INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA INFORMÁTICA




Aplicação de Machine Learning na identificação de clusters com cancro do reto em função de alterações metabólicas

SIMÃO PEDRO PEREIRA GOMES
Setembro de 2025

POLITÉCNICO
DO PORTO

ISEP INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO

Aplicação de Machine Learning na identificação de clusters com cancro do reto em função de alterações metabólicas

Simão Gomes

Dissertação para obtenção do Grau de Mestre em Engenharia Informática, Área de Especialização em Cibersegurança e Administração de Sistemas

Orientador: Professor Doutor José Tavares
Co-Orientador: Professora Doutora Isabel Praça

Porto, 21 de setembro de 2025

Declaração de Integridade

Declaro ter conduzido este trabalho académico com integridade.

Não plagiei ou apliquei qualquer forma de uso indevido de informações ou falsificação de resultados ao longo do processo que levou à sua elaboração.

Portanto, o trabalho apresentado neste documento é original e de minha autoria, não tendo sido utilizado anteriormente para nenhum outro fim. As exceções estão explicitamente reconhecidas na secção "Considerações éticas" do primeiro capítulo. Esta secção também declara como as ferramentas de IA foram utilizadas e para que finalidade.

Declaro ainda que tenho pleno conhecimento do Código de Conduta Ética do P.PORTO.

ISEP, Porto, 21 de setembro de 2025

Dedicatória

Dedico esta dissertação à minha família e a todas as pessoas que, de forma direta ou indireta, me apoiaram e incentivaram ao longo deste percurso. O vosso suporte foi imprescindível para a concretização deste objetivo.

Resumo

O cancro colorretal, o terceiro mais comum no mundo e a segunda principal causa de morte por cancro (Fernandes, Gollub e Brown 2022), apresenta uma resposta terapêutica bastante variável ao tratamento neoadjuvante (Roeder et al. 2020). Neste contexto, os perfis metabólicos têm sido alvo de estudo, revelando-se promissores como potenciais preditores da resposta terapêutica (F. Xu et al. 2023).

O presente trabalho analisou técnicas de aprendizagem não supervisionada para identificar padrões em perfis metabólicos de aminoácidos e acilcarnitinas, com o objetivo de melhorar a estratificação de doentes com cancro colorretal. Foram analisadas 4052 amostras de aminoácidos e 865 de acilcarnitinas da Unidade Local de Saúde de Santo António, organizadas em três conjuntos: população geral, doentes com cancro colorretal (CRC) e doentes CRC em estádio M0.

Os resultados demonstraram a capacidade de identificar *clusters* bem delimitados para os três conjuntos de dados. Observou-se uma associação moderada entre *clusters* e diagnósticos nos perfis de aminoácidos e uma associação forte nos perfis de acilcarnitinas. Adicionalmente, foram identificados metabólitos mais discriminantes em cada *cluster*. Contudo, a análise longitudinal realizada em doentes com múltiplas amostras não revelou associações significativas entre *clusters* e progressão clínica, sugerindo a necessidade de estudos futuros com amostras mais robustas.

Em suma, este trabalho reforça o potencial dos perfis metabólicos como ferramentas complementares de apoio ao diagnóstico e à estratificação de doentes com cancro colorretal.

Palavras-chave: *Machine Learning*, Aprendizagem Não Supervisionada, Cancro Colorretal, Perfis Metabólicos, Aminoácidos, Acilcarnitinas, *Clustering*, Biomarcadores

Abstract

Colorectal cancer is the third most common cancer worldwide and second leading cause of cancer-related deaths (Fernandes, Gollub e Brown 2022). This disease has a highly variable therapeutic response to neoadjuvant treatments (Roeder et al. 2020, and metabolic profiles have been studied as potential predictors of therapeutic response (F. Xu et al. 2023).

The present work analyzed unsupervised learning techniques to identify patterns in metabolic profiles of amino acids and acylcarnitines, aiming to improve the stratification of patients with colorectal cancer. A total of 4,052 amino acid samples and 865 acylcarnitine samples from Unidade Local de Saúde de Santo António were analyzed, organized into three datasets: the general population, patients with colorectal cancer (CRC), and CRC patients in the M0 stage.

The results demonstrated the ability to identify well-defined clusters for all three datasets. A moderate association was observed between clusters and diagnoses in amino acid profiles and a strong association in acylcarnitine profiles. Additionally, the most discriminant metabolites in each cluster were identified. However, a longitudinal analysis done on patients with multiple samples did not reveal any significant association between clusters and clinical progression, which suggests that future studies may need to use more robust samples.

In summary, this work reinforces the potential of metabolic profiles as complementary tools to support diagnosis and stratification of patients with colorectal cancer.

Keywords: Machine Learning, Unsupervised Learning, Colorectal Cancer,Metabolic Profiles, Amino acids, Acylcarnitines, Clustering, Biomarkers

Agradecimentos

Em primeiro lugar, expresso a minha profunda gratidão à minha família pelo apoio constante, compreensão e incentivo ao longo de todo este percurso. Sem o vosso suporte incondicional, a concretização deste projeto não teria sido possível.

De igual forma, quero manifestar o meu sincero agradecimento ao Professor Doutor José Tavares, que aceitou o desafio de orientar este trabalho. A sua disponibilidade, dedicação e o rigor científico que imprimiu ao desenvolvimento desta dissertação foram determinantes para o sucesso alcançado. Os seus conselhos valiosos constituíram uma fonte contínua de inspiração e orientação ao longo de todo o processo.

Deixo também um especial agradecimento à Professora Doutora Lúcia Lacerda, pela disponibilidade em participar em reuniões periódicas de validação dos resultados e pelo *feedback* rigoroso e construtivo que contribuiu de forma significativa para o desenvolvimento deste trabalho.

Agradeço igualmente a todos os profissionais do Instituto Superior de Engenharia do Porto, que tanto contribuíram para o meu crescimento pessoal e académico. As experiências vividas e os conhecimentos adquiridos durante este percurso foram determinantes para a consolidação das minhas competências e para a minha formação enquanto futuro profissional.

Por último, mas não menos importante, quero agradecer aos meus amigos e colegas de curso. A vossa amizade, companheirismo e constante apoio tornaram esta jornada académica mais leve, enriquecedora e significativa.

Conteúdo

Lista de Figuras **xv**

Lista de Tabelas **xvii**

Lista de Abreviações **xix**

Lista de Símbolos **xxi**

1 Introdução **1**
1.1 Contexto . . . 1
1.2 Problema . . . 2
1.3 Objetivos . . . 2
1.4 Interpretação Analítica, Crítica e Ética . . . 2
1.4.1 Interpretação analítica . . . 3
1.4.2 Interpretação crítica . . . 3
1.4.3 Interpretação ética . . . 4
1.5 Metodologia . . . 5
1.5.1 Metodologia de Pesquisa . . . 5
1.5.2 Metodologia para Desenvolvimento da Solução . . . 6
1.6 Uso de Inteligência Artificial . . . 7
1.7 Estrutura do Documento . . . 8

2 Revisão Literária **9**
2.1 Cancro Colorretal . . . 9
2.2 Perfis Metabólicos . . . 11
2.2.1 Perfis de Aminoácidos . . . 11
2.2.2 Perfis de acilcarnitinas . . . 12
2.3 *Machine Learning* . . . 12
2.4 Categorias de Aprendizagem em *Machine Learning* . . . 13
2.4.1 Aprendizagem Supervisionada . . . 13
2.4.2 Aprendizagem Não Supervisionada . . . 13
2.4.3 Aprendizagem Semi-Supervisionada . . . 14
2.4.4 Aprendizagem por Reforço . . . 14
2.5 Algoritmos de Aprendizagem não supervisionada . . . 15
2.5.1 Métodos de Deteção de *Outliers* . . . 15
Métodos Baseados em Estatística . . . 16
Métodos Baseados em Distância . . . 16
Métodos Baseados em Densidade . . . 16
Métodos Baseados em *Clustering* . . . 17
2.5.2 Algoritmos de *Clustering* . . . 17
Exclusive Clustering . . . 17

Overlapping Clustering . . . 17
Hierarchical Clustering . . . 18
Probabilistic Clustering . . . 18
2.6 Tecnologias para Desenvolvimento de *Machine Learning* . . . 18
2.6.1 *Scikit-learn* . . . 18
2.6.2 *MLlib* . . . 19
WEKA . . . 19
2.6.3 Comparação entre as diferentes bibliotecas . . . 20
2.7 Estudos de Aplicação de ML em Medicina . . . 20
2.8 Conclusões . . . 22

3 Análise e Preparação dos Dados **23**
3.1 *Dataset* . . . 23
3.2 Questões de Investigação . . . 27
3.3 Preparação dos Dados . . . 28
3.3.1 Normalização e Tratamento de Dados . . . 28
3.3.2 Seleção de *features* e Redução de dimensionalidade . . . 29
3.3.3 Deteção de *outliers* . . . 30

4 Modelação e Desenvolvimento Experimental **33**
4.1 Seleção dos Algoritmos . . . 33
4.2 Metodologia de Hiperparametrização . . . 35
4.3 *Clusters* para Perfis de Acilcarnitinas . . . 37
4.3.1 *Dataset* com dados de doentes com diversos diagnósticos . . . 37
4.3.2 *Dataset* com doentes com diagnóstico CRC . . . 40
4.3.3 *Dataset* com doentes com diagnóstico CRC em estádio M0 . . . 43
4.4 *Clusters* para Perfis de Aminoácidos . . . 45
4.4.1 *Dataset* com dados de doentes com diversos diagnósticos . . . 46
4.4.2 *Dataset* com doentes com diagnóstico CRC . . . 49
4.4.3 *Dataset* com doentes com diagnóstico CRC em estádio M0 . . . 52

5 Avaliação e Discussão dos Resultados **55**
5.1 Avaliação dos modelos . . . 55
5.2 Interpretação Biológica dos Resultados . . . 58
5.3 Implicações Clínicas . . . 59
5.4 Validação pelos Profissionais de saúde . . . 59
5.5 Limitações do estudo . . . 60
5.6 Sugestões para futuros trabalhos . . . 61
5.7 Discussão Final . . . 62

6 Conclusões **63**

Bibliografia **65**

Lista de Figuras

1.1 Fases do Modelo CRISP-DM (IBM 2021) 7

2.1 Comparação entre um cólon saudável e um cólon com pólipos, estruturas frequentemente associados com o desenvolvimento de cancro colorretal (Fonte: *O que é o cancro colorretal?* 2025) 10

3.1 Exemplo parcial do conjunto de dados 24

4.1 Distribuição das amostras após redução de dimensionalidade por *PCA*, com destaque para casos CRC, para ambos os algoritmos de *clustering*. 38
4.2 *Heatmap* com os 25 metabólitos com maior variabilidade entre *clusters* . . 39
4.3 Distribuição de diagnósticos por *clusters* obtidos pelos algoritmos *K-Means* e *GMM*. 40
4.4 Distribuição das amostras após redução de dimensionalidade por PCA para ambos os algoritmos de *clustering* no *dataset* CRC. 41
4.5 *Heatmap* com os 25 metabólitos com maior variabilidade entre *clusters* no *dataset* CRC 42
4.6 Distribuição das amostras após redução de dimensionalidade por *PCA* para ambos os algoritmos de *clustering* no *dataset* CRC M0. 44
4.7 *Heatmap* com os 25 metabólitos com maior variabilidade entre *clusters* no *dataset* CRC M0 45
4.8 Distribuição das amostras após redução de dimensionalidade por *PCA* para ambos os algoritmos de *clustering* no *dataset* de aminoácidos. 47
4.9 *Heatmap* com os 25 aminoácidos com maior variabilidade entre *clusters* . . 48
4.10 Distribuição dos diagnósticos por *cluster*, representada sob a forma de *heatmap*. 49
4.11 Distribuição das amostras após redução de dimensionalidade por *PCA* para ambos os algoritmos de *clustering* no *dataset* CRC. 51
4.12 *Heatmap* dos 25 aminoácidos com maior variabilidade entre *clusters*, obtidos a partir do *dataset* CRC, para os algoritmos *K-Means* e *GMM*. 52
4.13 Distribuição das amostras após redução de dimensionalidade por *PCA* para ambos os algoritmos de *clustering* no *dataset* de aminoácidos CRC M0. . . 53
4.14 *Heatmap* com os 25 aminoácidos com maior variabilidade entre *clusters* no *dataset* de aminoácidos CRC M0 54

5.1 *Heatmap* com valores normalizados para todos os conjuntos de dados e métricas de avaliação de agrupamento. 56

Lista de Tabelas

2.1 Comparação entre *Scikit-learn*, *MLlib* e *WEKA*. 20
2.2 Diferentes Publicações de Aplicação de *Machine Learning* em Saúde, e mais especificamente, Cancro Colorretal 21

3.1 Caracterização das *Features* Metabólicas do *Dataset* 24
3.2 Diagnósticos Presentes no *Dataset* de Aminoácidos 25
3.3 Caracterização das *Features* de Acilcarnitinas do *Dataset* 26
3.4 Diagnósticos Presentes no *Dataset* de Acilcarnitinas 27
3.5 Parâmetros utilizados no algoritmo *Isolation Forest* 31
3.6 Número total de amostras e número de *outliers* identificados em cada conjunto de dados analisado 31

4.1 Espaços de hiperparâmetros definidos para cada algoritmo de *clustering* . . 35
4.2 Configurações finais selecionadas para o *K-Means* e *GMM* 37
4.3 Distribuição das amostras por *cluster* para o *K-Means* e *GMM* 38
4.4 Comparação das métricas globais de qualidade do *clustering* para *K-Means* e *GMM* 38
4.5 Configurações finais selecionadas para o *K-Means* e *GMM* no *dataset* CRC 40
4.6 Distribuição das amostras por *cluster* para o *K-Means* e *GMM* no *dataset* CRC 41
4.7 Comparação das métricas globais de qualidade do *clustering* para *K-Means* e *GMM* no *dataset* CRC 41
4.8 Distribuição das amostras por *cluster* e estádio de CRC para *K-Means* e *GMM* 43
4.9 Configurações finais selecionadas para o *K-Means* e *GMM* no *dataset* CRC M0 43
4.10 Distribuição das amostras por *cluster* para o *K-Means* e *GMM* no *dataset* CRC M0 43
4.11 Comparação das métricas globais de qualidade do *clustering* para *K-Means* e *GMM* no dataset CRC M0 44
4.12 Taxa de progressão terapêutica por *cluster* para *K-Means* e *GMM* 45
4.13 Configurações finais selecionadas para o *K-Means* e *GMM* no *dataset* de aminoácidos 46
4.14 Distribuição das amostras por *cluster* para o *K-Means* e *GMM* no *dataset* de aminoácidos 46
4.15 Comparação das métricas globais de qualidade do *clustering* para *K-Means* e *GMM* no *dataset* de aminoácidos 47
4.16 Configurações finais selecionadas para o *K-Means* e *GMM* no *dataset* CRC 50
4.17 Distribuição das amostras por *cluster* para o *K-Means* e *GMM* no *dataset* CRC 50
4.18 Comparação das métricas globais de qualidade do *clustering* para *K-Means* e *GMM* no *dataset* CRC 51

4.19 Distribuição das amostras por *cluster* e estádio de CRC para os algoritmos *K-Means* e *GMM*. 52
4.20 Configurações finais selecionadas para o *K-Means* e *GMM* no *dataset* de aminoácidos CRC M0 53
4.21 Distribuição das amostras por *cluster* para *K-Means* e *GMM* no *dataset* de aminoácidos (CRC M0) 53
4.22 Comparação das métricas globais de qualidade do *clustering* para *K-Means* e *GMM* no dataset de aminoácidos CRC M0 54
4.23 Taxa de progressão terapêutica por *cluster* para *K-Means* e *GMM* no *dataset* de aminoácidos 54

Lista de Abreviações

BCAA	**B**ranched **C**hain **A**mino **A**cid
BSD	**B**erkeley **S**oftware **D**istribution
CPU	**C**entral **P**rocessing **U**nit
CRISP-DM	**C**ross **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining
DAM	**D**esvio **A**bsoluto **P**adrão
FCM	**F**uzzy **C**-**M**eans
GMM	**G**aussian **M**ixture **M**odels
GPU	**G**raphics **P**rocessing **U**nit
IA	**I**nteligência **A**rtificial
IPP	**I**nstituto **P**olitécnico do **P**orto
ISEP	**I**nstituto **S**uperior de **E**ngenharia do **P**orto
LASSO	**L**east **A**bsolute **S**hrinkage and **S**election **O**perator
LSTM	**L**ong **S**hort **T**erm **M**emory
ML	**M**achine **L**earning
PAM	**P**artition **A**round **M**edoids
PCA	**P**rincipal **C**omponent **A**nalysis
RFE	**R**ecursive **F**eature **E**limination
RMN	**R**essonância **M**agnética **N**uclear
SVM	**S**upport **V**ector **M**achine
t-SNE	**t**-**D**istributed **S**tochastic **N**eighbor **E**mbedding
ULSSA	**U**nidade **L**ocal de **S**aúde de **S**anto **A**ntónio
UMAP	**U**niform **M**anifold **A**pproximation and **P**rojection
VGG	**V**isual **G**eometry **G**roup
XAI	E**x**plainable **A**rtificial **I**ntelligence

Lista de Símbolos

μ	micro (prefixo do Sistema Internacional, 10^{-6})	Unidade de medida, por exemplo em μmol/L
mol/L	Moles por litro	Unidade de concentração de substância em solução

Capítulo 1

Introdução

Este capítulo serve como introdução à dissertação desenvolvida para o Mestrado em Engenharia de *Software* no Instituto Superior de Engenharia do Porto (ISEP). Começa com uma breve contextualização e descrição do problema, estabelecendo a base para o estudo. Seguidamente, são delineados os principais objetivos da investigação e é feita uma breve interpretação ética. O capítulo explica, depois, a metodologia e detalha o modo como a inteligência artificial e os Grandes Modelos de Linguagem (LLMs do inglês *Large Language Models*) foram utilizados ao longo do desenvolvimento da dissertação. Por fim, este capítulo é concluído com um resumo da estrutura geral do documento.

1.1 Contexto

O setor da saúde está a passar por uma transformação significativa, impulsionada pelo aumento dos custos e pela escassez de profissionais qualificados (Khan et al. 2023). Na área da oncologia, por exemplo, a despesa em cuidados oncológicos na Europa quase duplicou, passando de 52 mil milhões de euros, em 1995, para 103 mil milhões de euros, em 2018 (Hofmarcher et al. 2020). Paralelamente, o número de novos casos de cancro diagnosticados aumentou cerca de 50% durante o mesmo período (Hofmarcher et al. 2020).

Em resposta, a adoção de inteligência artificial (IA) e de análise de dados no setor da saúde está a expandir-se a um ritmo sem precedentes. No Reino Unido, por exemplo, os gastos com saúde aumentaram 4,5% em 2019 e 6,25% em 2020 (Whig et al. 2022). Associado a este crescimento dos gastos na saúde, verifica-se também um crescimento nos gastos em soluções digitais de saúde e inclusive do mercado de IA aplicada à saúde. De facto, estudos efetuados nos últimos anos estimavam um valor de mercado de 6,6 mil milhões de dólares no mercado de IA aplicada à saúde, em 2021, o que na altura representava uma taxa de crescimento anual composta de 40% (Bohr e Memarzadeh 2020).

Adicionalmente, espera-se que esta tendência se mantenha nos próximos anos, dado que as aplicações de IA têm o potencial de reduzir a carga de trabalho dos profissionais de saúde, melhorar a precisão diagnóstica e ampliar o conhecimento médico. Ao diminuir as taxas de erro e aumentar a precisão, a IA pode melhorar significativamente a tomada de decisões clínicas e apoiar recomendações médicas mais robustas e fundamentadas em evidências (Aung, D. C. Wong e Ting 2021).

É neste sentido que surge este projeto proposto com apoio de profissionais de saúde do Centro Hospitalar do Porto (Hospital de Santo António). Assim, e aproveitando estudos recentes, pretende-se explorar a aplicabilidade de algoritmos de aprendizagem não supervisionada como mecanismos de predição da resposta ao tratamento de cancro colorretal.

1.2 Problema

O cancro do reto é o terceiro cancro mais comum do mundo e o segundo cancro que mais leva à morte (Fernandes, Gollub e Brown 2022) . Adicionalmente, estudos mostram que a resposta terapêutica ao tratamento neoadjuvante desta doença pode variar bastante de paciente para paciente (Roeder et al. 2020). Como tal, torna-se essencial arranjar métodos que permitam identificar perfis que ajudem a determinar a agressividade tumoral e a resposta esperada ao tratamento.

É, neste sentido, que nos últimos anos têm surgido estudos que mostram que a deteção de alterações metabólicas pode servir como um mecanismo de predição da resposta do paciente ao tratamento (F. Xu et al. 2023).

Este estudo, em particular, pretende analisar perfis metabólicos de aminoácidos e acilcarnitinas em doentes com cancro do reto, de forma a associar estes perfis a uma resposta patológica e, assim, tentar identificar anomalias e *clusters* que permitam melhorar a predição da resposta ao tratamento neoadjuvante. Para fazer esta análise, serão utilizados dados clínicos fornecidos por profissionais de saúde da Unidade Local de Saúde de Santo António (ULSSA) que, adicionalmente, irão ajudar a validar os resultados obtidos.

1.3 Objetivos

Este trabalho tem como principal objetivo analisar um conjunto de dados com base no perfil de aminoácidos e acilcarnitinas de forma a detetar anomalias – instâncias diferentes da norma –, e fazer *clustering* – agrupar instâncias semelhantes em grupos distintos.

Para isso, é pretendido que sejam realizadas duas tarefas de aprendizagem não supervisionada. Aprendizagem não supervisionada é uma técnica de *Machine Learning* - área da Inteligência Artificial que estuda como solucionar problemas complexos e intuitivos -, na qual o modelo funciona de forma independente, sendo que este tipo de aprendizagem pode ser considerado como a identificação de padrões em dados que previamente não tinham sido identificados (Ghahramani 2004).

Resta, ainda, mencionar que a tarefa de *clustering* será um processo iterativo de validação com os profissionais de saúde, que disponibilizam os dados, por forma a validar os grupos a encontrar por técnicas de *Machine Learning*.

1.4 Interpretação Analítica, Crítica e Ética

A Secção de Interpretação Analítica, Crítica e Ética desempenha um papel central na estrutura desta tese, ao abordar questões fundamentais do estudo a ser feito nesta dissertação. Deste modo, são abordados pressupostos metodológicos do trabalho em causa, feitas considerações críticas ao projeto e destacadas algumas das implicações éticas do uso de Inteligência Artificial para a análise de dados metabólicos no contexto oncológico.

Assim, o olhar analítico permite explorar os limites e possibilidades do trabalho em questão. Já a análise crítica permite questionar as escolhas técnicas e metodológicas, promovendo uma visão equilibrada que considere vieses e desafios na aplicabilidade do projeto. Por fim, a interpretação ética permite que sejam destacados procedimentos necessários para um compromisso ético rigoroso.

Deste modo, esta secção contribui para, não só validar a robustez científica do estudo, mas também assegurar que as suas implicações são avaliadas de forma holística e que respeitam os princípios da integridade académica e responsabilidade social.

1.4.1 Interpretação analítica

Esta investigação tem como objetivo principal realizar uma análise dos perfis metabólicos de aminoácidos e acilcarnitinas em doentes com cancro do reto, visando associar essas características metabólicas à resposta dos doentes ao tratamento neoadjuvante. Este estudo fundamenta-se na crescente necessidade de integrar biomarcadores metabólicos como ferramentas promissoras para personalizar abordagens terapêuticas no contexto oncológico (Battini et al. 2017; Marengo e Robotti 2014). Enquadra-se, assim, num cenário em que a estratificação de doentes e a previsão de respostas ao tratamento são indispensáveis para otimizar resultados clínicos e reduzir efeitos adversos desnecessários, como a resistência ao tratamento (Tang et al. 2024).

A abordagem analítica deste estudo enfatiza a importância dos perfis metabólicos como fonte valiosa de informação para compreender a heterogeneidade biológica dos doentes. Nesse contexto, destaca-se o papel dos aminoácidos, não apenas como intermediários metabólicos, mas também como moduladores de vias metabólicas envolvidas no crescimento e na sobrevivência tumoral (Vettore, Westbrook e Tennant 2020). Além disso, são investigadas as acilcarnitinas devido à sua associação com a oxidação de ácidos gordos e a regulação da homeostase metabólica, tendo em conta a sua relevância para os mecanismos de adaptação metabólica em microambientes tumorais (S. Li, Gao e Jiang 2019).

A análise também inclui a exploração de *clusters* metabólicos, que podem refletir padrões biológicos distintos e identificar subgrupos de doentes com respostas diferenciadas ao tratamento neoadjuvante. Este processo exige uma avaliação rigorosa da robustez estatística na formação desses *clusters*, de modo a garantir a sua aplicabilidade como preditores confiáveis da resposta terapêutica.

Adicionalmente, é essencial considerar as limitações associadas à análise de perfis metabólicos. Apesar do elevado potencial desses perfis, a sua utilização na prática clínica requer a integração com outras camadas de dados, como os perfis genómicos (Hiller e Metallo 2013). Além disso, é necessário abordar questões éticas, como a privacidade e o consentimento informado para o uso de dados, aspectos detalhados na Secção 1.4.3.

Por fim, é crucial assegurar que este estudo seja conduzido de forma imparcial, mitigando vieses e garantindo representatividade amostral. Essas medidas são fundamentais para maximizar a aplicabilidade e a equidade dos resultados obtidos.

1.4.2 Interpretação crítica

A análise crítica da utilização de perfis metabólicos de aminoácidos e acilcarnitinas para prever a resposta ao tratamento neoadjuvante em doentes com cancro do reto exige um exame cuidadoso dos pressupostos subjacentes, possíveis enviesamentos e as implicações associadas a esta abordagem emergente.

Assim sendo, em primeiro lugar, é fundamental questionar o pressuposto que perfis metabólicos constituem indicadores precisos e abrangentes da resposta terapêutica. O metabolismo tumoral é influenciado por uma variedade de fatores tanto sistémicos como locais, entre os quais podemos mencionar a dieta, os perfis genómicos e os proteómicos (Davis e Milner

2004; Fernández, Cedrón e Molina 2020). Deste modo, a interpretação crítica leva a analisar a possibilidade de que ao isolar os perfis metabólicos se limita a capacidade preditiva dos modelos baseados exclusivamente nestes dados.

Adicionalmente, outro foco da lente crítica recai sobre a robustez na formação dos *clusters* e a sua associação à resposta ao tratamento neoadjuvante. Embora estas sejam ferramentas promissoras, a identificação destes subgrupos depende de escolhas técnicas como o número de *clusters* ou as métricas de distância (Patil e Baidari 2019; Shapcott 2024), pelo que é imperativo avaliar se estas escolhas refletem a heterogeneidade metabólica dos doentes.

Outro aspeto importante a considerar na interpretação crítica, são possíveis limitações relacionadas à integração dos perfis metabólicos. Embora exista a ideia de que estes possam ajudar na identificação e definição de terapias personalizadas para doentes, tendo em conta o seu perfil metabólico, questões como custo e aceitação clínica podem representar barreiras significativas para a integração clínica. Mais ainda, a extrapolação para populações de doentes com diferentes características demográficas, culturais e metabólicas deve ser examinada de forma cuidadosa, uma vez que pode limitar a aplicabilidade universal dos resultados (Trifonova et al. 2023).

Por fim, deve haver uma análise crítica e reflexiva sobre algumas das implicações éticas do uso de dados metabólicos na predição da resposta terapêutica ao tratamento neoadjuvante. Entre outros aspetos, deve-se refletir sobre potenciais riscos de estigmatização ou discriminação de grupos metabólicos que estejam associados a respostas desfavoráveis ao tratamento. Na secção 1.4.3 é feita uma interpretação ética mais detalhada do projeto em análise nesta Dissertação.

