INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA INFORMÁTICA




Aplicação de Machine Learning na identificação precoce da doença de Crohn e estratificação da sua gravidade com base em anomalias metabólicas.

SARA CARDOSO BORGES
Setembro de 2025

POLITÉCNICO
DO PORTO

ISEP INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO

Aplicação de Machine Learning na identificação precoce da doença de Crohn e estratificação da sua gravidade com base em anomalias metabólicas.

Sara Borges

Dissertação para obtenção do Grau de Mestre em Engenharia Informática, Área de Especialização em Engenharia Informática

Orientador: Professor Doutor José Reis Tavares
Co-Orientador: Professor Doutor Nuno Malheiro

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.

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

ISEP, Porto, 21 de setembro de 2025

Dedicatória

Agradeço ao meu orientador, Professor Doutor José Reis Tavares, e co-orientador, Professor Doutor Nuno Malheiro, pela orientação e apoio prestados ao longo da realização deste trabalho.

Ao meu namorado, por acreditar em mim mesmo quando a confiança me falha, por ser a minha maior motivação durante esta jornada e por nunca me deixar desistir.

Aos meus pais, por todo o amor incondicional, pela ajuda constante e pela confiança que depositam em mim, ajudando-me a tornar-me na melhor versão de mim mesma.

À minha irmã e eterna melhor amiga, pela presença constante e pela alegria que acrescenta todos os dias.

Aos meus colegas, pela partilha de conhecimento, pela amizade e por todos os momentos que tornaram esta caminhada mais bonita.

"Nunca deixes que os teus medos ocupem o lugar dos teus sonhos."
— Walt Disney

Resumo

Atualmente, com a quantidade crescente de dados disponíveis nos mais variados setores, tem emergido a necessidade de procurar abordagens tecnológicas capazes de analisar grandes volumes de dados. Assim, a *Machine Learning* (ML), um ramo da Inteligência Artificial (AI), tem-se tornado cada vez mais popular, permitindo aplicar algoritmos a conjuntos de dados com o intuito de prever resultados ou identificar relações entre as informações do *dataset*. As técnicas de ML têm vindo a ganhar popularidade do ramo da medicina, permitindo prever e monitorizar a evolução de determinadas doenças.

Este estudo pretende aplicar estratégias de ML a dados analíticos de casos com doença de Crohn, bem como de casos com doenças hereditárias do metabolismo, explorando o potencial dos perfis metabólicos como biomarcadores e a sua utilidade na estratificação de doentes. O estudo progrediu em várias etapas: pré-preparação dos dados; aplicação de modelos de classificação, avaliando a capacidade preditiva de diferentes algoritmos na distinção de diagnósticos; implementação de *clustering* no conjunto completo de amostras, visando a identificação de diferentes perfis metabólicos; e aplicação de *clustering* restrito a doentes com Crohn, analisando a heterogeneidade interna nesta população e sugerindo a existência de grupos mais predispostos a complicações, como o desenvolvimento de fístulas perianais.

Os resultado alcançados reforçam a importância da combinação da análise metabolómica com técnicas de ML, tanto na identificação de padrões metabólicos, como no potencial da utilização de biomarcadores no apoio do diagnóstico e acompanhamento da doença. Este trabalho contribui, assim, para evidenciar o papel da metabolómica e da Inteligência Artifical no avanço da medicina de precisão aplicada à Doença de Crohn.

Palavras-chave: Doença de Crohn, Fístulas Perianais, Perfis Metabólicos, Inteligência Artificial, Machine Learning

Abstract

Currently, with the growth of available data in a wide variety of sectors, there is an increasing demand for technological approaches capable of analysing substantial volumes of data. ML, a branch of AI, has become increasingly popular, making it possible to apply algorithms to data sets in order to predict results or identify relationships between information in the dataset. ML techniques have seen a surge in popularity within the medical field, with applications ranging from disease progression monitoring to predictive modelling.

The objective of this study is to implement ML methods on clinical data from patients diagnosed with Crohn's disease as well as patients with hereditary metabolic disorders, with the aim of investigating the potential of metabolic profiles as biomarkers and their efficacy in patient startification. The study progressed in several stages: firstly, data was prepared for analysis; secondly, classification models were applied, and the predictive capacity of different algorithms was evaluated in different diagnoses; thirdly, clustering was implemented on the complete set of samples to identify different metabolic profiles; and finally, clustering was applied exclusively to Crohn's disease patients, analysing the internal heterogeneity in this population and studying the possibility of the existence of subgroups with a higher predisposition to complications, such as the development of perianal fistulas.

The findings underscore the significance of integrating metabolomic analysis with machine learning methodologies, both in discerning metabolic patterns and in the prospective utilisation of biomarkers to facilitate diagnosis and disease monitoring. This work thus contributes to highlighting the role of metabolomics and artificial intelligence in advancing precision medicine applied to Crohn's disease.

Keywords: Crohn's Disease, Matabolomics, Metabolic Profiles, Artifical Intelligence, Machine Learning, Classification, Clustering, Biomarkers, Perianal Fistulas.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Acrónimos

AI	Inteligência Artificial.
ANOVA	*Analysis of Variance*.
AUC	Área sob a Curva.
BAIBA	Ácido β-aminoisobutírico.
BCAA	*Branched-Chain Amino Acids*.
CDC	*Centers for Disease Control and Prevention*.
CRISP-DM	Processo Padrão Interprofissional para Mineração de Dados.
DHM	Doenças Hereditárias do Metabolismo.
EHR	Registos Eletrónicos de Saúde.
EM	*Expectation-Maximization*.
EMT	Transição Epitelial-Mesenquimal.
FDR	*False Discovery Rate*.
GMM	*Gaussian Mixture Models*.
GPUs	*Graphics Processing Units*.
GWAS	*Genome-wide association study*.
IBD	Doença Inflamatória Intestinal.
kNN	*K-Nearest Neighbors*.
ML	*Machine Learning*.
MMPs	Metaloproteinases da matriz.
NCHS	*National Center for Health Statistics*.
NHANES	*National Health and Nutrition Examination Survey*.
NMF	Fatorização de matrizes não negativa.
PCA	Análise de Componentes Principais.
PCRAS	Proteína C-reativa de alta sensibilidade.
RGPD	Regulamento Geral de Proteção de Dados.
SMOTE	*Synthetic Minority Over-Sampling Technique*.

TGF-β	Fator de Crescimento Transformador Beta.
TNF-α	Fator de Necrose Tumoral Alfa.
TPUs	*Tensor Processing Units*.
uEVs	Vesículas extracelulares urinárias.
ULSSA	Unidade Local de Saúde de Santo António.

Capítulo 1

Introdução

No presente capítulo é apresentado o tema desta dissertação, elaborada no âmbito do mestrado de Engenharia Informática no Instituto Superior de Engenharia do Porto. Para o efeito, serão contemplados tópicos como o contexto e o problema em que se baseia a dissertação; os objetivos que se pretendem alcançar com o estudo realizado; a metodologia de trabalho utilizada e, por fim, a estrutura deste documento.

1.1 Contexto

A Inteligência Artificial (AI) é uma área bastante promissora, permitindo a conceção de soluções capazes de tornar as máquinas menos dependentes da intervenção humana. A sua aplicação possui como objetivos primordiais o aumento da eficiência dos processos e a minimização da introdução de erro humano, e tem-se demonstrado extremamente benéfica em várias áreas de negócio, como por exemplo no setor da saúde [1].

De forma a permitir o desenvolvimento de soluções para diferentes domínios, a Inteligência Artificial divide-se em múltiplos conceitos com funções distintas, sendo um deles a *Machine Learning* (ML). Focando-se no desenvolvimento de algoritmos e modelos de análise de dados, a ML permite conceder a habilidade de aprendizagem a sistemas computacionais, tornando-os capazes de formular previsões ou tomar decisões sem a necessidade de instruções explícitas.

Deste modo, esta ferramenta da AI, tem-se vindo a mostrar uma ótima aliada nos avanços da medicina, sendo que a sua aplicação é adequada em vários contextos diferentes da área [2]:

- suporta decisões clínicas quanto aos tratamentos a adotar (otimizando, também, a gestão de recursos hospitalares);
- possibilita o desenvolvimento de algoritmos capazes de prever o surgimento de determinada patologia podendo prever, também, de que forma evoluirá num dado doente;
- contribui para o aumento da qualidade de vida dos doentes, introduzindo metodologias de diagnóstico menos invasivas e mais rápidas.

O presente projeto integra-se ainda, no projeto de doutoramento da Doutora Ana Cristina Silva, *"Biomarcadores nas fístulas anais na Doença de Crohn: contribuição para o diagnóstico precoce, estratificação da gravidade da doença e previsibilidade da resposta à terapêutica"*, cuja orientação científica é assegurada pela Prof.ª Doutora Marisa Santos, MD, PhD (coordenadora da Unidade de Cirurgia Colorretal da Unidade Local de Saúde de Santo António (ULSSA) e do respetivo Centro de Referência do Tratamento do Cancro do Reto), pela

Doutora Paula Lago, MD (coordenadora da consulta multidisciplinar de Doença Inflamatória Intestinal no Serviço de Gastroenterologia da ULSSA) e pela Prof.ª Doutora Lúcia Lacerda, PhD (geneticista laboratorial da ULSSA e membro do Centro de Referência de Doenças Hereditárias do Metabolismo).

A análise de perfis metabólicos e a identificação de potenciais biomarcadores utilizando abordagens de ML, estão diretamente alinhadas com os objetivos propostos para o projeto de douturamento da Doutora Ana Cristina Silva, visando a melhoria dos processos de deteção precoce, estratificação da gravidade da doença e de previsão de resposta terapêutica em pacientes com fístulas anais associadas à Doença de Crohn. Os dados utilizados no desenvolvimento deste projeto foram obtidos do diagnóstico de:

- Doenças hereditárias do metabolismo pertencentes a casos estudados na Unidade de Bioquímica Genética, do Serviço de Genética Laboratorial, do Centro de Genética Médica Jacinto Magalhães, da Clínica de Genética e de Patologia da ULSSA;
- Doença de Crohn, de casos estudados na Unidade de Cirurgia Colorretal, do Serviço de Cirurgia Geral da ULSSA.

1.2 Descrição do problema

O número de doentes diagnosticados com Doença Inflamatória Intestinal (IBD) tem vindo a aumentar significativamente ao longo dos últimos anos, estimando-se que cerca de 7 milhões de pessoas em todo o mundo sejam portadoras desta patologia. A IBD caracteriza-se por ser um conjunto de dois tipos diferentes de inflamações intestinais, incluindo a Doença de Crohn e a Colite Ulcerosa. Enquanto a Doença de Crohn pode provocar inflamações em qualquer um dos componentes do tubo digestivo (desde a boca ao ânus), a Colite Ulcerosa apenas provoca inflamações e úlceras no intestino grosso [3]. Apesar de ainda não ter sido comprovada a causa exata do seu aparecimento, julga-se que estas patologias poderão ter origem numa disfunção do sistema imunitário [4].

De acordo com Fernando Magro et al. [5], que avaliou as despesas relacionadas com ações terapêuticas da Doença Inflamatória Intestinal no nosso país, estima-se que, por ano, o custo médio por doente seja de aproximadamente 6.075€, sendo que a grande maioria desse valor (cerca de 60%) é referente a terapias para a Doença de Crohn. Desta forma, existe a necessidade de melhorar os processos de decisão envolvidos no diagnóstico e tratamento da Doença de Crohn, tornando-os mais acessíveis aos utentes.

Um dos indicadores da Doença de Crohn são as fístulas anais[1], alertando para a presença da patologia num estado severo com risco elevado de destruição da região perianal, bem como da presença de tumores. Relativamente aos tratamentos existentes, o uso de células estaminais é bastante promissor, apresentando elevadas taxas de sucesso [7]. No entanto, esta solução apenas é viável para um grupo bem definido de indivíduos, uma vez que, para além de representar um custo bastante elevado, pode implicar algumas complicações como o surgimento de infeções ou a rejeição do tratamento pelo sistema imunitário do doente. Para além das células estaminais, o estudo de perfis metabólicos pode contribuir com padrões e marcadores relevantes e auxiliar no diagnóstico da doença, bem como no processo de decisão terapêutica.

[1] Fístulas Anais: Ligação anormal entre o canal anal e a superfície corporal [6].

1.3 Objetivos

O estudo exposto neste documento pretende analisar de que forma uma abordagem de Machine Learning pode trazer vantagens à monitorização da Doença de Crohn. Desta forma, o objetivo primordial desta dissertação passa pelo desenvolvimento e análise de algoritmos capazes de examinar um conjunto de informações clínicas e, através da aprendizagem automática, formular modelos de dados úteis para suportar a investigação de novas alternativas de diagnóstico e de previsão da evolução desta patologia. Para apoiar o desenvolvimento dos algoritmos de ML, serão utilizados dados reais fornecidos pela ULSSA, relativos a análises clínicas de utentes com Doença de Crohn e doentes com outras patologias. Assim, os objetivos propostos para este trabalho são:

- desenvolver algoritmos de classificação capazes de auxiliar na previsão da presença da Doença de Crohn, tendo como base o perfil metabólico dos doentes;
- aplicar algoritmos de aprendizagem não supervisionada com a finalidade de identificar diferentes grupos de resposta terapêutica com base no perfil metabólico.

