input para o modelo. O fluxo de doentes da especialidade de medicina interna desde janeiro de 2019 a junho de 2022.

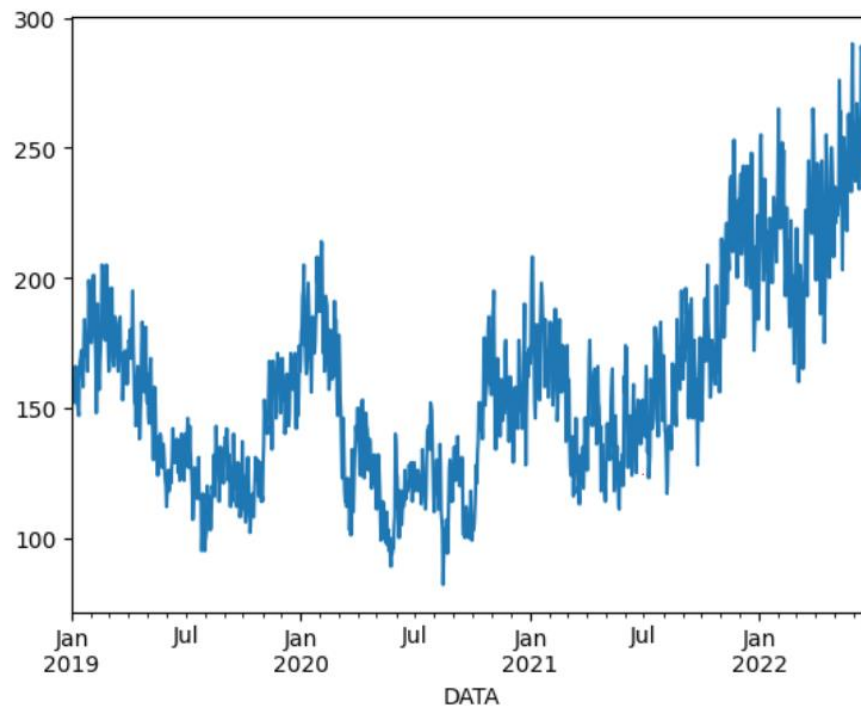


Figura 15 - Fluxo de doentes de Medicina Interna de 2019 a 2022

O objetivo passou por testar o modelo criando a partir deste conjunto de dados um *dataset* de treino, correspondente ao fluxo de doentes de janeiro de 2019 a maio de 2022 e um *dataset* de teste correspondente ao fluxo de doentes do mês de junho de 2022. Na figura 16 abaixo podemos ver a azul os dados correspondentes do *dataset* de treino e a laranja o *dataset* teste.

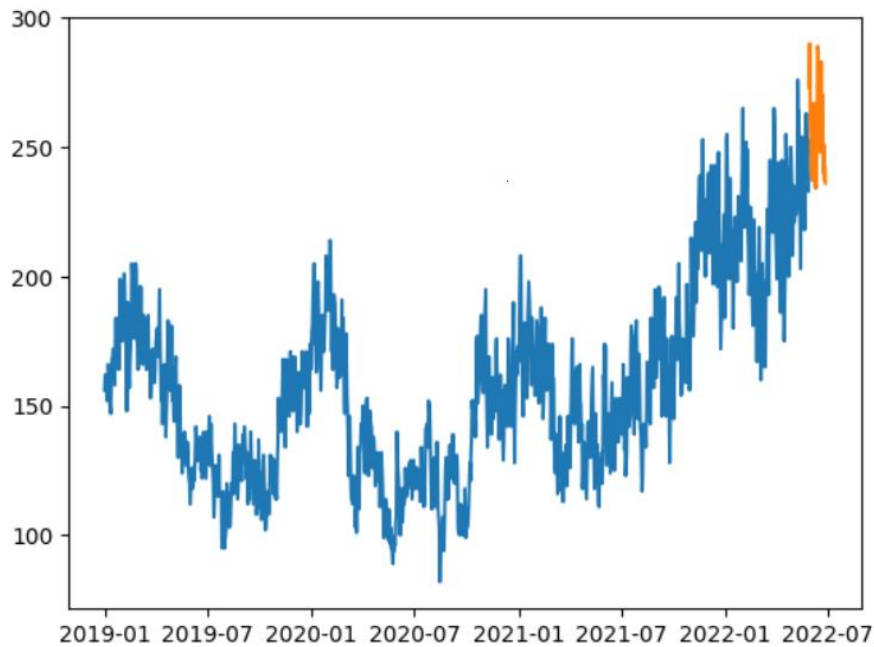


Figura 16 - Divisão dados treino e dados teste

Para avaliar a precisão da solução foram utilizadas três métricas entre as quais:

- *Mean Forecast Error*, Erro Médio da Previsão ou 'Bias', é tanto melhor quão mais próximo de 0 estiver, se for negativo significa que estamos perante um caso de *undercast* dos dados, se for positivo está a acontecer *Overfitting*.
- *Mean Squared Error*, ou Erro Médio Quadrado, medida que avalia os erros e desvios de maior magnitude atribuindo maior peso, de forma a combater anomalias e *outliers*.
- *Mean Absolute Error* ou Erro Médio Absoluto, o erro médio entre a diferença do valor previsto e o valor real.

A solução encontrada pelo modelo com menor AIC foi com os valores de $p = 4$, $d = 1$ e $q = 5$.

Os valores das métricas para avaliação foram:

- Erro Médio de Previsão = -0.22
- Erro Médio Quadrado = 200.8
- Erro Médio Absoluto = 11.7

Verificando estes valores podemos conferir que a previsão não sendo perfeita, teve sucesso, até porque apresentando um erro médio absoluto de 11 é considerado positivo tendo em conta que a escala do fluxo de doentes diário no mês de junho de 2022 ronda sempre as duas centenas, pelo que um erro de 11 pacientes na previsão não é muito significativo.