1.4.3 Interpretação ética

Na elaboração desta Dissertação para a obtenção do Grau de Mestre em Engenharia Informática, é crucial observar determinados princípios éticos que orientam tanto o rigor académico como a integridade da investigação.

Em primeiro lugar, é necessário cumprir o Código de Conduta do Instituto Politécnico do Porto (IPP) (IPP 2020). Em particular, destacam-se os seguintes princípios:

- O Artigo 6, alínea 2.8, sublinha a obrigatoriedade de citar e referenciar todas as fontes utilizadas, reconhecendo de forma transparente as ideias e afirmações de terceiros.
- O Artigo 8 reforça a necessidade de submissão de uma declaração formal de compromisso com a integridade e transparência.
- O Artigo 10 enfatiza a importância de citar de modo adequado e abrangente os trabalhos relevantes, apresentando os resultados e interpretações de forma clara, transparente e rigorosa.

Adicionalmente, considerando que este trabalho envolve Inteligência Artificial aplicada à área da saúde, algumas medidas éticas específicas devem ser adotadas para assegurar a privacidade e equidade no tratamento dos dados:

- Os dados utilizados devem ser anonimizados para garantir que nenhuma informação permita a identificação individual dos participantes.
- Em conformidade com o Regulamento Geral sobre a Proteção de Dados (RGPD), dever ser obtido o consentimento explícito dos doentes para o uso dos seus dados.

Estes devem ser informados de forma clara acerca dos dados a serem coletados, dos indivíduos e instituições com acesso a estes dados e dos objetivos do estudo.

- O conjunto de dados utilizado para o momento de treino do modelo de IA deve ser cuidadosamente selecionado para ser representativo da população-alvo (cancro do reto), minimizando vieses relacionados com género, idade ou outros fatores. Adicionalmente, os algoritmos devem ser avaliados para assegurar imparcialidade.
- Por fim, o modelo de IA deve ser concebido de modo a ser interpretável e a possibilitar que os profissionais de saúde compreendam as decisões automatizadas. Para esse efeito, deve-se garantir a documentação exaustiva do projeto e a realização de reuniões periódicas com os profissionais envolvidos.

1.5 Metodologia

A presente secção descreve os procedimentos adotados para a pesquisa e desenvolvimento da solução proposta nesta dissertação. Deste modo, inicialmente é detalhada a metodologia de pesquisa utilizada para a elaboração do Capítulo 2. Por fim, é detalhada a metodologia de Desenvolvimento da Solução, sendo que se optou pela metodologia *Cross-Industry Standard Process for Data Mining (CRISP-DM)*, visto que é uma metodologia amplamente reconhecida em projetos de mineração de dados.

1.5.1 Metodologia de Pesquisa

A pesquisa desenvolvida nesta dissertação visa explorar os perfis metabólicos de aminoácidos e acilcarnitinas em doentes com cancro do reto, associando esses padrões metabólicos à resposta patológica ao tratamento neoadjuvante [1]. Esta análise busca identificar potenciais anomalias e *clusters* metabólicos que possam aprimorar a predição da resposta ao tratamento. O método de pesquisa foi estruturado para assegurar a obtenção de dados robustos e confiáveis, sustentados por técnicas analíticas avançadas e fontes científicas relevantes.

Assim sendo, inicialmente foi realizada uma busca sistemática de literatura científica com o objetivo de compreender as bases metabólicas do cancro do reto e os métodos disponíveis para análise de perfis metabólicos. Deste modo, foram abordados temas como:

- Cancro do reto.
- Perfis Metabólicos.
- Tipos de Aprendizagem.
- Algoritmos de Aprendizagem Não Supervisionados.
- Exemplos de Estudos de Aplicação de IA em Medicina.
- Bibliotecas de *Machine Learning* em *Python*.
- Implantação de Modelos de IA

A revisão foi conduzida em repositórios reconhecidos, como *PubMed*, *ScienceDirect* e *Google Scholar*, utilizando palavras-chave como *"rectal cancer neoadjuvant response"*, *"metabolic profiling"*, *"amino acid profiling"* ou *"acylcarnitine profiling"*. Adicionalmente, foram

[1] **Tratamento Neoadjuvante:** consiste na aplicação de quimioterapia, radioterapia ou terapia-alvo antes de um procedimento cirúrgico ou um ciclo de radioterapia definitivo

priorizados artigos revisados por pares, publicados nos últimos dez anos, e que apresentassem relevância significativa para o campo.

1.5.2 Metodologia para Desenvolvimento da Solução

Para além da componente de investigação desenvolvida nesta dissertação, existe ainda uma vertente de desenvolvimento de uma solução. Esta solução tem como objetivo explorar os perfis metabólicos de aminoácidos e acilcarnitinas em doentes com cancro do reto, associando esses padrões metabólicos à resposta patológica ao tratamento neoadjuvante. Neste sentido, é fundamental definir a metodologia utilizada para a construção da solução pretendida.

Assim, no contexto deste projeto, optou-se por utilizar uma metodologia baseada no modelo **Cross Industry Standard Process For Data Mining** (CRISP-DM). Este é um modelo de mineração de dados que fornece uma visão geral do ciclo de vida de um projeto de mineração de dados, sendo este ciclo de vida dividido em seis fases (Schröer, Kruse e Gómez 2021; Wirth e Hipp 2000):

- **Compreensão do Negócio:** O objetivo principal é compreender o propósito do projeto e os requisitos do cliente, assegurando o alinhamento entre as expectativas e os resultados esperados. No contexto deste estudo, esta fase focou-se na formulação do problema de identificar perfis metabólicos para prever respostas ao tratamento neoadjuvante.
- **Compreensão dos Dados:** Inclui a exploração inicial dos dados fornecidos ou obtidos, permitindo uma análise detalhada para compreender a sua estrutura, qualidade e padrões relevantes.
- **Preparação dos Dados:** Consiste em transformar os dados brutos num formato adequado para os modelos, corrigindo inconsistências, eliminando erros e aplicando técnicas de pré-processamento.
- **Modelação:** Refere-se à seleção de técnicas e algoritmos de aprendizagem automática, bem como à definição dos respetivos parâmetros, resultando na criação de modelos que utilizam os dados previamente preparados. No âmbito deste estudo, esta fase inclui o desenvolvimento de algoritmos de aprendizagem não supervisionada para prever a resposta ao tratamento com base nos dados de perfis metabólicos.
- **Avaliação:** Envolve a análise do desempenho dos modelos desenvolvidos com base em critérios previamente definidos, assegurando que cumprem os objetivos do projeto.
- **Implantação:** Corresponde à apresentação e entrega dos modelos ao cliente, permitindo a sua aplicação prática e a criação de valor.

As diferentes fases do modelo CRISP-DM podem ser observadas na Figura 1.1.

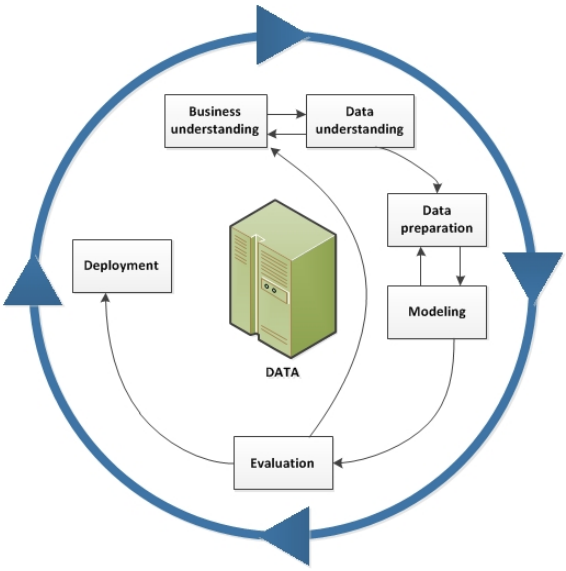


Figura 1.1: Fases do Modelo CRISP-DM (IBM 2021)

Adicionalmente, é importante referir que a sequência das fases não é rígida, pelo que as setas indicam apenas as dependências mais frequentes entre as diferentes fases. Além disso, o círculo externo representa a natureza cíclica da mineração de dados, dado que o processo de mineração de dados não termina após a implementação do modelo, uma vez que as informações obtidas durante essa fase podem originar novas questões de negócio (Wirth e Hipp 2000).

O modelo CRISP-DM foi escolhido devido aos vários benefícios que apresenta no contexto do presente trabalho (Saltz 2021):

- **Flexibilidade:** Permite adaptações ao longo do projeto, facilitando iterações e refinamentos.
- **Foco em Resultados Práticos:** Garante que os resultados atendam às necessidades do problema de negócio identificado.
- **Caráter Cíclico:** O modelo é compatível com a natureza iterativa da ciência de dados. No contexto deste trabalho, esta característica é especialmente relevante, pois a análise dos perfis metabólicos e a sua relação com as respostas patológicas podem exigir ciclos sucessivos de modelação, validação e refinamento dos modelos preditivos.

1.6 Uso de Inteligência Artificial

De modo a garantir um compromisso com a manutenção dos padrões de integridade académica, nesta secção é detalhada a forma como Grandes Modelos de Linguagem (*LLMs*) e outras ferramentas de Inteligência Artificial foram utilizadas ao longo do desenvolvimento desta dissertação.

Deste modo, ao longo desta dissertação foram utilizados o *Writefull* do *Latex* e o *ChatGPT* (em particular o modelo *Academic Assistant*) para o auxílio em termos de gramática, ortografia e uso de linguagem académica e para sugerir melhorias na clareza e coerência do texto. Adicionalmente, e na fase de desenvolvimento experimental, o *ChatGPT* foi utilizado

para depuração de funções complexas e clarificação de documentação técnica ou métodos existentes referenciados na literatura.

No entanto, importa salientar que todo o raciocínio científico, a definição da metodologia e a interpretação dos resultados e conclusões são contribuições independentes do autor.

1.7 Estrutura do Documento

Nesta secção, apresenta-se a organização geral da dissertação, destacando o conteúdo de cada capítulo.

A dissertação inicia-se com uma introdução ao tema, onde é feita a contextualização da área do projeto. Nesta parte, são descritos o problema em estudo, os objetivos a alcançar e a relevância do trabalho no contexto científico e prático. Adicionalmente, discutem-se os aspetos analíticos, críticos e éticos associados ao projeto. Inclui-se também a metodologia adotada — tanto na vertente de pesquisa como no desenvolvimento da solução — bem como o planeamento inicial do trabalho.

O capítulo seguinte apresenta o estado da arte, no qual são descritos os principais conceitos relacionados com o projeto. São abordados o cancro do reto, os perfis metabólicos (com destaque para aminoácidos e acilcarnitinas), os tipos de aprendizagem em *Machine Learning*, os algoritmos de aprendizagem não supervisionada e uma revisão da literatura que aplica *Machine Learning* em saúde para prognóstico, deteção de anomalias e agrupamento.

No terceiro capítulo realiza-se a análise do conjunto de dados, identificando a sua estrutura e a construção dos *subsets* utilizados na análise exploratória. Seguidamente, são apresentadas as questões de investigação que orientaram o projeto e detalham-se as etapas de pré-processamento e tratamento de dados efetuadas.

O quarto capítulo descreve a solução desenvolvida, incluindo os modelos de aprendizagem não supervisionada implementados para a identificação de *clusters*.

Já no quinto capítulo são discutidos e avaliados os resultados obtidos, com base em métricas de desempenho dos algoritmos de *clustering* e no *feedback* fornecido pelos profissionais de saúde da Unidade Local de Saúde de Santo António.

Por fim, o sexto capítulo apresenta as conclusões da dissertação, resumindo os objetivos alcançados, as principais contribuições do trabalho e apontando potenciais melhorias e linhas de investigação futura.

Capítulo 2

Revisão Literária

No segundo capítulo deste trabalho são descritos os principais conceitos relevantes para a compreensão deste. Assim, primeiramente é feita uma breve contextualização acerca do cancro do reto, explicando a sua relevância na sociedade atual, os principais fatores de risco e, por fim, alguns dos tratamentos viáveis.

De seguida, é feita uma breve introdução ao conceito de perfis metabólicos e, em particular, com principal incidência em perfis de aminoácidos e perfis de acilcarnitinas, visto que estes são os perfis metabólicos para os quais se pretende analisar a influência destes na resposta ao tratamento neoadjuvante.

Após esta introdução dos conceitos de saúde, é feita uma breve contextualização tecnológica. Inicialmente, são identificados os principais tipos de aprendizagem no contexto de inteligência artificial, e, após esta breve introdução a esses conceitos, são detalhados alguns dos principais algoritmos de aprendizagem não supervisionada.

Após isto, são analisadas algumas das possíveis ferramentas para o desenvolvimento da solução, sendo esta análise particularmente focada em bibliotecas de ML que poderiam ser usadas para o desenvolvimento da solução.

Por fim, é feita uma breve análise de trabalhos existentes que envolvem a integração de ML na área da saúde.

2.1 Cancro Colorretal

O cancro colorretal é uma doença também conhecida por cancro retal ou cancro do cólon dependendo da localização inicial do tumor. Assim, caso o cancro tenha início no cólon, chama-se cancro do cólon, já se tiver início no reto, chama-se cancro retal. Esta diferença na localização anatómica leva à necessidade de um tratamento diferente. Contudo, a biologia tumoral é a mesma (Medical Oncology 2016). Esta doença começa frequentemente com um pólipo, ou seja, um crescimento anómalo de tecido epitelial na parede do intestino, tal como representado na Figura 2.1.

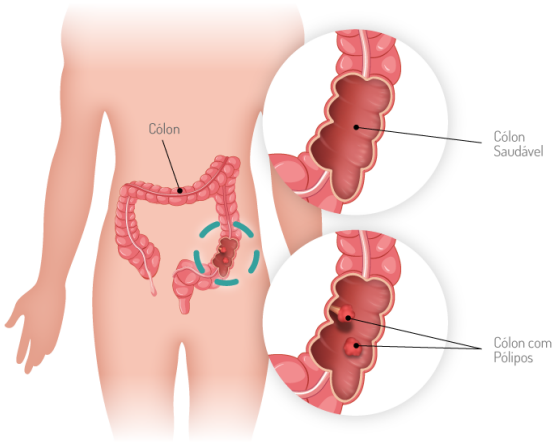


Figura 2.1: Comparação entre um cólon saudável e um cólon com pólipos, estruturas frequentemente associados com o desenvolvimento de cancro colorretal (Fonte: *O que é o cancro colorretal?* 2025)

Adicionalmente, a elevada prevalência global desta doença, associada a uma mortalidade significativa, reflete o impacto que exerce sobre a saúde pública. De facto, e de acordo com estudos realizados em 2020, este cancro correspondia a 10% (cerca de 1,93 milhões) dos casos de cancro identificados a nível mundial, e 9,4% (cerca de 9,96 milhões) de todas as mortes devido a doenças oncológicas resultavam de cancro retal (Xi e P. Xu 2021).

Mais ainda, previsões feitas tendo em conta projeções para o envelhecimento e crescimento da população indicam que está previsto que o número de casos de cancro do reto em 2040 atinja os 3,2 milhões de novos casos.

Estes dados evidenciam a relevância desta neoplastia tanto nos sistemas de saúde quanto no bem-estar das populações afetadas, reforçando a urgência em aprofundar os esforços de investigação científica para a prevenção, diagnóstico precoce e tratamento eficaz desta doença.

Assim sendo, um dos pontos que se deve analisar é a etiologia do cancro retal. Até ao momento, ainda não se sabe ao certo os motivos para a ocorrência do cancro colorretal. Contudo, alguns fatores de risco foram identificados, entre os quais se incluem (Hossain et al. 2022; Medical Oncology 2016; Xi e P. Xu 2021):

- Fatores de risco associado ao estilo de vida, como regime alimentar (por exemplo, um regime rico em carnes vermelhas), obesidade, sedentarismo ou tabagismo.
- Doenças inflamatórias intestinais, como a doença de Crohn.
- Historial familiar: quer devido a genes herdados (por exemplo, síndromes como Polipose Adenomatosa Familiar ou Síndrome de Lynch) quer devido a fatores ambientais partilhados.
- Envelhecimento.
- Historial prévio de cancro retal ou outro tipo de cancro no paciente.

Adicionalmente, cada um destes fatores pode estar relacionado com diferentes localizações anatómicas do cancro. Um exemplo é o tabagismo, que está associado a um maior risco de cancro de cólon proximal e retal, mas não de cancro de cólon distal (Xi e P. Xu 2021).

Por fim, relativamente ao tratamento, este evoluiu muito ao longo das últimas décadas, sendo que estratégias padrão de tratamento incluem cirurgia, quimioterapia, terapia biológica dirigida, radioterapia ou imunoterapia (Medical Oncology 2016). Contudo, o tratamento a ser aplicado varia dependendo do estágio do cancro e da localização do cancro, e estudos mostram que a resposta ao tratamento pode variar bastante de paciente para paciente (Roeder et al. 2020).

É neste sentido que se torna importante arranjar métodos capazes de prever a resposta esperada pelo paciente ao tratamento, sendo que estudos realizados na última década mostram que a deteção de alterações metabólicas pode servir como um mecanismo de predição da resposta do paciente ao tratamento (Hagland et al. 2013; Williams et al. 2013; F. Xu et al. 2023).

2.2 Perfis Metabólicos

A metabolómica, estudo de pequenas moléculas envolvidas nos processos biológicos, tem emergido como uma ferramenta para a caracterização e análise do metabolismo tumoral.

De facto, estudos elaborados nos últimos anos indicam que alterações nestes perfis de biomoléculas podem ser analisadas em doentes com cancro do reto para analisar a progressão da doença e a resposta terapêutica/resistência ao tratamento neoadjuvante (Tian et al. 2018; F. Xu et al. 2023).

Assim sendo, nesta secção é feita uma breve introdução aos dois perfis metabólicos que são o foco deste estudo: perfis de aminoácidos e perfis de acilcarnitinas.

2.2.1 Perfis de Aminoácidos

Os aminoácidos são compostos que possuem um grupo ácido ou carboxila (ligação simples COOH) e um grupo básico ou amina (ligação simples NH2), ambos ligados a um carbono alfa (αC) de um ácido orgânico (A. Blanco e G. Blanco 2022). Estes são as unidades estruturais básicas de proteínas e desempenham um papel fundamental no metabolismo celular, tanto como precursores biossintéticos (como hormonas, coenzimas, antibióticos, neurotransmissores) como também como fonte de energia (Milne e Kilian 2010).

Em tumores como o cancro do reto, o aumento das exigências metabólicas resulta em desequilíbrios nos níveis de aminoácidos, que podem ser detetados e quantificados através de técnicas metabolómicas, como a espectrometria de massa e a ressonância magnética nuclear (RMN). Em particular, algumas das alterações comuns que ocorrem são:

- **Metabolismo da Glutamina:** A glutamina é um substrato essencial para a proliferação de células tumorais, visto que funciona como uma fonte de azoto e carbono. Como tal, muitos tumores demonstram uma elevada dependência por glutamina e níveis elevados de glutamato. Estes dois aspetos têm sido associados à progressão tumoral e à resistência a tratamentos como a radioterapia (Yu et al. 2023).
- **Aminoácidos de Cadeia Ramificada:** A valina, a leucina e a isoleucina são frequentemente consumidas em excesso pelas células tumorais devido ao seu papel no suporte de vias bioenergéticas e biossintéticas. A super-expressão de BCAAs tem, deste modo, sido relacionada com prognósticos mais desfavoráveis para o tratamento de cancro (Pankevičiūtė-Bukauskienė 2024).

2.2.2 Perfis de acilcarnitinas

As acilcarnitinas são metabólitos formados pela ligação de ácidos gordos de diferentes comprimentos com a carnitina e, tal como os ácidos gordos, as acilcarnitinas podem ter diferentes tamanhos. As acilcarnitinas são fundamentais no metabolismo lipídico e energético, especialmente no transporte de ácidos gordos para a beta-oxidação mitocondrial (S. Li, Gao e Jiang 2019).

Estudos desenvolvidos nos últimos anos têm identificado acilcarnitinas como indicadores importantes em estudos metabólicos de diversas doenças, como doenças cardiovasculares, diabetes, depressão, problemas neurológicos e alguns cancros (Dambrova et al. 2022).

Deste modo, alterações nos perfis de acilcarnitinas refletem as adaptações metabólicas das células tumorais face às suas exigências energéticas e ao microambiente tumoral. Neste sentido, alguns estudos demonstram que:

- Existe um aumento dos níveis de acilcarnitina em células tumorais colorretais (Dambrova et al. 2022). Isto pode refletir um estado energético mais favorável destas células quando comparado com células não tumorais (Sánchez-Martínez et al. 2017).
- Perfis metabólicos de acilcarnitinas frequentemente revelam o acumulo de acilcarnitinas de cadeia longa, como a palmitoilcarnitina, indicando uma disfunção na oxidação mitocondrial (Al-Bakheit et al. 2016).

2.3 Machine Learning

No contexto deste projeto, tem-se o objetivo de analisar um conjunto de dados de perfis de aminoácidos e acilcarnitinas de forma a se detetar anomalias e agrupar perfis semelhantes por grupos. Para isto, é pretendido que sejam realizadas duas tarefas de aprendizagem não supervisionada. Como tal, e de forma a contextualizar este projeto, é essencial introduzir alguns conceitos da área de Inteligência Artificial, e particularmente no campo de aprendizagem de máquina (*Machine Learning*), que é a base metodológica deste trabalho.

Inteligência Artificial (IA) é um conceito proposto por John McCarthy, em 1955, na Universidade de *Darthmouth* e pode ser definida, de uma forma simplificada, como o ramo da ciência que visa criar sistemas ou máquinas capazes de simular aspetos da inteligência humana, como raciocínio, aprendizagem e tomada de decisão (Y. Xu et al. 2021; C. Zhang e Lu 2021). Esta é uma área multidisciplinar que integra campos como a ciência da computação, a estatística, a matemática, a biologia e a psicologia, entre outros.

Entre os diversos subcampos da IA, destaca-se a *Machine Learning* (ML). ML é, segundo Arthur Samuel, o campo de estudo que tem como objetivo permitir que sistemas aprendam sem a necessidade de programação específica (Mahesh 2020). Para isto, são utilizados algoritmos que recebem dados de entrada e prevêem os valores de saída, tendo em conta alguns parâmetros de aceitação. Sempre que novos dados são fornecidos ao algoritmo, este aprende e otimiza os seus processos, tornando-se mais preciso e obtendo resultados mais exatos (Peng e Bai 2019).

Atualmente, existem diversas categorias de ML. Na próxima secção desta dissertação, são apresentadas as principais categorias de ML, sendo que estas categorias são ainda associadas ao respetivo problema que visam resolver.

2.4 Categorias de Aprendizagem em Machine Learning

Adicionalmente, os modelos de ML podem ser categorizados com base no tipo de aprendizagem utilizada, refletindo o modo como os algoritmos interagem com os dados. Estes métodos são divididos em quatro categorias principais:

- Aprendizagem Supervisionada
- Aprendizagem Semi-Supervisionada
- Aprendizagem Não Supervisionada
- Aprendizagem por Reforço

De seguida, detalham-se as características de cada uma dessas categorias, estabelecendo a sua relação com os tipos de problemas que resolvem.

2.4.1 Aprendizagem Supervisionada

Aprendizagem Supervisionada refere-se ao uso de algoritmos em que se usam dados rotulados de forma a prever um tipo ou valor de dados novos (C. Zhang e Lu 2021). De facto, este tipo de aprendizagem corresponde a tarefas de aprendizagem de uma função que mapeia os dados de saída baseando-se em exemplos de pares de dados de entrada - dados de saída (Mahesh 2020).

Adicionalmente, modelos deste tipo podem ainda ser divididos em duas subcategorias: **classificação** e **regressão**.

Os algoritmos de classificação focam-se em problemas em que é necessário fazer previsão de valores de saída discretos, como, por exemplo, identificar se o animal numa foto é um gato ou um cão, pelo que o valor de saída ou é 1 ou 0 (C. Zhang e Lu 2021).

Já os algoritmos de regressão analisam problemas em que os valores de saída são valores contínuos. Um exemplo disto seria analisar o mercado imobiliário e fazer a previsão do preço de uma casa (Sah 2020).

Tendo em conta esta informação, é possível compreender que um dos problemas para este tipo de modelos é a necessidade de existência de uma grande quantidade de dados rotulados (Sah 2020). Uma vez que os dados fornecidos para este estudo não são rotulados, não é possível utilizar modelos deste tipo.

2.4.2 Aprendizagem Não Supervisionada

Aprendizagem Não Supervisionada aplica-se quando os dados a serem analisados, apenas existem na forma de dados de entrada, não existindo o valor de saída correspondente, ou seja, é responsabilidade da máquina determinar correlações e relações entre os dados disponíveis (Wakefield 2023; C. Zhang e Lu 2021).

Neste tipo de aprendizagem não existe nenhum operador e os algoritmos aprendem a partir da introdução de novos dados. Assim, quando novos dados são introduzidos, o algoritmo utiliza o seu conhecimento prévio para tentar reconhecer alguma estrutura ou classe nos dados (Mahesh 2020).

Os modelos de aprendizagem não supervisionada são amplamente utilizados em duas tarefas principais: **clustering** e **redução dimensional**.

Os algoritmos de *Clustering* analisam problemas em que há necessidade de agrupar dados com base em características semelhantes. Um exemplo de uma tarefa em que *clustering* é utilizado é a segmentação de clientes com base no seu comportamento de compras (Sah 2020). Contudo, é importante mencionar que esta técnica é utilizada em diversas áreas, desde *marketing* à biologia computacional ou medicina.

Já a redução dimensional envolve a simplificação dos dados ao reduzir o número de variáveis ou características, mantendo o máximo possível da informação relevante. Este processo é particularmente útil em conjuntos de dados de alta dimensionalidade, como imagens, dados genómicos ou séries temporais, ajudando a melhorar a eficiência dos algoritmos subsequentes (Badillo et al. 2020).

A principal vantagem dos modelos não supervisionados é, portanto, a sua capacidade de analisar grandes volumes de dados não rotulados, o que os torna úteis em situações onde seria inviável ou caro obter rótulos para todos os dados disponíveis. Contudo, este tipo de aprendizagem apresenta desafios, como a dificuldade de interpretar os resultados e a ausência de métricas diretas para avaliar a performance do modelo (Karsh 2023).

Tendo em conta que os dados fornecidos neste estudo não possuem rótulos, os modelos de aprendizagem não supervisionada tornam-se uma abordagem promissora para a análise inicial dos dados.

2.4.3 Aprendizagem Semi-Supervisionada

Aprendizagem Semi-Supervisionada, tal como o nome indica, pode ser definida por uma mistura entre aprendizagem supervisionada e não supervisionada. Assim, o modelo é ensinado com uma mistura de dados rotulados e dados não rotulados, sendo que, habitualmente, a quantidade de dados não rotulados é bastante superior ao número de dados rotulados (C. Zhang e Lu 2021).

Este tipo de aprendizagem tem como objetivo ultrapassar os principais problemas quer da aprendizagem supervisionada - por exemplo, o custo elevado da classificação de dados - quer da aprendizagem não supervisionada - dificuldade de interpretação dos dados. No entanto, este tipo de aprendizagem também apresenta as suas limitações, sendo que estudos mostram que este tipo de aprendizagem não escala bem, e visto que a quantidade de dados não rotulados é normalmente mais elevada, a tarefa de análise dos dados continua a ser desafiadora (Y. C. A. P. Reddy, Viswanath e B. E. Reddy 2018).

Deste modo, e apesar de ser uma ideia atrativa, a aprendizagem semi-supervisionada continua a não ser usada em aplicações práticas (Y. C. A. P. Reddy, Viswanath e B. E. Reddy 2018; C. Zhang e Lu 2021).

Assim sendo, e tendo em conta que este estudo é baseado em dados não rotulados, é possível concluir que este tipo de modelos não é aplicável ao problema em questão.