1.4 Perguntas de Investigação

A definição de questões de investigação é um passo essencial em qualquer trabalho científico, permitindo clarificar os objetivos a alcançar e refletir acerca da metodologia mais adequada. No caso específico deste estudo, dedicado à aplicação de técnicas de ML na análise de dados clínicos com foco na metabolómica dos doentes, a formulação de *research questions* assume particular relevância, estabelecendo uma ligação entre as hipóteses exploratórias e a validação dos resultados obtidos.

Neste contexto, sublinham-se as seguintes questões de investigação:

- **RQ1**- Que algoritmos de Classificação se demonstram mais eficientes a realizar a distinção entre doentes de Crohn e controlos?
- **RQ2**- De que forma as técnicas de ML podem ser aplicadas a dados clínicos e metabólicos de doentes, de modo a prever a presença da Doença de Crohn?
- **RQ3**- De que forma se podem aplicar algoritmos de *clustering* para identificar subgrupos de doentes de Crohn com perfis metabólicos diferentes?
- **RQ4**- Que perfis metabólicos poderão estar associados a complicações da Doença de Crohn, como o desenvolvimento de fístulas perianais?
- **RQ5**- Qual a utilidade do desenvolvimento de modelos de ML aplicados em tarefas de classificação e *clustering* enquanto ferramentas de apoio ao diagnóstico e de monitorização da Doença de Crohn?

As questões enunciadas servem como ponto de partida para toda a investigação, permitindo estruturar a análise de forma coerente e assegurar que os resultados alcançados são relevantes no contexto de investigação da Doença de Crohn.

1.5 Considerações Éticas

Para sustentar o desenvolvimento do trabalho proposto, foram disponibilizadas informações clínicas reais de doentes de Crohn e de outras doenças hereditárias do metabolismo acompanhados na ULSSA. A utilização deste tipo de dados para o desenvolvimento de algoritmos de ML exige uma atenção rigorosa de forma a garantir a ética no tratamento dos dados, priorizando a sua privacidade e proteção. Desta forma, as informações cedidas pelo hospital foram filtradas, descartando informações de caráter pessoal que pudessem ser utilizadas para identificar algum doente. Assim, todas as etapas do trabalho desenvolvido estão em conformidade com o Regulamento Geral de Proteção de Dados (RGPD), garantindo que os direitos de privacidade dos indivíduos são devidamente respeitados.

Para além disso, outro princípio fundamental do desenvolvimento deste projeto é a transparência. Assim, os modelos desenvolvidos foram documentados de forma clara, promovendo a sua compreensão facilitada por profissionais de saúde e, possivelmente, por outros investigadores. São, também, apontadas algumas limitações dos modelos desenvolvidos, reconhecendo a complexidade dos dados clínicos utilizados.

1.6 Metodologia

Para desenvolver o projeto em estudo, optou-se pela aplicação do **Processo Padrão Interprofissional para Mineração de Dados (CRISP-DM)**, por se tratar de uma *framework* capaz de fornecer uma abordagem estruturada para trabalhos que incluam análise de dados [8].

O seu funcionamento consiste nas seguintes etapas [9]:

- **Compreensão do Negócio:** A primeira etapa passa por compreender as necessidades do cliente, de modo a definir os objetivos e requisitos do trabalho;
- **Compreensão dos Dados:** Nesta fase, é feita uma análise dos dados existentes, avaliando se são suficientes e adequados para alcançar os objetivos do projeto;
- **Preparação dos dados:** A preparação dos dados consiste na sua seleção, limpeza e formatação, obtendo os *datasets* que serão utilizados.
- **Avaliação:** Nesta etapa, é feito o estudo de possíveis métodos para desenvolver a solução, culminando na escolha dos modelos mais adequados ao problema;
- **Implantação:** Por fim, é feita uma ponderação acerca de que forma os modelos desenvolvidos poderão ser utilizados pelo cliente.

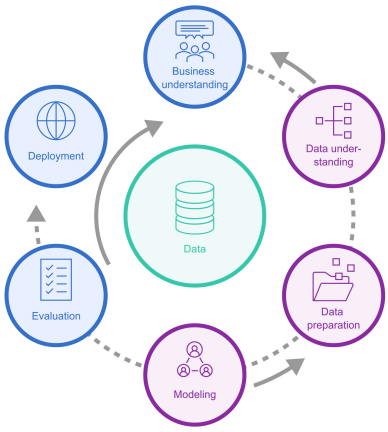

Figura 1.1: Fluxo do CRISP-DM [10]

1.7 Planeamento

De forma a fazer um planeamento adequado ao trabalho a desenvolver, recorreu-se a um diagrama de Gantt, apresentado na figura 1.2.

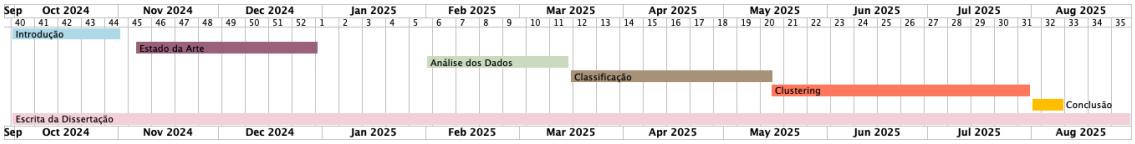

Figura 1.2: Diagrama de Gantt

Assim, a fase inicial do projeto passou pela análise e compreensão dos temas envolvidos na dissertação, incluindo estudos acerca dos trabalhos relacionados e das tecnologias aplicadas no projeto. De seguida, foi necessário o estudo do formato dos dados disponibilizados e do tratamento necessário para que pudessem ser processados por modelos de ML. Posteriormente, avançou-se para a etapa de desenvolvimento da solução, onde foram incluídas as fases de experimentação e avaliação dos modelos de Classificação e de *Clustering* implementados. Nesta etapa, foram realizadas reuniões com os profissionais de saúde da ULSSA envolvidos no projeto, de forma a alinhar o desenvolvimento do mesmo, com as suas necessidades e expectativas. Por fim, foi feita uma reflexão quanto ao trabalho desenvolvido como um todo, pesando a sua adequação para o problema em estudo.

1.8 Estrutura do Documento

O presente documento encontra-se repartido em cinco capítulos diferentes, de forma a permitir uma compreensão completa do contexto, análise, desenvolvimento e conclusões do estudo realizado:

- **Estado da Arte:**

  Neste capítulo, é apresentada uma análise detalhada dos principais temas abordados ao longo da dissertação. É, também, realizada uma revisão acerca dos principais trabalhos e estudos já existentes acerca de aplicações de Machine Learning no setor da saúde, explorando abordagens, métodos e soluções já existentes, proporcionando uma visão crítica do estado atual do tópico.

- **Pré-processamento dos Dados:**

  Este capítulo detalha o processo de análise dos dados fornecidos pela ULSSA, bem como a abordagem adotada para preparar os dados para que possam ser processados pelos modelos desenvolvidos.

- **Classificação:**

  Este capítulo foca-se na implementação e nos resultados obtidos durante a tarefa de Classificação, aplicando diferentes modelos às amostras de plasma e de urina disponibilizadas. Neste capítulo é ainda, realizada uma reflexão acerca dos resultados obtidos com as diferentes abordagens, comparando os modelos quanto à sua adequação à tarefa.

- **Clustering:**

  Neste capítulo é apresentada a análise das amostras com recurso a abordagens de *Clustering*, tendo como objetivo primordial, identificar perfis metabólicos associados à Doença de Crohn e, ainda, a possíveis complicações da mesma, de forma a auxiliar a previsão e monitorização da patologia. Para tal, a estratégia aplicada passou pela realização de duas tarefas distintas: *clustering* aplicado à totalidade das amostras, para identificar perfis associados à doença; e *clustering* somente nas observações de Crohn, de forma a investigar possíveis sub-perfis que possam estar relacionados a agravamentos na doença, como o surgimento de fístulas anais.

- **Conclusão:**

  Por fim, neste capítulo, é apresentada uma reflexão acerca do trabalho desenvolvido, sublinhando os objetivos cumpridos e possíveis melhorias a implementar no futuro.

Capítulo 2

Estado da Arte

No presente capítulo é analisado o estado da arte, expondo conceitos fundamentais para uma melhor compreensão do estudo realizado ao longo deste documento. Assim, e tendo em conta o tema desta dissertação, serão abordados os seguintes tópicos: Áreas de Estudo, Machine Learning (incluindo o estudo de diferentes ramos desta área), Aplicações de Machine Learning em *Healthcare* e, ainda, Tecnologias que poderão ser úteis para este trabalho.

2.1 Áreas de Estudo

Esta secção analisa as principais áreas de estudo incluídas no tema da dissertação, contribuindo para o entendimento do problema, destacando conceitos-chave.

2.1.1 Doença de Crohn

Juntamente com a **Colite Ulcerosa**, a **Doença de Crohn** é um dos principais componentes da **Doença Inflamatória Intestinal**, caracterizada por infeções persistentes a nível do sistema digestivo [11]. A Doença de Crohn qualifica-se como uma doença crónica inflamatória, capaz de afetar qualquer um dos componentes do tubo digestivo (desde a boca ao ânus), manifestando-se, por norma, através de dor e desconforto abdominal, febre e sinais clínicos de obstrução intestinal ou diarreia com presença de sangue. Sendo que a sua incidência e prevalência tendem a aumentar e, dada a sua natureza recorrente, esta patologia preocupa um grupo crescente de médicos e investigadores, focados na pesquisa da sua origem e no desenvolvimento de novos tratamentos. Atualmente, aponta-se para a existência de cerca de 7 milhões de casos de Doença de Crohn em todo o mundo, havendo uma tendência maior em países desenvolvidos, indicando uma possível ligação com as alterações das condições ambientais e, também, a nível da dieta da população [12].

Apesar de ainda não ter sido descoberta a origem exata da Doença de Crohn, acredita-se que pode ser resultante de interações complexas entre fatores genéticos e ambientais [13]:

- **Genética:**

  Ao longo do tempo, foram desenvolvidos inúmeros estudos com o objetivo de identificar fatores capazes de influenciar o surgimento e desenvolvimento da Doença Inflamatória Intestinal. Assim, *Genome-wide association study* (GWAS)[1] têm-se mostrado extremamente importantes na identificação de predisposições genéticas associadas à

[1] **GWAS:** abordagem de investigação que visa identificar variantes genéticas associadas ao risco de desenvolvimento de uma doença [14]

Doença de Crohn. Com o desenvolvimento deste tipo de investigação, foi possível identificar mais de 200 variantes genéticas associadas à patologia, como é o caso de mutações a nível do gene *NOD2*, que contribuem para o desenvolvimento da inflamações através de uma resposta imunológica desregulada. Este é apenas um dos vários exemplos de combinações genéticas que podem induzir comportamentos inflamatórios e complicações a nível intestinal [15].

- **Fatores Ambientais:**

  Vários estudos têm vindo a ser desenvolvidos com o fim de avaliar o impacto do ambiente circundante dos indivíduos no desenvolvimento de determinadas doenças. Resultados de estudos semelhantes com foco na IBD, apontam para um aumento global nas taxas de incidência da Doença de Crohn em grupos étnicos previamente menos afetados, como asiáticos e hispânicos, indicando uma crescente influência do estilo de vida no desenvolvimento da doença. Com a industrialização, por exemplo, surgiram várias alterações significativas nos hábitos dos cidadãos, incluindo o aumento do sedentarismo, da poluição no ar e do processamento alimentar. Todos estes fatores, aliados à crescente taxa de tabagismo (tanto ativo, como passivo), aumentam o risco de desenvolvimento da Doença de Crohn, através do impulsionamento de alterações na **microbiota intestinal**.

Embora ainda não tenha sido descoberta uma cura definitiva para a Doença de Crohn, existem diversos tratamentos que procuram proporcionar um estilo de vida normal aos doentes, controlando os sintomas e prevenindo possíveis complicações. Por norma, as terapias incluem a administração de antibióticos e de probióticos acompanhados por alterações na dieta e, também, no estilo de vida dos indivíduos, tendo em foco a diminuição dos níveis de stress (um dos fatores contribuintes para o agravamento da condição). Outra abordagem (ainda em avaliação) é o tratamento com recurso a células estaminais. Este tipo de tratamento tem-se mostrado promissor no contexto da doença, principalmente em casos em que se verifica a presença de fístulas perianais, uma complicação complexa da patologia. No entanto, para além de ainda ser necessária a condução de vários testes para provar a sua viabilidade, este tratamento apresenta um custo bastante elevado [16].

2.1.2 Fístulas Perianais na Doença de Crohn

Uma fístula anal (ou perianal) consiste no desenvolvimento de um canal anormal desde o interior do ânus até à superfície externa da pele. Por norma, o seu desenvolvimento é originado pela drenagem de infeções a nível das glândulas anais e a sua presença implica um risco acrescido de destruição do esfíncter e da zona perianal, bem como de neoplasia [17]. Um dos fatores de risco para o seu surgimento é o género, sendo que as mulheres apresentam um menor risco de desenvolvimento destas anomalias.

As fístulas anais são, também, uma complicação comum e debilitante da Doença de Crohn, representando um grande desafio no seu tratamento. Estima-se que cerca de 13,9% a 28,8% dos doentes possuam esta complicação, sendo que, para doentes de Crohn com presença de inflamações retais, a taxa de prevalência destas anomalias dispara para cerca de 92%. Apesar do risco de desenvolvimento de fístulas aumentar com a duração da doença, a sua deteção pode ocorrer em simultâneo com o diagnóstico da mesma, podendo, até, antecedê-lo. Desta forma, o estudo da sua origem e comportamento pode ser uma valiosa contribuição para o melhoramento dos processos de diagnóstico e tratamento da patologia [18].