De modo a visualizar a previsão realizada pelo modelo, estão representados na figura 17 os resultados do modelo, sendo a azul os dados reais e a laranja os dados previstos.

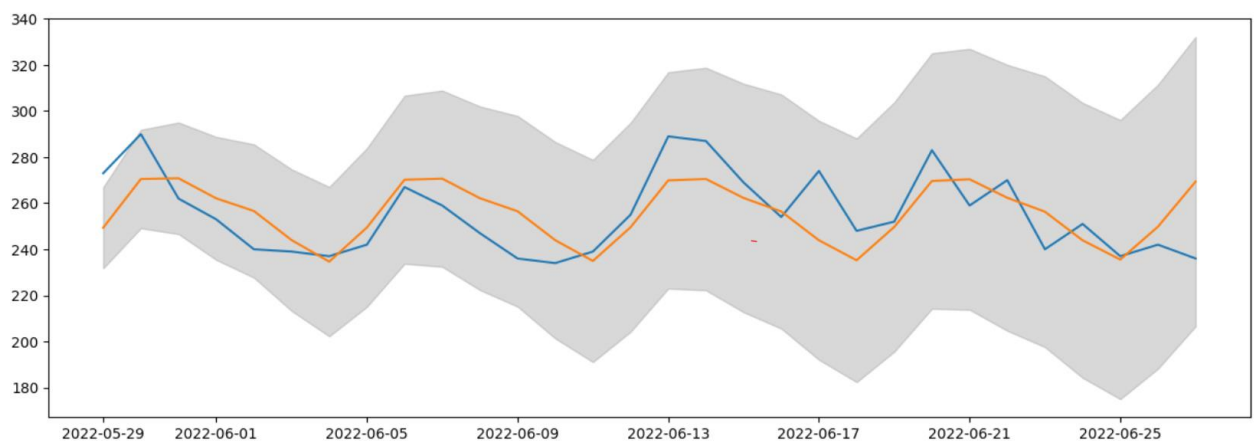


Figura 17 - Resultados da previsão

4.6 Modelos de Otimização

Após realizada a previsão do fluxo de doentes no Hospital, segue-se a etapa dos modelos ou algoritmos de otimização.

Na realização dos dois modelos de otimização, os algoritmos testados foram os mesmos, da maneira a que antes da descrição do desenvolvimento e formulação dos problemas seguem abaixo cada um dos algoritmos testados e o seu respetivo pseudocódigo.

Implementação dos modelos

Para otimizar tanto a atribuição de camas a doentes, como a atribuição horária dos enfermeiros foram então comparados quatro algoritmos de otimização diferentes, de forma a perceber qual

deles representaria um melhor desempenho. A configuração dos algoritmos foi adaptada de forma a considerar a data num dos parâmetros, uma vez que é uma variável de entrada na função de aptidão e essencial para definir o domínio diário do problema. Os algoritmos que foram considerados neste estudo estão apresentados abaixo assim como uma breve descrição dos mesmos.

Random Search (RS)

1. RS Pseudo Code

```
round = 0
best_satisfaction = 0
best solution TO NULL

While round < rounds
    solution = random solution between the domain values
    satisfaction = fitness function(date, solution)

    If satisfaction > best_satisfaction then
        best_satisfaction = satisfaction
        best solution = solution

    End If
    round ++
End While
Return best solution
```

Random Search ou algoritmos de pesquisa aleatória tem por base gerar soluções aleatórias durante um certo número de *rounds*. A cada round a solução gerada é avaliada, a função que no final de todos os *rounds* apresentar melhor score constitui a solução apresentada pelo algoritmo em questão.

Hill Climbing(HC)

2. HC Pseudo Code

```
Generate an initial solution s0,  
s = s0  
satisfaction = fitness(date, s)  
  
While termination criteria not met do  
    neighbor_solution = Generate_Neighbor(s, domain)  
    neighbor_satisfaction = fitness(date, neighbor_solution)  
  
    If neighbor_satisfaction > satisfaction then  
        s = neighbor_solution  
        satisfaction = neighbor_satisfaction  
    End If  
End While  
  
Return s
```

O *Hill Climbing* é um algoritmo de pesquisa local, baseado em incrementos e decrementos das soluções geradas, chamadas de "vizinhos", partindo de uma solução inicial aleatória ou previamente fornecida dentro do domínio. Na sua execução, cada vizinho gerado é analisado e avaliado através da *fitness function* quanto à sua qualidade, se um dos vizinhos gerados na iteração apresentar um resultado melhor passa a ser a solução atual. O algoritmo continua até que não haja possibilidade de melhoria nas soluções. Este algoritmo tende a encontrar soluções ótimas locais.

Simulated Annealing(SA)

3. SA Pseudo Code

```
Generate an initial solution s0,  
s = s0  
satisfaction0 = fitness(date, s)  
  
While temperature > 0.1 do  
  
    s1 = Generate(s0, domain)  
    satisfaction1 = fitness(date, s1)  
  
    If Accept(satisfaction0, satisfaction1, temperature) then  
        satisfaction0 = satisfaction1  
        s0 = s1  
    End If  
    temperature = temperature * cooling  
End While  
Return s0
```

O *Simulated Annealing* é um método análogo à termodinâmica inspirado no processo de arrefecimento de metais. Este algoritmo consiste numa sequência de iterações onde ocorre uma diminuição progressiva da temperatura, inicializando-a num valor elevado onde cada alteração é aceite, sendo esta probabilidade de aceitação cada vez menor à medida que a temperatura diminui de acordo com uma taxa de arrefecimento. O método de pesquisa é idêntico ao algoritmo HC baseado na transformação de soluções através de incrementos à solução anterior, a grande diferença é que através da aceitação ou não das soluções esta abordagem evita ficar preso a mínimos ou máximos locais.