2.4.4 Aprendizagem por Reforço

O último tipo de aprendizagem abordado nesta dissertação é a de aprendizagem por reforço. Este é um método de aprendizagem em que o modelo é desenvolvido a partir de um sistema de recompensas.

Assim, diferentemente da aprendizagem não supervisionada ou da aprendizagem supervisionada, este tipo de aprendizagem não depende de um conjunto fixo de dados de treino. Em

vez disso, o modelo aprende ao interagir com o ambiente, no qual toma decisões e recebe *feedback* na forma de recompensas ou penalizações (C. Zhang e Lu 2021). Deste modo, os sistemas de aprendizagem por reforço consistem em três componentes principais (Y. Li 2022):

- **Agente:** O modelo ou algoritmo que toma decisões e aprende com o ambiente.
- **Ambiente:** O sistema ou contexto com o qual o agente interage, fornecendo *feedback* em forma de estados e recompensas.
- **Recompensa:** O sinal de *feedback* recebido pelo agente após cada ação, indicando se a ação tomada foi boa ou má.

Deste modo, neste tipo de modelos, o agente interage com o ambiente ao longo do tempo e tenta aprender, a partir de uma estratégia de tentativa e erro, quais são as políticas que maximizam a recompensa.

A aprendizagem por reforço é particularmente poderosa em problemas onde as ações têm consequências a longo prazo e as recompensas não são imediatas. Alguns dos cenários em que este tipo de aprendizagem é aplicado são os modelos de linguagem natural, jogos ou sistemas de recomendação.

Contudo, também apresenta desafios significativos, como o facto de que alguns problemas requerem simulações complexas e demoradas para que o agente aprenda eficientemente, e é difícil equilibrar a necessidade de o agente explorar novas ações e utilizar as estratégias já conhecidas para maximizar a recompensa. Por fim, o treino pode apresentar uma instabilidade elevada quando os ambientes são complexos ou são ambientes com recompensas esparsas (Y. Li 2022; A. Wong et al. 2023).

Tendo em conta que para o problema em questão não existe o conceito de recompensa ou ambiente interativo, pode-se concluir que este tipo de aprendizagem não é o indicado para o problema a ser analisado nesta dissertação.

2.5 Algoritmos de Aprendizagem não supervisionada

Tal como visto em 2.4, o tipo de aprendizagem indicado para a tarefa de analisar os perfis metabólicos de aminoácidos e acilcarnitinas de modo a identificar grupos e *outliers* é a aprendizagem não supervisionada. Como tal, ao longo desta secção são apresentados alguns dos principais algoritmos de aprendizagem não supervisionada, sendo estes divididos em métodos para *clustering* e métodos para detecção de anomalias.

2.5.1 Métodos de Deteção de Outliers

Detetar *outliers* é uma componente crítica de aprendizagem não supervisionada. Ao contrário de métodos supervisionados, onde há rótulos claros que ajudam na distinção entre dados normais ou anómalos, em métodos de aprendizagem não supervisionada, a identificação de *outliers* é feita a partir da análise de padrões intrínsecos nos dados analisados.

Nesta secção, introduzem-se alguns dos principais tipos de métodos de identificação de anomalias, classificados em métodos baseados em estatística, densidade, proximidade, *clustering* (Smiti 2020).

Métodos Baseados em Estatística

Métodos de deteção de *outliers* estatísticos, também conhecidos por métodos de deteção baseados em distribuição, utilizam propriedades estatísticas dos dados para identificar observações que se desviam significativamente da distribuição geral. Em geral, estes métodos assumem que os dados seguem uma distribuição específica, como a normal, e classificam como *outliers* os pontos que estão em regiões de baixa probabilidade dessa distribuição (Smiti 2020).

Apesar da simplicidade destes métodos estatísticos, estes apresentam algumas limitações, sendo a principal o facto de que não são aplicáveis quando a distribuição não é conhecida (Smiti 2020).

Métodos Baseados em Distância

Os métodos baseados em distância são métodos em que a existência de *outliers* é detetada a partir do cálculo das distâncias entre os diferentes dados e considerando diversas métricas de distância. Abordagens de "vizinho mais próximo" são normalmente as mais usadas (Smiti 2020).

Métodos baseados em distância são utilizados em diversas áreas atualmente, entre as quais se podem destacar, por exemplo, a área da saúde, de reconhecimento de padrões ou das finanças.

Relativamente às suas principais vantagens, os métodos baseados em métricas de distância têm como principais características o facto de (Wang, Bah e Hammad 2019):

- Não assumirem uma distribuição específica dos dados.
- Serem mais eficazes em dados multidimensionais que métodos estatísticos.

No entanto, apresentam algumas limitações (Smiti 2020; Wang, Bah e Hammad 2019):

- Alto custo computacional, especialmente em grandes conjuntos de dados.
- Podem ter dificuldades com dados de alta dimensionalidade devido à "maldição da dimensionalidade"[1].
- A maioria dos métodos baseados em distância existentes têm dificuldades em lidar com fluxos de dados (*data streams*), uma vez que é complicado calcular distâncias para dados em fluxo.

Métodos Baseados em Densidade

Outro tipo de métodos de identificação de *outliers* são os métodos baseados em densidade. Nestes, a presença de um *outlier* é detetada quando a sua densidade local é diferente da dos seus vizinhos. Este método segue a premissa de que regiões de baixa densidade são mais propensas a conter *outliers*, enquanto pontos normais tendem a residir em regiões de alta densidade (J. Zhang e Y. Yang 2023).

Neste sentido, algumas das principais vantagens deste tipo de métodos, quando comparados com, por exemplo, métodos estatísticos ou baseados em distância, são que apresentam maior

[1] **Maldição de Dimensionalidade:** Termo introduzido por Bellman para descrever o problema causado pelo aumento exponencial no volume associado à adição de dimensões extras ao espaço euclidiano (Eamonn e Mueen 2017).

performance, identificando *outliers* que não tinham sido identificados por outros modelos (Wang, Bah e Hammad 2019).

No entanto, este tipo de método torna-se bastante dispendioso em termos computacionais rapidamente, pelo que, mais uma vez, não é ideal para grandes conjuntos de dados e apresenta problemas a lidar eficientemente com *data streams*. Mais ainda, estes métodos são extremamente sensíveis aos parâmetros de configuração com o número de vizinhos a considerar e podem apresentar piores resultados em dados com regiões de densidade variáveis. (Smiti 2020; Wang, Bah e Hammad 2019).

Métodos Baseados em Clustering

A última categoria de métodos detalhada nesta secção é a de métodos baseados em *clustering*. Este tipo de técnicas depende do processo de identificar diferentes *clusters* e os objetos/dados que não pertencerem a nenhum *cluster* são considerados *outliers* (Smiti 2020).

Métodos deste tipo têm a vantagem de se tratar de métodos não supervisionados, o que os torna bastante úteis na identificação de *outliers* em *data streams*. Isto deve-se a que, após o processo de aprendizagem, novos dados podem ser introduzidos e testados, o que torna este modelo incremental e bastante adaptável (Wang, Bah e Hammad 2019).

No entanto, tal como todos os métodos, métodos de deteção de *outliers* baseados em *clustering* também apresentam algumas limitações. Em primeiro lugar, de uma forma geral, estes métodos são bastante sensíveis aos parâmetros de inicialização, como o número de *clusters*. Adicionalmente, e visto que a maioria das técnicas de *clustering* necessita de bastantes parâmetros, é bastante complicado escolhê-los corretamente (Smiti 2020).

2.5.2 Algoritmos de Clustering

Como mencionado na Secção 2.4.2, os métodos de *clustering* têm como principal objetivo agrupar dados com base em características semelhantes. No contexto deste projeto, esses métodos têm como objetivo identificar padrões específicos associados a diferentes estados da doença e ao resultado do tratamento neoadjuvante.

Neste sentido, é importante analisar alguns dos principais tipos e métodos de *clustering*, sendo essencial realçar que o objetivo desta secção é introduzir o leitor a técnicas de *clustering* utilizadas em Aprendizagem Não Supervisionada. Portanto, nesta secção, são apenas introduzidos alguns algoritmos, mas outros podem vir a ser utilizados no desenvolvimento deste projeto.

Exclusive Clustering

Clustering exclusivo é uma forma de agrupar dados em que estes podem apenas pertencer a exatamente um *cluster*. Já a sobreposição de *cluster* é uma estratégia em que dados podem ser atribuídos a múltiplos grupos. Assim sendo, um exemplo de *exclusive clustering* é a definição de grupos com base em idade ou sexo (Sozuer 2015).

Overlapping Clustering

A sobreposição de *cluster* é uma estratégia em que dados podem ser atribuídos a múltiplos *clusters*. Um exemplo de *overlapping clustering* é agrupar doente tendo em conta a categoria

de doenças - uma vez que cada pessoa pode ter várias doenças simultaneamente (Sozuer 2015).

Hierarchical Clustering

O *Hierarchical Clustering* é uma estratégia de *clustering* em que os dados são organizados numa estrutura hierárquica. Esta estrutura é construída com base em relações de similaridade entre os dados, o que permite que *clusters* sejam analisados com diferentes níveis de granularidade (Z. Zhang et al. 2017). Adicionalmente, e ao contrário de métodos como o *K-Means*, este tipo de *clustering* não necessita que o número de *clusters* seja predefinido, o que torna este algoritmo uma ferramenta versátil para análise em que o número de *clusters* é desconhecido.

O *Hierachical Clustering* pode adicionalmente ser classificado de duas formas (Noble 2024):

- **Aglomerativo (Bottom-Up):** Começa com cada dado como um *cluster* individual e, em cada iteração, combina os dois *clusters* mais semelhantes, terminando quando todos os dados estão agrupados num único *cluster*.
- **Divisivo (Top-Down):** Começa com todos os dados em um único *cluster* e, iterativamente, divide-os em grupos menores com base nas suas diferenças.

Este tipo de método destaca-se por ser prático, eficiente, facilmente interpretado e, tal como já mencionado, destaca-se por o número de *clusters* não ter de ser predefinido (T. Yang, Ren e Zhou 2018). Deste modo, este método já foi usado em diversas áreas, desde análise de redes à pesquisa clínica e bioinformática.

No entanto, este tipo de modelo também apresenta limitações. Algumas das limitações do *Hierarchical Clustering* são a sensibilidade a ruídos e *outliers*, a rigidez do mecanismo (não é possível corrigir posteriormente decisões erradas) e a má performance quando a análise está a ser feita para conjuntos de dados grandes (Papakyriakou e Barbounakis 2022).

Probabilistic Clustering

O *Probablistic Clustering* é uma forma de agrupar dados baseada na probabilidade de cada dado pertencer a diferentes *clusters*. Esta abordagem é amplamente utilizada em situações onde a incerteza é uma característica intrínseca dos dados (Carrasco 2024).

2.6 Tecnologias para Desenvolvimento de Machine Learning

Atualmente, o desenvolvimento de soluções de *Machine Learning* requer um conjunto robusto de ferramentas e tecnologias que permitam a implementação de algoritmos complexos, processamento eficiente de dados e visualização dos resultados. Nesta secção são apresentadas algumas destas ferramentas e é feita uma breve comparação entre as mesmas.

2.6.1 Scikit-learn

SciKit-learn é uma biblioteca [2] para processamento de dados em *Python* que implementa diversos métodos de classificação, *clustering*, de regressão e outros algoritmos de *Machine Learning* (Gevorkyan et al. 2019).

[2] **Biblioteca:** no contexto de *software*, bibliotecas são coleções de componentes de *software* que fornecem funcionalidades específicas (Sterling, Anderson e Brodowicz 2018)

Esta é uma biblioteca de código aberto que é bastante popular pela sua *API* intuitiva e documentação detalhada. Entre os algoritmos implementados, destacam-se alguns algoritmos de *clustering* como o *K-Means*, o *Agglomerative clustering* ou o *HDBSCAN*, e algoritmos de deteção de anomalias como o *LOF* ou o *Isolation Forest*. Além disso, a biblioteca oferece suporte a operações importantes, como pré-processamento de dados (normalização, codificação categórica), seleção de atributos e otimização de parâmetros (scikitLearn 2024).

Adicionalmente, a biblioteca é compatível com outras ferramentas populares do ecossistema *Python*, como *NumPy*, *pandas* e *matplotlib*, o que amplia as suas funcionalidades.

No entanto, esta biblioteca também apresenta algumas limitações em que se destacam o facto de não ser ideal para projetos de *Deep Learning* , poder ter baixa eficiência e ser lenta em projetos de larga escala ou com grandes volumes de dados (Cantu 2023).

2.6.2 MLlib

Outra biblioteca que foi considerada para o desenvolvimento deste projeto foi a *MLlib*, que é uma biblioteca do ecossistema *Apache Spark* e que foi desenvolvida como uma solução para *Machine Learning* em ambientes de computação distribuída (Reichert 2023).

Tal como o *Scikit-learn*, esta biblioteca disponibiliza a implementação de algoritmos como o *K-Means* e o *GMM*, mas esta biblioteca tem como principal objetivo possibilitar o tratamento de dados de grande escala e ultrapassar limitações de memória e processamento de bibliotecas como *Scikit-learn* (Abdulwahid 2025; Spark 2024).

No entanto, a utilização do *MLlib* apresenta maior complexidade e requer maior conhecimento, uma vez que exige conhecimento de estruturas específicas do *Apache Spark*, e domínio de conceitos de paralelização de dados e configurações de ambientes distribuídos, pelo que esta não é a ferramenta ideal para projetos de pequenas dimensões e com conjuntos de dados de pequena a média dimensão (Abdulwahid 2025).

WEKA

O *WEKA* é um *software* de código aberto utilizado para mineração de dados e *machine learning*, bastante reconhecido na área académica devido à sua flexibilidade e facilidade de uso (GeeksForGeeks 2025).

Este *software* apresenta uma coleção extensa de algoritmos de aprendizagem supervisionada e não supervisionada, onde se incluem métodos de *clustering*, como *K-means* e o *Hierarchical Clustering* (GeeksForGeeks 2025).

Adicionalmente, uma das principais vantagens do *Weka* é que apresenta uma interface gráfica que permite aplicar algoritmos de forma direta sem necessidade de programação direta, mas possui ainda uma API em Java para utilizadores mais avançados (Ballesteros 2023; GeeksForGeeks 2025).

No entanto, e como qualquer ferramenta, o *Weka* também apresenta limitações, tais como o facto de não ser ideal para conjuntos de dados de grandes dimensões e apresentar pior desempenho computacional que soluções mais modernas como *Scikit-learn* e o *MLlib* (GeeksForGeeks 2025).

2.6.3 Comparação entre as diferentes bibliotecas

Nesta secção é feita uma breve reflexão acerca das ferramentas apresentadas anteriormente. Assim sendo, na Tabela 2.1 é feita uma síntese das principais características mencionadas.

Critério	**Scikit-learn**	**MLlib**	**WEKA**
Propósito	ML	Big Data & ML	ML
Facilidade	Alta	Média	Alta
Desempenho	CPU	Distribuído	CPU
Curva de Aprendizagem	Baixa	Alta	Baixa
Complexidade	Média	Alta	Baixa
Escala ideal	Pequeno/Médio	Muito grande	Pequeno/Médio

Tabela 2.1: Comparação entre *Scikit-learn*, *MLlib* e *WEKA*.

A análise comparativa apresentada na Tabela 2.1 evidencia que as três bibliotecas possuem perfis distintos de utilização.

Assim sendo, o *Scikit-learn* destaca-se pela simplicidade, documentação abrangente e curva de aprendizagem reduzida, o que o torna ideal para problemas de pequena a média escala, como é o caso do presente estudo.

Por outro lado, o *MLlib (Spark)* oferece maior escalabilidade e é mais apropriado para contextos de *Big Data*, ainda que à custa de uma maior complexidade de implementação.

Por fim, o *WEKA*, apresenta a vantagem de uma interface gráfica intuitiva, mas revela limitações de desempenho quando comparado com soluções modernas baseadas em *Python*.

Assim, considerando a dimensão dos dados utilizados e a necessidade de um equilíbrio entre facilidade de uso e desempenho computacional, optou-se pelo *Scikit-learn*, uma vez que se trata de uma opção mais adequada para as tarefas deste projeto.

2.7 Estudos de Aplicação de ML em Medicina

Tal como mencionado, no Capítulo 1 esta dissertação tem como objetivo analisar perfis metabólicos de aminoácidos e acilcarnitinas em doentes com cancro do reto, de forma a associar estes perfis a uma resposta patológica e, assim, tentar identificar anomalias e *clusters* que permitam melhorar a predição da resposta ao tratamento neoadjuvante. O uso de Inteligência Artificial tem tido um crescimento em diversas áreas e, em particular, na área da Saúde, onde a sua utilização permite reduzir a carga de trabalho dos profissionais de saúde e melhorar a precisão do diagnóstico (Aung, D. C. Wong e Ting 2021).

Deste modo, é importante fazer uma revisão de alguns dos trabalhos que foram feitos na área da saúde e, mais especificamente, para a doença do cancro colorretal, e que utilizam Inteligência Artificial. Neste sentido, na Tabela 2.2 é apresentada uma sintetização de alguns estudos, destacando a área de aplicação e o tipo de aprendizagem e algoritmos usados.

Publicação	Área de Investigação	Alguns dos Algoritmos Usados
Smyth et al. 2024	Estratificação e Análise de Sobrevivência de doentes de Cancro Colorretal	*K-Means*, *Hierarchical Clustering* *C-Means* *Pam Clustering*
Liu et al. 2024	Caracterização do Microambiente Imune e Prognóstico em doentes com Cancro Colorretal	*LASSO Regression*, *SVM-RFE*, *Stepwise Cox Regression*
Bychkov et al. 2018	Previsão de Sobrevivência de doentes com Cancro Colorretal Usando Redes Neuronais	*VGG-16* *LSTM (Long Short-Term Memory)* *Logistic Regression*, *Naïve Bayes*, *SVM*
Kather et al. 2019	Prognóstico de Sobrevivência em doentes com Cancro Colorretal a partir de Lâminas Histológicas	*VGG19*, *ResNet50*, *AlexNet*, *GoogLeNet*, *SqueezeNet*
Chen et al. 2023	Seleção de Características para Previsão de Resposta à Quimiorradiação Neoadjuvante em Cancro do Reto Avançado Localmente	*K-Means* *Random Forest* *Silhouette Coefficient*

Tabela 2.2: Diferentes Publicações de Aplicação de *Machine Learning* em Saúde, e mais especificamente, Cancro Colorretal

Assim, através da análise dos estudos apresentados na Tabela 2.2 é possível observar uma diversidade significativa nas abordagens de *Machine Learning* aplicadas ao cancro colorretal. Estes trabalhos demonstram uma variedade não só em termos de modelos de aprendizagem diferentes, como de áreas de investigação - agrupamento de doentes, previsões de sobrevivência ou resposta ao tratamento. Dependendo do objetivo do estudo, cada tipo de aprendizagem pode demonstrar vantagens específicas. Em particular, nota-se que:

- **Estratificação e agrupamento de doentes:** Estudos como Smyth et al. 2024 e Chen et al. 2023 utilizam algoritmos de aprendizagem não supervisionada, como *K-Means* e *Pam Clustering*, para identificar padrões e *clusters* de doentes. Estas abordagens são relevantes para esta dissertação, dado o foco na identificação de padrões metabólicos que podem prever respostas ao tratamento.
- **Prognóstico e previsão de sobrevivência:** Métodos supervisionados e de *Deep Learning* são predominantes nesta área, como observado em Liu et al. 2024 e Bychkov et al. 2018. Modelos como *LSTM* e *ResNet* destacam-se pela capacidade de lidar com dados complexos, como imagens e séries temporais.

- **Caracterização de microambientes tumorais:** Estudos como Liu et al. 2024 ilustram a aplicabilidade de algoritmos como *LASSO Regression* na identificação de biomarcadores relevantes.

Como tal, estes trabalhos confirmam o potencial e viabilidade do uso de técnicas de *Machine Learning* na análise de dados clínicos relacionados com o cancro colorretal.

2.8 Conclusões

Nesta secção são apresentadas algumas das conclusões que se podem efetuar, tendo em conta a informação apresentada neste capítulo de Revisão Literária.

Assim, primeiro de tudo, esta investigação permitiu obter uma melhor compreensão acerca do cancro colorretal. Entre estes dados, é importante destacar a prevalência global do cancro colorretal e a variabilidade das respostas ao tratamento, o que dificulta a aplicação do tratamento ideal a cada paciente (Roeder et al. 2020; Xi e P. Xu 2021). Estes dados realçam a relevância da realização de estudos como o desta dissertação.

Adicionalmente, a investigação feita ao longo deste capítulo permitiu concluir que o estudo de perfis metabólicos e o uso de ML em áreas de saúde como a resposta ao tratamento do cancro colorretal têm vindo a ganhar relevância nos últimos anos (Badillo et al. 2020; Tian et al. 2018). Mais ainda, estes estudos e outros estudos mencionados ao longo deste capítulo destacam a variabilidade de métodos de aprendizagem que podem ser utilizados dependendo da área de investigação e dos objetivos do estudo.

No contexto deste trabalho, é pretendida a exploração de métodos de aprendizagem não supervisionada e, em particular, de métodos de *clustering* e deteção de *outliers* de modo a identificar padrões em perfis metabólicos que possam prever a resposta de um paciente ao tratamento neoadjuvante. Neste sentido, estudos como Smyth et al. 2024 e Chen et al. 2023 demonstram a possibilidade de uso de diversos algoritmos como o *K-Means*, o *C-Means* e o *Pam Clustering* para a identificação de padrões e grupos de doentes com cancro colorretal.

Por fim, definiu-se a biblioteca *Scikit-Learn* como a principal ferramenta a ser utilizada no desenvolvimento deste trabalho. Esta escolha deve-se à sua vasta gama de algoritmos de *clustering* e deteção de anomalias, além de oferecer uma documentação abrangente e uma curva de aprendizagem acessível, fatores que facilitam a implementação no contexto desta dissertação.

Capítulo 3

Análise e Preparação dos Dados

Neste capítulo apresentam-se os dados utilizados na realização das tarefas de *clustering* desenvolvidas ao longo deste projeto.

Numa primeira fase, descreve-se o *dataset* principal, bem como os *subsets* construídos com o objetivo de apoiar a análise exploratória de dados. Posteriormente, são detalhadas as etapas de pré-processamento realizadas, que incluíram a deteção de anomalias, a redução da dimensionalidade dos perfis metabólicos e a seleção das *features* relevantes para a aplicação das técnicas de *clustering*.

3.1 Dataset

Para uma melhor compreensão do conjunto de dados, esta secção apresenta a estrutura dos ficheiros fornecidos. Estes dados foram fornecidos pela Unidade Local de Saúde de Santo António e incluem informações de diagnóstico em duas áreas distintas: doenças hereditárias do metabolismo, provenientes do Serviço de Genética Laboratorial, e cancro do reto localmente avançado, da Unidade de Cirurgia Colorretal. O segundo conjunto de dados foi recolhido no âmbito do projeto de doutoramento do Dr. Pedro Brandão, com orientação dos Professores Doutores Marisa Santos, António Araújo e Lúcia Lacerda.

Cada linha do ficheiro corresponde a uma amostra, identificada por um código único (*SAMPLE_ID*), e contém variáveis demográficas como o sexo (codificado como 0 para masculino e 1 para feminino) e a idade do doente. Adicionalmente, são fornecidas informações clínicas sobre o diagnóstico. O *dataset* inclui amostras de indivíduos com diferentes patologias e, no caso dos doentes com cancro colorretal (CRC), estão ainda disponíveis dados sobre o estádio clínico - pré-quimiorradioterapia (M0), pós-quimiorradioterapia (M1), pós-cirurgia (M2), vigilância (M3) ou recidiva (M4).

Cada amostra integra ainda medidas metabólicas resultantes de análises laboratoriais, organizadas em dois grupos principais:

- **Aminoácidos:** concentrações de aminoácidos proteinogénicos e não proteinogénicos (*GLU, GLN, ASP, SER, THR, VAL, ILE, LEU, LYS, PHE, TYR, ARG*, entre outros), bem como derivados como *TAU, HYP, ORN, CYSTA, AAA*, e ainda biomarcadores específicos como *1MHIS, 3MHIS*.
- **Acilcarnitinas:** medidas associadas à carnitina total e aos seus derivados acilados (*C2, C3, C4, C5, C16, C18*, entre outros), incluindo variantes dicarboxílicas como *C3DC, C4DC, C5DC*. Estas variáveis encontram-se relacionadas com o metabolismo energético e lipídico.

Na Figura 3.1 apresenta-se um exemplo parcial dos dados disponibilizados.

N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL
2	C14	C16	C18	C3DC	C40H	C4DC	C5:1	C50H	C6DC	C8:1	C10:1	C10:2	C12:1	C14:1	C14:1OH	C14:2	C140H	C16:1	C160H	C18:1	C18:1OH	C18:2	C180H	DIAG_ID
047	0.146	1.783	1.034	0.404	0.032	0.027	0.0	1.809	0.039	0.059	0.039	0.036	0.094	0.186	0.0	0.0	0.017	0.173	0.021	2.083	0.031	0.435	0.018	CRC
021	0.075	1.196	0.409	0.07	0.0	0.436	0.0	0.276	0.0	0.0	0.069	0.0	0.214	0.081	0.011	0.015	0.009	0.062	0.014	1.599	0.0	0.331	0.0	CRC
015	0.067	1.059	0.587	0.065	0.0	2.502	0.0	1.867	0.134	0.063	0.119	0.125	0.129	0.113	0.019	0.101	0.017	0.067	0.025	1.059	0.027	0.258	0.0	CRC
023	0.063	1.04	0.571	0.0	0.0	2.379	0.0	1.772	0.035	0.047	0.202	0.207	0.17	0.134	0.016	0.106	0.014	0.077	0.02	1.313	0.02	0.264	0.015	CRC
0	0.064	0.525	0.248	0.0	0.0	0.058	0.0	0.0	0.0	0.061	0.104	0.0	0.22	0.102	0.0	0.024	0.0	0.044	0.014	0.732	0.006	0.257	0.006	CRC
023	0.0	0.92	0.424	0.025	0.0	0.0	0.0	0.169	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.006	0.009	0.074	0.017	1.05	0.013	0.223	0.0	CRC
075	0.045	0.845	0.39	0.0	0.0	0.676	0.0	0.42	0.099	0.0	0.298	0.0	0.192	0.137	0.0	0.04	0.0	0.071	0.007	1.04	0.01	0.369	0.0	CRC
013	0.045	0.593	0.456	0.0	0.0	0.388	0.0	0.164	0.0	0.0	0.0	0.0	0.121	0.04	0.005	0.01	0.003	0.029	0.007	0.913	0.0	0.361	0.005	CRC
014	0.043	0.591	0.278	0.0	0.0	0.0	0.0	0.111	0.0	0.036	0.0	0.0	0.019	0.028	0.0	0.003	0.0	0.047	0.0	0.896	0.009	0.359	0.006	CRC

Figura 3.1: Exemplo parcial do conjunto de dados

De forma a caracterizar detalhadamente as variáveis analisadas, as *features* metabólicas de aminoácidos encontram-se descritas na Tabela 3.1.

Tabela 3.1: Caracterização das *Features* Metabólicas do *Dataset*

Abreviatura	**Nome Completo**	**Unidade**
TAU	Taurina	μmol/L
ASP	Ácido Aspártico	μmol/L
HYP	Hidroxiprolina	μmol/L
THR	Treonina	μmol/L
SER	Serina	μmol/L
ASN	Asparagina	μmol/L
GLU	Ácido Glutâmico	μmol/L
GLN	Glutamina	μmol/L
AAA	Ácido α-Aminoadípico	μmol/L
PRO	Prolina	μmol/L
GLY	Glicina	μmol/L
ALA	Alanina	μmol/L
CIT	Citrulina	μmol/L
ABU	Ácido α-Aminobutírico	μmol/L
VAL	Valina	μmol/L
CYS2	Cistina	μmol/L
MET	Metionina	μmol/L
CYSTA	Cistationina	μmol/L
ILE	Isoleucina	μmol/L
LEU	Leucina	μmol/L
TYR	Tirosina	μmol/L
PHE	Fenilalanina	μmol/L
ORN	Ornitina	μmol/L
LYS	Lisina	μmol/L
1MHIS	1-Metil-Histidina	μmol/L
HIS	Histidina	μmol/L
3MHIS	3-Metil-Histidina	μmol/L
ARG	Arginina	μmol/L

Os dados de aminoácidos apresentam ainda os diagnósticos listados na Tabela 3.2.