Patogénese

A patogénese de fístulas perianais na Doença de Crohn é um processo complexo dependente de diversos fatores, envolvendo inflamações intestinais crónicas, defeitos epiteliais e disbiose microbiana [2] [18].

A presença de inflamação crónica característica da Doença de Crohn, conduz a um aumento de citocinas inflamatórias, como:

- **Fator de Necrose Tumoral Alfa (TNF-α):** incluído no grupo das glicoproteínas[3] , este fator foi, inicialmente reconhecido pela sua capacidade de impulsionar o desenvolvimento de necrose[4] tumoral. No entanto, investigações recentes alertaram para a contribuição patológica deste componente em doenças autoimunes [22].
- **Fator de Crescimento Transformador Beta (TGF-β):** caraterizado como uma citocina imunossupressora, este fator funciona como um regulador que desempenha diversas funções, sendo capaz de inibir ou estimular a proliferação celular [23]. O componente ativo do TGF-β liga-se ao seu recetor e regula as reações imunitárias da mucosa intestinal através da sinalização de TGF-β. Em doentes diagnosticados com IBD, foi possível observar um comportamento intestinal desregulado deste componente, ocorrendo sinalizações de TGF-β desajustadas, que podem contribuir para o agravamento da condição [24].

A ação de ambas as citocinas, em conjunto com os defeitos epiteliais, pode conduzir à Transição Epitelial-Mesenquimal (EMT). Este é um processo biológico segundo o qual as células epiteliais, tipicamente dispostas em camadas compactas e aderidas umas às outras, perdem as suas características, adquirindo traços mesenquimais, como maior mobilidade e capacidade invasiva. A EMT pode contribuir para o agravamento da Doença de Crohn, sendo que contribui para a disfunção da barreira intestinal, ao promover a perda da integridade do epitélio intestinal. Consequentemente, as células epiteliais intestinais perdem a sua capacidade adesiva, o que pode proporcionar o aumento da permeabilidade intestinal, facilitando a entrada de microorganismos na mucosa, promovendo a agravação de infeções e o desenvolvimento de fístulas [25] [26].

Adicionalmente, a disbiose microbiana também parece contribuir para a formação das fístulas [18]:

- a presença de inflamações intestinais pode ser originada e mantida por bactérias e pelos seus componentes, como o peptidoglicano [5], contribuindo para a formação de fístulas.
- as Metaloproteinases da matriz (MMPs) (enzimas envolvidas na degradação da matriz extracelular), costumam surgir em grandes quantidades nos canais fistulosos e contribuem no processo de danificação dos tecidos, de inflamação e, posteriormente, de formação de fístulas.

[2] **Disbiose Microbiana:** Indicador da existência de um desiquilíbrio entre os microorganismos do corpo humano [19].

[3] **Glicoproteína:** Proteína conjugada com cadeias de hidratos de carbono mais curtas e ramificadas, conhecidas como oligossacáridos [20].

[4] **Necrose:** Lesão celular irreversível, podendo provocar eventuais mortes celulares originadas por processos patológicos [21].

[5] **Peptidoglicano:** polímero constituído por açúcares extistente na membrana da maioria das bactérias [27].

Tratamento

O processo de tratamento de fístulas anais no contexto da Doença de Crohn é complexo e dependente de múltiplos fatores, como a localização, a severidade dos sintomas, a quantidade de canais fistulosos e o histórico de cirurgias locais anteriores, por exemplo. Dependendo dos casos, o tratamento pode incluir intervenções médicas, cirúrgicas ou uma combinação de ambas as abordagens [28]:

- **Terapia Médica:**

  Existem diversos medicamentos que podem ser administrados para tratar as fístulas perianais na Doença de Crohn:

  - **Antibióticos**: a combinação de fármacos como a ciprofloxacina e o metronidazol é uma das abordagens mais comuns, sendo que o metronidazol possui uma ação antimicrobiana e a ciprofloxacina é eficaz contra organismos gram-negativos [6].
  - **Imunomoduladores** [7]: vários estudos têm vindo a ser conduzidos, de forma a avaliar a ação deste tipo de medicamentos, que têm demonstrado uma excelente eficácia no tratamento de canais fistulosos no contexto da Doença de Crohn.

- **Terapia Cirúrgica:** a intervenção cirúrgica é frequentemente necessária no tratamento das fístulas na Doença de Crohn, nomeadamente no caso de fístulas complexas e na presença de doença retal ativa.

Em doentes de Crohn com tendência fistulosa, é essencial a análise detalhada do caso, sendo que a colaboração entre gastroenterologistas e cirurgiões é fundamental no processo de decisão da abordagem adequada a adotar.

2.1.3 Perfis Metabólicos

Os perfis metabólicos representam o conjunto completo de metabólitos presentes num organismo, tecido, célula ou fluido num determinado momento, refletindo a atividade metabólica e o estado fisiológico do sistema. A análise destes perfis potencia a identificação e avaliação da quantidade de metabólitos presentes num organismo, permitindo analisar de que forma o organismo reage sob determinadas condições específicas, como doenças, situações de stress ou intervenções terapêuticas. O recurso à análise de perfis metabólicos é uma técnica amplamente adotada, uma vez que pode facilitar processos como diagnósticos clínicos e monitorizações de tratamentos [31].

O presente estudo utilizou algoritmos de ML para extrair informações de análises clínicas, constituídas por uma variedade de aminoácidos. Os aminoácidos são moléculas que entram no processo de formação das proteínas e desempenham papéis importantes em diversos processos biológicos como na reparação de tecidos, na imunidade e no metabolismo energético [32]. Estes compostos têm uma grande relevância no contexto da Doença de Crohn, sobretudo em casos onde existe formação de fístulas anais, uma vez que contribuem para o processo de resposta inflamatória, cicatrização e manutenção da função intestinal.

[6] **Organismos gram-negativos:** bactérias que possuem uma estrutura única na sua parede celular. Destacam-se, por norma, por possuírem uma resistência elevada a antibióticos [29].

[7] **Imunomoduladores:** medicamentos com a capacidade de alterar o funcionamento do sistema imunitário, tornando-o mais eficiente. A sua administração tem-se mostrado bastante útil no tratamento de doenças autoimunes e cancerígenas [30].

Entre os aminoácidos relevantes no processo, destacam-se a glutamina e a arginina, uma vez que possuem funções imprescindíveis em processos inflamatórios e de cicatrização: a glutamina desempenha um papel essencial na regeneração de células intestinais; enquanto que a arginina participa na produção de óxido nítrico, intermediário na modulação da inflamação e na cicatrização dos tecidos [33].

Embora as investigações existentes acerca da Doença de Crohn realcem a glutamina e a arginina, esta dissertação irá explorar outros aminoácidos que possam ser relevantes na caracterização da patologia, através da aplicação de estratégias de *clustering*.

2.2 Machine Learning

Machine Learning (ML) é um ramo da AI focado no desenvolvimento de algoritmos e modelos de análise que permitem aos computadores aprender padrões e tomar decisões sem que sejam necessárias instruções específicas. Assim, ao contrário da programação tradicional (dependente de instruções *hard-coded*), esta vertente possui a capacidade de, através da análise de grandes *datasets*, descobrir padrões e relações existentes entre as informações analisadas [34].

Os algoritmos de ML organizam-se em diferentes categorias, tendo em conta o resultado que se pretende obter com a aplicação do algoritmo. Entre os tipos mais comuns de algoritmos, incluem-se [35]:

- **Aprendizagem Supervisionada:** os algoritmos geram uma função capaz de mapear um conjunto de dados de entrada rotulados nos *outputs* pretendidos. O uso deste tipo de algoritmos é comum na resolução de problemas de classificação.
- **Aprendizagem Não Supervisionada:** permite analisar e agrupar um conjunto de *inputs* não rotulados.
- **Aprendizagem Semi-Supervisionada:** permite analisar e classificar um conjunto de dados contendo entradas rotuladas e não rotuladas.
- **Aprendizagem por Reforço:** os algoritmos evoluem e aprendem como atuar com base na observação do mundo. Cada ação tem impacto no ambiente, que fornece um *feedback* necessário para orientar o processo de aprendizagem.

De seguida, são abordadas com mais detalhe, as categorias com mais relevância no contexto do trabalho a desenvolver.

2.2.1 Aprendizagem Supervisionada

A Aprendizagem Supervisionada é um ramo da ML no qual os algoritmos são treinados com dados rotulados para aprender a associar *inputs* a *outputs* específicos. Esta é uma técnica bastante aplicada em tarefas como classificação, regressão e previsão, sendo particularmente útil na resolução de problemas do mundo real, como na deteção de *spam*, análise de imagens médicas e planeamento cirúrgico [36].

Nas secções que se seguem, são detalhados alguns modelos de aprendizagem supervisionada.

Regressão Logística

A Regressão Logística é uma técnica de estatística aplicada para modelar a probabilidade de um resultado binário com base em pelo menos uma variável de previsão. Enquanto que a Regressão Linear prevê valores contínuos, a Regressão Logística faz a conversão dos seus resultados lineares aplicando a função logística (2.1) para limitar as previsões entre 0 e 1, representado probabilidades [37]. Assim, a Regressão Logística pode ser considerada uma extensão da Regressão Linear.

$$f(x) = \frac{1}{1 + e^{-x}} \tag{2.1}$$

Este é um método bastante útil para desenvolver soluções para problemas de classificação, podendo auxiliar na previsão da probabilidade de uma determinada condição médica, que é um dos objetivos do trabalho abordado nesta dissertação.

Árvores de Decisão

As árvores de decisão são um tipo de modelo preditivo usado no contexto de ML para realizar tarefas de classificação e regressão. O seu funcionamento passa pela divisão recursiva de um conjunto de dados em subconjuntos menores, baseando-se no valor das variáveis de *input*. Como resultado da sua aprendizagem, obtém-se um grafo em forma de árvore, contendo nós de decisão e folhas. Enquanto que a raiz representa o *dataset* inicial, as folhas representam os *outputs* finais ou as previsões elaboradas pelo modelo [38].

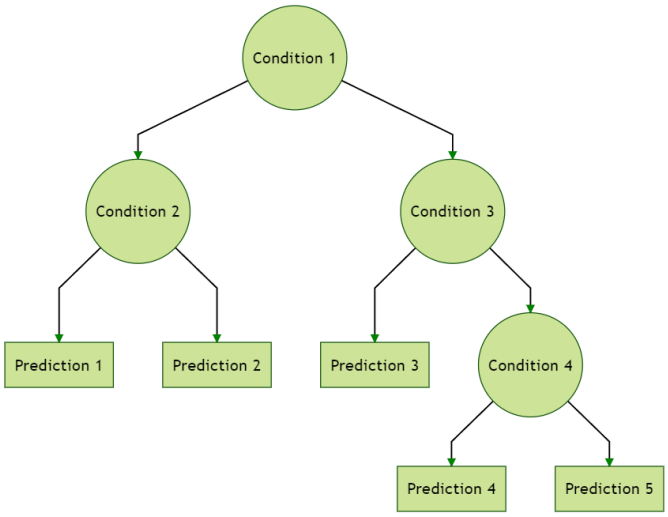

Figura 2.1: Exemplo de árvore de decisão [39]

K-Vizinhos Mais Próximos

O algoritmo k-vizinhos mais próximos (*K-Nearest Neighbors* (kNN)) é um método de aprendizagem supervisionada utilizado em tarefas de classificação e regressão. Tratando-se de um algoritmo não paramétrico, o kNN não infere nenhuma distribuição específica para os dados. Assim, o kNN classifica um novo caso através da distância entre o ponto a classificar e os k pontos do conjunto de treino que se encontram mais próximos. Após serem encontrados os vizinhos mais próximos, a previsão para o novo ponto de dados é feita com base nos rótulos ou valores dos vizinhos [40].

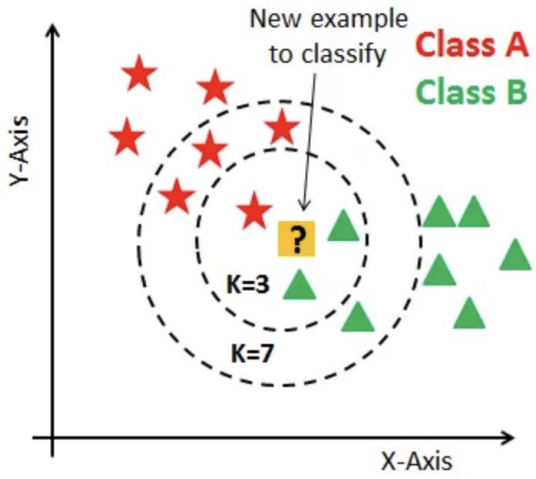

Figura 2.2: Exemplo do algoritmo kNN [41]

No exemplo da figura 2.2, pretende-se aplicar o algoritmo para classificar um determinado objeto de forma a concluir se pertence à classe das estrelas ou à classe dos triângulos. No entanto, dependendo do valor de k, o algoritmo fornece conclusões diferentes:

- para **k=3**, em que se consideram apenas os três pontos mais próximos, existem dois triângulos e apenas uma estrela, classificando o objeto como um triângulo;
- para **k=7**, considerando os sete pontos mais próximos, existem três triângulos e quatro estrelas, classificando o objeto como uma estrela.