Genetic Algorithm(GA)

4. GA Pseudo Code

```
Initialize population
elite = elitism * Population-Size

For i in range number_generations

    fitness(date, individual) for each individual in population
    Select the elite with the best result
    Insert elite into new_population

    While new population-size < Population-Size

        IF probability of mutation
            Mutate elite individual -> new individual

        ELSE
            Crossover elite individuals -> new individual

        Insert individual into new population
    End While
End For

Return last generation best individual
```

O Algoritmo Genético baseia-se na teoria da evolução de Charles Darwin. Atua sobre um conjunto de soluções possíveis denominadas indivíduos de uma população. O algoritmo parte de uma população inicial e são efetuadas transformações em cada um dos indivíduos através de mutação ou cruzamento/*cross-over*, cuja decisão é baseada numa probabilidade previamente definida. Os melhores indivíduos são escolhidos de acordo com um número de elitismo para

gerar a nova população. O algoritmo termina quando se atinge o número desejado de populações/gerações e se atinge o melhor indivíduo da última população.

De notar que, de acordo com vários estudos, como o efetuado na Universidade de Neuchâtel, na Suíça, verificou-se que estes algoritmos apresentam melhores resultados com populações maiores do que com o aumento do número máximo de gerações.

Otimização da alocação errada de doentes-camas no Hospital

Para o modelo de otimização de camas apenas precisávamos dos dados do fluxo de doentes e de saber a quantidade de camas disponíveis no Hospital. De acordo com o fluxo de doentes diário o modelo de otimização tem em conta as necessidades de cada uma das Especialidades e tenta alocar as camas de forma que cada especialidade tenha recursos para responder às suas necessidades, minimizando assim a transferência de doentes e a alocação errada dos mesmos. Aumentando dos doentes físicos, diminuindo os responsáveis e previstos.

Antes de aplicar os algoritmos de otimização, foi necessário definir:

- i) O domínio ou espaço de solução do problema;
- ii) como as soluções serão representadas;
- iii) como definir as restrições;
- iv) como avaliar a eficiência das soluções apresentadas. Assim, é imperativo definir o domínio do problema e a função objetivo.

O domínio do problema define a forma como o conjunto de soluções é representado tendo em conta as restrições do problema. O domínio corresponde ao número mínimo e máximo de camas disponíveis em cada especialidade de modo que a solução seja sempre possível no espaço do problema desde que não ultrapasse o número máximo de camas necessárias para o dia.

De forma a otimizar o desempenho dos algoritmos e diminuir a probabilidade de subutilização de recursos, foi criado um "domínio dinâmico" em função do dia em questão. Ao contrário do que é habitualmente entendido como uma variável fixa no problema, neste problema o domínio pode variar consoante as necessidades diárias das diferentes especialidades do hospital. Por outras palavras, para cada dia, o domínio pode variar entre 0 e o número de doentes sob a responsabilidade de cada especialidade.

A solução foi representada numa matriz/ *array* de 1 dimensão, com 36 elementos, cada um correspondendo a um número inteiro de camas a atribuir numa dada especialidade para o dia em questão.

Para avaliar um conjunto solução é necessário elaborar uma *fitness function* ou função objetivo. Esta função atribui uma determinada pontuação à solução a avaliar, tendo em conta um conjunto de condições específicas do contexto do problema. O objetivo passa minimizar ou maximizar a pontuação, consoante o problema.

Apresentada no Pseudocódigo 1, a função objetivo tenta maximizar o número de atribuições corretas de camas, conhecendo as necessidades diárias de cada especialidade, minimizando o número de camas vazias ou não utilizadas. Isto leva a maximizar o número de doentes corretamente atribuídos a uma cama, ou seja, maximizar os doentes físicos em vez dos doentes responsáveis

No modelo de otimização de camas de maneira a avaliar uma solução, foi atribuído então sucesso se:

- O limite de camas do Hospital fosse respeitado – restrição;
- Uma especialidade não apresentasse camas disponíveis;
- Uma especialidade tivesse na mesma o número de camas necessário para responder á suas necessidades;
- Caso não estivesse igualdado o número de camas necessário, o quão mais próximo desse número estivesse;

5. Fitness Function Pseudo Code – Bed Problem

```
satisfaction = 0
solution = solution input
max_bed = 551

If sum(solution) > max beds then
    satisfaction = -100

For each specialty in solution do

    If beds_needed == beds in solution and beds available == 0 then
        satisfaction += 2

    Else if beds_needed == beds in solution and beds available > 0 then
        satisfaction += 1

    Else if beds_needed not equal to beds in solution:
        satisfaction -= |beds_in_solution – beds_needed| * 0.1

End If
End For
Return satisfaction
```

Posteriormente e antes de implementar os modelos foi necessário definir quais os critérios de avaliação de cada modelo de maneira a avaliar a execução de cada um e perceber qual o que apresentava melhor performance.

Avaliação

Para então avaliar o desempenho de cada modelo, foram efetuadas 10 execuções de cada um deles. Para cada uma delas foi registrado o score obtido e o tempo de computação necessário. Nesta experiência, todos os modelos foram executados com dados baseados na data de 5 de maio de 2021, tendo em conta que o domínio corresponde a uma variável que mudaria dia após dia, dependendo das necessidades reais do hospital.

A Tabela 2 apresenta as métricas utilizadas para a avaliação dos algoritmos, bem como uma breve descrição.