Tabela 3.2: Diagnósticos Presentes no *Dataset* de Aminoácidos

Código	**Diagnóstico**
0	WD (Sem Diagnóstico)
1	Acidúria 3-Hidroxi-3-metilglutárica
2	Acidúria Argininossuccínica
3	Acidúria Glutárica Tipo I
4	Acidúria Glutárica Tipo II
5	Acidúria Metilmalónica por Défice da Mutase
6	Défice do Metabolismo Intracelular da Vitamina B12
7	Acidúria Propiónica
8	Alcaptonúria
9	Argininemia
10	Cistinúria - heterozigotia tipo B
11	Défice de Piruvato Desidrogenase
12	Citrulinemia Tipo I
13	Citrulinemia Tipo II
14	Défice de CPS
15	Défice de Metionina Adenosil Transferase
16	Défice de OTC
17	Défice em GAMT
18	Défice em MCAD
19	Hiperfenilalaninemia
21	Galactosemia
22	Glicogenose Tipo Ia
23	Hiperargininemia
25	Hiperglicinemia sem Cetose
26	Hipermetioninemia
27	Hiperornitinemia com Atrofia Girata
28	Hipofosfatásia
29	Homocistinúria Clássica
30	Iminodipeptiduria/Défice Prolidase
31	Intolerância à Frutose
32	Leucinose
35	Tirosinemia Tipo I
36	Tirosinemia Tipo II
37	3-Metilcrotonilglicinúria
40	Síndrome Dorfman-Chanarin
42	Défice VLCAD
44	CLN variante 3
48	Défice de Glicerol Quinase
49	GM2/Sandhoff
50	Glicogenose Tipo V
51	Síndrome HHH
52	Hiperprolinemia
54	Niemann-Pick Tipo B
55	Niemann-Pick Tipo C

Continua na próxima página...

Tabela 3.2 (continuação)

Código	Diagnóstico
57	Tirosinemia Tipo III
99	CRC (Cancro Colorretal)

Já as *features* metabólicas de acilcarnitinas encontram-se descritas na Tabela 3.3.

Tabela 3.3: Caracterização das *Features* de Acilcarnitinas do *Dataset*

Abreviatura	Nome Completo	Unidade
CARNITINE	Carnitina (livre)	μmol/L
C2	Acetilcarnitina	μmol/L
C3	Propionilcarnitina	μmol/L
C4	Butirilcarnitina	μmol/L
C5	Valerilcarnitina	μmol/L
C5DC	Glutarilcarnitina	μmol/L
C6	Hexanoilcarnitina	μmol/L
C8	Octanoilcarnitina	μmol/L
C10	Decanoilcarnitina	μmol/L
C12	Dodecanoilcarnitina	μmol/L
C14	Tetradecanoilcarnitina	μmol/L
C16	Palmitoilcarnitina	μmol/L
C18	Estearoilcarnitina	μmol/L
C3DC	Malonilcarnitina	μmol/L
C4OH	3-Hidroxibutirilcarnitina	μmol/L
C4DC	Succinilcarnitina	μmol/L
C5:1	Tiglilcarnitina	μmol/L
C5OH	3-Hidroxivalerilcarnitina	μmol/L
C6DC	Adipilcarnitina	μmol/L
C8:1	Octenoilcarnitina	μmol/L
C10:1	Decenoilcarnitina	μmol/L
C10:2	Decadienoilcarnitina	μmol/L
C12:1	Dodecenoilcarnitina	μmol/L
C14:1	Tetradecenoilcarnitina	μmol/L
C14:1OH	3-Hidroxitetradecenoilcarnitina	μmol/L
C14:2	Tetradecadienoilcarnitina	μmol/L
C14OH	3-Hidroxitetradecanoilcarnitina	μmol/L
C16:1	Palmitoleílcarnitina	μmol/L
C16OH	3-Hidroxipalmitoilcarnitina	μmol/L
C18:1	Oleílcarnitina	μmol/L
C18:1OH	3-Hidroxioleílcarnitina	μmol/L
C18:2	Linoleílcarnitina	μmol/L
C18OH	3-Hidroxiestearoilcarnitina	μmol/L

Os diagnósticos correspondentes ao conjunto de dados de acilcarnitinas estão listados na Tabela 3.4.

Tabela 3.4: Diagnósticos Presentes no *Dataset* de Acilcarnitinas

Código	Diagnóstico
WD	Sem Diagnóstico
CRC	Cancro Colorretal
MCAD	Deficiência da Acil-CoA Desidrogenase de Cadeia Média
AG1	Acidúria Glutárica Tipo I
LCHAD	Deficiência da 3-Hidroxiacil-CoA Desidrogenase de Cadeia Longa
AG2	Acidúria Glutárica Tipo II
VLCAD	Deficiência da Acil-CoA Desidrogenase de Cadeia Muito Longa
AMM	Acidúria Metilmalónica
3-HMG	Acidúria 3-Hidroxi-3-Metilglutárica
3-MCC	Deficiência da 3-Metilcrotonil-CoA Carboxilase
AMM_MUT	Acidúria Metilmalónica por Deficiência da Mutase
OTC	Deficiência da Ornitina Transcarbamilase
CUD	Deficiência da Captação de Carnitina

Tendo em consideração a estrutura descrita, foram delineados três estudos distintos de *clustering*:

- Análise global, considerando a totalidade dos dados disponíveis.
- Análise restrita a doentes com cancro colorretal (CRC), selecionado a partir da filtragem da coluna `Diag_Id`.
- Análise focada em doentes no estádio M0, selecionados a partir da coluna `M_Treatment`.

Para garantir que o processo de agrupamento se baseasse exclusivamente nos perfis metabólicos, foram excluídas do conjunto de variáveis utilizadas para *clustering* as colunas de identificação da amostra (*SAMPLE_ID*), as variáveis demográficas (*SEX*, *AGE*) e os campos relativos ao diagnóstico e estádio clínico (*M_TREATMENT*, *DIAG_ID* e *DIAG_NUM*). Estas variáveis foram, no entanto, utilizadas posteriormente para avaliação e interpretação dos resultados obtidos.

3.2 Questões de Investigação

Na Seção 3.1 foram delineados três estudos a realizar a partir dos dados fornecidos pela Unidade Local de Saúde de Santo António. As questões de investigação foram definidas de modo a refletir os objetivos da investigação e a aplicação de métodos de *clustering* aos perfis de aminoácidos e acilcarnitinas.

O propósito destas questões é compreender os padrões de organização metabólica e avaliar a sua relevância na caracterização de doentes com cancro do reto. Assim, apresentam-se as seguintes questões de investigação:

- **Q1:** Qual a estrutura de agrupamento dos perfis metabólicos de aminoácidos e acilcarnitinas na totalidade da população analisada?
  - Pretende-se descrever a organização global dos dados e verificar se emergem agrupamentos naturais, independentemente do diagnóstico clínico no momento da recolha.

- **Q2:** De que modo os doentes com cancro colorretal (CRC) se distribuem em função dos seus perfis metabólicos quando considerados separadamente?
  - Procura-se identificar padrões metabólicos característicos do grupo CRC, distinguilos de outras patologias incluídas no *dataset* e explorar a eventual existência de subgrupos metabolicamente distintos.
- **Q3:** Que subgrupos metabólicos podem ser identificados entre os doentes em estádio M0, e que associação apresentam com a evolução clínica?
  - Esta questão visa compreender se os perfis metabólicos permitem identificar subgrupos prognósticos ou relacionados com diferentes respostas ao tratamento.

3.3 Preparação dos Dados

Com o objetivo de assegurar a qualidade e a fiabilidade dos resultados obtidos através dos métodos de *clustering*, tornou-se necessário proceder a uma adequada preparação dos dados (Maharana, Mondal e Nemade 2022). Para tal, antes da aplicação dos algoritmos de *clustering*, foi realizada uma etapa de pré-processamento que incluiu diferentes procedimentos, tais como a normalização dos dados, o tratamento de valores ausentes, a deteção de *outliers*, a seleção de características e a redução da dimensionalidade.

3.3.1 Normalização e Tratamento de Dados

Um passo fundamental do pré-processamento consistiu na normalização dos dados. Considerando que, conforme descrito na Seção 3.1, apenas os dados metabólicos foram utilizados no processo de *clustering*, apenas estes foram submetidos à normalização, dado que as demais variáveis já se encontravam normalizadas devido à existência de análises prévias realizadas com o mesmo conjunto de dados.

Para a normalização dos dados metabólicos, foi aplicada a transformação *Z-score*, implementada através do método *StandardScaler* da biblioteca *Scikit-learn* (Scikit-learn 2025). Esta técnica ajusta os dados de modo a apresentarem média nula e desvio padrão unitário, de acordo com a seguinte equação matemática:

$$z = \frac{x - \mu}{\sigma} \tag{3.1}$$

onde x corresponde ao valor original, μ à média e σ ao desvio padrão da variável.

Este procedimento assegura que todas as características metabólicas contribuem de forma equilibrada para os algoritmos de *clustering*, evitando que variáveis com diferentes escalas dominem o processo de análise.

Para além da normalização, foi também realizado o tratamento de valores ausentes e não finitos. No caso dos dados utilizados para a avaliação dos resultados, estes valores foram substituídos pela mediana, pelos seguintes motivos (Sparkl 2025):

- A mediana é menos sensível à presença de valores extremos, e portanto, é mais robusta do que a média.
- A substituição pela mediana preserva as propriedades centrais da distribuição de cada metabólito, evitando a introdução de vieses artificiais nos dados.

Adicionalmente, foi efetuada a correção de inconsistências na variável *M_ Treatment*, uma vez que algumas amostras apresentavam valores que não correspondiam às categorias esperadas (`M0`, `M1`, `M2`, `M3`, `M4`). Como tal, nesses casos, os dados foram ajustados para as categorias corretas. Por exemplo, uma das amostras apresentava o valor `M2?`, sendo que este foi devidamente convertido para `M2`, de modo a garantir que os dados utilizados na avaliação estavam corretos.

3.3.2 Seleção de features e Redução de dimensionalidade

Tal como referido no Capítulo 1, neste projeto são analisados perfis metabólicos de aminoácidos e acilcarnitinas. Estes dados apresentam elevada complexidade e multidimensionalidade, o que constitui um desafio adicional para algoritmos de *clustering* e de deteção de anomalias.

Com o objetivo de otimizar os resultados obtidos, a etapa de pré-processamento integrou técnicas de seleção de *features* e de redução de dimensionalidade. Estas técnicas permitem representar os dados com um número reduzido de dimensões, preservando, contudo, as propriedades essenciais do conjunto original. Para tal, eliminam-se atributos irrelevantes, redundantes ou ruidosos, resultando em modelos mais compactos e interpretáveis (Murel e Kavlakoglu 2024).

No contexto deste trabalho, estas abordagens revelam-se particularmente relevantes, uma vez que contribuem para a melhoria da interpretabilidade dos dados, aumentam a eficiência computacional dos algoritmos de *clustering* e reduzem o risco de sobreajuste (Obi 2019). Ainda assim, apresentam limitações, como a possibilidade de perda de informação e a complexidade associada à escolha dos parâmetros para os algoritmos de redução de dimensionalidade.

Relativamente à seleção de *features*, estas podem ser classificadas em três grandes categorias (Jha 2024):

- Métodos de filtro: avaliam a relevância das variáveis com base em medidas estatísticas, aplicando a seleção antes da utilização de algoritmos de aprendizagem;
- Métodos de envolvimento (*Wrapper*): recorrem a algoritmos de aprendizagem para medir o impacto de cada variável no desempenho do modelo, selecionando as mais relevantes;
- Métodos embutidos (*Embedded*): realizam a seleção das variáveis durante o próprio processo de treino, integrando a avaliação da sua relevância no ajuste do modelo.

No presente estudo, a seleção de *features* foi conduzida em duas etapas. Na primeira, aplicou-se um filtro de variância através do método `VarianceThreshold` da biblioteca `scikit-learn`, com limiar de 0.05. O valor de 0.05 foi determinado através de análise exploratória em que foi identificado que este era o ponto de equilíbrio entre a eliminação de ruído e a preservação de informação biologicamente relevante. Esta abordagem elimina variáveis de baixa variabilidade, dado que estas pouco contribuem para a diferenciação de grupos e para a deteção de padrões anómalos.

De seguida, calculou-se a matriz de correlação de *Pearson* entre todos os pares de variáveis. Sempre que duas *features* apresentavam correlação superior a 0.85, era removida a de menor variância, assumindo-se que a de maior variância contém informação mais discriminativa. O valor de 0.85 foi definido com base numa análise exploratória, sendo que observou-se que, a partir deste limiar, surgiam redundâncias significativas entre variáveis com funções

metabólicas semelhantes. Adicionalmente, garantiu-se que o conjunto final incluía pelo menos oito metabólitos, de forma a assegurar a robustez da análise.

Importa ainda referir que, com exceção do conjunto de perfis de acilcarnitinas de doentes com CRC no estádio M0 - onde as acilcarnitinas `C10` e `C40H` foram removidas por correlação excessiva -, nos restantes conjuntos de dados não foram eliminados metabólitos. Note-se que colunas relativas a metadados, dados demográficos e caracterização clínica (estádio e diagnóstico) já tinham sido previamente excluídas.

Após a seleção de características, foi aplicada uma técnica de redução de dimensionalidade com dois objetivos principais: facilitar a visualização dos resultados e reduzir o ruído computacional. De forma geral, estas técnicas podem ser divididas em dois grupos (Associate 2015):

- Técnicas lineares: transformam os dados através de combinações lineares das variáveis originais, preservando a proporcionalidade entre elas. São particularmente eficazes quando as relações subjacentes são predominantemente lineares;
- Técnicas não lineares: capturam relações complexas e não lineares, permitindo representações mais flexíveis dos dados. Apesar da sua maior expressividade, implicam maior custo computacional e menor interpretabilidade.

No âmbito deste projeto, a principal técnica utilizada foi a Análise de Componentes Principais (PCA), que identifica as direções (componentes principais) ao longo das quais os dados apresentam maior variabilidade, ordenando-as por importância decrescente (Jollife e Cadima 2016). Após uma análise experimental, definiu-se `n_components = 2`, valor que apresentou melhores resultados nas métricas de *clustering* (detalhadas nos Capítulos 4 e 5).

Complementarmente, foram também aplicadas técnicas não lineares, nomeadamente t-SNE (*t-distributed Stochastic Neighbor Embedding*) e UMAP (*Uniform Manifold Approximation and Projection*), utilizadas sobretudo para auxiliar na visualização dos dados nos gráficos apresentados no Capítulo 5.

3.3.3 Deteção de outliers

Tal como mencionado, detetar *outliers* trata-se de um componente essencial na aprendizagem não supervisionada, uma vez que nestes métodos não existem rótulos claros que ajudam na distinção entre dados normais e anómalos.

Assim, e antes da aplicação dos algoritmos de *clusters* aos dados fornecidos pela Unidade Local de Saúde de Santo António, foi analisada a existência de *outliers* no conjunto de dados fornecidos pelos profissionais de saúde.

Para isto, foi utilizado o algoritmo *Isolation Forest* da biblioteca *Scikit-Learn*. Este é um algoritmo de aprendizagem não supervisionada para deteção de anomalias que funciona através da divisão dos dados em árvores de decisão para isolar os pontos. Como as anomalias normalmente diferem bastante dos restantes dos dados, *outliers* normalmente estão isolados em árvores com caminhos mais curtos (Dhiraj, Skelton e Mukherjee 2025).

Considerando as suas características de implementação, o algoritmo demonstra elevada escalabilidade e eficiência em contextos de grande dimensionalidade, como os perfis metabólicos analisados neste estudo. Destaca-se ainda pela boa performance em subconjuntos reduzidos de dados (situação também verificada no presente trabalho) e pela maior robustez face aos efeitos de *masking* (anomalias ocultas em *clusters* densos) e *swamping* (anomalias próximas

de pontos normais), superando, nesse sentido, outros métodos de deteção de anomalias (Dhiraj, Skelton e Mukherjee 2025; Yousef, Feng e Jelinek 2024).

Na Tabela 3.5, é possível observar a parametrização usada para o algoritmo de *Isolation Forest*.

Tabela 3.5: Parâmetros utilizados no algoritmo *Isolation Forest*

Parâmetro	**Valor**
`contamination`	0.01
`random_state`	42
`n_estimators`	200
`max_samples`	auto
`n_jobs`	1

Note-se que o parâmetro de contaminação foi definido de forma conservadora, decisão tomada após discussão com profissionais de saúde, tendo em vista evitar a exclusão de variações biológicas potencialmente relevantes. Adicionalmente, ficou acordado que os *outliers* identificados não seriam removidos da análise de *clustering*, mas mantidos para posterior revisão detalhada por especialistas da área, cujo conhecimento do domínio dos dados é mais aprofundado.

Por fim, na Tabela 3.6 são identificados o número de amostras e de *outliers* para cada um dos seis conjuntos de dados analisados.

Tabela 3.6: Número total de amostras e número de *outliers* identificados em cada conjunto de dados analisado

Descrição	**Amostras Totais**	**Outliers**
Acilcarnitinas - todos os dados	865	9
Acilcarnitinas - doentes com diagnóstico de CRC	254	3
Acilcarnitinas - doentes com diagnóstico de CRC e estádio M0	99	1
Aminoácidos - todos os dados	4052	41
Aminoácidos - doentes com diagnóstico de CRC	273	1
Aminoácidos - doentes com diagnóstico de CRC e estádio M0	99	1

Capítulo 4

Modelação e Desenvolvimento Experimental

No capítulo quatro apresentam-se os modelos e o processo de desenvolvimento experimental adotado no âmbito deste trabalho.

Assim, inicialmente, são descritos os algoritmos utilizados neste trabalho e é justificado o motivo pelo qual estes foram escolhidos e como se enquadram nos objetivos do projeto.

De seguida, é descrita a metodologia de hiperparametrização utilizada no contexto deste projeto, detalhando as estratégias usadas para garantir o desempenho ótimo dos modelos.

Por fim, são apresentados e analisados os *clusters* obtidos para os diferentes conjuntos de perfis metabólicos analisados neste projeto. Para tal, e como mencionado na Seção 3.1 foram utilizados conjuntos de dados específicos de perfis de acilcarnitinas e aminoácidos, tanto em populações heterogéneas de doentes como em subconjuntos com diagnóstico de cancro colorretal (CRC) e incluindo ainda a análise particular de doentes em estádio M0.

4.1 Seleção dos Algoritmos

A seleção dos algoritmos de *clustering* a utilizar neste projeto foi uma das principais etapas desta investigação exploratória, uma vez que uma boa seleção dos algoritmos é essencial para a obtenção de resultados satisfatórios (Wegmann et al. 2021).

No contexto deste projeto, a seleção de algoritmos foi, então, feita de modo a refletir as particularidades dos dados metabólicos e atingir os objetivos da investigação:

- Priorizou-se a inclusão de algoritmos de diferentes famílias metodológicas, de modo a garantir uma abordagem abrangente (R. Xu e Wunsch 2005).
- Privilegiaram-se algoritmos que permitissem o ajuste fino de parâmetros, de modo a poder facilitar a otimização dos parâmetros de forma automática.
- Tentou-se utilizar algoritmos que fossem interpretáveis no contexto clínico, de modo a permitir a existência de uma discussão objetiva dos resultados com os profissionais de saúde e obter *feedback* quanto à relevância e aplicabilidade dos resultados.

Tendo em conta estes critérios, foram então selecionados cinco algoritmos:

- *K-means*: O *K-Means* é um dos algoritmos mais conhecidos de *Exclusive Clustering*, em que os dados são classificados e agrupados tendo em conta as suas similaridades e cada novo elemento é adicionado ao *cluster* do qual se encontra mais próximo do

centro (Ahmed, Seraj e Islam 2020). Este é um algoritmo bastante eficiente, com elevada simplicidade conceptual, pelo que foi considerado adequado para a análise experimental elaborada neste projeto. Este algoritmo é bastante sensível à inicialização (Fränti e Sieranoja 2019), pelo que se tentou mitigar esta limitação através de múltiplas execuções com parâmetros de inicialização diferentes.

- *Clustering Hierárquico Aglomerativo*: O *Clustering* Hierárquico aglomerativo é um algoritmo de *clustering* hierárquico, em que a criação de uma hierarquia de *clusters* é feita a partir de uma *abordagem bottom-up*, ou seja, cada elemento começa num *cluster* individual e sucessivamente os *clusters* mais próximos são fundidos (Oti e Olusola 2024). Entre alguns dos motivos de se considerar este algoritmo na fase de experimentação deve-se ao facto de este gerar um dendograma [1], o que facilita uma análise e interpretação visual da estrutura dos dados e permite ainda o número ideal de *clusters* (Boyko e Tkachyk 2023).

- *Density-Based Spatial Clustering* (*DBSCAN*): o *DBSCAN* é um algoritmo baseado em densidade, em que pontos de alta densidade de vizinhos são agrupados e pontos em regiões esparsas são identificados como *outliers*/*noise points* (Schubert et al. 2017). Assim sendo, decidiu-se usar este algoritmo devido à sua robustez a *outliers*, o que pode ser comum em dados metabólicos, quer devido à variabilidade biológica, quer a erros técnicos na recolha dos valores. Mais uma vez este algoritmo apresenta alguma sensibilidade na definição dos seus parâmetros, pelo que, de forma a mitigar esta limitação, foram efetuadas múltiplas execuções com inicializações de parâmetros diferentes (Ramadan et al. 2022).

- *Gaussian Mixture Model* (*GMM*) : é um algoritmo do tipo *Probabilistic Clustering* e é identificado como um modelo de mistura, ou seja, um modelo que é constituído por um número indeterminado de funções de distribuição de probabilidade. No caso do *GMM*, este assume que cada *cluster* segue uma distribuição Gaussiana e utiliza métodos estatísticos para ajustar os parâmetros das distribuições (Deng e Han 2018). Como tal, escolheu-se este modelo, uma vez que permite a modelação da incerteza/atribuição de probabilidade de uma amostra pertencer a um *cluster*, algo que é bastante comum em *datasets* com dados metabólicos/médicos (Nguyen 2024).

- *Hierarchical DBSCAN* (*HDBSCAN*): O *HDBSCAN* é de certa forma uma extensão do *DBSCAN* que ultrapassa algumas das limitações do algoritmo original e permite, por exemplo, identificar *clusters* com densidades variáveis e construir uma hierarquia de densidades (Ramadan et al. 2022). De facto, optou-se por utilizar este algoritmo na fase de experimentação, devido à sua robustez a *outliers*, a que se juntou o facto de ser mais robusto que *DBSCAN* na escolha dos parâmetros (McInnes e Healy 2017).

Por fim, e para concluir, é importante mencionar que outro fator que levou à consideração inicial destes algoritmos foi o facto de em grande parte já terem sido usados nos estudos apresentados na Seção 2.7.

[1] **Dendograma**: é uma representação visual que representa a disposição dos *clusters* e as suas relações. Nestas representações, a altura dos ramos representa a dissimilaridade na qual os grupos se fundem, pelo que alturas mais baixas indicam *clusters* unidos e como tal mais parecidos (Chia 2025).

4.2 Metodologia de Hiperparametrização

Em estudos de *clustering*, a otimização de parâmetros é uma etapa essencial, uma vez que a escolha destes hiperparâmetros influencia diretamente a coesão, a separação e a interpretabilidade dos grupos formados pelos algoritmos (Fränti e Sieranoja 2019; Ramadan et al. 2022; Rodriguez et al. 2019).

Assim, e de forma a tentar limitar a sensibilidade dos algoritmos à parametrização, foi implementada uma abordagem de *hyperparameter tuning* baseada em uma *grid search* (Ogunsanya, Isichei e Desai 2023). Contudo, e de forma a tentar maximizar a eficiência computacional, o número de combinações totais a serem testadas foi limitado a cinco mil por algoritmo. Como tal, em caso de o número de combinações de parâmetros ser maior do que cinco mil, eram selecionadas cinco mil combinações de forma aleatória.

Na Tabela 4.1 são apresentados os espaços de parametrização definidos para cada algoritmo.

Tabela 4.1: Espaços de hiperparâmetros definidos para cada algoritmo de *clustering*

Algoritmo	**Parâmetro**	**Valores Testados**
K-means	Número de *clusters* (k)	3, 4, 5, 6, 7, 8, 9, 10
	Método de inicialização	*k-means++, random*
	Número de inicializações	10, 20, 30, 40, 50, 100, auto
	Máximo de iterações	300, 500
	Algoritmo de implementação	*Lloyd, Elkan*
DBSCAN	*Epsilon* (ε)	0.2, 0.3, 0.4, 0.5, 0.7, 1.0, 1.5, 2.0, 3.0
	Amostras mínimas	3, 5, 10, 15, 20, 30
	Métrica de distância	euclidiana, *Manhattan, Chebyshev, cityblock*
	Algoritmo de busca	auto, ball_tree, kd_tree, brute
	Tamanho da folha	10, 20, 30, 40, 50
Clustering Hierárquico Aglomerativo	Número de *clusters* (k)	3, 4, 5, 6, 7, 8, 9, 10
	Critério de ligação	*ward*, complete, *average*
	Métrica de distância	euclidiana, *Manhattan*[1]
Gaussian Mixture Model	Número de componentes (k)	3, 4, 5, 6, 7, 8, 9, 10
	Tipo de covariância	*full, tied, diag, spherical*
	Número de inicializações	1, 5, 10
	Parâmetros de inicialização	*kmeans, random*
	Regularização da covariância	1e-6, 1e-4, 1e-2
	Máximo de iterações	100, 200, 300
HDBSCAN	Tamanho mínimo do *cluster*	5, 10, 15, 20, 30, 40, 50
	Amostras mínimas	*None*, 1, 5, 10, 15
	Método de seleção	*eom, leaf*
	Métrica de distância	euclidiana, *Manhattan*
	Parâmetro *alpha* (α)	0.5, 1.0, 1.5
	Gerar árvore de *spanning*	*True, False*
	Dados de predição	*True, False*

De forma a validar cada uma das combinações, foram calculadas diversas métricas (Kalimara 2023):

[1] Para o critério *Ward*, apenas a métrica euclidiana é compatível; para *complete* e *average*, ambas as métricas são válidas.

- *Coeficiente de Silhueta*: esta métrica mede a semelhança entre um objeto e o seu grupo. O valor obtido varia entre 1 e -1, sendo que quanto mais próximo de 1 for o valor obtido, melhor as amostras estão bem agrupadas e pelo contrário, quanto mais próximo de -1 for, mais provável que a amostra tenha sido atribuída ao *cluster* errado.
- *Índice de Davies-Bouldin*: Por sua vez, o Índice *Davis-Bouldin* é uma métrica que pretende avaliar a separação dos *clusters*. Esta é uma métrica cujo valor pode variar entre 0 e o infinito positivo, sendo que quanto mais próximo de 0 o valor obtido for, melhor - uma vez que isto indica que os grupos estão bem separados.
- *Índice de Calinski-Harabasz*: O *Índice* de *Calinski-Harabasz* é uma métrica que mede simultaneamente a dispersão dos *clusters* (a distância entre *clusters*) e dispersão dentro dos *clusters* (a densidade de cada *cluster* internamente). O resultado desta métrica pode variar de 0 até ao infinito positivo, mas, neste caso, quanto maior o valor obtido, melhor é o resultado.