Assim, a determinação do valor ideal de k é crucial para garantir o equilíbrio entre a previsão e a generalização do modelo, existindo diferentes formas de o fazer.

Random Forest

O *Random Forest* é um algoritmo que se fundamenta na construção de múltiplas árvores de decisão. Segundo esta técnica, cada árvore é treinada sobre um subconjunto de dados e utiliza uma amostra aleatória de características em cada divisão dos nós. O seu processo de construção assenta em duas ideologias fundamentais [42]:

1. **Bagging (Bootstrap Aggregating):** múltiplas amostras são removidas com reposição do conjunto de dados de treino e utilizadas para gerar diferentes Árvores de Decisão, reduzindo a variância e o *overfitting*.

2. **Seleção aleatória de atributos:** em cada partição, o algoritmo considera somente um subconjunto aleatório de características, introduzindo diversidade nas árvores geradas.

O resultado final é obtido da agregação dos resultados das árvores, de forma a produzir previsões mais robustas e reduzir a variância associada à modelação com uma única árvore [43, 44].

Sendo um algoritmo capaz de lidar com dados heterogéneos, resistir ao *overftting* e de incluir medidas de importância das características, o *Random Forest* tem sido bastante requisitado em contextos biomédicos, incluindo problemas de classificação de doenças [45].

2.2.2 Aprendizagem Não Supervisionada

A aprendizagem não supervisionada é uma abordagem de ML em que não existem exemplos de treino rotulados como orientação externa. Uma vez que este método opera sem orientação, a construção dos modelos pode tornar-se mais desafiante. No entanto, esta é uma técnica bastante relevante no processamento de conteúdos multimédia, permitindo o *clustering* de dados na ausência de rótulos de classe [46]. Esta é uma abordagem relevante para o projeto proposto, uma vez que um dos objetivos passa pela identificação de grupos de resposta terapêutica, que podem ser encontrados tirando partido de técnicas de *clustering* e de redução de características.

Nas secções que se seguem, são apresentadas algumas técnicas de aprendizagem não supervisionada que poderão ser úteis para o desenvolvimento do trabalho.

Clustering

O *clustering*, ou agrupamento, é uma abordagem de aprendizagem não supervisionada que inclui a organização de dados em grupos (*clusters*), baseados em semelhanças entre os diversos pontos de dados. Existem várias técnicas de *clustering*, possuindo cada uma o seu próprio método de agrupamento de dados [47]:

- **K-Means:**

  O *k-means* tem o objetivo de minimizar a distância entre pontos dentro de um *cluster* e maximizar a distância entre os centros geométricos dos *clusters*. Este é um algoritmo relativamente simples de implementar e interpretar, sendo, por isso, bastante popular em aplicações na medicina. No entanto, possui algumas desvantagens que devem ser consideradas:

  - a quantidade de *clusters* (**k**) tem de ser determinada *a priori*, sendo que influencia significativamente os resultados;
  - o *k-means* é sensível a *outliers* (pontos de dados significativamente diferentes dos restantes pontos do conjunto), podendo ocorrer uma distorção da posição dos centros e originar atribuições de *clusters* incertas.

- **Gaussian Mixture Models (GMM):**

  O GMM é um algoritmo de *clustering* pertencente à família dos métodos probabilísticos. Por oposição às abordagens determinísticas, como o *K-Means*, este algoritmo assume que os dados são gerados por uma combinação de várias distribuições gaussianas, onde cada uma representa um *cluster* [48]. O GMM avalia os parâmetros das

distribuições que melhor se adaptam aos dados, recorrendo, por norma, ao algoritmo de **Expectation-Maximization (EM)** que alterna [49]:

1. **Expectation (E-step):** cálculo da probabilidade de cada observação pertencer a cada *cluster*, com base nos parâmetros atuais;
2. **Maximization (M-step):** atualização dos parâmetros das distribuições gaussianas (média, variância, peso), de modo a maximizar a probabilidade de veracidade dos dados.

A atribuição probabilística de cada observação aos respetivos *clusters*, constitui um dos principais atributos do GMM. Este aspeto permite conferir maior flexibilidade em relação a algoritmos como o *K-Means*, onde a atribuição é mais rígida.

Entre as vantagens do GMM, destacam-se a capacidade de modelar grupos com formas elípticas, em vez de esféricas (como no *K-Means*) e a maior robustez em contextos onde ocorra sobreposição de grupos. No entanto, este algoritmo apresenta algumas limitações, como a necessidade de definir previamente a quantidade de componentes gaussianas e a possível convergência para máximos locais, o que está fortemente interligado à inicialização [50].

Redução de Características

A Redução de Características é um método que pretende transformar um *dataset* de grande dimensão, num *dataset* menor, de forma a manter a lógica e o significado dos dados o mais intactos possível. A representação dos dados sob a forma de uma dimensão menor, permite uma análise, um processamento e uma visualização simplificados. A Redução de Características pode ser conseguida através de duas formas [51]:

- **Seleção de Características:** esta técnica realiza a seleção das características com maior importância e relevância de um conjunto de dados, excluindo os atributos redundantes e irrelevantes. O seu objetivo é a construção de um grupo de características o mais pequeno possível, preservando as informações essenciais do *dataset* original.
- **Extração de Características:** este é um procedimento no qual são criadas novas características a partir do conjunto de dados original, preservando as informações mais pertinentes e reduzindo a dimensão do conjunto.

2.3 Trabalhos Relacionados

Nesta secção são explorados alguns trabalhos existentes que se relacionam com o objetivo do projeto exposto neste documento.

2.3.1 Construction and Validation of a Risk Prediction Model for Acute Asthma Exacerbations based on Machine Learning

Este estudo foi apresentado na **Conferência Internacional de 2024 sobre Cuidados de Saúde Inteligentes e Dispositivos Inteligentes Vestíveis (SHWID 2024)** e aplica a base de dados *National Health and Nutrition Examination Survey* (NHANES) 2017–2018 [8] para

[8] **NHANES:** programa conduzido pelo *National Center for Health Statistics* (NCHS), pertencente ao *Centers for Disease Control and Prevention* (CDC), e que fornece dados sobre a saúde e nutrição dos habitantes dos EUA.

construir um modelo cuja função é a previsão do risco de complicações agudas de asma. Para tal, vários parâmetros foram analisados, incluindo o histórico clínico, o nível de atividade física, o estilo de vida, dados demográficos e marcadores laboratoriais. Embora este estudo não se foque de forma explícita nos perfis metabólicos, utiliza variáveis como a hemoglobina glicada e a Proteína C-reativa de alta sensibilidade (PCRAS), que são indicadores do estado metabólico dos indivíduos. Desta forma, este trabalho considera, em parte, a importância destes biomarcadores na previsão do agravamento da asma.

No desenvolvimento deste projeto, foram aplicados algoritmos de regressão logística para realizar a identificação dos fatores de alto risco. Após esta etapa, as variáveis identificadas foram aplicadas na construção do modelo preditivo. Os resultados obtidos demonstraram que o modelo possui um bom desempenho preditivo, possuindo uma Área sob a Curva (AUC) de 0.67 para o conjunto de validação e de 0.64 para o conjunto de treino. Deste modo, os autores concluíram que o modelo desenvolvido pode ser um bom poderoso aliado para os processos de avaliação clínicos, permitindo a identificação de doentes de alto risco [52].

2.3.2 Prognosis of Clinical Outcomes with Temporal Patterns and Experiences with One Class Feature Selection

Na mesma conferência (SHWID 2024), foi apresentado um estudo que analisou Vesículas extracelulares urinárias (uEVs) de doentes com nefropatia diabética para identificar RNAs que se expressam de uma forma significativamente diferente. Para tal, os autores aplicaram diferentes algoritmos de classificação e de seleção de características, de forma a apontar os RNAs mais preditivos da nefropatia diabética.

Como resultado, o trabalho desenvolvido revelou que as uEVs possuem uma enorme quantidade de informação genética, podendo servir como biomarcadores não invasivos capazes de auxiliar o processo de diagnóstico da doença. Para além disso, os algoritmos desenvolvidos identificaram diferentes perfis de RNA com diferenças significativas, como é o caso dos **RNAs MYO1C** e do **SP100 mRNA**, que também podem ser aplicados como biomarcadores de diagnóstico para a nefropatia diabética [52].

2.3.3 Identification of Functional Microbial Modules Through Network-Based Analysis of Meta-Microbial Features Using Matrix Factorization

O objetivo deste estudo passa pelo desenvolvimento de uma nova abordagem baseada na fatorização de matrizes para identificar módulos microbianos funcionais através de dados de abundância de microbiomas. O trabalho foi desenvolvido pela necessidade, identificada pelos autores, de atualizar os métodos existentes de análise da microbiota, que possuem uma capacidade limitada de considerar características dos micro-organismos. Embora não utilize algoritmos de ML supervisionados tradicionais, este trabalho aplica a técnica de Fatorização de matrizes não negativa (NMF), um algoritmo de aprendizagem não supervisionada, para extrair padrões ocultos a partir dos dados em análise.

Após a avaliação do modelo desenvolvido, os autores concluíram que o NMF é adequado para auxiliar o processo de identificação de módulos microbianos funcionais através de dados de abundância de microbiomas. A eficácia do modelo é comprovada pelos autores através da aplicação do modelo tanto a dados de microbiotas humanos de diferentes localizações do corpo, como a dados de microbiotas relacionadas com determinadas doenças. Desta forma, o trabalho representa um grande avanço na análise de dados de microbiomas, facilitando

a compreensão de interações microbianas complexas e das suas possíveis implicações em contextos saudáveis e de doença [53].

2.3.4 Machine Learning for Clinical Outcome Prediction

Este artigo aborda o aumento da influência dos modelos de ML na previsão de resultados clínicos através de Registos Eletrónicos de Saúde (EHR). Para tal, os autores analisam todo o circuito típico de ML, desde o pré-processamento dos dados, até à avaliação dos modelos desenvolvidos, comparando diferentes técnicas de seleção de caraterísticas e abordando vários métodos de previsão, incluindo técnicas de regressão, de *deep-learning* e de *clustering*.

Apesar de os autores do estudo não terem alcançado nenhuma conclusão específica acerca das abordagens mais adequadas para auxiliar os processos de previsão clínica, foram identificadas algumas áreas importantes suscetíveis a melhorias :

- **Imprecisão dos rótulos dos resultados:** os investigadores aconselham a substituição da utilização dos rótulos binários pela análise da sobrevivência[9], permitindo modelar a progressão gradual da danificação do estado de saúde de um utente, levando a previsões mais específicas.
- **Modelos preditivos personalizados:** os autores defendem o desenvolvimento de modelos adaptados a cada utente, tendo em conta fatores específicos de cada indivíduo e da doença, podendo aprimorar significativamente a capacidade de previsão dos modelos.
- **Modelos de aprendizagem gerais:** também é aconselhado o investimento em desenvolver modelos multifacetados, capazes de realizar tarefas de diagnóstico e prognóstico clínico através da aplicação de dados heterogéneos de EHR para fornecer uma visão ampla do estado de saúde dos doentes.

Desta forma, este artigo revela considerações de úteis para o desenvolvimento de trabalhos de ML orientados à área da medicina.

2.4 Bibliotecas de Machine Learning

Para desenvolver e implementar algoritmos eficientes de ML, existem várias bibliotecas disponíveis que podem auxiliar nesses processos. Nas secções seguintes, serão abordadas algumas das opções existentes, terminando com uma análise comparativa das respetivas vantagens e limitações.

2.4.1 Scikit-learn

A Scikit-learn é uma biblioteca *open-source* em Python, cuja função é fornecer ferramentas eficientes de análise de dados, de pré-processamento e de construção de modelos, suportando um vasto leque de algoritmos de aprendizagem supervisionada e não supervisionada. A sua aplicação pode implicar um valioso contributo no suporte a tarefas de classificação, regressão, *clustering*, redução das características e de avaliação de modelos. Este é um recurso abundantemente utilizado por cientistas e investigadores em trabalhos de ML, por

[9] **Análise da sobrevivência:** técnica estatística focada no estudo do tempo desde o instante inicial até à ocorrência de um determinado evento [54].

facilitar a integração com todo o ecossistema de Python e por ser compatível com diversas bibliotecas, como [55]:

- **NumPy (Numerical Python):** biblioteca de Python aplicada para realizar tarefas de computação numérica.
- **SciPy (Scientific Python):** biblioteca construída sob o *NumPy*, cuja função é suportar a computação científica e técnica.
- **Matplotlib:** biblioteca utilizada para criar visualizações estáticas, interativas e animadas, sendo capaz de gerar uma grande variedade de gráficos, desde gráficos simples de linhas até visualizações mais complexas.

2.4.2 TensorFlow

O TensorFlow é uma biblioteca *open-source* desenvolvida pelo Google para auziliar o desenvolvimento de métodos de ML em larga escala. Esta ferramenta possui a capacidade de simplificar cálculos frequentemente complexos em gráficos de simples interpretação. É, também, capaz de mapear parcialmente os gráficos gerados para máquinas num *cluster* ou para processadores de uma única máquina. Algumas das suas funcionalidades mais importantes incluem interface simples e intuitiva, ambiente de visualização interativo, prototipagem rápida, entre outras. O TensorFlow foi desenvolvido de forma a suportar uma vasta gama de algoritmos de ML e de *deep-learning* [56].