Tabela 2 - Métricas

Metric	Description
Score Médio	Média de Scores obtida nas 10 execuções
Tempo de execução	Média dos tempos de execução
Desvio padrão(STD)	Desvio padrão dos scores obtidos

Num cenário perfeito, a atribuição de camas corresponde às necessidades diárias destas especialidades e o número de camas não utilizadas tende para 0. A pontuação máxima obtida seria de 2 pontos por cada uma das 36 especialidades, ou seja, 2×36 , o que significa uma pontuação máxima de 72 pontos. É necessário ter em conta que, para atingir esta pontuação, o hospital não se poderia encontrar num estado de sobre/sub capacidade. Todas as camas deveriam estar ocupadas e atribuídas às especialidades corretamente, o que não se verifica em cerca de 99% das ocorrências.

O desvio padrão do tempo de execução não foi registado uma vez que se verificaram diferenças insignificantes entre execuções para o mesmo algoritmo, o que poderá ser resultado das condições computacionais no ambiente de desenvolvimento.

Definições dos algoritmos

Para executar os algoritmos considerados neste estudo foi necessário definir os seus parâmetros iniciais de entrada, exceto para o HC uma vez que não é necessário. Na tabela 3 são apresentados os valores utilizados para a implementação de cada técnica.

Tabela 3 - Parâmetros de input

Model	Parameters
Random Search	<i>Rounds</i> -> 10000
Hill Climbing	None
Simulated Annealing	<i>Cooling Rate</i> -> 0.95 <i>Temperature</i> -> 10000000000
Genetic Algorithm	<i>Population-size</i> -> 400 <i>Number_Generations</i> -> 25 <i>Probability of Mutation</i> -> 0.2 <i>Elitism</i> -> 0.2

O número de gerações definido para o algoritmo genético foi obtido através de várias execuções até chegar à estabilização da função objetivo, conforme representado na Figura 18, onde podemos observar que a partir de cerca das 20 a 25 gerações não se verifica nenhum aumento do score obtido, o que permite também reduzir o tempo de computação necessário.

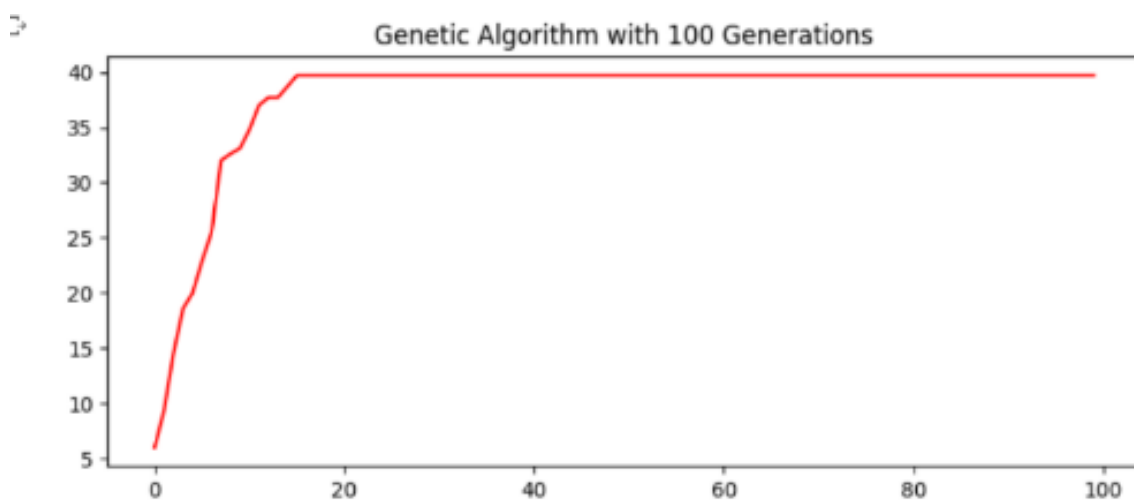


Figura 18 -. Score obtido pelo GA durante as 100 gerações

Resultados e análise crítica

Passou-se então depois de tudo definido para a execução dos algoritmos de otimização onde o desempenho de cada um dos algoritmos foi registado considerando um total de 10 execuções, conforme apresentado na tabela 4.

Tabela 4 - Resultados Obtidos

Model	MS	STD	ART
Random Search	13.78	1.86	10 minutos
Hill Climbing	43.2	2.97	14 minutos
Simulated Annealing	-11.9	7.52	Menos de 1 minuto
Genetic Algorithm	42.2	1.88	6 minutos

A partir destes resultados, podemos ver que tanto o *Hill Climbing* como o Algoritmo Genético se destacam positivamente em termos de desempenho. Numa comparação justa, podemos dizer que o HC, apesar de obter as melhores pontuações médias, também regista um desvio padrão mais elevado. Isto significa que num conjunto maior de execuções, a média pode ser alterada pela variância das soluções apresentadas. O GA, por outro lado, apresenta quase a mesma pontuação média, com uma diferença de 1 ponto e um desvio padrão significativamente menor,

promovendo maior consistência dos resultados a longo prazo e maior fiabilidade. Relativamente ao tempo de computação, o GA apresenta menos de metade do HC.

Pelo lado negativo, o *Simulated Annealing* não apresenta uma única solução viável, no entanto, este facto deve-se aos parâmetros escolhidos na sua implementação. Mais concretamente a temperatura inicial utilizada, que apesar de ser elevada, é um parâmetro que deve ser alterado consoante o cenário. Para se fazer uma comparação justa com os outros algoritmos, teria de se descobrir qual o valor ótimo a utilizar para este parâmetro, o que é uma tarefa bastante complexa.

O algoritmo *Random Search* não apresentou surpresas no seu comportamento, uma vez que as soluções são geradas através de probabilidades aleatórias.