No entanto, individualmente, estas métricas possuem algumas limitações. Por exemplo, o *Coeficiente de Silhueta* é sensível a *outliers*, tipo de dados que podem ser comuns em perfis metabólicos e tende a favorecer *clusters* esféricos. Já o *Índice de Calinski-Harabasz* tende a sobrestimar o número de *clusters* quando existe alguma sobreposição dos mesmos. Por fim, o *Índice de Davies-Bouldin* tende a penalizar modelos de maior variabilidade interna - o que pode ser justificável em perfis metabólicos.

Como tal, de modo a reduzir a influência das limitações de cada uma das métricas, foi ainda considerada uma métrica combinada composta pela soma dos seguintes elementos:

- O valor do *Coeficiente de Silhueta* normalizado para o intervalo [0,1], através da equação:

$$\text{Silhueta}_{\text{norm}} = \frac{\text{Silhueta} + 1}{2} \tag{4.1}$$

- O valor do *Índice de Calinski-Harabasz* foi normalizado para o intervalo [0,1] através de:

$$\text{Calinski}_{\text{norm}} = \frac{2}{1 + \exp\left(-\frac{\ln(1+\text{Calinski})}{10}\right)} - 1 \tag{4.2}$$

- O valor do *Índice de Davis-Bouldin* foi invertido (uma vez que valores mais baixos são melhores) e escalado ao intervalo [0,1] com:

$$\text{Davies}_{\text{norm}} = \frac{1}{1 + \frac{\text{Davies}}{2}} \tag{4.3}$$

Por fim, e com o objetivo de lidar com *datasets* de diferentes dimensões, foi introduzida uma ponderação que valoriza o *Coeficiente de Silhueta* em *datasets* de menores dimensões e simultaneamente, foram penalizados os casos em que os *clusters* apresentam tamanhos bastante desproporcionais.

No caso da avaliação de algoritmos como o *DBSCAN* e o *HDBSCAN*, foi ainda considerado o ruído identificado, sendo atribuída uma penalização adicional a resultados com níveis elevados de ruído.

Tendo tudo isto em conta, após a análise dos resultados, concluiu-se que, de forma geral, os algoritmos *GMM* e *K-means* apresentaram desempenhos superiores. Assim, são estes os algoritmos detalhados nas Seções 4.3 e 4.4, bem como no Capítulo 5.

4.3 Clusters para Perfis de Acilcarnitinas

Tal como detalhado no Seção 4.1, para este estudo foram inicialmente analisados cinco algoritmos, mas após o processo de hiperparametrização, os cinco algoritmos iniciais foram filtrados para apenas dois que foram usados para o restante do estudo. Assim, nesta secção são apresentados os resultados da aplicação do *K-Means* e *GMM* no agrupamento das amostras de perfis metabólicos de acilcarnitinas. Estes resultados são apresentados para os três cenários identificados na Seção 3.1 :

1. Conjunto total de doentes com diferentes diagnósticos.
2. Subconjunto restrito a doentes com CRC.
3. Subconjunto de doentes com CRC em estádio M0.

Deste modo, para cada um dos cenários é apresentada a estrutura dos *clusters*, as assinaturas metabólicas associadas, a relação com as características clínicas e a avaliação quantitativa da qualidade do *clustering*.

4.3.1 Dataset com dados de doentes com diversos diagnósticos

O primeiro *dataset* estudado foi um *dataset* com perfis de acilcarnitinas com doentes com diversos diagnósticos. Este é um *dataset* com 865 amostras.

Após a fase de hiperparametrização (Seção 4.2), na qual se testaram 5000 combinações para o *GMM* e 1680 para o *K-Means*, foi utilizada a métrica combinada para selecionar as configurações finais utilizadas. Na Tabela 4.2 são apresentadas estas configurações.

Tabela 4.2: Configurações finais selecionadas para o *K-Means* e *GMM*

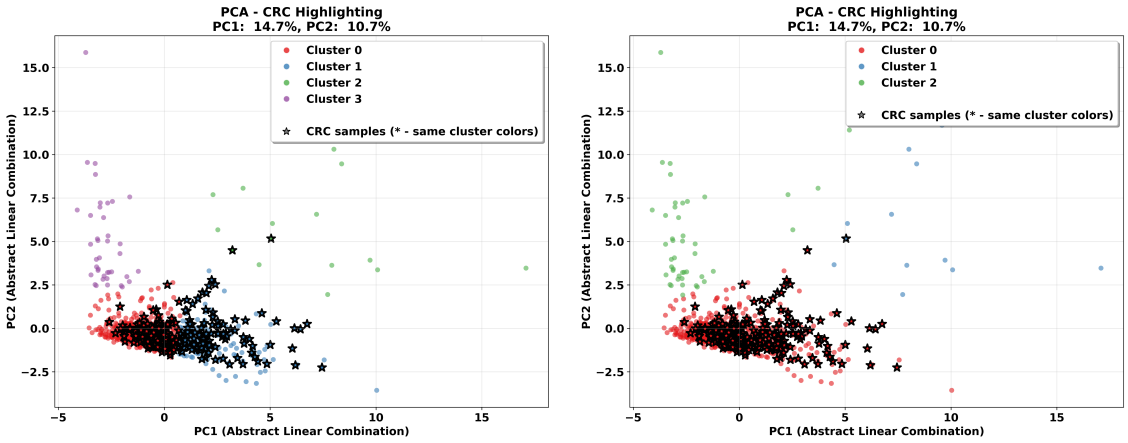
Algoritmo	**Configuração Final**
K-Means	$k = 4$, algoritmo = *lloyd*, *init* = *random*, *max_iter* = 500
GMM	$k = 3$, *covariance_type* = *tied*, *init_params* = *random*, *max_iter* = 200

Adicionalmente, e desde logo, é importante referir que o *K-means* produziu quatro *clusters* com distribuições assimétricas e o *GMM*, por sua vez, identificou três *clusters* com distribuições distintas. Na tabela Tabela 4.3 encontra-se a distribuição de amostras por *clusters* para ambos os algoritmos.

Tabela 4.3: Distribuição das amostras por *cluster* para o *K-Means* e *GMM*

Cluster	K-Means	GMM
0	585 (67.6%)	807 (93.3%)
1	222 (25.7%)	13 (1.5%)
2	19 (2.2%)	45 (5.2%)
3	39 (4.5%)	–

Adicionalmente, e de forma a facilitar a visualização dos resultados, foram desenvolvidos alguns gráficos através do uso do *PCA*. Na Figura 4.1 encontra-se a figura com os gráficos obtidos por ambos os algoritmos. Note-se que nestes gráficos, as amostras com diagnóstico *CRC* estão destacadas de forma a permitir analisar-se facilmente a sua distribuição pelos *clusters*.



(a) *K-Means* (b) *GMM*

Figura 4.1: Distribuição das amostras após redução de dimensionalidade por *PCA*, com destaque para casos CRC, para ambos os algoritmos de *clustering*.

Relativamente ao *clustering*, é importante destacar os resultados das métricas de avaliação. Tal como mencionado na Seção 4.2, para esta avaliação foram utilizadas diversas métricas, pelo que na Tabela 4.4 encontram-se os resultados obtidos para cada um dos algoritmos na sua respetiva configuração.

Tabela 4.4: Comparação das métricas globais de qualidade do *clustering* para *K-Means* e *GMM*

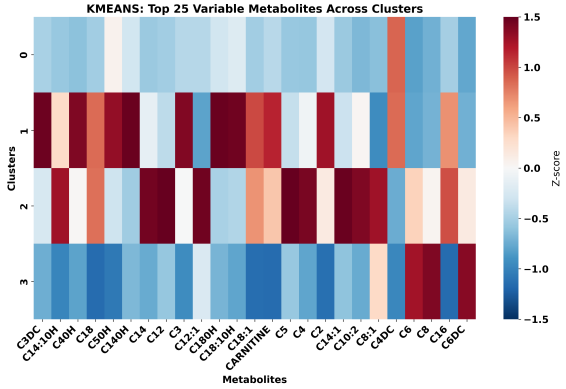
Métrica	K-Means	GMM
Coeficiente de Silhueta	0.507	0.636
Índice de Calinski–Harabasz	641.3	315.6
Índice de Davies–Bouldin	0.715	0.734
Métrica combinada	0.582	0.572

Assim, com os resultados das métricas, observa-se que o *K-means* apresentou melhores valores no *Índice de Calinski-Harabasz*, métrica combinada e *Índice de Davies-Bouldin*, mas o *GMM* apresentou um *Coeficiente de Silhueta* mais elevado.

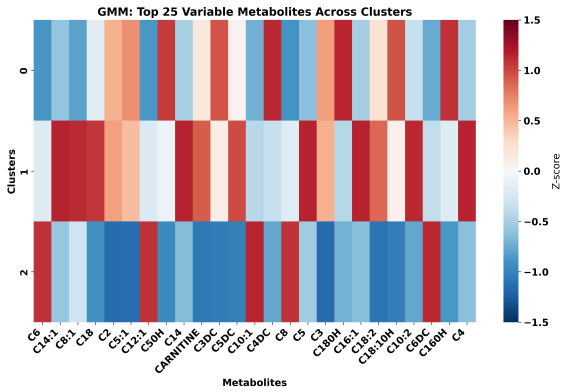
Para complementar o estudo de *clustering*, foram ainda analisadas as assinaturas metabólicas [2] de cada *cluster* e foi neste caso em que estavam a ser utilizados dados de diversos diagnósticos que foi feito um estudo da associação do *cluster* da amostra e o seu diagnóstico.

Relativamente às assinaturas metabólicas, foram observados padrões possivelmente relevantes. Por exemplo, no `Cluster 2` do *GMM* e `Cluster 3` do *K-Means*, verificou-se um aumento do valor de `8`, `C6DC` e `C6`, e no `Cluster 1` do *GMM* e `Cluster 2` do *K-Means*, houve um aumento de `C14:1`, `C5` e `C4`.

Na Figura 4.2 apresenta-se uma representação visual destes dados através de um *heatmap* que identifica os metabólitos com maior variabilidade entre *clusters*, destacando as diferenças metabólicas mais relevantes. Assim, no eixo horizontal são indicadas as acilcarnitinas (selecionadas com base na maior variabilidade entre *clusters*) e no eixo vertical são identificados os *clusters*. Por fim, as *cores* indicam o valor padronizado do *z-score* [3] de cada metabólito para cada *cluster*.



(a) *K-Means*



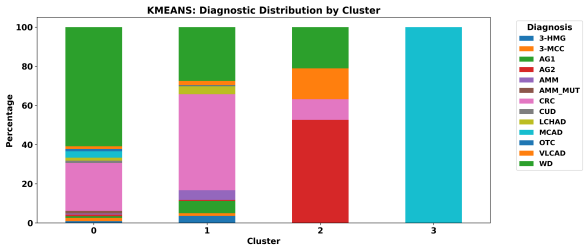
(b) *GMM*

Figura 4.2: *Heatmap* com os 25 metabólitos com maior variabilidade entre *clusters*

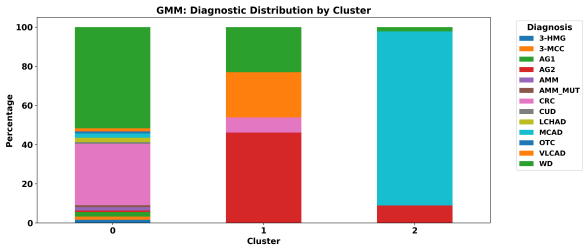
Por fim, foi feita uma análise entre os resultados obtidos e o diagnóstico identificado. Deste modo, na Figura 4.3 é possível ver a divisão de diagnóstico por cada *cluster* para cada um dos algoritmos.

[2] Para fazer esta análise foi calculado o valor médio de cada metabólito em cada *cluster* e a esse valor médio foi subtraído o valor da média global do mesmo metabólito em todas as amostras.

[3] **Z-score**: uma medida estatística que quantifica a distância de um ponto e a média global do *dataset* (Nevil, Kindness e Velasquez 2025).



(a) *K-Means*



(b) *GMM*

Figura 4.3: Distribuição de diagnósticos por *clusters* obtidos pelos algoritmos *K-Means* e *GMM*.

Para complementar esta análise, foi ainda feito um teste de independência *Qui-Quadrado* e o cálculo do V de *Cramer* [4]. Para validar as associações entre os diagnósticos e os *clusters*. Esta análise do *Qui-Quadrado* evidenciou associações significativas entre *clusters* e diagnósticos ($p < 0.001$), e com *V de Cramer* de 0.64 para o *GMM* e de 0.62 para o *K-means*.

4.3.2 Dataset com doentes com diagnóstico CRC

Após o estudo do *dataset* de amostras de doentes com diversos diagnósticos, foi repetido o processo para um subconjunto com apenas perfis de acilcarnitinas de doentes com diagnóstico CRC. Este é um subconjunto menor, com apenas 254 amostras.

Para este, foram mais uma vez testadas as mesmas combinações apresentadas na Subsecção 4.3.1. No entanto, as melhores configurações obtidas foram diferentes, sendo que estas são apresentadas na Tabela 4.5.

Tabela 4.5: Configurações finais selecionadas para o *K-Means* e *GMM* no *dataset* CRC

Algoritmo	**Configuração Final**
K-Means	$k = 3$, algoritmo = *elkan*, *init* = *k-means++*, *max_iter* = 500
GMM	$k = 3$, *covariance_type* = *spherical*, *init_params* = *kmeans*, *max_iter* = 300

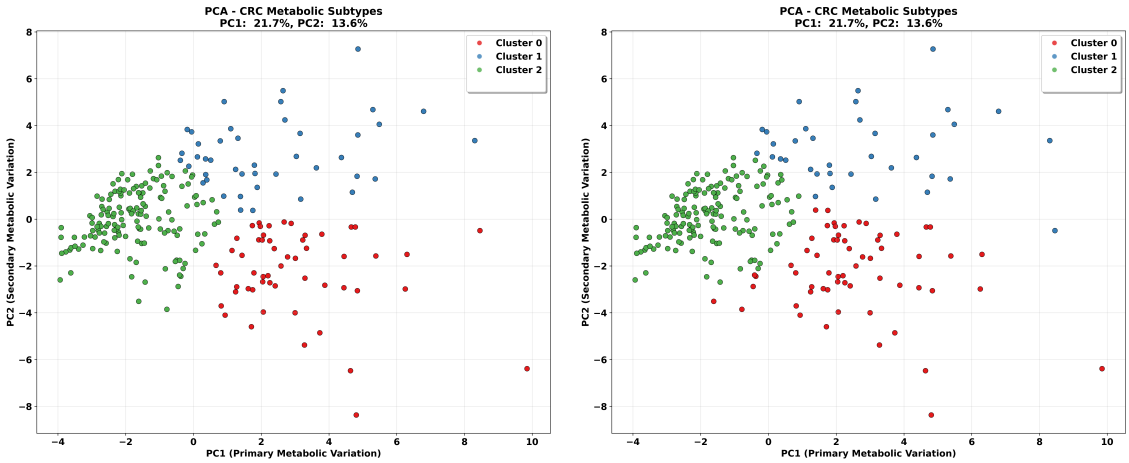
[4] **V de Cramer**: é uma medida do tamanho do efeito para o teste *Qui-Quadrado* de independência que mede a intensidade de associação entre dois campos categóricos. O *V de Cramer* varia entre 0 (sem associação) e 1 (associação perfeita), sendo que valores acima de 0,5 indicam uma associação forte (IBM 2025a)

Neste caso, destaca-se que tanto o *GMM* como o *K-means* produziram três *clusters* com distribuições assimétricas. Os dados desta distribuição das amostras dos *clusters* para ambos os algoritmos podem ser observados na Tabela 4.6,

Tabela 4.6: Distribuição das amostras por *cluster* para o *K-Means* e *GMM* no *dataset* CRC

Cluster	K-Means	GMM
0	54 (21.3%)	60 (23.6%)
1	44 (17.3%)	37 (14.6%)
2	156 (61.4%)	157 (61.8%)

Adicionalmente, e para mais facilmente visualizar e interpretar os resultados dos algoritmos de *clustering*, foram desenvolvidos alguns *gráficos*, em que se incluem gráficos da representação do *clustering* com *PCA*. Os gráficos para cada um dos algoritmos encontram-se na Figura 4.4.



(a) *K-Means* (b) *GMM*

Figura 4.4: Distribuição das amostras após redução de dimensionalidade por PCA para ambos os algoritmos de *clustering* no *dataset* CRC.

Para validar os resultados foram tal como o *dataset* apresentado na Subseção 4.3.1 diversas métricas de avaliação de *clusters*. Na tabela Tabela 4.7 são apresentados esses resultados.

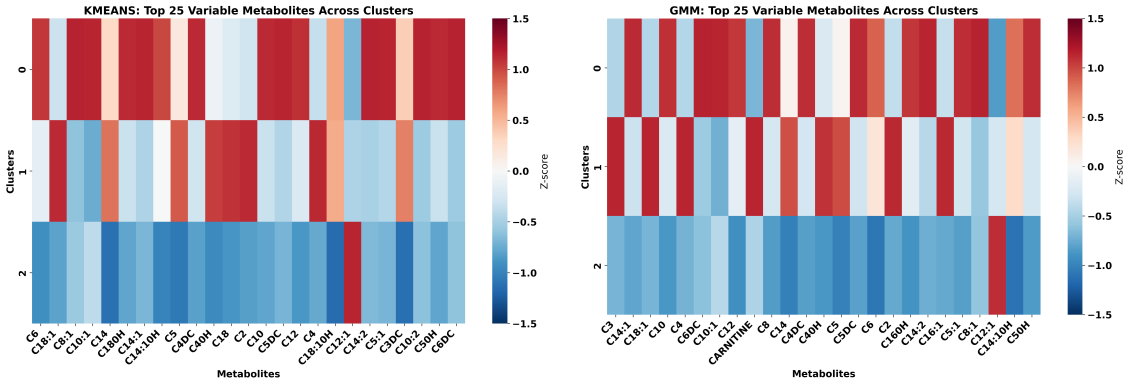
Tabela 4.7: Comparação das métricas globais de qualidade do *clustering* para *K-Means* e *GMM* no *dataset* CRC

Métrica	K-Means	GMM
Coeficiente de Silhueta	0.477	0.479
Índice de Calinski–Harabasz	206.5	200.9
Índice de Davies–Bouldin	0.855	0.841
Métrica combinada	0.607	0.607

Observa-se que ambos os algoritmos apresentam resultados bastante semelhantes, apresentando mesmo o mesmo resultado na métrica combinada e apenas variações ligeiras nas restantes métricas.

Complementarmente ao estudo de *clustering* para os dados de acilcarnitinas *CRC*, foram analisados os padrões de assinatura metabólica de cada *cluster* e foi feito um estudo acerca da distribuição de estágios de CRC por cada *cluster*.

Relativamente ao estudo da assinatura metabólica, foram observados padrões consistentes em ambos os algoritmos. Entre estes, destaca-se que, em ambos os algoritmos, nos *Clusters* `0` foram assinalados aumentos ligeiros nas acilcarnitinas `C4DC` e `C6`, e nos *Clusters* `1` de `C18:1` e `C4`. De forma a simplificar a visualização destes dados, na Figura 4.5 apresenta-se uma representação visual dos mesmos, indicando os metabólitos com maior variabilidade entre *clusters*.



(a) *K-Means* (b) *GMM*

Figura 4.5: *Heatmap* com os 25 metabólitos com maior variabilidade entre *clusters* no *dataset* CRC

Por fim, e de forma a concluir, foi analisada a distribuição de amostras por *cluster* com base no seu estádio de CRC, de forma a tentar avaliar a existência de uma associação entre a progressão da doença e a segmentação obtida pelos algoritmos. Os resultados encontram-se resumidos na Tabela 4.8 e observa-se que ambos os algoritmos apresentam resultados globalmente consistentes, embora as proporções variem ligeiramente. Adicionalmente, realça-se que, por um lado, os doentes no estádio `M0` e `M2` se encontram maioritariamente agrupados no *Cluster* 2, e que, por outro, o estádio `M1` apresenta uma divisão mais equilibrada entre os *Clusters* 0 e 2, enquanto os estádios mais avançados (`M3` e `M4`) apresentam amostras muito reduzidas, estando estas essencialmente concentradas no *Cluster* 0 ou 2.

Tabela 4.8: Distribuição das amostras por *cluster* e estádio de CRC para *K-Means* e *GMM*

Algoritmo	Cluster	M0	M1	M2	M3	M4	Total
GMM	0	26	13	16	3	0	58
	1	15	7	12	0	0	34
	2	58	28	33	2	1	122
K-Means	0	22	12	16	2	0	52
	1	15	11	14	0	0	40
	2	62	25	31	3	1	122

4.3.3 Dataset com doentes com diagnóstico CRC em estádio M0

O último conjunto de dados de *acilcarnitinas* analisado foi um subconjunto em que apenas existiam amostras de doentes com diagnóstico CRC em estádio M0. Como tal, o processo de estudo utilizado foi o mesmo que já foi apresentado nas Secções 4.3.1 e 4.3.2. Note-se ainda que este é o conjunto de dados de acilcarnitinas mais restrito, sendo constituído por apenas 99 amostras.

As configurações finais selecionadas para este *dataset* são apresentadas na Tabela 4.9.

Tabela 4.9: Configurações finais selecionadas para o *K-Means* e *GMM* no *dataset* CRC M0

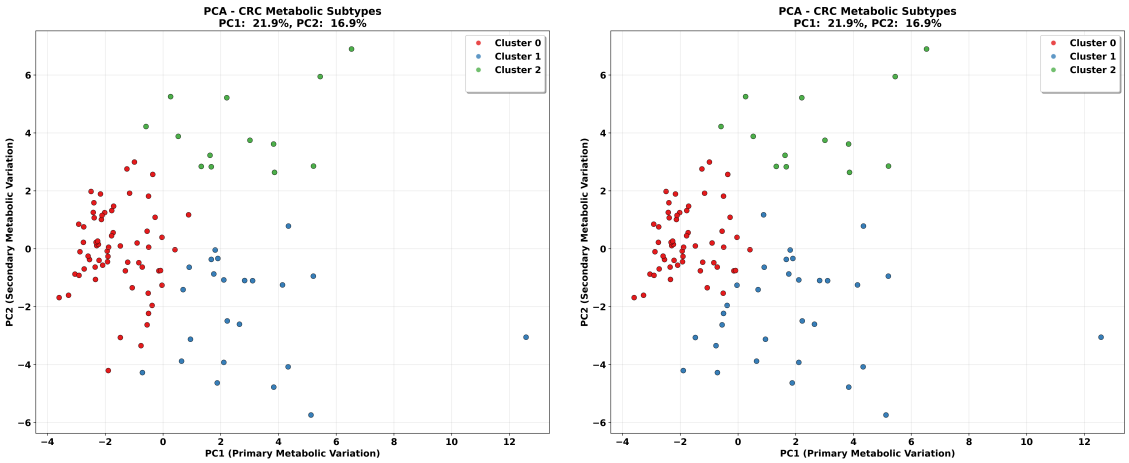
Algoritmo	Configuração Final
K-Means	$k = 3$, algoritmo = *lloyd*, *init* = *k-means++*, *max_iter* = 500
GMM	$k = 3$, *covariance_type* = *diag*, *init_params* = *kmeans*, *max_iter* = 100

Adicionalmente, e tal como nos resultados apresentados na Subseção 4.3.2, destaca-se que ambos os algoritmos convergiram para uma solução de três *clusters* e, mais uma vez, a distribuição das amostras pelos *clusters* é detalhada na Tabela 4.10.

Tabela 4.10: Distribuição das amostras por *cluster* para o *K-Means* e *GMM* no *dataset* CRC M0

Cluster	K-Means	GMM
0	63 (63.6%)	55 (55.6%)
1	23 (23.2%)	31 (31.3%)
2	13 (13.1%)	13 (13.1%)

De forma a simplificar a interpretação dos resultados, é ainda apresentado o gráfico da Figura 4.6, com a representação visual dos *clusters*, obtidos através de uma redução de dimensionalidade com o *PCA*. Estas visualizações evidenciam, não só a separação entre os *clusters*, como também a elevada consistência entre os resultados dos dois algoritmos.



(a) *K-Means* (b) *GMM*

Figura 4.6: Distribuição das amostras após redução de dimensionalidade por *PCA* para ambos os algoritmos de *clustering* no *dataset* CRC M0.

Adicionalmente, e de forma a avaliar os resultados de *clustering* obtidos, foram utilizadas as métricas já estabelecidas e referidas. Na Tabela 4.11 são detalhados os resultados de cada um dos algoritmos em cada uma das métricas utilizadas. Estas evidenciam uma performance muito semelhante entre ambos os algoritmos, mas com o *K-means* a apresentar resultados ligeiramente mais favoráveis em todas as métricas.

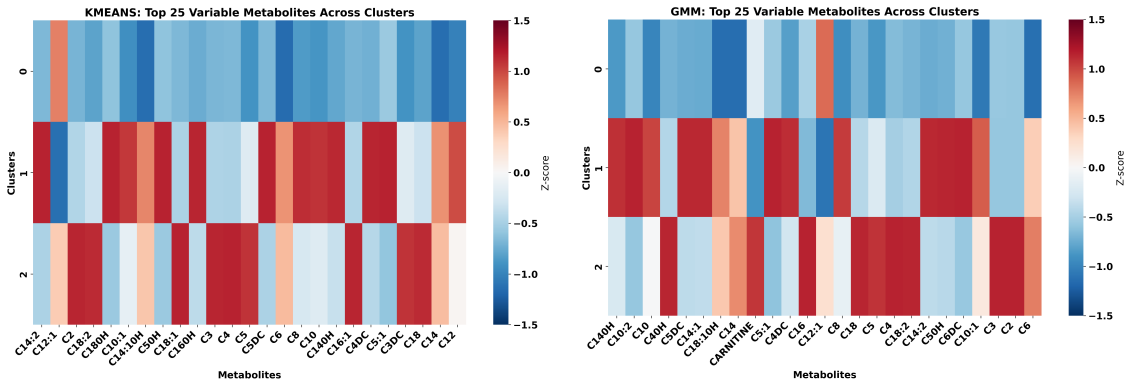
Tabela 4.11: Comparação das métricas globais de qualidade do *clustering* para *K-Means* e *GMM* no dataset CRC M0

Métrica	**K-Means**	**GMM**
Coeficiente de Silhueta	0.478	0.450
Índice de Calinski–Harabasz	75.9	71.2
Índice de Davies–Bouldin	0.794	0.842
Métrica combinada	0.621	0.620

Para complementar o estudo de *clustering* foram ainda analisadas as assinaturas metabólicas de cada *cluster* e foi feita uma breve análise longitudinal [5].

Relativamente às assinaturas metabólicas, foram identificados padrões semelhantes em ambos os algoritmos. Assim, por exemplo, em ambos os algoritmos o *Cluster* 1 caracteriza-se por uma redução do C18:1 e do C3, e o *Cluster* 2 por um aumento dos níveis de C18, C4 e C5. Na Figura 4.7 apresenta-se a representação visual destes padrões metabólicos.

[5] **Análise Longitudinal**: um estudo que segue o que acontece a determinadas variáveis ao longo do tempo (Cherry 2023).



(a) *K-Means* (b) *GMM*

Figura 4.7: *Heatmap* com os 25 metabólitos com maior variabilidade entre *clusters* no *dataset* CRC M0

Por fim, foi elaborada uma análise longitudinal. Um dos objetivos que se pretendia atingir com este trabalho era identificar se os grupos gerados pelos algoritmos evidenciam padrões na progressão da doença ou não - ou seja, se, por exemplo, a maioria dos elementos do *Cluster* 1 representavam casos de doentes que tiveram de passar por quimioradioterapia e cirurgia, ou se apresentavam apenas casos de doentes que fizeram cirurgia e não passaram por quimioradioterapia. Uma forma de fazer isto foi fazer um estudo longitudinal em que se comparavam diferentes amostras do mesmo doente [6]. Assim, primeiramente destaca-se que existem 37 doentes com múltiplas amostras. Nesses doentes, as transições mais frequentes ocorreram entre os estádios M0 e M1 (33 transições) e M1 e M2 (20 transições), o que sugere uma progressão natural da doença. Não menos importante, verificou-se que o *Cluster* 0 apresenta uma maior taxa de progressão de `M1` para `M2` (60,9%) e o *Cluster* 2 apresenta uma tendência de progressão de `M0` para `M2` mas menor transição para `M2` (33,3%). Estes dados estão resumidos na Tabela 4.12, que apresenta os resultados obtidos para ambos os algoritmos.