2.4.3 PyTorch

O PyTorch é uma biblioteca de ML baseada em Python, desenvolvida pelo laboratório de pesquisa de AI do Facebook. É uma ferramenta focada no auxílio de tarefas de *deep-learning*, permitindo o desenvolvimento de modelos de forma flexível e rápida e fornecendo uma interface de simples interpretação para análise dos resultados obtidos. Uma das suas principais particularidades é a sua abordagem de Gráfico Computacional Dinâmico, permitindo aos programadores definir e modificar modelos durante a sua execução [57].

2.4.4 Comparação

A tabela 2.1 compara as bibliotecas de ML anteriormente apontadas, realizando um balanço das vantagens e desvantagens de cada uma, de forma a auxiliar o processo de escolha das bibliotecas a aplicar no projeto.

Biblioteca	Vantagens	Limitações
Scikit-learn	- **Simplicidade:** interface simples e intuitiva; - **Ampla gama de algoritmos:** vasta coleção de algoritmos implementados de ML; - **Integração com ecossistema de Python:** compatibilidade com outras bibliotecas de Python para manipulação e análise de dados; - **Documentação:** documentação extensa e bem escrita.	- **Limitação a algoritmos clássicos:** principalmente direcionado para algoritmos d ML clássicos, não fornecendo suporte para *deep-learning*; - **Não utiliza grafos computacionais:** a sua execução é imperativa, sem representação de dependências em grafo, o que pode limitar a otimização automática e o aproveitamento efeciente de *Graphics Processing Units* (GPUs)/*Tensor Processing Units* (TPUs); - **Dificuldade no desenvolvimento de modelos personalizados:** falta de flexibilidade para a criação de modelos personalizados, sendo mais adequados para algoritmos pré-implementados.
TensorFlow	- **Estrutura de dados:** aplicação da estrutura de dados tensor, que funciona como um array multidimensional ; - **Cálculos Paralelos:** a sua arquitetura baseada no princípio de fluxo de dados simplifica a aplicação de cálculos paralelos em CPUs *multi-core*; - **Popularidade:** sendo mais popular que o PyTorch, possui uma maior rede de suporte.	- **Grafo computacional estático:** o modelo deve ser definido previamente como um grafo completo de operações, permitindo, por um lado, a sua otimização eficiente em GPUs/TPUs, mas reduzindo a sua flexibilidade durante o desenvolvimento; - **Utilização:** embora possua APIs em *Python*, a implementação de baixo nível é em C++, podendo exigir maior inspeção no código;
PyTorch	- **Grafo Computacional Dinâmico:** permite a manipulação do grafo em tempo real, sendo possível a adição e remoção de nós durante a sua execução; - **Integração com Python:** sintaxe e funcionamento muito semelhantes ao Python padrão; - **Adequado para investigação:** a simplicidade na experimentação com diferentes arquiteturas facilita o processo de investigação.	- **Tempo:** tempos de aprendizagem significativamente superiores do que os de outras bilbiotecas; - **Utilização desafiante:** compreensão de conceitos de grafos computacionais pode exigir um esforço adicional na aplicação da biblioteca; - **Suporte:** menor popularidade comparativamente a outras bilbiotecas semelhantes (como é o caso do TensorFlow), existindo uma menor rede de suporte.

Tabela 2.1: Vantagens e limitações das bibliotecas de ML (adaptado de [58])

Capítulo 3

Pré-processamento dos Dados

Este capítulo descreve as etapas realizadas no contexto da preparação dos dados, desenvolvimento dos modelo e avaliação dos resultados obtidos.

3.1 Descrição do Dataset

Os dados utilizados neste projeto pertencem a análises clínicas realizadas a dois grupos distintos de doentes: doentes com Doença de Crohn e doentes com Doenças Hereditárias do Metabolismo (DHM).

Estes dados foram disponibilizados em formato digital, resultantes de análises clínicas de plasma e de urina realizadas em equipamento médico especializado. Para cada amostra biológica, o equipamento exportou dois ficheiros de texto livre (com a extensão *.ESTD-Conc*), contendo os valores numéricos obtidos nas medições para diferentes aminoácidos. Estes ficheiros apresentavam um formato pouco estruturado, dificultando o seu tratamento computacional direto. Assim, procedeu-se à sua conversão para um formato tabular mais compatível (*.csv*), permitindo a sua utilização em processos de análise estatística e computacional de forma mais eficiente e reprodutível.

O *dataset* original estava estruturado em duas pastas principais, cada uma contendo múltiplos subdiretórios, correspondentes a pares de amostras de um único doente. Cada par era constituído por dois ficheiros, correspondentes a medições realizadas em comprimentos de onda distintos, identificados pelo sufixo **'-440nm'** e **'-570nm'**. Estes ficheiros registavam a concentração de diferentes aminoácidos e outros compostos químicos detetados no organismo do doente.

Nas tabelas 3.1 e 3.4, são apresentados extratos das amostras de plasma para ambos os comprimentos de onda de um determinado doente. Note-se que os *sample IDs* são apresentados no formato "*XXXX-P*", de forma a garantir anonimato das amostras.

Date	Time	Sample Id	PPS	TAU	PEA	UREIA	ASP	...
02/2022	08	XXXXX-P	3	97	1	5414	6	...

Tabela 3.1: Exemplo de uma amostra de 570*nm* (parcial)

Date	Time	Sample Id	Sulfocisteina	HYP	PRO
02/2022	08	XXXXX-P	3	97	1

Tabela 3.2: Exemplo de uma amostra de 440*nm*

Adicionalmente, dado que as amostras poderiam corresponder a análises sanguíneas (plasma) ou a análises de urina, o conjunto de dados disponibilizado encontrava-se organizado segundo a sua origem biológica. No entanto, esta informação não se encontrava explicitamente indicada nos ficheiros fornecidos. Para garantir a correta distinção entre amostras de plasma e de urina, recorreu-se à validação através dos identificadores das amostras (*sample IDs*), cuja nomenclatura incluía sufixos específicos que permitiam distinguir a sua origem biológica:

- Identificadores terminados em **-P** ou **SJ** correspondiam a análises de plasma;
- Identificadores terminados em **-U** ou **-U[0-9]** correspondiam a análises de urina.

De forma a automatizar esta tarefa, foi desenvolvido um *script* em *Python* com o objetivo de analisar os sufixos dos *sample IDs* e atribuir a cada amostra a sua origem biológica.

3.2 Uniformização dos Dados

Adicionalmente, após a análise estrutural dos dados disponibilizados, foi desenvolvido um *script* em *Python* com o propósito de consolidar e normalizar as amostras clínicas, de forma a torná-las adequadas para a posterior aplicação de modelos de aprendizagem automática. O foco incidiu sobre amostras de plasma e urina, com medições realizadas a dois comprimentos de onda distintos (440 nm e 570 nm).

Com o apoio da equipa médica envolvida no projeto, foi possível identificar compostos biológicos considerados irrelevantes para os objetivos da investigação. Esta colaboração permitiu uma filtragem inicial criteriosa, resultando na exclusão de variáveis não pertinentes e, consequentemente, na redução da dimensionalidade do conjunto de dados. Esta abordagem teve como finalidade facilitar a análise de dados e, ainda, garantir o foco nos compostos de maior interesse clínico.

Como exemplo, no conjunto de dados correspondente ao comprimento de onda de 570 nm, foram removidas colunas descritivas e compostos menos relevantes, incluindo identificadores técnicos, parâmetros do equipamento de medição e alguns metabolitos de baixa significância clínica:

```
drop_columns_plasma_570 = ['Date', 'Time', 'File Name', 'Method
    Name', 'User Name', 'Vial', 'Volume', 'Autosampler Program', '
    Injection Source', 'PIB', 'Homocys', 'Car', 'Ethan', 'Urea', '
    NLEU', 'Amm', 'Gaba', 'Pea', 'allo', 'Ans', 'PPS']

df_570nm = df_570nm.drop(columns=drop_columns_plasma_570, errors=
    'ignore')
```

Listing 3.1: Remoção de colunas sem relevância (*dataset* de 570nm)

Além da redução de variáveis, foi implementada uma rotina de validação de colunas, com o objetivo de assegurar a consistência dos tipos de dados, garantindo que todas as variáveis, à exceção do identificador da amostra (*Sample Id*), fossem tratadas como valores numéricos.

Durante o processo de uniformização, foram identificadas diversas inconsistências na nomenclatura dos compostos químicos, resultantes da utilização de diferentes equipamentos na extração das amostras. Para mitigar esse problema, foi elaborado um mapeamento, fazendo corresponder os nomes originais e uma nomenclatura padronizada, promovendo assim

a coerência semântica entre ficheiros. A tabela 3.4 apresenta a correspondência aplicada no mapeamento das variáveis.

Nome Original	Nome padronizado
SAR	Sarc
CIT	Citr
ABU	Aaba
CYS2	Cys
BALA	B-ala
BAIB	Baiba
HYL	Hylys
HYP	Hypro
THR	Thr
3MHIS	3-Mhis

Tabela 3.3: Exemplo de mapeamento dos compostos químicos

Posteriormente, procedeu-se à fusão dos dados recolhidos nos dois comprimentos de onda, integrando as medições associadas a cada amostra individual (identificadas com o mesmo *Sample Id*) numa única entrada no *dataset* a ser utilizado nos treinos e testes dos modelos. Esta integração foi essencial para assegurar que cada amostra fosse representada uma única vez no conjunto final, evitando duplicações e promovendo uma estrutura coesa para os modelos de ML.

Com vista à preparação para a fase de classificação supervisionada, foi ainda adicionada uma coluna binária designada por *Diagnosis*, que representa a variável-alvo dos modelos. Esta variável indica se a amostra pertence a um doente com diagnóstico de Doença de Crohn (valor 1) ou se corresponde a um doente diagnosticado com DHM (valor 0).

Como resultado deste processo de pré-processamento, foram gerados dois conjuntos de dados finais: um ficheiro para as amostras de plasma e outro para as amostras de urina. Estes ficheiros contêm as medições padronizadas, estruturadas e anotadas, prontas para as fases seguintes de análise exploratória, modelação e avaliação. Na tabela 3.4, é apresentado um excerto do *dataset* final das amostras de plasma:

Sample Id	HYP	PRO	Phser	Tau	Asp	Thr	Asn	...	Diagnosis
XXXXXX1SJ	5	212	1.0	85	5	146	90	...	1
XXXX1-rep-P	0	118	2.0	60	4	82	53	...	1
XXXX1-P	8	125	0.0	129	12	113	67	...	1
GBXXXX2SJ	9	193	2.0	73	2	126	72	...	1
XXXX2-P	40	170	0.0	83	10	74	47	...	0
GBXXXX3SJ	9	214	0.0	62	2	137	71	...	0
XXXX2-P-REP	38	174	0.0	70	4	118	64	...	0
GBXXXXX3SJ	98	668	13.0	647	30	391	306	...	0

Tabela 3.4: Exemplo parcial do *dataset* final (plasma)

Este processo foi crucial para garantir a qualidade dos dados de entrada, reduzir o ruído dos dados e aumentar o desempenho dos algoritmos aplicados nas fases seguintes do projeto.

3.3 Distribuição das amostras

De forma a obter uma visão mais clara da composição do *dataset*, foi realizada uma análise exploratória sobre a distribuição das amostras em função dos diferentes diagnósticos clínicos atribuídos aos doentes.

O gráfico apresentado na figura 3.1 compara a quantidade de amostras de plasma e urina, estratificadas de acordo com o diagnóstico: **Doença de Crohn (1)** e **Outras patologias (0)**.

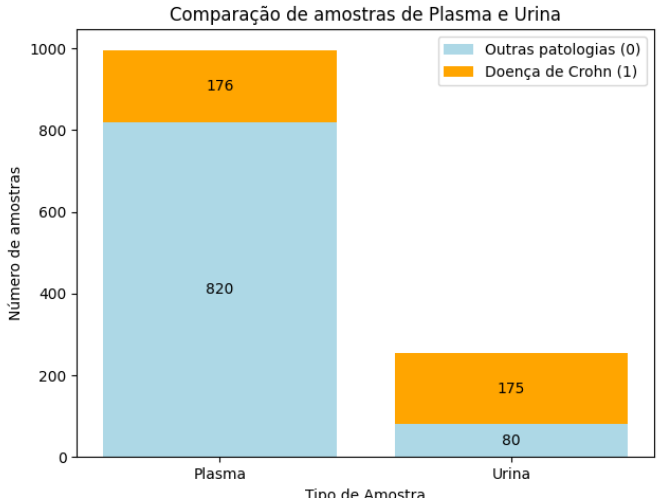

Figura 3.1: Distribuição das amostras por tipo de amostra e por diagnóstico

Por análise do gráfico, observa-se um desequilíbrio considerável entre os dois tipos de amostras: no plasma, predominam amostras com outras patologias (**820**), enquanto **176** correspondem a casos de Crohn; já na urina, valida-se o oposto: **175** amostras associadas à Doença de Crohn e apenas **80** a outras patologias. Este desequilíbrio pode degradar a capacidade de generalização dos modelos em cenários de aprendizagem supervisionada. De forma a mitigar este efeito, consideraram-se técnicas de balanceamento de classes: *oversampling*, que aumenta a representatividade da classe minoritária e *undersampling*, que reduz o número de exemplos da classe maioritária. A literatura demonstra a utilidade de ambas as abordagens em contextos desequilibrados e discute as suas possíveis implicações na estimação e avaliação do desempenho [59]. As técnicas de balanceamento utilizadas no desenvolvimento do presente trabalho serão descritas com maior detalhe no capítulo da Classificação (4).