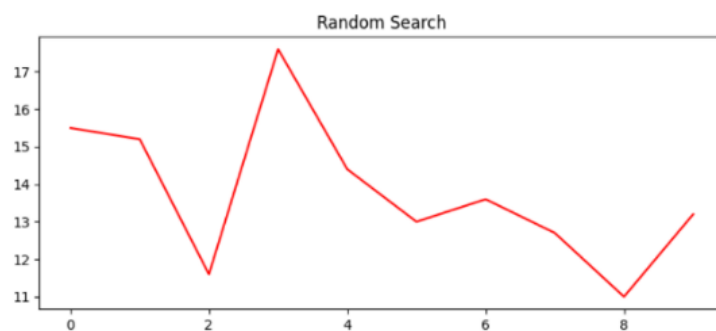


Figura 19 - Execução do Random Search

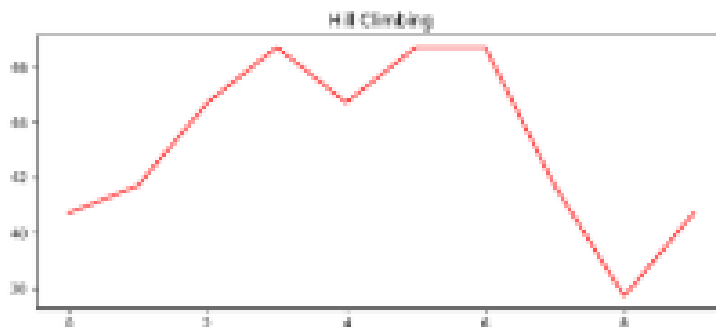


Figura 20 - Execução do Hill Climbing

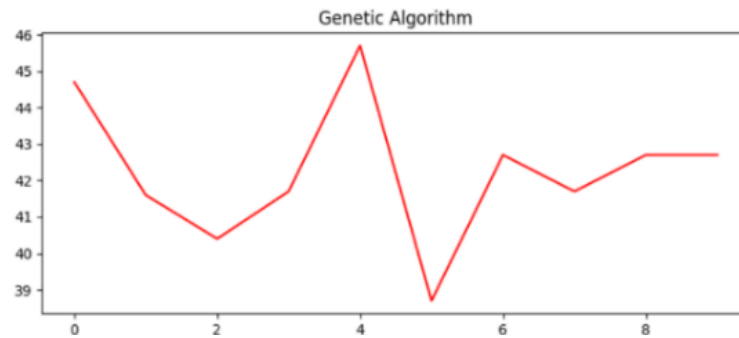


Figura 21 - Execução do Algoritmo Genético

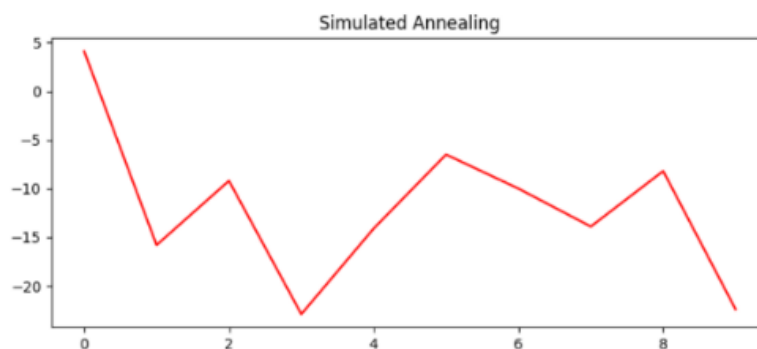


Figura 22 - Execução do Simulated Annealing

As figuras 19 a 22 representam os resultados obtidos para as 10 execuções efetuadas para cada algoritmo.

Na data utilizada para realizar as 10 execuções, havia um caso de sobrelotação com mais 53 doentes do que camas a disponibilizar. Este facto aliado à atribuição errada de camas levou ao registo de cerca de 92 doentes alocados a especialidades erradas. Quando comparamos os resultados apresentados pelo HC e pelo GA, apesar de cumprirem com as atuais limitações de camas, não foi necessária a alocação errada de doentes. O número de doentes físicos foi maximizado de 459 para 551 e a classe responsável foi minimizada de 92 doentes para 0.

De modo conclusivo, foi revelado um desempenho positivo de dois algoritmos: o *Hill Climbing* e o Algoritmo Genético. Ambos, avaliados ao longo de 10 execuções, obtiveram pontuações médias idênticas de 43,2 e 42,2, respetivamente. Tiveram um impacto positivo na gestão de camas quando comparados com a situação atual, recuperando menos 92 alocações erradas e aumentando a classe de doentes físicos de 459 para 551. No que respeita ao desvio-padrão, o

GA destaca-se, com um valor de 1,88 face ao valor de 2,97 obtido pelo HC, sendo assim mais consistente. Relativamente ao tempo de execução, a diferença torna-se significativa uma vez que o GA apresenta menos de metade do HC. É seguro, portanto dizer que, a partir deste estudo o melhor algoritmo a ser utilizado na alocação de camas para promover o aumento da eficiência e desempenho dos hospitais para com seus pacientes é o Algoritmo Genético.

Otimização da distribuição horária dos Enfermeiros

Após a realização do problema da alocação errada de camas e doentes no Hospital seguiu-se então a resolução da alocação horário semanal dos enfermeiros com vista à minimização das horas extra.

Neste problema foi realizada uma abordagem teórica de resolução, isto é, não existiu implementação nem desenvolvimento, foi apenas percebido o problema, definidas as suas limitações e objetivos e explicada a estratégia a seguir para uma posterior resolução.

Para tentar perceber se era ou não possível chegar a uma solução ótima ou que pelo menos apresentasse melhorias significativas face à situação atual a abordagem em termos de modelos testados foi diferente e reduzida. Como está suportado através de estudo realizados anteriormente onde já foram testados a aplicação de algoritmos de otimização, mais concretamente os GA ou algoritmos genéticos, para gestão dos recursos humanos na indústria hospitalar e estes registaram sucesso, destacando-se também de outros modelos e *frameworks* utilizadas, como aconteceu na resolução do problema anterior, neste estudo foi tomado apenas como hipótese a aplicação desta classe de algoritmos de modo a utilizar o tempo e os recursos de uma forma mais eficiente.