Progressão	**Cluster 0**		**Cluster 1**		**Cluster 2**	
	K-Means	*GMM*	*K-Means*	*GMM*	*K-Means*	*GMM*
Total doentes	24	23	7	8	6	6
`M0 → M1`	21 (87,5%)	20 (87,0%)	6 (85,7%)	7 (87,5%)	6 (100,0%)	6 (100,0%)
`M1 → M2`	15 (62,5%)	14 (60,9%)	3 (42,9%)	4 (50,0%)	2 (33,3%)	2 (33,3%)
`M0 → M2`	3 (12,5%)	3 (13,0%)	1 (14,3%)	1 (12,5%)	0 (0,0%)	0 (0,0%)

Tabela 4.12: Taxa de progressão terapêutica por *cluster* para *K-Means* e *GMM*

4.4 Clusters para Perfis de Aminoácidos

Esta secção segue a mesma estrutura apresentada na Seção 4.3, mas aplicada aos dados relativos aos perfis de aminoácidos. Tal como no caso das acilcarnitinas, são analisados três cenários distintos, definidos na Seção 3.1.

[6] Os identificadores das amostras são compostos por um identificador único seguido de uma data. Amostras com o mesmo identificador representam amostras do mesmo doente recolhidas em datas diferentes

Após o processo de hiperparametrização, foram novamente selecionados os algoritmos *K-Means* e *GMM* para a análise final. Para cada cenário, descrevem-se:

- A estrutura dos *clusters* obtidos.
- Os resultados das métricas de avaliação da qualidade do *clustering*.
- As assinaturas metabólicas características de cada grupo.

Adicionalmente, no primeiro conjunto de dados (população com diagnósticos diversos) avalia-se a associação entre os *clusters* e o diagnóstico clínico. Já no conjunto de dados restrito a doentes com CRC em estádio M0, é efetuada uma análise longitudinal para investigar possíveis relações entre os grupos identificados e padrões de progressão da doença.

4.4.1 Dataset com dados de doentes com diversos diagnósticos

Tal como mencionado para além do estudo de perfis metabólicos, este projeto analisa ainda *datasets* com perfis metabólicos de aminoácidos. Neste sentido, o primeiro *dataset* deste tipo a ser analisado foi um *dataset* com doentes com diferentes diagnósticos, sendo que este *dataset* é composto por 4052 amostras.

Assim, mais uma vez, começou-se pelo processo de hiperparametrização, sendo que, após este processo, foram obtidas as seguintes configurações para cada um dos algoritmos.

Tabela 4.13: Configurações finais selecionadas para o *K-Means* e *GMM* no *dataset* de aminoácidos

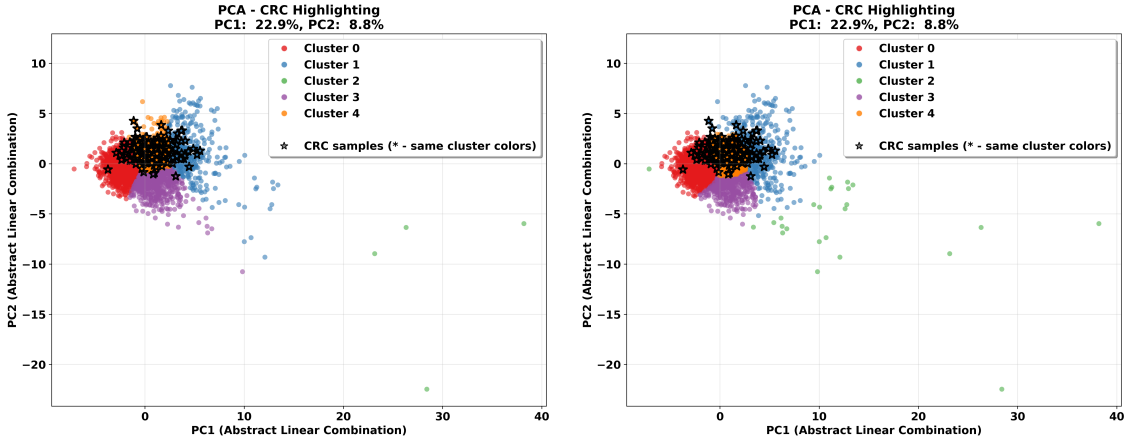
Algoritmo	**Configuração Final**
K-Means	k=5k=5 k=5, algoritmo = *lloyd*, *init* = *k-means++*, *max_iter* = 300
GMM	k=5k=5 k=5, *covariance_type* = *spherical*, *init_params* = *kmeans*, *max_iter* = 200

Destaca-se que para ambos os algoritmos a melhor configuração apresenta cinco *clusters*. Na tabela Tabela 4.14 é apresentada a distribuição das amostras por cada um dos cinco *clusters* para ambos os algoritmos.

Tabela 4.14: Distribuição das amostras por *cluster* para o *K-Means* e *GMM* no *dataset* de aminoácidos

Cluster	**K-Means**	**GMM**
0	1682 (41.5%)	1870 (46.1%)
1	465 (11.5%)	473 (11.7%)
2	4 (0.1%)	26 (0.6%)
3	607 (15.0%)	388 (9.6%)
4	1294 (31.9%)	1295 (32.0%)

Para facilitar esta visualização, é ainda apresentado um gráfico com redução de dimensionalidade através do *PCA* na Figura 4.8



(a) *K-Means* (b) *GMM*

Figura 4.8: Distribuição das amostras após redução de dimensionalidade por *PCA* para ambos os algoritmos de *clustering* no *dataset* de aminoácidos.

De modo a concluir a tarefa de *clustering*, resta mencionar os resultados nas métricas de avaliação previamente mencionadas na Seção 4.2. Como tal, na Tabela 4.15 encontram-se os resultados nas diferentes métricas obtidos para cada um dos algoritmos na sua respetiva configuração.

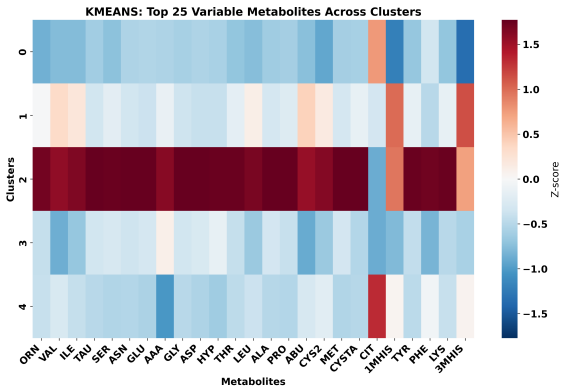
Tabela 4.15: Comparação das métricas globais de qualidade do *clustering* para *K-Means* e *GMM* no *dataset* de aminoácidos

Métrica	**K-Means**	**GMM**
Coeficiente de Silhueta	0.379	0.378
Índice de Calinski–Harabasz	2251.1	2006.1
Índice de Davies–Bouldin	0.790	0.921
Métrica combinada	0.589	0.577

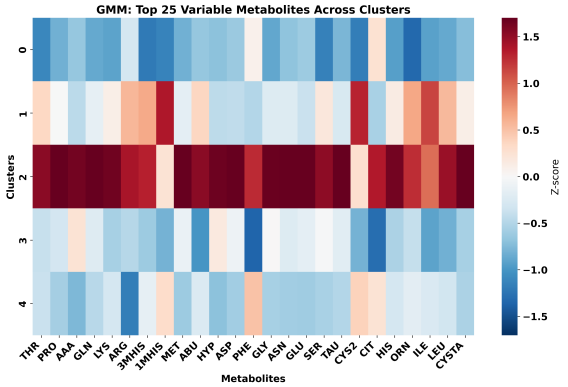
Estes resultados evidenciam que ambos os algoritmos apresentam resultados bastante semelhantes, mas o *K-Means* apresenta resultados superiores para os *Índices de Calinski-Harabasz* e *Davies-Bouldin* (e, como tal, também para a métrica combinada).

Para complementar este estudo, foram ainda analisadas as assinaturas metabólicas de cada *cluster*, e uma vez que os dados apresentavam diversos diagnósticos, foi feito um estudo acerca da associação do diagnóstico com o *cluster* ao qual tinha sido atribuído.

Quanto às assinaturas metabólicas, evidenciaram-se variações relevantes nos diferentes *clusters* e para ambos os algoritmos de aminoácidos como a valina, a prolina e a alanina. Na Figura 4.9 apresenta-se uma representação visual destes dados, indicando os aminoácidos com maior variabilidade entre *clusters*.



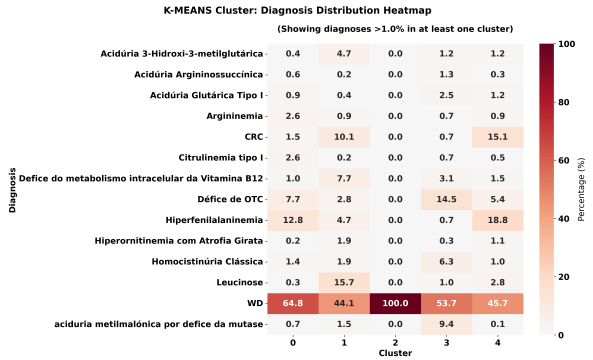
(a) *K-Means*



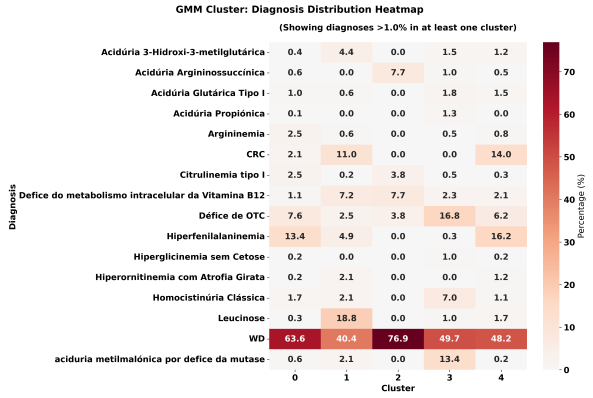
(b) *GMM*

Figura 4.9: *Heatmap* com os 25 aminoácidos com maior variabilidade entre *clusters*

Por fim, foi feita uma análise entre os resultados obtidos e o diagnóstico identificado. Deste modo, na Figura 4.10 é possível ver a divisão de diagnóstico por cada *cluster* para cada um dos algoritmos. Neste caso, devido ao grande número de diagnósticos possíveis, foi decidido utilizar-se um *heatmap* para representar a distribuição ao invés de um gráfico de barras.



(a) *K-Means*



(b) *GMM*

Figura 4.10: Distribuição dos diagnósticos por *cluster*, representada sob a forma de *heatmap*.

Com base nestes dados e aplicando um teste de independência do *Qui-Quadrado* juntamente com o cálculo do V de *Cramer*, avaliou-se a associação entre diagnósticos e *clusters*. Assim, foram encontradas associações estatisticamente significativas entre os *clusters* e os diagnósticos. No entanto, o V de *Cramer* apresentou valores de apenas 0,32 para o *GMM* e 0,30 para o *K-Means*, o que corresponde a uma associação de intensidade moderada.

4.4.2 Dataset com doentes com diagnóstico CRC

Após a análise do *dataset* com dados de doentes com diversos diagnósticos, foi feito o mesmo estudo para um subconjunto de dados de perfis de aminoácidos de doentes com diagnóstico de CRC. Este é um subconjunto de 273 amostras, e , mais uma vez, através do estudo de hiperparametrização, foram obtidas as configurações indicadas na Tabela 4.16 para cada algoritmo.

Tabela 4.16: Configurações finais selecionadas para o *K-Means* e *GMM* no *dataset* CRC

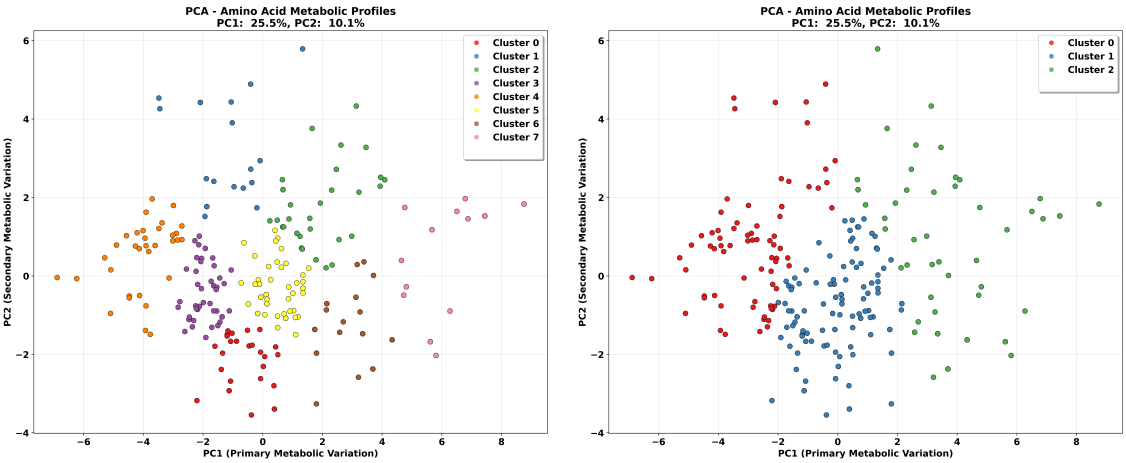
Algoritmo	**Configuração Final**
K-Means	k=8k=8 k=8, algoritmo = *elkan*, *init* = *random*, *max_iter* = 500
GMM	k=3k=3 k=3, *covariance_type* = *spherical*, *init_params* = *random*, *max_iter* = 200

No entanto, neste caso, e contrariamente aos restantes estudos, os dois algoritmos convergiram para resultados bastante diferentes. De facto, no *K-Means* foi selecionada uma solução de oito *clusters* com distribuições relativamente equilibradas, enquanto o *GMM* identificou três *clusters* com distribuições bastante assimétricas. Na tabela Tabela 4.17 encontram-se os dados das distribuições das amostras dos *clusters* para ambos os algoritmos.

Tabela 4.17: Distribuição das amostras por *cluster* para o *K-Means* e *GMM* no *dataset* CRC

Cluster	**K-Means**	**GMM**
0	33 (12,1%)	82 (30,0%)
1	21 (7,7%)	132 (48,4%)
2	37 (13,6%)	59 (21,6%)
3	51 (18,7%)	—
4	33 (12,1%)	—
5	52 (19,0%)	—
6	30 (11,0%)	—
7	16 (5,9%)	—

Adicionalmente, e para facilitar a visualização destas diferenças, na Figura 4.11 pode-se observar um gráfico com os diferentes *clusters* de ambos os algoritmos.



(a) *K-Means* (b) *GMM*

Figura 4.11: Distribuição das amostras após redução de dimensionalidade por *PCA* para ambos os algoritmos de *clustering* no *dataset* CRC.

Relativamente à avaliação dos resultados do *clustering* na Tabela 4.18 encontram-se os valores obtidos por cada um dos *clusters* nas diferentes métricas de avaliação.

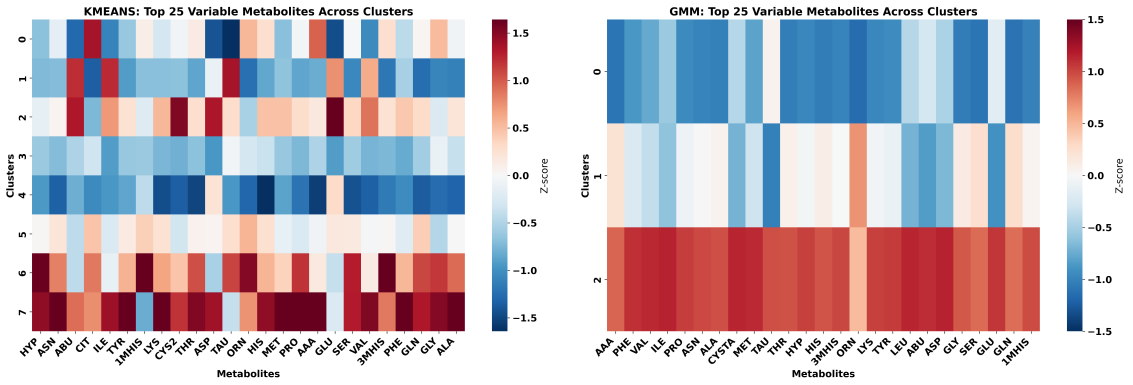
Tabela 4.18: Comparação das métricas globais de qualidade do *clustering* para *K-Means* e *GMM* no *dataset* CRC

Métrica	**K-Means**	**GMM**
Coeficiente de Silhueta	0,368	0,344
Índice de Calinski–Harabasz	223,7	198,0
Índice de Davies–Bouldin	0,817	0,986
Métrica combinada	0,600	0,586

Como tal, observa-se que o *K-Means* apresenta resultados ligeiramente superiores em todas as métricas avaliadas.

Por fim, e para complementar o *clustering*, foram feitos dois estudos complementares.

O primeiro foi um estudo das assinaturas metabólicas de cada *clusters*, e mais uma vez foram evidenciadas variações de aminoácidos como a valina, a prolina e a alanina em ambos os algoritmos, tal como representado na Figura 4.12.



(a) *K-Means* (b) *GMM*

Figura 4.12: *Heatmap* dos 25 aminoácidos com maior variabilidade entre *clusters*, obtidos a partir do *dataset* CRC, para os algoritmos *K-Means* e *GMM*.

O segundo, e visto que este é um *dataset* de amostras de CRC apenas, trata-se de um estudo acerca da distribuição de amostras por *cluster* em função do estádio do CRC. Na tabela Tabela 4.19 apresenta-se esta distribuição para ambos os algoritmos.

Tabela 4.19: Distribuição das amostras por *cluster* e estádio de CRC para os algoritmos *K-Means* e *GMM*.

Algoritmo	Cluster	M0	M1	M2	M3	M4	Total
K-Means	0	16	2	6	1	0	25
	1	7	8	2	0	0	17
	2	8	9	10	2	0	29
	3	17	9	15	0	1	42
	4	13	14	5	0	0	32
	5	25	4	11	1	0	41
	6	9	0	6	1	0	16
	7	4	2	6	0	0	12
GMM	0	31	26	14	0	1	72
	1	54	13	30	2	0	99
	2	14	9	17	3	0	43

Observa-se que, no caso do *K-Means*, o estádio `M0` é mais frequente nos *clusters* `0`, `3` e `5` e que o estádio *M2* encontra-se mais representado nos *Clusters* 3, 5 e 2. Já no caso do *GMM*, o estádio *M0* é mais frequente no *Cluster* `1` e o estádio `M2` se distribui pelos três *clusters*.

4.4.3 Dataset com doentes com diagnóstico CRC em estádio M0

O último conjunto de dados analisado trata-se de um conjunto de perfis de aminoácidos de doentes com diagnóstico de CRC em estádio M0, pelo que este se trata do conjunto de dados mais restrito, com apenas 99 amostras. Para estudar o *clustering* deste conjunto de dados, foram utilizados o *GMM* e o *K-Means*, sendo as configurações utilizadas e obtidas após o processo de hiperparametrização explicado na Seção 4.2 e identificadas na Tabela 4.20.

Tabela 4.20: Configurações finais selecionadas para o *K-Means* e *GMM* no *dataset* de aminoácidos CRC M0

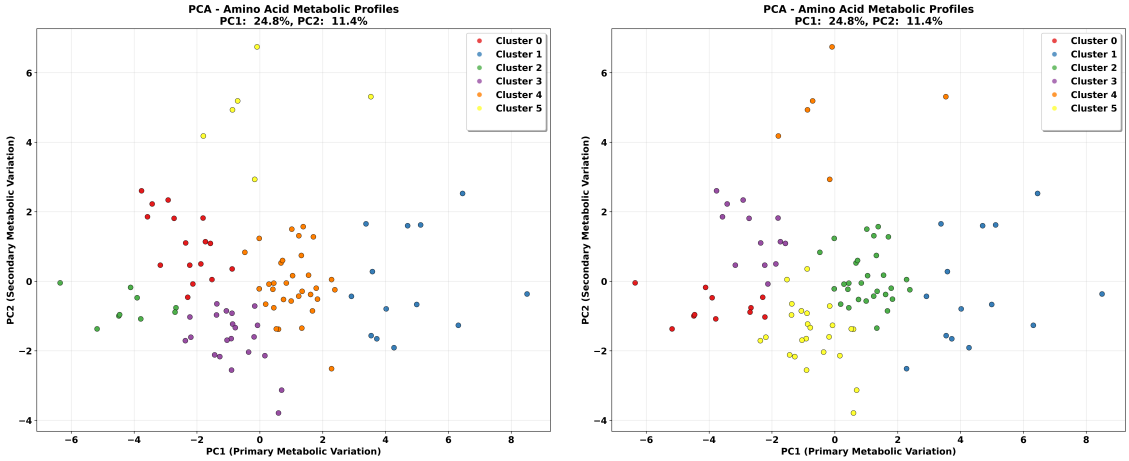
Algoritmo	**Configuração Final**
K-Means	$k = 6$, algoritmo = *lloyd*, *init* = *random*, *max_iter* = 500
GMM	$k = 6$, *covariance_type* = *diag*, *init_params* = *kmeans*, *max_iter* = 100

Ambos os algoritmos voltaram a apresentar resultados semelhantes, tendo ambos identificado seis *clusters*. A distribuição das amostras pelos *clusters* é detalhada na Tabela 4.21.

Tabela 4.21: Distribuição das amostras por *cluster* para *K-Means* e *GMM* no *dataset* de aminoácidos (CRC M0)

Cluster	**K-Means**	**GMM**
0	16 (16,2%)	11 (11,1%)
1	13 (13,1%)	14 (14,1%)
2	9 (9,1%)	30 (30,3%)
3	22 (22,2%)	13 (13,1%)
4	33 (33,3%)	6 (6,1%)
5	6 (6,1%)	25 (25,3%)

De forma a facilitar a visualização destes *clusters*, na Figura 4.13 são apresentados gráficos para cada um dos algoritmos, obtidos através da redução de dimensionalidade com o *PCA*.



(a) *K-Means* (b) *GMM*

Figura 4.13: Distribuição das amostras após redução de dimensionalidade por *PCA* para ambos os algoritmos de *clustering* no *dataset* de aminoácidos CRC M0.

Mais uma vez, e para avaliar os *clusters* obtidos, e tal como os outros estudos já apresentados, foram calculadas as métricas de avaliação. Na Tabela 4.22 são detalhados os resultados de cada um dos algoritmos em cada uma das métricas utilizadas, mas estas evidenciam um

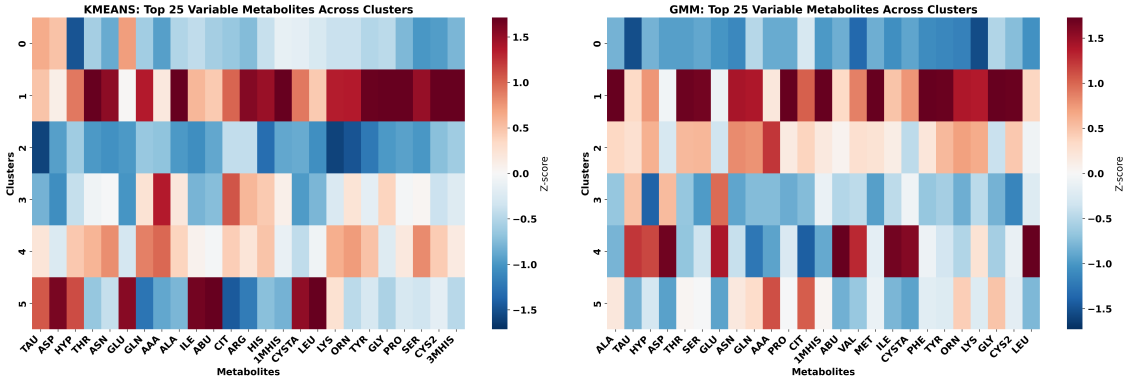
resultado muito semelhante para ambos os algoritmos, com um desempenho ligeiramente melhor do *K-Means*.

Tabela 4.22: Comparação das métricas globais de qualidade do *clustering* para *K-Means* e *GMM* no dataset de aminoácidos CRC M0

Métrica	**K-Means**	**GMM**
Coeficiente de Silhueta	0.398	0.382
Índice de Calinski–Harabasz	80.8	78.1
Índice deDavies–Bouldin	0.780	0.818
Métrica combinada	0.610	0.606

Por fim, e tal como para o estudo apresentado na Subseção 4.3.3, foram identificados os padrões metabólicos e foi feita uma breve análise longitudinal.

Relativamente às assinaturas metabólicas, obtiveram-se resultados bastante semelhantes aos das Secções 4.4.1 e 4.4.2, tal como apresentado na Figura 4.14.



(a) *K-Means* (b) *GMM*

Figura 4.14: *Heatmap* com os 25 aminoácidos com maior variabilidade entre *clusters* no *dataset* de aminoácidos CRC M0

Já quanto à análise longitudinal, esta seguiu a mesma metodologia aplicada às acilcarnitinas e foram utilizados os mesmos 37 doentes com múltiplas amostras, sendo que os resultados obtidos encontram-se resumidos na Tabela 4.23.

Progressão	**Cluster 0**		**Cluster 1**		**Cluster 2**		**Cluster 3**		**Cluster 4**		**Cluster 5**	
	K-Means	*GMM*	*K-Means*	*GMM*	*K-Means*	*GMM*	*K-Means*	*GMM*	*K-Means*	*GMM*	*K-Means*	*GMM*
Total doentes	5	3	4	4	3	10	8	4	12	2	5	14
M0 → M1	4 (80,0%)	3 (100%)	3 (75,0%)	3 (75,0%)	3 (100%)	9 (90,0%)	7 (87,5%)	3 (75,0%)	11 (91,7%)	2 (100%)	5 (100%)	12 (85,7%)
M1 → M2	2 (40,0%)	2 (66,7%)	2 (50,0%)	2 (50,0%)	2 (66,7%)	6 (60,0%)	5 (62,5%)	2 (50,0%)	7 (58,3%)	1 (50,0%)	2 (40,0%)	8 (57,1%)
M0 → M2	0 (0,0%)	0 (0,0%)	1 (25,0%)	1 (25,0%)	0 (0,0%)	1 (10,0%)	1 (12,5%)	1 (25,0%)	2 (16,7%)	0 (0,0%)	0 (0,0%)	2 (14,3%)

Tabela 4.23: Taxa de progressão terapêutica por *cluster* para *K-Means* e *GMM* no *dataset* de aminoácidos

Capítulo 5

Avaliação e Discussão dos Resultados

No capítulo Capítulo 5 são apresentados e discutidos os resultados obtidos nos algoritmos de aprendizagem não supervisionada sobre os perfis metabólicos de aminoácidos e acilcarnitinas e que foram apresentados no Capítulo 4.

Assim, inicialmente são avaliadas as métricas de desempenho dos algoritmos de *clustering* testados e, após isto, são interpretadas as implicações clínicas com base nos *clusters* identificados. De forma a complementar este processo de avaliação, é ainda partilhada uma reflexão crítica dos resultados, com base no *feedback* dos profissionais de saúde, e analisadas as implicações práticas deste trabalho.

Por fim, e para concluir este capítulo, são apresentadas algumas limitações do estudo e sugestões para trabalhos futuros.