Para além disso, a diferença entre a quantidade de amostras de plasma (**996**) e a quantidade de amostras de urina (**255**), requer um tratamento personalizado a cada tipo de amostragem, sendo que as características predominantes num tipo de amostra pode diferir das selecionadas no outro tipo.

3.4 Feature Selection

A seleção de variáveis foi utilizada para filtrar preditores pouco informativos ou redundantes. Esta etapa reduz o risco de sobreajuste, melhora a eficiência computacional e aumenta a interpretabilidade, ao destacar um subconjunto reduzido de preditores com maior capacidade discriminativa [60]. É descrita neste capítulo por constituir um procedimento de pré-processamento; contudo, a sua avaliação empírica e o respetivo impacto no desempenho são apresentados no capítulo 4 - Classificação. Note-se que esta técnica foi aplicada apenas nos cenários de classificação, tendo sido comparada com a abordagem sem seleção de variáveis.

Para esta etapa, foram realizados os seguintes passos:

1. Remoção inicial de componentes biológicos apontados como irrelevantes pela equipa médica (conforme o apresentado na listagem 3.2);
2. Treino de um classificador **Decision Tree Classifier**, cujos coeficientes de importância permitiram identificar os atributos mais significativos.

Para a seleção de variáveis (*feature selection*), optou-se pela aplicação de um *Decision Tree Classifier*, uma vez que este modelo se destaca pela sua capacidade de fornecer métricas de importância das variáveis. Esta característica permite a identificação e hierarquização dos atributos mais relevantes para o problema. A utilidade das árvores de decisão com a finalidade de selecionar variáveis, nomeadamente em contextos "*ómicos*"[1], é sustentada em diversos estudos que realçam a sua simplicidade, robustez e eficácia na descoberta de biomarcadores [62].

Para selecionar as características, mantiveram-se apenas aquelas com importância superior a um limiar definido de 0.01. No *DecisionTreeClassifier* do *scikit-learn*, a importância de cada preditor corresponde à redução total de impureza normalizada, de modo a que a soma das importâncias resulte em 1. Assim, 0.01 corresponde a um contributo inferior a 1% do total e funciona como patamar de corte para eliminar preditores residuais sem penalizar o desempenho. A documentação do *scikit-learn* descreve esta definição e suporta o uso de limiares explícitos em seleção baseada no modelo, legitimando a adoção de um valor percentual simples e interpretável neste contexto [63, 64]. Devido à disparidade nas quantidades de amostras de plasma e de urina, o processo de seleção de características foi feito individualmente para cada tipo de amostra.

3.4.1 Feature Selection nas amostras de plasma

Após a seleção de características nas amostras de plasma, o número de *features* diminiu de **34 para 23**, sendo que a característica marcada com maior importância foi o **Glutamato (Glu)**, com uma classificação de **0.223**. A figura 3.2 e a tabela 3.5 apresentam os resultados obtidos da *feature selection* aplicada às amostras de plasma.

[1] **Ómicos:** conjunto de campos de estudo da biologia que utilizam técnicas de alto rendimento e bioinformática para analisar um conjunto de dados biológicos num organismo ou num sistema orgânico. A metabolómica é uma das áreas incluídas neste contexto [61].

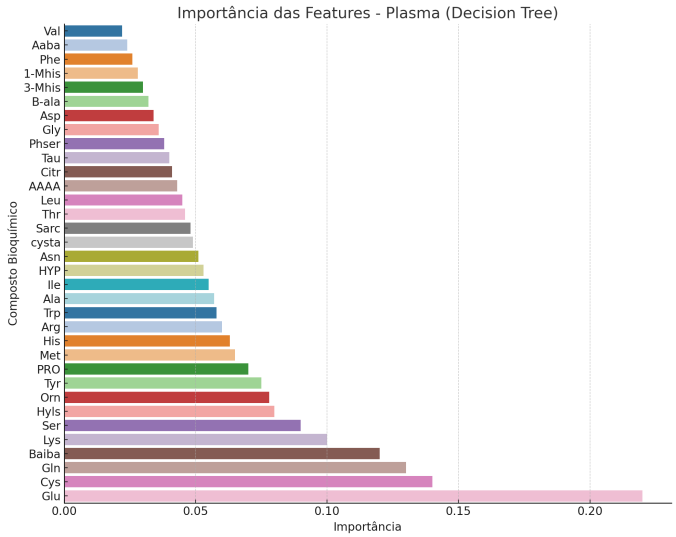

Figura 3.2: Importância das *features* das amostras de plasma de acordo com o modelo

Feature	**Importância**
Glu	0.223
Cys	0.101
Gln	0.078
Baiba	0.075
Lys	0.056
Ser	0.038
Hylys	0.034
Orn	0.030
Tyr	0.029
PRO	0.028
Met	0.025
His	0.024
Arg	0.024
Trp	0.023
Ala	0.023
Ile	0.023
HYP	0.022
Asn	0.021
cysta	0.018
Sarc	0.017
Thr	0.016
Leu	0.012
AAAA	0.011

Tabela 3.5: Importância das *features* selecionadas nas amostras de plasma

3.4.2 Feature Selection nas amostras de urina

Após a seleção de características nas amostras de urina, o número de *features* diminiu de **34 para 17**, sendo que a característica marcada com maior importância foi o **Glutamina (Gln)**, com uma classificação de **0.362**. Na figura 3.3 e na tabela 3.6 estão expostos os resultados da *feature selection* nas amostras de urina.

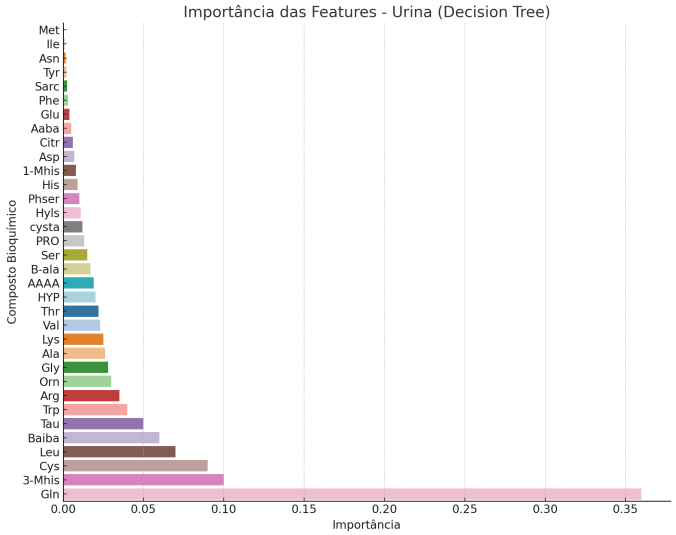

Figura 3.3: Importância das *features* das amostras de urina de acordo com o modelo

Feature	Importância
Gln	0.362
3-Mhis	0.093
Cys	0.078
Leu	0.065
Baiba	0.056
Tau	0.048
Trp	0.048
Arg	0.036
Orn	0.031
Gly	0.025
Ala	0.025
Lys	0.023
Val	0.022
Thr	0.016
HYP	0.016
AAAA	0.015
B-ala	0.014

Tabela 3.6: Importância das *features* selecionadas nas amostras de urina

Capítulo 4

Classificação

Este capítulo descreve a aplicação de algoritmos de aprendizagem automática na tarefa de classificação de amostras de plasma e urina com o objetivo de efetuar uma classificação binária: determinar, para cada amostra, se o paciente é doente de Crohn (classe positiva) ou se não é detentor da Doença de Crohn, sendo esta última categoria abrangente a doentes com DHM. As análises foram conduzidas separadamente por origem biológica (plasma e urina) e incluíram pré-processamento (normalização), estratégias de balanceamento de classes e comparação de múltiplos algoritmos supervisionados. A seleção, quando aplicável, seguiu o procedimento descrito no capítulo 3 - Pré-processamento dos Dados e foi avaliada pelo seu impacto no desempenho dos modelos implementados.

Embora não se inclua nos objetivos definidos para este trabalho, reconhece-se que o problema poderia ser formulado noutros moldes, analisando por exemplo, a classificação multi-caso (discriminar patologias específicas como a Doença de Crohn, outras DHM e condições não inflamatórias). No entanto, não estando incluídas nas metas definidas, estas variantes podem ser apontadas como trabalho futuro. Assim, o presente capítulo foca-se na decisão binária Crohn *vs.* Não-Crohn.

4.1 Pré-Processamento

Para garantir a eficácia dos modelos, foi realizado um pré-processamento específico à tarefa de classificação, incluindo a normalização das variáveis e técnicas de balanceamento de classes.

4.1.1 Balanceamento de classes

O conjunto de dados mostrou-se ligeiramente desequilibrado entre as classes, com predominância de diagnósticos de outras patologias.

Este desequilíbrio pode conduzir ao favorecimento da classe maioritária, afetando negativamente métricas como o *Recall* da classe minoritária (no caso, Diagnóstico de Crohn). De forma a mitigar este efeito e para estudar o impacto na performance, foram testadas três abordagens distintas:

- **Utilização dos dados originais**

  Os dados foram utilizados sem qualquer alteração na proporção entre classes, servindo de base na comparação com as restantes técnicas.

- **Oversampling**

  Foi aplicado o método *Synthetic Minority Over-Sampling Technique* (SMOTE), que gera amostras sintéticas da classe minoritária com base em combinações dos seus vizinhos mais próximos.

```
from imblearn.over_sampling import SMOTE

smote = SMOTE(random_state=42)
X_resampled, y_resampled = smote.fit_resample(X_train,
    y_train)
```

Listing 4.1: Aplicação do método SMOTE

- **Undersampling**

  A classe dominante (*Diagnosis = 0*) foi reduzida aleatoriamente para igualar o número de observações da classe minoritária (*Diagnosis = 1*). Esta técnica tem o objetivo de reduzir a tendência para o diagnóstico maioritário sem a introdução dos dados sintéticos, embora possa conduzir à perda de informação.

```
from imblearn.under_sampling import RandomUnderSampler

rus = RandomUnderSampler(random_state=42)
X_resampled, y_resampled = rus.fit_resample(X_train, y_train)
```

Listing 4.2: Aplicação de *Undersampling*

Cada uma destas técnicas foi testada separadamente, de forma a validar o seu impacto na performance dos modelos de classificação. A avaliação dos modelos foi feita com os mesmo parâmetros de treino e validação, permitindo fazer uma comparação direta.

Na secção da Avaliação dos Modelos 4.3, serão apresentados os resultados da aplicação destas estratégias e o seu impacto nas métricas de classificação.

4.1.2 Normalização dos Dados

Devido à diversidade nos componentes das amostras de plasma e urina, os dados apresentavam escalas de grandeza significativamente distintas. Enquanto determinados compostos surgem em concentrações próximas dos milhares, outros estão representados por valores de magnitude bastante inferior. Esta discrepância pode comprometer o desempenho de algoritmos de ML sensíveis à escala, como o kNN, cujos cálculos são baseados em distâncias euclidianas.

De forma a contornar este problema, foi aplicada uma técnica de normalização designada *Standard Scaler*, fornecida pela biblioteca *scikit-learn*. Esta é uma abordagem amplamente aplicada em tarefas de aprendizagem supervisionada e permite transformar os dados de modo a que cada variável possua média zero e desvio padrão igual a um [65].

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

Listing 4.3: Aplicação do *Standard Scaler*

Neste método, cada característica x é transformada segundo a fórmula da normalização padronizada 4.1:

$$x' = \frac{x - \mu}{\sigma} \tag{4.1}$$

onde:

- μ representa a média da variável;
- σ corresponde ao desvio padrão.

Esta transformação garante que todas as variáveis contribuam de igual forma para o processo de aprendizagem, erradicando discrepâncias provocadas por diferentes ordens de magnitude entre compostos.

4.2 Modelos Aplicados

Para classificar doentes com Doença de Crohn com base nas amostras disponíveis, foram aplicados algoritmos de aprendizagem supervisionada: *Decision Tree*, *Random Forest*, Regressão Logística e kNN. Estes algoritmos foram apresentados no Capítulo 2, no qual se discutiu a sua adequação ao problema.

O desenho experimental baseou-se na validação cruzada *stratified k-fold* com 5 dobras (*folds*), com mistura e semente fixa (`StratifiedKFold(n_splits=5, shuffle=True, random_state=42)`). A escolha de 5 *folds* resulta de um compromisso entre a variância das estimativas e o custo computacional: no plasma, cada iteração foi treinada em $\approx 80\%$ dos dados (≈ 797 amostras) e validada em $\approx 20\%$ (≈ 199 amostras); na classificação aplicada às amostras de urina, os modelos foram treinados em ≈ 204 observações e validados em ≈ 51. Estes tamanhos de validação são suficientemente grandes para estimar métricas de desempenho com estabilidade, sobretudo na urina, onde a utilização de *folds* demasiado pequenos poderia aumentar a variância das estimativas [66].