O desafio em si consistia, a partir de um conjunto de dados correspondentes à alocação real dos enfermeiros no Hospital e tendo em conta a previsão do fluxo de doentes, aplicar um modelo de otimização que fosse capaz de entregar uma planificação horária dos respetivos, cumprindo certas restrições e objetivos de maneira que minimizar as horas extra e maximizasse a eficiência na utilização destes recursos.

Restrições apresentadas:

- São necessárias em média 4.2 horas de cuidado enfermeiro por cada doente
- Existem três turnos (Manhã – Tarde – Noite)
- O turno da Manhã tem de representar cerca de 50% do tempo laboral de Enfermaria
- O turno da tarde tem de representar 30% do tempo de Enfermaria
- O turno da noite tem de representar 20% do tempo de Enfermaria
- O turno da manhã tem a duração de 6 horas (8:30h – 14:30h)
- O turno da tarde tem a duração de 6 horas (14:30h – 20:30h)
- O turno da noite tem a duração de 12 horas (20:30 – 8:30)
- São necessários no mínimo 6 enfermeiros por turno
- Existem 45 Enfermeiros no Departamento
- Cada enfermeiro trabalha 35 horas semanais mais horas extras consoante as necessidades

A partir de estas restrições foi possível inferir que o tempo necessário para cuidado de enfermaria num departamento é igual a $4.2 \times \text{número de doentes responsáveis previstos}$. Significa isto também que para determinar o número de horas de cuidado de enfermaria por turno basta multiplicar o valor resultante da multiplicação pelas percentagens de tempo laboral de cada um dos turnos.

Como a distribuição horária é feita por dias e por turnos dentro da mesma semana, a abordagem escolhida foi de otimizar os Turnos de um dia separadamente, otimizando a distribuição horária de cada turno teremos a solução ótima, se assim for possível, para o dia em questão. Para tal, é necessário ter em conta que com o limite de horas extra dedicadas da para uma determinada semana, podemos deduzir o número de horas extra dedicados a cada um dos Turnos. Por exemplo, na semana a otimizar foi definido um limite de horas extra semanais de 7 horas por cada enfermeiro, isto é o mesmo que dizer que cada enfermeiro terá de fazer é média mais 1 hora por dia. Aplicando o mesmo princípio utilizado para saber a quantidade horas laborais a realizar por turno, multiplicando o limite diário de horas extra pelas percentagens de cada turno temos o número de horas extra que cada enfermeiro pode realizar dentro desse turno.

É importante referir que na abordagem realizada foi tido como pressuposto de solução que foi definido um número limite de horas extra semanais para os enfermeiros, o algoritmo com base no mesmo tentava minimizar as tais horas extra laboradas e quantidade de vezes que o limite era ultrapassado. Este limite podia ser calculado através do tempo necessário para cuidado de enfermagem da semana em questão, tendo sempre em conta que apenas existiam 45 Enfermeiros disponíveis o que podia ser manifestamente pouco para as necessidades do Hospital, nas datas a testar.

Os dados fornecidos para este estudo foram os mesmos adicionando ainda um *dataset* contendo informação acerca da alocação horária dos enfermeiros no departamento deste do final do mês de janeiro de 2021 até ao final do mês de fevereiro de 2021. Destes dados podíamos retirar informação como:

- Departamento Médico
- Serviço do Departamento de cada Enfermeiro
- Quadro de departamento alocado a cada Enfermeiro e respetivo código
- Data de início da semana
- Data diária
- Turno de cada Enfermeiro
- Nome
- Código de Identificação

Tendo em conta os dados fornecidos, as restrições e os objetivos do problema em questão, a fase seguinte consistiu em definir o domínio ou espaço solução com base nessas informações. Tal como aconteceu com o problema da alocação de camas, a distribuição horária também seguiu uma abordagem dinâmica do domínio. Ou seja, mais uma vez, o domínio não é uma variável fixa e constante, mas sim um valor ou intervalo de valores que varia diariamente e de turno para turno, uma vez que as limitações de horas extra foram aplicadas aos turnos, conforme mencionado anteriormente. O domínio para cada turno de um determinado dia seria compreendido entre 0 e o limite de horas extra do turno em questão, calculado através do limite de horas extra imposto para a semana.

A solução era então apresentada por turno e consistia num *array* de 45 posições, onde cada posição estava associada a um enfermeiro específico e às horas extra que lhe foram atribuídas nesse mesmo turno. Elaborando isto para cada turno de todos os dias da semana, obtemos a distribuição horária das horas extras realizadas para uma determinada semana. Dessa forma, é possível ter uma visão clara e organizada da distribuição das horas extras entre os enfermeiros e garantir que as restrições impostas sejam cumpridas.

Além disso, esta abordagem dinâmica do domínio permite uma maior flexibilidade na gestão das horas extras, permitindo ajustes em tempo real de acordo com as necessidades e restrições impostas. Isso garante que a distribuição das horas extras seja feita de forma justa e equilibrada entre os enfermeiros, evitando sobrecarga de trabalho e garantindo o cumprimento das normas estabelecidas. Através desta solução, é possível gerir eficientemente a distribuição das horas extras e garantir o bem-estar dos enfermeiros e a qualidade do atendimento prestado aos pacientes.

Após realizada a interpretação e proposta de resolução seria necessário elaborar uma função objetivo à semelhança do problema da afetação de camas seguido da aplicação do algoritmo genético em todos os turnos de uma determinada semana. Posteriormente uma comparação entre a solução atual encontrada pelo Hospital e a solução encontrada pelo algoritmo de otimização, tendo deste modo perceber se é ou não pertinente a aplicação deste tipo de algoritmos sobre este contexto e também quais as componentes a serem melhoradas para uma melhor execução dos mesmos.