5.1 Avaliação dos modelos

Ao longo do Capítulo 4 foi detalhado o processo utilizado para avaliar os modelos de *clustering* desenvolvidos para avaliar os perfis metabólicos de aminoácidos e acilcarnitinas. Como tal, e como já referenciado, esta avaliação foi feita com base em métricas quantitativas, como o *Coeficiente de Silhueta*, o *Índice de Calinski-Harabasz* e o *Índice de Davies-Bouldin*, métricas que são bastante referenciadas em estudos de *clustering*, bem como uma métrica combinada das três métricas mencionadas que tinha como objetivo reduzir a influência das limitações de cada métrica de forma individual.

Os resultados obtidos para cada uma das métricas foram detalhados no Capítulo 4 nas Tabelas 4.4, 4.7, 4.11, 4.15, 4.18 e 4.22. Contudo, de forma a resumir estes dados, na Figura 5.1 é apresentado um *heatmap* com os valores normalizados de cada uma das métricas para cada um dos *datasets* analisados.

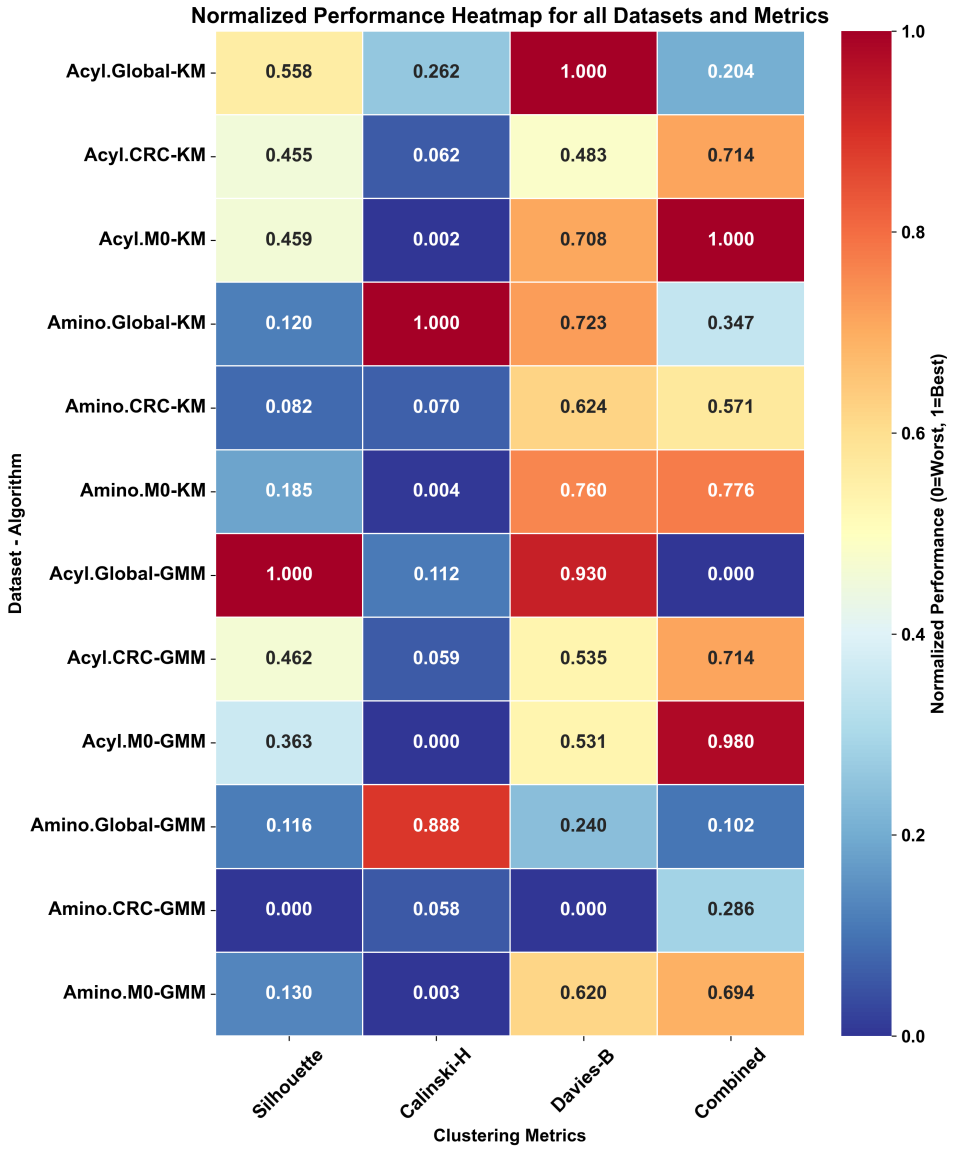


Figura 5.1: *Heatmap* com valores normalizados para todos os conjuntos de dados e métricas de avaliação de agrupamento.

Assim, analisando as tabelas e a Figura 5.1, no caso dos perfis metabólicos de acilcarnitinas, são observados os seguintes padrões em cada um dos *datasets* utilizados:

- *Dataset* com diversos diagnósticos: Neste conjunto de dados, o *GMM* apresentou um melhor *Coeficiente de Silhueta* - 0.636 em vez de 0.507, o que indica que apresentou uma melhor coesão interna dos *clusters*. No entanto, o *K-Means* apresentou valores bastante superiores no *Índice de Calinski-Harabasz* - 641.3 contra os 315.6 do *GMM* -, e um *Índice de Davies-Bouldin* ligeiramente melhor - 0.715 ao invés dos 0.734 do *GMM* - o que sugere uma melhor separação entre grupos e menor sobreposição.
- *Dataset* de doentes CRC: Ambos algoritmos apresentam resultados bastante semelhantes, sendo que o *K-Means* apresenta melhor *Índice de Calinski-Harabasz* - 206.5 ao invés de 200.9 do *GMM* - o que indica que os *clusters* identificados no *K-Means* obtiveram melhor separação, mas o *GMM* apresentou melhor *Coeficiente de Silhueta*

- 0.479 ao invés dos 0.477 no *K-Means* - e *Índice de Davies-Bouldin* - 0.841 no *GMM* e 0.855 no *K-Means* -, o que indica que os *clusters* do *GMM* têm uma melhor coesão interna.

- *Dataset* de doentes CRC em estágio M0: Neste *dataset* o *K-Means* apresentou melhores resultados em todas as métricas, mais uma vez, o que indica que os *clusters* gerados com este modelo encontram-se melhor definidos e apresentam melhor separação e coesão interna que os *clusters* identificados pelo *GMM*.

Já para os perfis de aminoácidos, observa-se um padrão semelhante ao identificado nos dados dos perfis metabólicos de acilcarnitinas:

- *Dataset* com diversos diagnósticos: para este conjunto de dados o *K-Means* foi melhor em todas as métricas utilizadas, destacando-se particularmente no *Índice de Calinski-Harabasz* - em que apresentou um valor de 2251.1 e o *GMM* apenas de 2006.1 - e no *Índice de Davies-Bouldin* - 0.790 no *K-Means* e 0.921 no *GMM*. Isto indica que o *K-Means* formou *clusters* mais compactos e bem separados. É importante realçar que o *Coeficiente de Silhueta* foi relativamente baixo para ambos os algoritmos, indicando que existe alguma sobreposição entre alguns grupos e que os *clusters* não estão fortemente separados.

- *Dataset de doentes CRC*: mais uma vez, para este conjunto de dados, o *K-Means* apresentou melhores resultados em todas as métricas, mas o *Coeficiente de Silhueta* foi relativamente baixo para ambos os algoritmos, pelo que as conclusões são bastante semelhantes às do *dataset* para perfis de aminoácidos com vários diagnósticos.

- *Dataset* de doentes CRC em estágio M0: o *K-Means* foi novamente o algoritmo com melhor resultado, repetindo-se os padrões observados nos conjuntos anteriores.

Para concluir esta secção, resta apenas mencionar dois detalhes-chave.

Em primeiro lugar, é importante mencionar que, de uma forma geral, à medida que os grupos foram restringidos de populações mais heterogéneas para populações mais homogéneas de doentes de cancro colorretal e doentes de cancro colorretal em estágio M0, evidenciou-se uma redução dos valores das métricas, o que demonstra uma maior dificuldade em identificar *clusters* nestes conjuntos com diferenças metabólicas mais ligeiras.

De seguida, nos perfis de aminoácidos, os valores de *Coeficiente de Silhueta* foram bastante inferiores aos obtidos para os perfis de acilcarnitinas, o que sugere uma maior complexidade na identificação de padrões distintivos nestes perfis. Contudo, os valores do *Índice de Calinski-Harabasz* foram, maioritariamente, superiores nos *datasets* de aminoácidos, o que indica uma melhor separação global entre *clusters*, mesmo que internamente não tão coesos.

Em suma, com base neste estudo, o *K-Means* foi identificado como o algoritmo com melhor desempenho global, apresentando *clusters* com qualidade razoável a forte nos *datasets* mais heterogéneos e qualidade mais fraca nos subconjuntos mais restritos e homogéneos.

Mais ainda, e embora ambos os algoritmos tenham tido desempenhos semelhantes, a análise das métricas revela diferenças relevantes. De facto, o *K-Means* tende a apresentar um *Índice de Calinski-Harabasz* mais elevado, o que sugere uma segmentação mais clara dos doentes, e o *GMM* alcança melhores valores no *Coeficiente de Silhueta* e *Índice de Davies-Bouldin*, o que indica uma maior consistência interna dos grupos. Neste sentido, o *K-Means* apresenta-se como o algoritmo mais adequado para o objetivo de identificação de subgrupos bem demarcados.

5.2 Interpretação Biológica dos Resultados

Para além do estudo de *clustering*, ao longo deste projeto, tentou-se fazer uma análise da relevância biológica dos resultados obtidos. É neste sentido que, ao longo do Capítulo 4, são apresentados diferentes *heatmaps* com os valores dos metabólitos com maior variabilidade em cada conjunto de dados e estudos acerca da correlação do *cluster* e do diagnóstico (no caso dos *datasets* com dados de vários diagnósticos) ou ainda estudo acerca da distribuição de estágios do CRC com base no *cluster* da amostra ou análises longitudinais para os doentes em estágio M0.

Com base nisto, em relação aos perfis de acilcarnitinas, foi possível observar que no *dataset* de CRC existem níveis mais elevados de `C2` e `C18:1` no *Cluster* `1`, o que, de acordo com a literatura atual, pode indicar uma maior resistência ao tratamento, visto que alterações na β-oxidação mitocondrial têm sido associadas a uma maior resistência ao tratamento (Y. J. Li et al. 2022).

Em contraste, o *Cluster* `0` mostrou níveis reduzidos destes mesmos metabólitos, sugerindo a presença de um subgrupo com um perfil metabólico potencialmente menos agressivo e com menor disfunção mitocondrial.

Já no caso do *dataset* de doentes em estágio M0, o *cluster* com níveis mais elevados de acilarnitinas de cadeia longa como o *C16* e *C18* é o *Cluster* `2`. Contudo, a partir da análise longitudinal, verificou-se que este grupo não apresenta qualquer progressão direta de `M0` - pré-quimioradioterapia - para `M2` - pós-cirurgia - e apresenta ainda menor progressão de `M1` - pós-quimioradioterapia - para `M2`, o que indica que este aumento de valores não está necessariamente associado a uma má resposta à quimioradioterapia, sendo potencialmente necessária a integração dos marcadores dos perfis metabólicos de acilcarnitinas com outros marcadores clínicos ou moleculares.

Por fim, quanto ao *dataset* de vários diagnósticos para perfis de acilcarnitinas, foi obtido um *V de Crammer* superior a 0.60 para ambos os algoritmos, o que se trata de uma associação forte entre o diagnóstico e *cluster* identificado, pelo que estes resultados reforçam o potencial uso destes metabólitos como marcadores no processo de diagnóstico diferencial.

Já relativamente aos perfis de aminoácidos, a análise do *dataset* de doentes em estágio `M0` também revelou padrões metabólicos com possíveis implicações clínicas relevantes. Entre estes, por exemplo, o *Cluster* `5` do *K-Means* e o *Cluster* `4` do *GMM* apresentam um aumento de aminoácidos essenciais e não essenciais como *ILE*, *LEU* e *CYSTA*, que estão, por vezes, associados a uma maior proliferação celular e, consequentemente, maior taxa de crescimento do cancro (Akbay et al. 2024; Jiménez-Alonso e López-Lázaro 2023). No entanto, este *cluster* não apresentou nenhuma amostra longitudinal com progressão direta de `M0` para `M2` e apresentou uma das progressões mais baixas entre *clusters* de `M1` para `M2`, pelo que não é possível concluir que existe uma associação direta entre estes padrões metabólicos e uma progressão clínica mais agressiva.

Finalmente, foi calculado o *V de Crammer* para a associação entre diagnósticos e *cluster* para o conjunto de dados de vários diagnósticos de perfis de aminoácidos. E, neste caso, foi obtido de cerca de 0.30 para ambos os algoritmos. Este trata-se de um resultado bastante pior do que o obtido para os perfis de acilcarnitinas, mas ainda indica uma associação moderada e que deve ser analisada entre as duas variáveis.

Em suma, ambos os resultados indicam padrões metabólicos diferenciados para os diferentes *clusters*. Adicionalmente, foi identificada uma associação moderada ou forte entre

diagnósticos e *clusters* para ambos os perfis, o que indica que estes marcadores podem ser importantes no processo de diagnóstico. No entanto, não foi encontrada uma associação direta entre a progressão clínica e os perfis metabólicos obtidos, pelo que se concluiu que pode ser necessária a integração destes dados com outros dados clínicos de forma a obter uma melhor estratificação dos doentes e uma melhor compreensão do prognóstico.

5.3 Implicações Clínicas

Os resultados deste estudo apresentam algumas implicações clínicas relevantes. Assim, a identificação de padrões metabólicos em subgrupos de doentes através de técnicas de *clustering* e aplicada a perfis de acilcarnitinas e aminoácidos poderá facilitar a estratificação dos doentes, permitindo:

- Identificar metabólitos que possam estar mais associados a uma maior agressividade tumoral ou a respostas diferenciadas ao tratamento neoadjuvante, tal como sugerido pelos níveis alterados de, por exemplo, *C16*, *C18*, *LEU* e *ILE*.
- Apoiar o processo de diagnóstico, tal como sugerido pelas associações fortes entre diagnósticos e *clusters* obtidos nos *datasets* com vários diagnósticos.

Contudo, para que estas abordagens sejam efetivamente úteis para a prática clínica, será essencial garantir:

- A validação externa dos modelos com mais dados de diferentes Centros de Investigação ao invés de utilizar-se apenas dados da Unidade Local de Saúde de Santo António.
- O desenvolvimento de uma interface que permita aos profissionais de saúde interagir, aplicar e compreender os resultados do modelo em tempo real.
- A garantia de que os modelos desenvolvidos são transparentes e interpretáveis, de modo a assegurar a confiança dos profissionais de saúde.

Neste sentido, o presente trabalho apresenta um contributo inicial para a caracterização e uso não supervisionado de perfis metabólicos para o processo de diagnóstico e decisão terapêutica. Contudo, ainda existe trabalho a ser realizado em projetos futuros para concluir o trabalho feito e garantir a sua relevância e aplicabilidade em contexto real.

5.4 Validação pelos Profissionais de saúde

De modo a aumentar a qualidade do trabalho realizado, e tal como mencionado na Seção 1.3, foi efetuada a validação dos resultados obtidos com profissionais de saúde, e em particular médicos e investigadores da Unidade Local de Saúde de Santo António. Esta validação foi conduzida através de reuniões periódicas durante as quais foram apresentados:

- Relatórios detalhados contendo a descrição dos *clusters* identificados, os metabólitos mais discriminantes em cada grupo e estatísticas da distribuição das características clínicas por *cluster*.
- Visualizações gráficas dos resultados, incluindo *heatmaps* dos metabólitos mais discriminantes, representações do *clustering* a partir de métodos de redução de dimensionalidade (*PCA*, *t-SNE*, *LDA*), bem como gráficos baseados em rácios de metabólitos de vias metabólicas específicas, como a β-oxidação.

Nestas reuniões, os profissionais de saúde forneceram *feedback* qualitativo sobre os grupos identificados.

Assim, de um modo geral, o *feedback* recebido ao longo destas reuniões foi bastante positivo, e este estudo permitiu aos profissionais de saúde identificarem nos resultados padrões que consideraram relevantes e, em alguns casos, inesperados, o que reforça o valor deste tipo de abordagem exploratória. No entanto, devido a limitações de tempo, a análise por parte da equipa clínica ainda não está concluída, sendo este um processo que está previsto ser continuado ao longo dos próximos meses.

A partir deste *feedback*, foi ainda identificada uma limitação que se trata da necessidade de desenvolvimento de um modelo de IA explicável, de forma a facilitar a interpretação dos resultados e garantir que estes modelos são auditáveis.

5.5 Limitações do estudo

Apesar dos resultados promissores obtidos ao longo deste projeto, é importante reconhecer as limitações do mesmo, quer a nível técnico e metodológico, como a nível clínico. A identificação e compreensão destas limitações permitem contextualizar os resultados, evitar que sejam feitas generalizações incorretas e que se orientem futuras possibilidades de investigação.

Como tal, a primeira limitação identificada e a mais significativa trata-se do baixo número de amostras, particularmente nos subgrupos de doentes com cancro colorretal e doentes com cancro colorretal em estágio M0. De facto, o *dataset* mais restrito contém um total de noventa e nove amostras. Este baixo número de amostras pode comprometer a robustez estatística do modelo, comprometendo a capacidade de generalizar os resultados obtidos.

Adicionalmente, e ainda relativamente aos *datasets* utilizados, todos os dados utilizados neste estudo provêm da Unidade Local de Saúde de Santo António, o que pode introduzir vieses relacionados com protocolos e práticas clínicas utilizados na recolha de dados.

Outra limitação, que já foi brevemente mencionada ao longo deste Capítulo 5, trata-se do facto de que esta análise focou-se exclusivamente na análise de perfis metabólicos de aminoácidos e não integrou outros dados clínicos potencialmente relevantes, como um histórico clínico detalhado, imagens de ressonância magnética, entre outros. Esta limitação reduz a capacidade de interpretar os *clusters* gerados pelos modelos dentro de um contexto clínico real dos doentes.

Do ponto de vista técnico, a utilização de *PCA* para redução de dimensionalidade, embora tenha melhorado a interpretabilidade e reduzido o ruído, poderá ter conduzido à perda de informação relevante. Mais ainda, devido a limitações computacionais, a hiperparametrização foi restringida a 5.000 combinações, o que pode ter impedido a exploração de configurações com melhor desempenho.

Por fim, a análise longitudinal feita com base nos subconjuntos de dados de doentes com CRC no estágio M0 também se apresenta limitada. De facto, este conjunto de dados apresenta apenas trinta e sete doentes, o que restringe a robustez da análise da progressão da doença. Mais ainda, praticamente não existem dados de amostras em estágios mais adiantados (como o estágio `M3` e `M4`), o que limita ainda mais a interpretação desta análise.

5.6 Sugestões para futuros trabalhos

Tendo em conta os resultados obtidos e as limitações identificadas na Seção 5.5, fica evidente que existem oportunidades para expandir o trabalho realizado neste projeto e continuar a aplicação do mesmo em contextos mais amplos. Deste modo, nestas secções são apresentadas algumas sugestões para trabalhos futuros, quer com o objetivo de explorar novas direções metodológicas, como para melhorar a aplicabilidade clínica deste estudo.

Assim, primeiramente, e de modo a aumentar a robustez dos modelos, será importante aumentar o conjunto de dados utilizados. Em particular, deve-se tentar que:

- Sejam recolhidas mais amostras clínicas e, em especial, de subgrupos menos representados (amostras de CRC e, particularmente, de CRC em estádios avançados como `M3` e `M4`).
- Sejam incluídos dados de mais centros hospitalares de forma a reduzir a existência de vieses devido à forma como os dados são recolhidos e a fatores geográficos e demográficos.
- Sejam recolhidas mais amostras longitudinais (amostras do mesmo paciente em diferentes períodos de tempo), de modo a permitir uma análise a evolução dos perfis metabólicos ao longo do tempo e a forma como estes respondem ao tratamento.

Adicionalmente, e com o objetivo de tornar os modelos mais robustos, destaca-se que futuros estudos poderão beneficiar de uma abordagem mais abrangente que integre diversos tipos de dados clínicos e moleculares, como dados genómicos, imagens médicas e outros dados clínicos, como a existência de múltiplas condições, terapias anteriores, entre outros.

Mais relacionado com aspetos técnicos, recomenda-se que futuros estudos explorem outras opções técnicas que poderão ser úteis, tais como:

- Métodos de *clustering ensemble*, nos quais se combinam múltiplos algoritmos de *clustering* para aumentar a robustez e estabilidade dos grupos identificados.
- Abordagens de *deep learning*, como *Autoencoders* e *Deep Embedded Clustering*, visto que estas técnicas são mais adequadas para dados de alta dimensionalidade como os perfis metabólicos analisados.
- Técnicas de *Explainable AI* [1], de modo a criar modelos mais interpretáveis e que permitam aos profissionais de saúde entender as decisões e aumentar a sua confiança nos resultados.

Por fim, devem ainda ser tomados passos para que estes modelos sejam mais facilmente aplicados em contexto hospitalar. Entre estas medidas encontram-se, por exemplo, o desenvolvimento de uma aplicação *web* que permita o *upload* de amostras de perfis metabólicos de acilcarnitinas ou aminoácidos, a atribuição dos mesmos a um *cluster* e que forneça interpretações visuais dos resultados (por exemplo, identifique os principais metabólitos discriminantes, com *heatmaps* como os apresentados na Figura 4.2). Esta ferramenta permitiria, simultaneamente, que os modelos fossem testados em tempo real e que os dados fossem estruturados de forma mais consistente, o que facilitaria a sua integração clínica e melhoraria a avaliação do impacto da sua utilização no desempenho.

[1] **Explainable AI (xAI)**: conjunto de processos e métodos que permitem a humanos compreender e confiar em resultados e dados de saída criados por algoritmos de *Machine Learning* (xAI) (IBM 2025b).

5.7 Discussão Final

Neste capítulo, foram analisados e discutidos os resultados obtidos ao longo deste estudo.

Assim, a análise quantitativa revelou que o algoritmo *K-Means* apresentou, de uma forma geral, um desempenho superior nas diferentes métricas de avaliação dos *clusters*. Esta superioridade foi particularmente evidente em conjuntos com amostras de diferentes diagnósticos, mas em subconjuntos mais restritos com amostras de doentes com cancro colorretal ou cancro colorretal no estádio M0, os resultados foram piores, o que reflete a dificuldade de separação entre grupos mais homógeneos. Esta tendência foi observada tanto em perfis de acilcarnitinas como de aminoácidos, mas os perfis de acilcarnitinas demonstraram mesmo assim uma melhor definição dos *clusters*.

Já quanto às implicações biológicas e clínicas, este estudo mostrou que alguns *clusters* estavam associados a variações metabólicas relevantes como *C16*, *C18*, *LEU* e *ILE*. Estes são referidos na literatura como potenciais marcadores de agressividade tumoral e resposta ao tratamento neoadjuvante, mas no contexto deste estudo não foi identificada uma associação clara entre esses padrões metabólicos e a progressão clínica.

Adicionalmente, foi feito um trabalho de validação iterativa com o apoio dos profissionais de saúde, sendo que esta validação permitiu identificar que existe uma necessidade de aumentar a interpretabilidade e explicabilidade dos modelos. A par desta limitação, foram identificadas outras limitações, como o número reduzido de amostras, a proveniência única dos dados, o que restringe a generalização dos resultados. Estas limitações devem ser tidas em conta em futuros trabalhos.

Em suma, os resultados obtidos reforçam a relevância dos perfis metabólicos como fonte de informação biomédica, mas também evidenciam a necessidade de abordagens mais integradas, interpretáveis e orientadas para o contexto clínico. Como tal, foi possível identificar padrões metabólicos distintos em doentes com cancro colorretal através de técnicas de *clustering*, cumprindo o objetivo de caracterização dos perfis de aminoácidos e acilcarnitinas. Estes perfis revelaram associações com características clínicas relevantes, apoiando a hipótese de que tais marcadores podem contribuir para a predição da resposta ao tratamento neoadjuvante. Adicionalmente, a validação junto de profissionais de saúde, embora limitada nesta fase, forneceu *feedback* valioso quanto à relevância e aplicabilidade dos resultados. Por fim, as limitações discutidas reforçam a necessidade de trabalhos futuros, mas não invalidam que os objetivos da dissertação tenham sido globalmente alcançados.

Capítulo 6

Conclusões

O trabalho apresentado nesta dissertação teve como objetivo analisar perfis metabólicos de aminoácidos e acilcarnitinas, com o intuito de identificar padrões e potenciais grupos clínicos, através do uso de técnicas de *clustering*.

Deste modo, na Seção 3.2 foram identificadas algumas questões de investigação que serviram de orientação para o desenvolvimento do estudo e permitiram estruturar a componente metodológica e a discussão dos resultados.

Assim, na primeira questão de investigação, procurava-se compreender a estrutura de agrupamentos dos perfis metabólicos na totalidade da população em estudo - incluindo diversos diagnósticos. Nos perfis de acilcarnitinas com 865 amostras, o *K-Means* identificou quatro *clusters* e o *GMM* três, com distribuições assimétricas. Já para os perfis de aminoácidos com 4052 amostras, ambos os algoritmos identificaram cinco *clusters*. Adicionalmente, no caso dos perfis de acilcarnitinas foi identificada uma forte associação entre *clusters* e diagnósticos clínicos (para o *GMM* foi obtido um *V de Crammer* de 0.64 e para o *K-Means* um valor de 0.62), e para os aminoácidos foi identificada uma associação moderada (com valores de 0.32 e 0.30, respetivamente). Estes padrões confirmam que existem agrupamentos naturais que refletem diferenças patológicas subjacentes, independentemente do conhecimento prévio do diagnóstico.

No que se refere à Questão 2, esta era direcionada para a análise de doentes com cancro colorretal (CRC) e os resultados mostraram que estes perfis não constituem um grupo homógeneo do ponto de vista metabólico. Pelo contrário, no caso dos perfis de acilcarnitinas, foram identificados três *clusters* bem definidos tanto através do uso do *K-Means* e do *GMM*, e para os perfis de aminoácidos foram identificados oito *clusters* pelo *K-Means* e o *GMM* convergiu para três. Adicionalmente, a análise dos metabólitos mais discriminantes permitiu identificar que alguns grupos apresentavam aumentos em acilcarnitinas de cadeia longa (*C16* e *C18*), e outros de aminoácidos como a valina, prolina, alanina, leucina e isoleucina, o que sugere alterações nas vias de β-oxidação e metabolismo proteico. Estas são alterações associadas na literatura a cancros mais agressivos e resistentes ao tratamento, mas ao longo deste estudo não foi possível validar uma associação entre estes grupos e a gravidade da patologia. Contudo, este resultado é relevante, pois indica que o cancro colorretal pode apresentar assinaturas metabólicas variadas, potencialmente relacionadas com a heterogeneidade da doença e com diferentes trajetórias clínicas.

A última questão apresentada na Seção 3.2 centrava-se na análise de doentes com cancro colorretal em estádio `M0` e pretendia analisar a existência de subgrupos metabólicos que permitissem identificar uma associação entre os perfis e a evolução clínica. Nesta análise, para acilcarnitinas foram identificados três *clusters* por ambos os algoritmos, e nos aminoácidos

seis. Contudo, a análise longitudinal de 37 doentes com múltiplas amostras não revelou associações claras entre os *clusters* identificados e a evolução clínica.

Tendo em conta os resultados, reconhecem-se assim algumas limitações que condicionam a generalização dos mesmos. Entre estas incluem-se o número reduzido de amostras utilizadas, particularmente nos subgrupos mais específicos, como conjunto de amostras de doentes com cancro colorretal no estágio M0 ou doentes com múltiplas amostras recolhidas para análise longitudinal. Adicionalmente, os dados foram provenientes de uma única fonte de dados (Unidade Local de Saúde de Santo António), o que pode introduzir vieses e focou-se exclusivamente em perfis metabólicos, não integrando outros dados clínicos que poderiam ser relevantes, como imagiologia ou histopatologia.