A regressão logística foi selecionada por ser adequada para problemas de classificação binária, como é o caso. Para implementar um modelo fundamentado neste algoritmo, recorreu-se à classe `LogisticRegression` da biblioteca *scikit-learn* e o modelo foi, depois, aplicado às amostras de plasma e de urina, com `LogisticRegression(max_iter=1000, solver='lbfgs', penalty='12', random_state=42)`. Definiu-se `max_iter=1000` para garantir que o algoritmo de otimização (`'lbfgs'`) dispunha de iterações suficientes para atingir o mínimo da função de perda, evitando paragens prematuras por limite de iterações. Adicionalmente, adotou-se o *solver* `'lbfgs'` com penalização `L2` por se tratar de uma combinação padronizada, robusta e estável, controlando a complexidade do modelo [67].

As árvores de decisão foram também aplicadas com recurso à *scikit-learn*, utilizando `DecisionTreeClassifier(random_state=42)`, mantendo os restantes parâmetros por omissão. As árvores de decisão são úteis neste tipo de contexto para captar relações não lineares e interações entre atributos [63].

Da mesma forma, recorreu-se à *scikit-learn* para utilizar a classe `KNeighborsClassifier` para implementar um modelo de classificação baseado no kNN. O número de vizinhos ($k = 5$) foi otimizado por validação cruzada, procurando o melhor compromisso entre viés e variância

[68]. O kNN explora padrões de proximidade entre amostras e oferece uma alternativa complementar a classificadores baseados em regras ou *ensembles*[1].

Por fim, utilizou-se `RandomForestClassifier(n_estimators=100, random_state=42)` para implementar um modelo baseado no *Random Forest*. O valor `n_estimators=100` reflete um compromisso entre desempenho e custo computacional: aumentos no número de árvores tendem a melhorar o modelo até certo ponto, após o qual os ganhos são reduzidos para o custo envolvido [43]. O algoritmo foi usado como classificador (resultados na Secção 4.3) e para avaliar a importância dos atributos, através de *feature selection*:

- **Feature Selection nas amostras de Plasma**

  A análise da importância dos atributos aplicados ao conjunto de amostras de plasma resultou na identificação de um conjunto reduzido de compostos com elevada relevância para a distinção entre diagnósticos. A figura 4.1 ilustra os dez atributos mais importantes, com destaque para a **Citrulina** (**Citr**), cuja contribuição foi consistentemente superior.

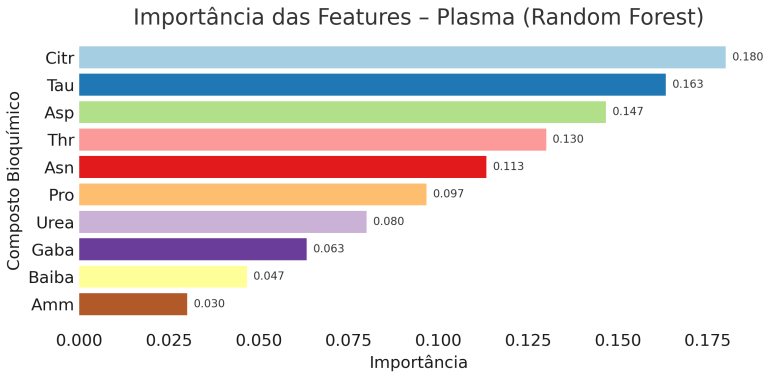

Figura 4.1: Importância das *features* das amostras de plasma de acordo com o modelo *Random Forest*

- **Feature Selection nas amostras de Urina**

  Semelhantemente, a aplicação do *Random Forest* ao conjunto de amostras de urina permitiu identificar os compostos mais relevantes para a tarefa de classificação. A figura 4.2 apresenta os dez atributos com maior importância, de entre os quais se destacam a **Serina** (**Ser**), a **Glutamina** (**Glu**) e a **Histidina (His)**, sugerindo um perfil distinto do observado no plasma.

[1]Modelos *ensemble* combinam múltiplos classificadores de base para aumentar a robustez e reduzir a variância dos resultados [69].

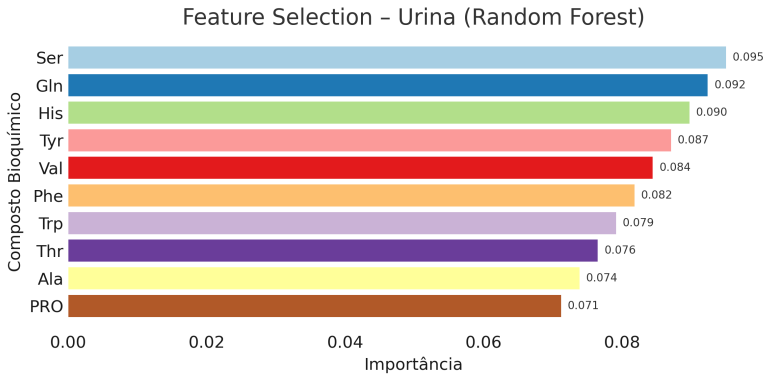

Figura 4.2: Importância das *features* das amostras urina de acordo com o modelo *Random Forest*

As diferenças observadas entre os componentes considerados mais relevantes nos resultados do algoritmo *Decision Tree* e no *Random Forest* podem ser justificadas pela natureza agregadora do *Random Forest*, tendendo a atenuar variações locais e a favorecer atributos com impacto mais consistente ao longo do conjunto de dados.

Apesar da utilização deste algoritmo para analisar a importância de atributos, importa salientar que esta técnica não foi aplicada para remover nem selecionar *features* nas abordagens com recurso aos restantes algoritmos.

4.3 Avaliação dos Modelos

Após a implementação dos modelos de classificação, procedeu-se à sua validação sistemática com o objetivo de quantificar a capacidade preditiva de cada técnica. De forma a comparar o desempenho dos modelos sob diferentes condições experimentais, a avaliação considerou diferentes tipos de técnicas de balanceamento de classes, bem como a aplicação do *dataset* completo ou com um subconjunto reduzido de características baseadas no resultado da *feature selection*.

4.3.1 Separação dos Dados

Visando a garantia da robustez da avaliação dos modelos, o conjunto de dados foi dividido em dois subconjuntos mutuamente exclusivos:

- 80% das amostras foram alocadas à etapa de treino dos modelos;
- os restantes 20% foram reservados para a fase de testes, avaliando a performance do modelo gerado.

Os dados de treino foram submetidos ao processo de *K-fold Cross Validation*, segundo o qual o conjunto de treino é dividido em subconjuntos de tamanho semelhante, mantendo, mais uma vez, a proporção das classes em cada *fold* (subconjunto). Segundo esta abordagem, o modelo é treinado k vezes, utilizando, em cada iteração, k - 1 *folds* para treino e o subconjunto restante para teste. Após serem realizadas k iterações, todas as amostras do conjunto foram utilizadas uma vez para teste e k - 1 vezes para treino. A performance global é calculada pela média das métricas obtidas em cada uma das iterações.

Neste estudo, foi aplicada a variante *Stratified K-folds* para preservar a proporção original de cada diagnóstico nas amostras disponibilizadas. Esta é uma medida fundamental para casos de distribuição assimétrica das classes, sendo que garante a representatividade adequada da classe minoritária tanto na preparação como na avaliação dos modelos [70].

Para esta etapa, definiu-se que a quantidade de subconjuntos mais adequadas seria cinco (k = 5), por resultar do equilíbrio entre a estabilidade da estimativa de desempenho e a variabilidade entre divisões, procurando assegurar uma quantidade suficiente de dados para treino e avaliação representativa em cada iteração. A aplicação de valores mais reduzidos (como k = 2) pode culminar em estimativas enviesadas, enquanto que a adoção de valores demasiado elevados tende a apresentar uma variância elevada e custos computacionais desnecessários [71].

4.3.2 Métricas de Avaliação

De forma a avaliar a performance dos modelos desenvolvidos, foram utilizadas métricas padronizadas que permitem quantificar diferentes dimensões do desempenho preditivo.

Como base para o cálculo das métricas, foi aplicada a matriz de confusão, resumindo as decisões tomadas por cada modelo quanto aos rótulos reais das amostras. Sendo o presente estudo uma tarefa de classificação binária, as métricas utilizadas foram as seguintes:

- **Verdadeiros Positivos (TP):** casos corretamente identificados como Doença de Crohn;
- **Falsos Positivos (FP):** casos indevidamente identificados como Crohn;
- **Verdadeiros Negativos (TN):** casos corretamente identificados como outras patologias;
- **Falsos Negativos (FN):** casos de Crohn que o modelo não conseguiu identificar.

A matriz de confusão é adequada a este contexto, uma vez que permite a interpretar o comportamento do modelo face a diferentes tipos de erro, sendo particularmente útil em conjuntos de dados desequilibrados, como é o caso [72].

Com base nos resultados obtidos na matriz de confusão, foi possível derivar métricas fundamentais para a avaliação da performance de modelos de classificação:

- **Accuracy:** representa a proporção de classificações corretas (positivas e negativas) face à quantidade total de observações avaliadas. Apesar de ser uma métrica intuitiva e de simples interpretação, a *accuracy* pode conduzir a erros em cenários de classes desbalanceadas, tendendo a favorecer a classe maioritária [70].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.2}$$

- **Precisão:** avalia a proporção de classificações positivas corretas entre todas as predições positivas feitas pelo modelo. Esta é uma métrica particularmente útil para avaliar o grau de confiança das predições positivas. No caso do presente estudo, esta métrica avalia a probabilidade de um diagnóstico de Crohn atribuído pelo modelo corresponder, efetivamente, a um caso verdadeiro [72].

$$Precision = \frac{TP}{TP + FP} \tag{4.3}$$

- **Recall (Sensibilidade):** diz respeito à capacidade do modelo de identificar corretamente todos os casos positivos reais. Esta métrica pode ser importante em contextos clínicos, uma vez que os falsos negativos podem resultar em falhas de diagnóstico, por exemplo [73].

$$Recall = \frac{TP}{TP + FN} \tag{4.4}$$

- **F1-score:** representa a média harmónica entre a precisão e a sensibilidade (*recall*), dando mais importância aos valores mais baixos. Esta métrica é particularmente recomendada em problemas com classes desequilibradas ou quando existe a necessidade de detetar a classe em minoria com fiabilidade, como é o caso[70].

$$F1 = 2 * \frac{Precision * Recall}{Precision\ + Recall} \tag{4.5}$$

Todas estas métricas foram aplicadas para avaliar a performance dos modelos durante a etapa de validação como de avaliação final nos dados de teste. A sua utilização fornece uma análise completa quanto à eficácia dos modelos, podendo ser extremamente útil na avaliação de sistemas de apoio à decisão clínica [74].

4.4 Resultados da Classificação

Após a aplicação dos diferentes modelos de classificação, elaborou-se a avaliação do seu desempenho utilizando as métricas previamente selecionadas (apresentadas na secção anterior 4.3.2). A performance dos modelos foi validada em três cenários distintos: dados desbalanceados, dados com *oversampling* via SMOTE e dados com *undersampling* da classe maioritária (doentes com outras patologias). Para além disso, analisou-se, para cada cenário, o desempenho dos modelos com o conjunto completo de características e com o conjunto reduzido após a aplicação da *feature selection*, mantendo as mesmas condições de treino e teste.

4.4.1 Resultados sem aplicação de técnicas de balanceamento

Neste primeiro cenário, os modelos foram treinados e testados sem alteração da distribuição original das classes. Esta abordagem permite analisar de que modo cada modelo lidou com os dados desequilibrados e pode servir de referência para avaliar o impacto das técnicas de balanceamento. As tabelas 4.1 e 4.2 apresentam os resultados obtidos para as amostras de plasma e de urina, respetivamente.

Resultados da classificação sem balanceamento para amostras de plasma								
Modelo	**Accuracy**		**Precision**		**Recall**		**F1-score**	
	(s/ FS)	**(c/ FS)**	**(s/ FS)**	**(c/ FS)**	**(s/ FS)**	**(c/ FS)**	**(s/ FS)**	**(c/ FS)**
Decision Tree	0.83	0.81	0.52	0.48	0.60	0.55	0.56	0.51
Random Forest	0.90	0.89	0.94	0.90	0.46	0.43	0.61	0.57
KNN	0.88	0.89	0.70	0.75	0.61	0.59	0.65	0.65
Regressão Logística	0.88	0.87	0.73	0.71	0.54	0.52	0.61	0.58

Tabela 4.1: Comparação de modelos com dados desbalanceados para amostras de plasma

Os resultados obtidos indicam que, para as amostras de plasma, o kNN apresentou o melhor compromisso global de *f1-score* (0.65), independentemente da aplicação de *feature selection*, embora com valores moderados de *recall* (0.59 - 0.61) e *precision* elevada (0.70 - 0.75). O *Random Forest* apresentou resultados de *f1-score* inferiores (0.57 - 0.61) por combinar *precision* elevada com *recall* reduzido. Por outro lado, o *Decision Tree* foi o modelo com o pior desempenho, com o níveis de *f1-score* mais baixos (0.51 - 0.56), apesar de apresentar *accuracies* aceitáveis.

Resultados da classificação sem balanceamento para amostras de urina								
Modelo	**Accuracy**		**Precision**		**Recall**		**F1-score**	
	(s/ FS)	**(c/ FS)**	**(s/ FS)**	**(c/ FS)**	**(s/ FS)**	**(c/ FS)**	**(s/ FS)**	**(c/ FS)**
Decision Tree	0.76	0.73	0.83	0.81	0.82	0.89	0.82	0.80
Random Forest	0.80	0.78	0.82	0.81	0.91	0.89	0.86	0.85
KNN	0.79	0.73	0.79	0.76	0.94	0.90	0.86	0.82
Regressão Logística	0.78	0.76	0.79	0.77	0.93	0.94	0.85	0.83

Tabela 4.2: Comparação de modelos com dados desbalanceados para amostras de urina

Nas amostras de urina, tanto o *Random Forest* como o kNN apresentaram desempenhos elevados: no cenário sem *feature selection* ambos os modelos alcançaram valores de *f1-score* de 0.86, sendo ambos opções adequadas para a tarefa de classificação neste cenário. No entanto, todos os modelos se mantiveram estáveis neste cenário, tendo obtido valores satisfatórios de *f1-score*.