5. CONCLUSÃO

Finalizado o projeto podemos concluir que a realização do mesmo foi bem-sucedida uma vez que todos os objetivos propostos foram atingidos e ainda foi possível realizar um extra para modelação da otimização das horas extra por parte do Enfermeiros. De modo a compreender melhor o trabalho realizado seguem abaixo enumerados os objetivos propostos e os resultados atingidos.

Objetivos esperados:

- A definição de um conjunto de modelos preditivos sobre o fluxo dos doentes;
- A definição de um conjunto de medidas de eficiência para avaliar cenários de disponibilidade de camas e afetação de recursos;
- A definição de modelos de otimização da disponibilidade de camas e afetação de recursos humanos com base nos modelos preditivos desenvolvidos nos pontos anterior, os recursos disponíveis e na programação de cirurgias;

Resultados atingidos:

- Desenvolvimento e implementação de um modelo preditivo do fluxo dos doentes
- Definição de um conjunto de métricas e medidas de avaliação para cenários de disponibilidade de camas e afetação de recursos;
- Desenvolvimento e implementação de vários modelos de otimização das camas no Hospital;
- Benchmarking dos modelos realizados e devidas ilações;
- Definição e modelação de um modelo de otimização visando minimizar o recurso da hora extra por parte dos Enfermeiros;
- Questão de investigação inicialmente formulada, respondida.

Quanto aos resultados atingidos foram exatamente os esperados sendo estes a promoção da implementação de um sistema de ABI que permita, em tempo real, otimizar a gestão e planeamento de camas e do fluxo de doentes, sendo que os modelos em questão se encontram prontos a serem implementados.

A implementação dos algoritmos testados neste estudo, distinguindo o algoritmo genético, pode levar a uma diminuição do desperdício de recursos ao permitir que as unidades hospitalares tenham a oportunidade de enfrentar a sobrelotação e sejam capazes de responder às suas necessidades diárias. Isto leva a uma gestão e planeamento de camas muito melhor, onde todos os pacientes têm a possibilidade de ser alocados à sua unidade / especialidade hospitalar correta, aumentando a qualidade dos serviços médicos prestados aos pacientes. Com isto é possível dizer, relativamente à questão de investigação proposta, é possível obter uma melhoria efetiva na alocação de camas conseguindo minimizar a distribuição errada de doentes com recurso à junção de técnicas de inteligência artificial, preditivas e de otimização.

No decorrer desta dissertação foi também realizado um artigo abordando o problema da alocação errada de camas nos hospitais, intitulado de “Better Medical Efficiency by means of Hospital Bed Management Optimization – a Comparison of Artificial Intelligence Techniques”. Este artigo foi aceite para publicação para conferência de Inteligência Artificial - EPIA 2023.

Para estudos futuros, aconselha-se não só a um estudo mais aprofundado sobre algoritmos dependentes da otimização de parâmetros de entrada, como é o caso do AG e principalmente do *Simulated Annealing*, que com a otimização dos parâmetros de *input* podem apresentar resultados muito bons, mas também, dar também seguimento a uma implementação e resolução adequada ao problema de minimização das horas extra dos Enfermeiros numa unidade hospitalar.

6. REFERÊNCIAS BIBLIOGRÁFICAS

- [1] K. Peffers, T. Tuunanen, ... M. R.-J. of, and undefined 2007, "A design science research methodology for information systems research," *Taylor & Francis*, vol. 24, no. 3, pp. 45–77, Dec. 2007, doi: 10.2753/MIS0742-1222240302.
- [2] C. Schröer, F. Kruse, J. G.-P. C. Science, and undefined 2021, "A systematic literature review on applying CRISP-DM process model," *Elsevier*, Accessed: Mar. 23, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1877050921002416>
- [3] R. Wirth and J. Hipp, "CRISP-DM: Towards a Standard Process Model for Data Mining."
- [4] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 526–534. doi: 10.1016/j.procs.2021.01.199.
- [5] Z. Michalewicz, M. Schmidt, M. Michalewicz, and C. Chiriac, *Adaptive Business Intelligence*. Springer, 2007. Accessed: Jun. 27, 2023. [Online]. Available: <https://hdl.handle.net/2440/36909>
- [6] J. Lopes, T. Guimarães, and M. F. Santos, "Adaptive business intelligence: A new architectural approach," in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 540–545. doi: 10.1016/j.procs.2020.10.075.
- [7] J. Lopes, J. Braga, and M. F. Santos, "Adaptive business intelligence platform and its contribution as a support in the evolution of hospital 4.0," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 905–910. doi: 10.1016/j.procs.2021.04.016.
- [8] R. Goldschmidt, E. Passos, and E. Bezerra, "Data Mining," 2015, Accessed: Mar. 23, 2022. [Online]. Available: [https://books.google.com/books?hl=pt-PT&lr=&id=HN6sCQAAQBAJ&oi=fnd&pg=PT4&dq=Goldschmidt,+R.,+Passos,+E.,+%26+Bezerra,+E.,+\(2015\).+Data+Mining.+Elsevier+Brasil.&ots=3tuN_KdNO3&sig=03I9282PjN7heyIkzPYAEnGK7Yw](https://books.google.com/books?hl=pt-PT&lr=&id=HN6sCQAAQBAJ&oi=fnd&pg=PT4&dq=Goldschmidt,+R.,+Passos,+E.,+%26+Bezerra,+E.,+(2015).+Data+Mining.+Elsevier+Brasil.&ots=3tuN_KdNO3&sig=03I9282PjN7heyIkzPYAEnGK7Yw)
- [9] K. R.-I. J. of A. R. in and undefined 2012, "Data mining techniques," *Citeseer*, vol. 2, no. 10, p. 2277, 2012, Accessed: Mar. 23, 2022. [Online]. Available: <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.898.6657&rep=rep1&type=pdf>