No entanto, reconhece-se que este trabalho apresenta algumas contribuições significativas, uma vez que permitiu demonstrar que algoritmos de aprendizagem não supervisionada têm o potencial para identificar padrões clinicamente relevantes em dados reais, e permitiu a identificação de alguns metabólitos discriminantes em perfis de aminoácidos e acilcarnitinas em doentes com cancro colorretal.

Como tal, futuras investigações podem partir deste projeto e continuar o seu desenvolvimento, a partir do uso de um conjunto de dados mais robustos em que os dados sejam fornecidos por múltiplos centros hospitalares ou sejam integrados outros tipos de dados, como imagens de ressonância magnética ou o histórico clínico dos doentes. Do ponto de vista técnico, é sugerido que se explorem algoritmos de *deep learning* e técnicas de *clustering ensemble*, uma vez que estas poderão melhorar a qualidade dos agrupamentos.

Adicionalmente, e visto que este é um trabalho que integra duas áreas diferentes e exige o contributo de profissionais com conhecimentos técnicos diferentes, sugere-se a utilização de modelos de IA explicáveis, uma vez que poderão ajudar a aumentar a confiança dos profissionais de saúde nos resultados.

Por último, é sugerido o desenvolvimento de uma plataforma *web* ou um módulo clínico que permita que os profissionais de saúde usem estes modelos em tempo real, o que facilitará a avaliação do impacto da utilização dos mesmos.

Para além dos resultados obtidos e das sugestões técnicas apresentadas, é ainda importante refletir acerca do impacto deste trabalho.

Assim, de uma forma geral, considera-se que este trabalho demonstra o potencial da utilização de *Machine Learning* na área da oncologia e, em particular, na investigação do cancro colorretal.

Embora os desafios identificados exijam que seja feita uma investigação adicional, os resultados obtidos demonstram o potencial dos perfis metabólicos de acilcarnitinas e aminoácidos como fonte de informação clínica valiosa.

Deste modo, conclui-se que os objetivos iniciais foram globalmente alcançados, tendo sido possível identificar *clusters* metabólicos em doentes com cancro colorretal. Estes resultados foram validados pelos profissionais de saúde e servem como uma base para futuros estudos realizados na mesma área.

Bibliografia

Abdulwahid, Ahmed (fev. de 2025). *Scikit-Learn vs. PyTorch vs. Spark: The Ultimate Battle | by Ahmed Abdulwahid | Medium*. url: https://medium.com/@ahmedabdulwahid.data/scikit-learn-vs-pytorch-vs-spark-the-ultimate-battle-%EF%B8%8F-c3f5ea6845e9 (acedido em 03/08/2025).

Ahmed, Mohiuddin, Raihan Seraj e Syed Mohammed Shamsul Islam (ago. de 2020). *The k-means algorithm: A comprehensive survey and performance evaluation*. doi: 10.3390/electronics9081295.

Akbay, Burkitkan et al. (nov. de 2024). *Double-Edge Effects of Leucine on Cancer Cells*. doi: 10.3390/biom14111401.

Associate, Subu Surendran (2015). *A Review of Various Linear and Non Linear Dimensionality Reduction Techniques*. url: www.ijcsit.com.

Aung, Yuri Y.M., David C.S. Wong e Daniel S.W. Ting (set. de 2021). *The promise of artificial intelligence: A review of the opportunities and challenges of artificial intelligence in healthcare*. doi: 10.1093/bmb/ldab016.

Badillo, Solveig et al. (abr. de 2020). «An Introduction to Machine Learning». Em: *Clinical Pharmacology and Therapeutics* 107 (4), pp. 871–885. issn: 15326535. doi: 10.1002/cpt.1796.

Al-Bakheit, Ala'a et al. (out. de 2016). «Accumulation of Palmitoylcarnitine and Its Effect on Pro-Inflammatory Pathways and Calcium Influx in Prostate Cancer». Em: *Prostate* 76 (14), pp. 1326–1337. issn: 10970045. doi: 10.1002/pros.23222.

Ballesteros, John (out. de 2023). *Weka: How to learn Machine Learning for Non-Experts | by John R. Ballesteros | Medium*. url: https://medium.com/@jrballesteros/weka-how-to-learn-machine-learning-for-non-experts-70e9767b08b2 (acedido em 03/08/2025).

Battini, S. et al. (mar. de 2017). «Metabolomics approaches in pancreatic adenocarcinoma: Tumor metabolism profiling predicts clinical outcome of patients». Em: *BMC Medicine* 15 (1). issn: 17417015. doi: 10.1186/s12916-017-0810-z.

Blanco, Antonio e Gustavo Blanco (jan. de 2022). «Proteins». Em: *Medical Biochemistry*, pp. 21–75. doi: 10.1016/B978-0-323-91599-1.00004-3. url: https://linkinghub.elsevier.com/retrieve/pii/B9780323915991000043.

Bohr, Adam e Kaveh Memarzadeh (jan. de 2020). «The rise of artificial intelligence in healthcare applications». Em: Elsevier, pp. 25–60. isbn: 9780128184387. doi: 10.1016/B978-0-12-818438-7.00002-2.

Boyko, Nataliya e O. Tkachyk (mai. de 2023). «Hierarchical clustering algorithm for dendrogram construction and cluster counting». Em: *Informatics and mathematical methods in simulation* 13, pp. 5–15. doi: 10.15276/imms.v13.no1-2.5.

Bychkov, Dmitrii et al. (dez. de 2018). «Deep learning based tissue analysis predicts outcome in colorectal cancer». Em: *Scientific Reports* 8 (1). issn: 20452322. doi: 10.1038/s41598-018-21758-3.

Cantu, Jesús (jun. de 2023). *A Guide to Choosing the Right Python Machine Learning Library | by Jesús Cantú | Medium*. url: https://medium.com/@jesus.cantu217/a-guide-to-

choosing-the-right-python-machine-learning-library-27a3d556526e (acedido em 13/12/2024).

Carrasco, Oscar (fev. de 2024). *Gaussian Mixture Model Explained | Built In*. url: https://builtin.com/articles/gaussian-mixture-model (acedido em 08/12/2024).

Chen, Hao et al. (dez. de 2023). «Feature selection based on unsupervised clustering evaluation for predicting neoadjuvant chemoradiation response for patients with locally advanced rectal cancer». Em: *Physics in Medicine and Biology* 68 (23). issn: 13616560. doi: 10.1088/1361-6560/ad0d46.

Cherry, Kendra (dez. de 2023). *What Is a Longitudinal Study?* url: https://www.verywellmind.com/what-is-longitudinal-research-2795335 (acedido em 11/08/2025).

Chia, Austin (abr. de 2025). *Clusterização hierárquica: Visão geral do conceito com exemplos | DataCamp*. url: https://www.datacamp.com/pt/tutorial/hierarchical-clustering?dc_referrer=https%5C%3A%5C%2F%5C%2Fwww.google.com%5C%2F (acedido em 06/08/2025).

Dambrova, Maija et al. (jul. de 2022). *Acylcarnitines: Nomenclature, Biomarkers, Therapeutic Potential, Drug Targets, and Clinical Trials*. doi: 10.1124/pharmrev.121.000408.

Davis, Cindy D. e John Milner (jul. de 2004). *Frontiers in nutrigenomics, proteomics, metabolomics and cancer prevention*. doi: 10.1016/j.mrfmmm.2004.01.012.

Deng, Hongbo e Jiawei Han (2018). *Chapter 3 Probabilistic Models for Clustering*.

Dhiraj, K, James Skelton e Shaoni Mukherjee (ago. de 2025). *Anomaly Detection in Python with Isolation Forest | DigitalOcean*. url: https://www.digitalocean.com/community/tutorials/anomaly-detection-isolation-forest (acedido em 05/08/2025).

Eamonn, Abdullah Keogh e Mueen (2017). «Curse of Dimensionality». Em: ed. por Geoffrey I Sammut Claude e Webb. Springer US, pp. 314–315. isbn: 978-1-4899-7687-1. doi: 10.1007/978-1-4899-7687-1_192. url: https://doi.org/10.1007/978-1-4899-7687-1_192.

Fernandes, Maria Clara, Marc J. Gollub e Gina Brown (ago. de 2022). «The importance of MRI for rectal cancer evaluation». Em: *Surgical Oncology* 43, p. 101739. issn: 0960-7404. doi: 10.1016/J.SURONC.2022.101739.

Fernández, Lara P., Marta Gómez de Cedrón e Ana Ramírez de Molina (out. de 2020). *Alterations of Lipid Metabolism in Cancer: Implications in Prognosis and Treatment*. doi: 10.3389/fonc.2020.577420.

Fränti, Pasi e Sami Sieranoja (set. de 2019). «How much can k-means be improved by using better initialization and repeats?» Em: *Pattern Recognition* 93, pp. 95–112. issn: 00313203. doi: 10.1016/j.patcog.2019.04.014.

GeeksForGeeks (jul. de 2025). *Introduction to Weka: Key Features and Applications - GeeksforGeeks*. url: https://www.geeksforgeeks.org/machine-learning/introduction-to-weka-key-features-and-applications/#introduction-to-weka-in-data-mining (acedido em 03/08/2025).

Gevorkyan, Migran N. et al. (2019). *Review and comparative analysis of machine learning libraries for machine learning*. doi: 10.22363/2658-4670-2019-27-4-305-315.

Ghahramani, Zoubin (2004). «Unsupervised Learning». Em: ed. por Olivier Bousquet, Ulrike von Luxburg e Gunnar Rätsch. Springer Berlin Heidelberg, pp. 72–112. isbn: 978-3-540-28650-9. doi: 10.1007/978-3-540-28650-9_5. url: https://doi.org/10.1007/978-3-540-28650-9_5.

Hagland, Hanne R. et al. (jun. de 2013). *Molecular pathways and cellular metabolism in colorectal cancer*. doi: 10.1159/000347166.

Hiller, Karsten e Christian M. Metallo (fev. de 2013). *Profiling metabolic networks to study cancer metabolism*. doi: 10.1016/j.copbio.2012.11.001.

Hofmarcher, Thomas et al. (abr. de 2020). «The cost of cancer in Europe 2018». Em: *European Journal of Cancer* 129, pp. 41–49. issn: 18790852. doi: 10.1016/j.ejca.2020.01.011.

Hossain, Md Sanower et al. (abr. de 2022). *Colorectal Cancer: A Review of Carcinogenesis, Global Epidemiology, Current Challenges, Risk Factors, Preventive and Treatment Strategies*. doi: 10.3390/cancers14071732.

IBM (ago. de 2021). *CRISP-DM Help Overview - IBM Documentation*. url: https://www.ibm.com/docs/sr/spss-modeler/saas?topic=dm-crisp-help-overview (acedido em 16/11/2024).

– (jul. de 2025a). *Cramér's V - IBM Documentation*. url: https://www.ibm.com/docs/en/cognos-analytics/12.0.x?topic=terms-cramrs-v (acedido em 10/08/2025).

– (2025b). *What is Explainable AI (XAI)? | IBM*. url: https://www.ibm.com/think/topics/explainable-ai (acedido em 21/08/2025).

IPP (nov. de 2020). *Diário da República, 2.ª série PARTE E Artigo 2.º*.

Jha, Nirajan (mai. de 2024). *Understanding Feature Selection Techniques in Machine Learning | by NIRAJAN JHA | Medium*. url: https://medium.com/@nirajan_DataAnalyst/understanding-feature-selection-techniques-in-machine-learning-02e2642ef63e (acedido em 05/08/2025).

Jiménez-Alonso, Julio José e Miguel López-Lázaro (jul. de 2023). *Dietary Manipulation of Amino Acids for Cancer Therapy*. doi: 10.3390/nu15132879.

Jollife, Ian T. e Jorge Cadima (abr. de 2016). *Principal component analysis: A review and recent developments*. doi: 10.1098/rsta.2015.0202.

Kalimara (abr. de 2023). *Métricas de Agrupamento: Coeficiente de Silhueta, Índice de Davies-Bouldin e Índice de Calinski-Harabasz | by Kalimara | Medium*. url: https://medium.com/@kalimarapeleteiro/m%C3%A9tricas-de-agrupamento-coeficiente-de-silhueta-%C3%ADndice-de-davies-bouldin-e-%C3%ADndice-de-9462b87ce676 (acedido em 07/08/2025).

Karsh, Patrick (set. de 2023). *Challenges of Unsupervised Learning: Machine Learning Basics | by Patrick Karsh | Medium*. url: https://patrickkarsh.medium.com/challenges-of-unsupervised-learning-machine-learning-basics-b8025044be1f (acedido em 07/12/2024).

Kather, Jakob Nikolas et al. (2019). «Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study». Em: *PLoS Medicine* 16 (1). issn: 15491676. doi: 10.1371/journal.pmed.1002730.

Khan, Bangul et al. (set. de 2023). *Drawbacks of Artificial Intelligence and Their Potential Solutions in the Healthcare Sector*. doi: 10.1007/s44174-023-00063-2.

Li, Shangfu, Dan Gao e Yuyang Jiang (fev. de 2019). *Function, detection and alteration of acylcarnitine metabolism in hepatocellular carcinoma*. doi: 10.3390/metabo9020036.

Li, Yi Jia et al. (mai. de 2022). «Fatty acid oxidation protects cancer cells from apoptosis by increasing mitochondrial membrane lipids». Em: *Cell Reports* 39 (9). issn: 22111247. doi: 10.1016/j.celrep.2022.110870.

Li, Yuxi (fev. de 2022). «Reinforcement Learning in Practice: Opportunities and Challenges». Em: arXiv: 2202.11296 [cs.LG]. url: http://arxiv.org/abs/2202.11296.

Liu, Xianqiang et al. (nov. de 2024). «Multi-Algorithm-Integrated Tertiary Lymphoid Structure Gene Signature for Immune Landscape Characterization and Prognosis in Colorectal Cancer Patients». Em: *Biomedicines* 12 (11). issn: 22279059. doi: 10.3390/biomedicines12112644.

Maharana, Kiran, Surajit Mondal e Bhushankumar Nemade (jun. de 2022). «A review: Data pre-processing and data augmentation techniques». Em: *Global Transitions Proceedings* 3 (1), pp. 91–99. issn: 2666285X. doi: 10.1016/j.gltp.2022.04.020.

Mahesh, Batta (jan. de 2020). «Machine Learning Algorithms - A Review». Em: *International Journal of Science and Research (IJSR)* 9 (1), pp. 381–386. doi: 10.21275/art20203995.

Marengo, Emilio e Elisa Robotti (out. de 2014). *Biomarkers for pancreatic cancer: Recent achievements in proteomics and genomics through classical and multivariate statistical methods*. doi: 10.3748/wjg.v20.i37.13325.

McInnes, Leland e John Healy (mai. de 2017). *Accelerated Hierarchical Density Clustering*. doi: 10.1109/ICDMW.2017.12. url: http://arxiv.org/abs/1705.07321%20http://dx.doi.org/10.1109/ICDMW.2017.12.

Medical Oncology, European Society for (2016). *Cancro Colorretal Um Guia para o Doente*. Rel. téc. European Society for Medical Oncology. url: www.anticancerfund.orgwww.esmo.org.

Milne, P. J. e G. Kilian (jan. de 2010). «The Properties, Formation, and Biological Activity of 2,5-Diketopiperazines». Em: *Comprehensive Natural Products II: Chemistry and Biology* 5, pp. 657–698. doi: 10.1016/B978-008045382-8.00716-4.

Murel, Jacob e Eda Kavlakoglu (jan. de 2024). *What is Dimensionality Reduction? | IBM*. url: https://www.ibm.com/topics/dimensionality-reduction.

Nevil, Scott, David Kindness e Vikki Velasquez (jul. de 2025). *Z-Score: Meaning and Formula*. url: https://www.investopedia.com/terms/z/zscore.asp (acedido em 10/08/2025).

Nguyen, Lea (abr. de 2024). *Clustering for recognizing medical patterns: Gaussian Mixture Models explained » Lamarr-Blog*. url: https://lamarr-institute.org/blog/clustering-gaussian-mixture-models/ (acedido em 06/08/2025).

Noble, Joshua (2024). *What is Hierarchical Clustering? | IBM*. url: https://www.ibm.com/think/topics/hierarchical-clustering (acedido em 08/12/2024).

O que é o cancro colorretal? (2025). Cancro-Online. url: https://www.cancro-online.pt/cancro-colorretal/informacao-basica/o-que-e-o-cancro-colorretal/ (acedido em 04/08/2025).

Obi, Benjamin Tayo (2019). *Feature Selection and Dimensionality Reduction Using Covariance Matrix Plot | by Benjamin Obi Tayo Ph.D. | Towards AI*. url: https://pub.towardsai.net/feature-selection-and-dimensionality-reduction-using-covariance-matrix-plot-b4c7498abd07.

Ogunsanya, Michael, Joan Isichei e Salil Desai (2023). «Grid search hyperparameter tuning in additive manufacturing processes». Em: *Manufacturing Letters* 35. 51st SME North American Manufacturing Research Conference (NAMRC 51), pp. 1031–1042. issn: 2213-8463. doi: https://doi.org/10.1016/j.mfglet.2023.08.056. url: https://www.sciencedirect.com/science/article/pii/S221384632300113X.

Oti, Eric e Michael Olusola (jun. de 2024). «Overview of Agglomerative Hierarchical Clustering Methods». Em: *British Journal of Computer, Networking and Information Technology* 7, pp. 14–23. doi: 10.52589/BJCNIT-CV9POOGW.

Pankevičiūtė-Bukauskienė, Monika (2024). *HALLMARKS OF AMINO ACID METABOLISM IN BREAST CANCER CELLS*.

Papakyriakou, Dimitrios e Ioannis S. Barbounakis (jan. de 2022). «Data Mining Methods: A Review». Em: *International Journal of Computer Applications* 183 (48), pp. 5–19. doi: 10.5120/ijca2022921884.

Patil, Channamma e Ishwar Baidari (jun. de 2019). «Estimating the Optimal Number of Clusters k in a Dataset Using Data Depth». Em: *Data Science and Engineering* 4 (2), pp. 132–140. issn: 23641541. doi: 10.1007/s41019-019-0091-y.

Peng, Hao e Xiaoli Bai (dez. de 2019). «Comparative evaluation of three machine learning algorithms on improving orbit prediction accuracy». Em: *Astrodynamics* 3 (4), pp. 325–343. issn: 25220098. doi: 10.1007/s42064-018-0055-4.

Ramadan, Hassan Sayed et al. (2022). «A HEURISTIC NOVEL APPROACH FOR DETERMINATION OF OPTIMAL EPSILON FOR DBSCAN CLUSTERING ALGORITHM». Em: *Journal of Theoretical and Applied Information Technology* 15 (7). issn: 1817-3195. url: www.jatit.org.

Reddy, Y C A Padmanabha, P Viswanath e B Eswara Reddy (fev. de 2018). «Semi-supervised learning: a brief review». Em: *International Journal of Engineering & Technology* 7 (1.8), p. 81. doi: 10.14419/ijet.v7i1.8.9977.

Reichert, Ingo Junior (abr. de 2023). *Apache Spark Machine Learning: Como Funciona o MLlib | by Ingo Reichert Junior | Medium*. url: https://medium.com/@ingoreichertjr/apache-spark-machine-learning-como-funciona-o-mllib-2d9a800b7052 (acedido em 03/08/2025).

Rodriguez, Mayra Z. et al. (jan. de 2019). «Clustering algorithms: A comparative approach». Em: *PLoS ONE* 14 (1). issn: 19326203. doi: 10.1371/journal.pone.0210236.

Roeder, F et al. (2020). «Recent advances in (chemo-)radiation therapy for rectal cancer: a comprehensive review». Em: *Radiation Oncology* 15 (1), p. 262. issn: 1748-717X. doi: 10.1186/s13014-020-01695-0. url: https://doi.org/10.1186/s13014-020-01695-0.

Sah, Shagan (jul. de 2020). *Machine Learning: A Review of Learning Types*. doi: 10.20944/preprints202007.0230.v1. url: https://www.preprints.org/manuscript/202007.0230/v1.

Saltz, Jeffrey S. (2021). «CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps». Em: *Proceedings - 2021 IEEE International Conference on Big Data, Big Data 2021*. Institute of Electrical e Electronics Engineers Inc., pp. 2337–2344. isbn: 9781665439022. doi: 10.1109/BigData52589.2021.9671634.

Sánchez-Martínez, Ruth et al. (dez. de 2017). «Complementary ACSL isoforms contribute to a non-Warburg advantageous energetic status characterizing invasive colon cancer cells». Em: *Scientific Reports* 7 (1). issn: 20452322. doi: 10.1038/s41598-017-11612-3.

Schröer, Christoph, Felix Kruse e Jorge Marx Gómez (2021). «A systematic literature review on applying CRISP-DM process model». Em: *Procedia Computer Science*. Vol. 181. Elsevier B.V., pp. 526–534. doi: 10.1016/j.procs.2021.01.199.

Schubert, Erich et al. (jul. de 2017). «DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN». Em: *ACM Transactions on Database Systems* 42 (3). issn: 15574644. doi: 10.1145/3068335.

Scikit-learn (2025). *StandardScaler — scikit-learn 1.7.1 documentation*. url: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html (acedido em 04/08/2025).

scikitLearn (2024). *User Guide — scikit-learn 1.6.0 documentation*. url: https://scikit-learn.org/stable/user_guide.html (acedido em 13/12/2024).

Shapcott, Zoe (abr. de 2024). «An Investigation into Distance Measures in Cluster Analysis». Em: url: http://arxiv.org/abs/2404.13664.

Smiti, Abir (nov. de 2020). *A critical overview of outlier detection methods*. doi: 10.1016/j.cosrev.2020.100306.

Smyth, Joshua et al. (nov. de 2024). «Microbiome-Based Colon Cancer Patient Stratification and Survival Analysis». Em: *Cancer Medicine* 13 (22). issn: 20457634. doi: 10.1002/cam4.70434.

Sozuer, Secil (set. de 2015). «SURVIVABLE FIBER OPTICAL NETWORK DESIGN». Tese de doutoramento. Lehigh University.

Spark, Apache (2024). *Clustering - Spark 4.0.0 Documentation*. url: https://spark.apache.org/docs/latest/ml-clustering.html (acedido em 03/08/2025).

Sparkl (2025). *Revision Notes - Outliers & Resistant Measures | Exploring One-Variable Data | Statistics | Collegeboard AP | Sparkl*. url: https://www.sparkl.me/learn/collegeboard-ap/statistics/outliers-resistant-measures/revision-notes/413 (acedido em 04/08/2025).

Sterling, Thomas, Matthew Anderson e Maciej Brodowicz (2018). «Libraries». Em: *High Performance Computing*, pp. 313–345. doi: 10.1016/B978-0-12-420158-3.00010-1. url: https://linkinghub.elsevier.com/retrieve/pii/B9780124201583000101.

Tang, Tingxi et al. (jul. de 2024). «Plasma Metabolic Profiles-Based Prediction of Induction Chemotherapy Efficacy in Nasopharyngeal Carcinoma: Results of a Bidirectional Clinical Trial». Em: *Clinical Cancer Research* 30 (14), pp. 2925–2936. issn: 15573265. doi: 10.1158/1078-0432.CCR-23-3608.

Tian, Yanhua et al. (mai. de 2018). «Prediction of chemotherapeutic efficacy in non–small cell lung cancer by serum metabolomic profiling». Em: *Clinical Cancer Research* 24 (9), pp. 2100–2109. issn: 15573265. doi: 10.1158/1078-0432.CCR-17-2855.

Trifonova, Oxana P. et al. (jan. de 2023). *Current State and Future Perspectives on Personalized Metabolomics*. doi: 10.3390/metabo13010067.

Vettore, Lisa, Rebecca L. Westbrook e Daniel A. Tennant (jan. de 2020). *New aspects of amino acid metabolism in cancer*. doi: 10.1038/s41416-019-0620-5.

Wakefield, Katrina (jun. de 2023). *A guide to machine learning algorithms and their applications | SAS*. url: https://www.sas.com/cs_cz/insights/articles/analytics/machine-learning-algorithms-guide.html (acedido em 07/12/2024).

Wang, Hongzhi, Mohamed Jaward Bah e Mohamed Hammad (2019). «Progress in Outlier Detection Techniques: A Survey». Em: *IEEE Access* 7, pp. 107964–108000. issn: 21693536. doi: 10.1109/ACCESS.2019.2932769.

Wegmann, Marc et al. (2021). «A review of systematic selection of clustering algorithms and their evaluation». Em: *CoRR* abs/2106.12792. arXiv: 2106.12792. url: https://arxiv.org/abs/2106.12792.

Whig, Vandana et al. (2022). «An Empirical Analysis of Artificial Intelligence (AI) as a Growth Engine for the Healthcare Sector». Em: *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering, ICACITE 2022*. Institute of Electrical e Electronics Engineers Inc., pp. 2454–2457. isbn: 9781665437899. doi: 10.1109/ICACITE53722.2022.9823607.

Williams, Michael D. et al. (jun. de 2013). *Metabolomics of colorectal cancer: Past and current analytical platforms*. doi: 10.1007/s00216-013-6777-5.

Wirth, Rüdiger e Jochen Hipp (2000). *CRISP-DM: Towards a Standard Process Model for Data Mining*.

Wong, Annie et al. (jun. de 2023). «Deep multiagent reinforcement learning: challenges and directions». Em: *Artificial Intelligence Review* 56 (6), pp. 5023–5056. issn: 15737462. doi: 10.1007/s10462-022-10299-x.

Xi, Yue e Pengfei Xu (out. de 2021). *Global colorectal cancer burden in 2020 and projections to 2040*. doi: 10.1016/j.tranon.2021.101174.

Xu, Fen et al. (2023). «World Journal of Clinical Cases Characteristics of amino acid metabolism in colorectal cancer Specialty type: Oncology Provenance and peer review: Peer-review model: Single blind Peer-review report's scientific quality classification Grade A (Excellent): 0 Grade B (Very good): B Grade C (Good): C Grade D (Fair): 0 Grade E (Poor): 0». Em: *World J Clin Cases* 11 (27), pp. 6318–6326. issn: 2307-8960. doi: 10.12998/wjcc.v11.i27.6318. url: https://www.f6publishing.com.

Xu, Rui e Donald Wunsch (mai. de 2005). *Survey of clustering algorithms*. doi: 10.1109/TNN.2005.845141.

Xu, Yongjun et al. (nov. de 2021). *Artificial intelligence: A powerful paradigm for scientific research*. doi: 10.1016/j.xinn.2021.100179.

Yang, Ting, Minglun Ren e Kaile Zhou (ago. de 2018). *Identifying household electricity consumption patterns: A case study of Kunshan, China*. doi: 10.1016/j.rser.2018.04.037.

Yousef, Hibba, Samuel F. Feng e Herbert F. Jelinek (dez. de 2024). «Exploratory risk prediction of type II diabetes with isolation forests and novel biomarkers». Em: *Scientific Reports* 14 (1). issn: 20452322. doi: 10.1038/s41598-024-65044-x.

Yu, Yilin et al. (2023). *Novel insight into metabolic reprogrammming in cancer radioresistance: A promising therapeutic target in radiotherapy*. doi: 10.7150/ijbs.79928.

Zhang, Caiming e Yang Lu (set. de 2021). «Study on artificial intelligence: The state of the art and future prospects». Em: *Journal of Industrial Information Integration* 23, p. 100224. issn: 2452-414X. doi: 10.1016/J.JII.2021.100224.

Zhang, J e Y Yang (2023). «Density-Distance Outlier Detection Algorithm Based on Natural Neighborhood». Em: *axioms* 12, p. 425. doi: 10.3390/axioms. url: https://doi.org/10.3390/axioms12050425.

Zhang, Zhongheng et al. (fev. de 2017). «Hierarchical cluster analysis in clinical research with heterogeneous study population: Highlighting its visualization with R». Em: *Annals of Translational Medicine* 5 (4). issn: 23055847. doi: 10.21037/atm.2017.02.05.