- [10] H. Koh, G. T.-J. of healthcare information management, and undefined 2011, "Data mining applications in healthcare," *Citeseer*, Accessed: Mar. 23, 2022. [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.92.3184&rep=rep1&type=pdf>
- [11] J. Santos-Pereira, L. Gruenwald, and J. Bernardino, "Top data mining tools for the healthcare industry," *Journal of King Saud University - Computer and Information Sciences*. King Saud bin Abdulaziz University, 2021. doi: 10.1016/j.jksuci.2021.06.002.
- [12] I. el Naqa and M. J. Murphy, "What Is Machine Learning?," *Machine Learning in Radiation Oncology*, pp. 3–11, 2015, doi: 10.1007/978-3-319-18305-3_1.
- [13] T. A.-N. advances in machine learning and undefined 2010, "Types of machine learning algorithms," *books.google.com*, Accessed: Mar. 23, 2022. [Online]. Available: [https://books.google.com/books?hl=pt-PT&lr=&id=XAqhDwAAQBAJ&oi=fnd&pg=PA19&dq=Ayodele,+T.+O.+\(2010\).+Types+of+machine+learning+algorithms.+New+advances+in+machine+learning,+3,+19-48.&ots=r2Jq7UDjKm&sig=cEk02QrQZQB4WRnozEZ80pM1FOc](https://books.google.com/books?hl=pt-PT&lr=&id=XAqhDwAAQBAJ&oi=fnd&pg=PA19&dq=Ayodele,+T.+O.+(2010).+Types+of+machine+learning+algorithms.+New+advances+in+machine+learning,+3,+19-48.&ots=r2Jq7UDjKm&sig=cEk02QrQZQB4WRnozEZ80pM1FOc)
- [14] M. Mohri, A. Rostamizadeh, and A. Talwalkar, "Foundations of machine learning," 2018, Accessed: Mar. 23, 2022. [Online]. Available: [https://books.google.com/books?hl=pt-PT&lr=&id=dWB9DwAAQBAJ&oi=fnd&pg=PR5&dq=%5D+Mohri,+M.,+Rostamizadeh,+A.,+%26+Talwalkar,+A.+\(2018\).+Foundations+of+machine+learning.+MIT+press.+&ots=AyrXURw3n2&sig=UdHidqK-fmo-RK2GO3fTtgpLhIU](https://books.google.com/books?hl=pt-PT&lr=&id=dWB9DwAAQBAJ&oi=fnd&pg=PR5&dq=%5D+Mohri,+M.,+Rostamizadeh,+A.,+%26+Talwalkar,+A.+(2018).+Foundations+of+machine+learning.+MIT+press.+&ots=AyrXURw3n2&sig=UdHidqK-fmo-RK2GO3fTtgpLhIU)
- [15] B. Liu, "Supervised Learning," in *Web Data Mining*, Springer Berlin Heidelberg, 2011, pp. 63–132. doi: 10.1007/978-3-642-19460-3_3.
- [16] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," *ACM International Conference Proceeding Series*, vol. 148, pp. 161–168, 2006, doi: 10.1145/1143844.1143865.
- [17] P. Dayan, M. Sahani, G. D.-T. M. encyclopedia of the, and undefined 1999, "Unsupervised learning," *math.princeton.edu*, Accessed: Mar. 25, 2022. [Online]. Available: <https://web.math.princeton.edu/~sswang/developmental-diaschisis-references/dun99b.pdf>
- [18] B. M.-I. J. of S. and R. (IJSR) and undefined 2020, "Machine learning algorithms-a review," *researchgate.net*, 2019, doi: 10.21275/ART20203995.

- [19] Z. Ghahramani, "Unsupervised learning," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 3176, pp. 72–112, 2004, doi: 10.1007/978-3-540-28650-9_5.
- [20] L. Pack Kaelbling, M. L. Littman, A. W. Moore, and S. Hall, "Reinforcement Learning: A Survey," 1996.
- [21] S. Kaushik *et al.*, "AI in Healthcare: Time-Series Forecasting Using Statistical, Neural, and Ensemble Architectures," *Front Big Data*, vol. 3, Mar. 2020, doi: 10.3389/fdata.2020.00004.
- [22] F. Harrou, F. Kadri, Y. Sun, and S. Khadraoui, "Monitoring patient flow in a hospital emergency department: ARMA-based nonparametric GLRT scheme," *Health Informatics J*, vol. 27, no. 2, 2021, doi: 10.1177/14604582211021649.
- [23] E. Talbi, T. M.-[1993] P. of the Twenty-sixth, and undefined 1993, "Hill-climbing, simulated annealing and genetic algorithms: a comparative study and application to the mapping problem," *ieeexplore.ieee.org*, Accessed: Apr. 06, 2022. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/284069/>
- [24] S. P. Brooks, B. J. T. Morgan, and B. J. T. Morgant, "Optimization Using Simulated Annealing," 1995.
- [25] J. Shapiro, "Genetic algorithms in machine learning," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 2049 LNAI, pp. 146–168, 2001, doi: 10.1007/3-540-44673-7_7.
- [26] P. Ross, D. C.-A. workshop on evolutionary computing, and undefined 1995, "Comparing genetic algorithms, simulated annealing, and stochastic hillclimbing on timetabling problems," *Springer*, Accessed: Apr. 20, 2022. [Online]. Available: https://link.springer.com/chapter/10.1007/3-540-60469-3_28
- [27] A. Apornak, S. Raissi, A. Keramati, and K. Khalili-Damghani, "Human resources optimization in hospital emergency using the genetic algorithm approach," *Int J Healthc Manag*, vol. 14, no. 4, pp. 1441–1448, 2021, doi: 10.1080/20479700.2020.1763236.