4.4.2 Resultados com Oversampling

No segundo cenário da avaliação, foi introduzida a técnica de *oversampling* com SMOTE, tendo como objetivo o aumento da representatividade da classe minoritária, no caso os diagnósticos de Crohn, através da geração sintética de novas amostras. As tabelas 4.3 e 4.4 apresentam os resultados obtidos com a utilização desta estratégia.

Resultados da classificação com *oversampling* para amostras de plasma								
Modelo	**Accuracy**		**Precision**		**Recall**		**F1-score**	
	(s/ FS)	**(c/ FS)**	**(s/ FS)**	**(c/ FS)**	**(s/ FS)**	**(c/ FS)**	**(s/ FS)**	**(c/ FS)**
Decision Tree	0.81	0.80	0.48	0.44	0.61	0.34	0.54	0.48
Random Forest	0.90	0.90	0.79	0.78	0.67	0.65	0.72	0.71
KNN	0.77	0.78	0.43	0.45	0.95	0.90	0.59	0.78
Regressão Logística	0.85	0.84	0.54	0.53	0.85	0.82	0.66	0.64

Tabela 4.3: Comparação de modelos com *oversampling* para amostras de plasma

Para as amostras de plasma, o *oversampling* permitiu aumentar consideravelmente o desempenho do *Random Forest*, que atingiu o seu pico com valores de *f1-score* de 0.72 no cenário sem *feature selection* e de 0.71 no cenário com seleção de variáveis, conjugando ainda *precision* alta (0.78 - 0.79) e valores satisfatórios de *recall* (0.65 - 0.67), representando uma opção sólida para obter uma sensibilidade elevada e controlar, simultaneamente, falsos positivos. Neste cenário, a Regressão Logística também se destacou, com valores de *f1-score* entre 0.66 e 0.64. Novamente, à semelhança do cenário sem balanceamento de classes, o *Decision Tree* manteve-se o modelo com o pior desempenho.

Nas amostras de urina, o melhor desempenho foi alcançado pelo *Random Forest* sem a aplicação de *feature selection*, alcançando um *f1-score* de 0.86. De seguida, a Regressão

Resultados da classificação com *oversampling* para amostras de urina								
Modelo	**Accuracy**		**Precision**		**Recall**		**F1-score**	
	(s/ FS)	**(c/ FS)**	**(s/ FS)**	**(c/ FS)**	**(s/ FS)**	**(c/ FS)**	**(s/ FS)**	**(c/ FS)**
Decision Tree	0.76	0.71	0.82	0.82	0.83	0.86	0.83	0.78
Random Forest	0.81	0.78	0.85	0.82	0.88	0.86	0.86	0.84
KNN	0.74	0.72	0.88	0.83	0.74	0.75	0.80	0.79
Regressão Logística	0.79	0.79	0.84	0.82	0.86	0.88	0.81	0.60

Tabela 4.4: Comparação de modelos com *oversampling* para amostras de urina

Logística também atingiu resultados relevantes, com *f1-score* de 0.85 e *recall* e precisão de 0.86 e 0.88, respetivamente. O pior desempenho foi registado no cenário com aplicação do *Decision Tree* e com *feature selection* (*f1-score* de 0.78).

4.4.3 Resultados com Undersampling

No último cenário, recorreu-se à técnica de *undersampling*, removendo-se, de forma aleatória, amostras da classe maioritária até que existam na mesma quantidade que as amostras da classe minoritária. As tabelas 4.5 e 4.6 demonstram os resultados obtidos neste cenário.

Resultados da classificação com *undersampling* para amostras de **plasma**								
Modelo	**Accuracy**		**Precision**		**Recall**		**F1-score**	
	(s/ FS)	**(c/ FS)**	**(s/ FS)**	**(c/ FS)**	**(s/ FS)**	**(c/ FS)**	**(s/ FS)**	**(c/ FS)**
Decision Tree	0.72	0.73	0.37	0.37	0.70	0.66	0.48	0.47
Random Forest	0.85	0.85	0.56	0.57	0.81	0.81	0.66	0.67
KNN	0.70	0.73	0.37	0.40	0.94	0.94	0.53	0.56
Regressão Logística	0.82	0.83	0.51	0.51	0.82	0.84	0.62	0.63

Tabela 4.5: Comparação de modelos com *undersampling* para amostras de plasma

Na classificação das amostras de plasma com aplicação da técnica de *undersampling*, o *Random Forest* foi o modelo mais adequado em termos de *f1-score*, atingindo 0.67 (sem *feature selection*) e 0.66 (com *feature selection*), conjungando sensibilidade elevada (0.81) com precisão moderada (0.56–0.57). Assim, este modelo representa um equilíbrio entre a deteção de casos de Doença de Crohn e controlo de falsos positivos. Por outro lado, o kNN maximizou a sua sensibilidade (0.94), à custa de uma precisão mais reduzida (com valores entre 0.37 e 0.40), resultando num *f1-score* inferior. Mais uma vez, as árvores de decisão voltaram a ser o modelo com pior desempenho (*f1-score* com valores de 0.47 e 0.48), apesar de apresentarem ligeiras melhorias na sensibilidade comparativamente ao cenário sem técnicas de balanceamento.

Resultados da classificação com *undersampling* para amostras de **urina**								
Modelo	**Accuracy**		**Precision**		**Recall**		**F1-score**	
	(s/ FS)	**(c/ FS)**	**(s/ FS)**	**(c/ FS)**	**(s/ FS)**	**(c/ FS)**	**(s/ FS)**	**(c/ FS)**
Decision Tree	0.73	0.71	0.85	0.84	0.74	0.72	0.79	0.77
Random Forest	0.79	0.78	0.85	0.87	0.84	0.80	0.84	0.83
KNN	0.71	0.70	0.80	0.79	0.77	0.77	0.78	0.78
Regressão Logística	0.78	0.75	0.85	0.81	0.84	0.82	0.84	0.82

Tabela 4.6: Comparação de modelos com *undersampling* para amostras de urina

Nas amostras de urina, obtiveram-se resultados competitivos: tanto a Regressão Logística, como o *Random Forest* obtiveram *f1-score* de 0.84 (no cenário sem *feature selection*). O kNN obteve *f1-score* de 0,78 (e *recall* de 0,77), e o *Decision Tree* apresentou os piores resultados no cenário com *feature selection*, com *f1-score* de 0.77, refletindo maior perda de informação relevante quando se remove amostras da classe dominante.

4.5 Comparação dos Modelos

Após a organização dos resultados obtidos para os diferentes cenários, procedeu-se à comparação do desempenho dos modelos aplicados durante a tarefa de classificação, com o propósito de identificar o modelo que melhor se adequa ao problema em estudo. O gráfico da figura 4.4 representa os diferentes valores médios de *f1-score* obtidos para cada cenário de classificação realizado com as amostras de plasma.

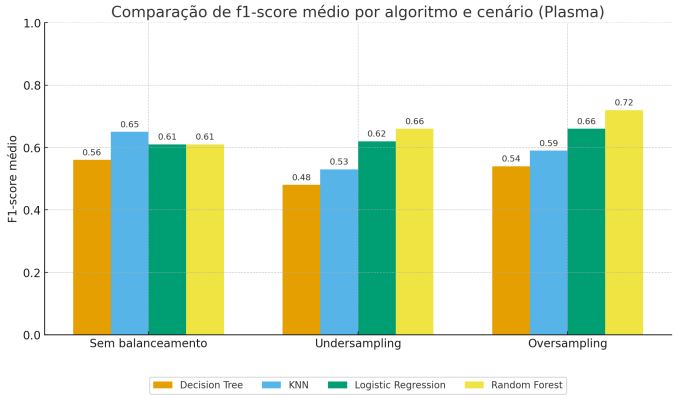

Figura 4.3: Comparação dos valores médios de F1-*score* obtidos com as amostras de plasma

Em termos gerais, o *Random Forest* evidenciou-se como o modelo mais robusto e estável face às variações de cenário. Apresentou a melhor combinação de precisão, *recall* e *f1-score*, sendo que a sua aplicação com a estratégia de *oversampling* via SMOTE resultou no melhor desempenho de classificação nas amostras de plasma (*f1-score* de 0.72).

A Regressão logística revelou um desempenho consistente e equilibrado entre cenários, destacando-se como alternativa fiável: na abordagem com *oversampling* obteve *f1- score* médio de 0.66, mantendo valores competitivos, também, nas experiências com *undersampling* (*f1-score* médio de 0.62) e sem balanceamento (*f1-score* médio 0.61).

O kNN evidenciou-se sobretudo no cenário sem balanceamento, onde registou o *f1-score* mais elevado dessa experiência (0,65). No entanto, mostrou maior sensibilidade às estratégias de equilíbrio de classes, exibindo variações mais acentuadas do desempenho entre *undersampling* e *oversampling*.

Por contraste, o *Decision Tree* apresentou, de forma consistente, os *f1-scores* mais baixos entre os modelos avaliados, o que limita a sua adequação para a classificação de Crohn *vs.* Não-Crohn no conjunto de plasma.

No caso da classificação aplicada às amostras de urina, obtiveram-se desempenhos globalmente modestos e superiores do observado no plasma, conforme o apresentado na figura 4.4.

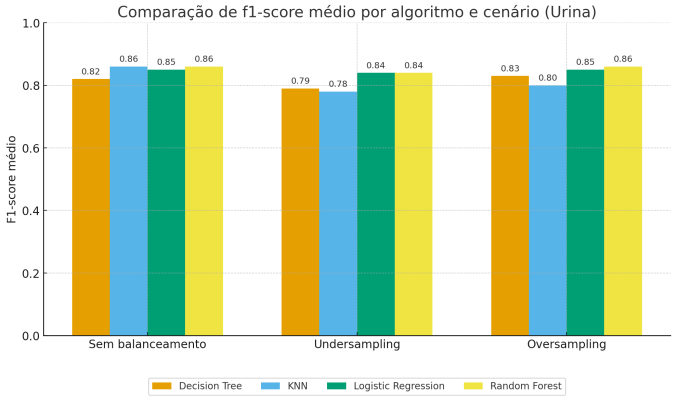

Figura 4.4: Comparação dos valores médios de F1-*score* obtidos com as amostras de urina

Em termos gerais, o *Random Forest* mostrou-se o modelo mais robusto e estável nos dados de urina, mantendo *f1-score* elevado em todos os cenários.

A Regressão Logística apresentou um perfil consistente e equilibrado entre cenários, com *f1-score* de 0.85 nos cenários sem balanceamento de classes e *oversampling* e *f1-score* = 0.84 no cenário de *undersampling*. Assim, este algoritmo constitui uma alternativa fiável quando se pretende estabilidade sem perda apreciável de desempenho.

O kNN sobressaiu no cenário sem balanceamento (*f1-score* médio de 0.86), mas revelou-se mais sensível às estratégias de reequilíbrio: com *f1-score* de cerca de 0.78–0.80.

Em contraste, a *Decision Tree* registou os valores mais baixos entre os modelos (tipicamente 0.79–0.83, consoante o cenário), evidenciando menor capacidade discriminativa neste conjunto.

Tanto nas amostras de plasma como de urina, não se verificaram ganhos sistemáticos de *f1-score* com a aplicação de *feature selection*, tendo-se até, verificado uma ligeira degradação do desempenho. Assim, a sua utilidade neste conjunto de dados revelou-se sobretudo interpretativa, nomeadamente para a identificação de varáveis preditoras.

Do ponto de vista clínico, modelos com elevada precisão — como o *Random Forest* (sobretudo na urina e no plasma com *oversampling*) — são relevantes para reduzir falsos positivos e, consequentemente, intervenções desnecessárias. No entanto, em cenários de deteção, a sensibilidade assume prioridade: a *Regressão Logística* e o kNN (em particular na urina, e o kNN com reamostragem no plasma) evidenciaram níveis de *recall* elevados.

Quanto à adequação dos modelos, os resultados diferenciam-se entre fluidos: no plasma, o melhor desempenho resultou da combinação de *Random Forest* com *SMOTE*, alcançando *f1-score* médio de 0.72. Este valor representa um desempenho moderado, útil como ferramenta de apoio, mas insuficiente para aplicação como teste isolado. Já na urina, o *Random Forest* e o kNN atingiram *f1-score* médios de 0.85–0.86, demonstrando desempenhos robustos e clinicamente promissores, ainda que condicionados à necessidade de validações externas. A discrepância de resultados poderá estar associada à diferente representatividade de doentes de Crohn em cada matriz: na urina, a proporção mais elevada de doentes de Crohn poderá ter contribuído para uma melhor capacidade discriminativa dos modelos, enquanto no plasma a menor presença desta classe limitou a sua performance.

Capítulo 5

Clustering

O *clustering* é uma abordagem de aprendizagem não supervisionada que tem como objetivo organizar um conjunto de observações em diferentes grupos (*clusters*), de forma a que os elementos de um mesmo grupo apresentem um grau de semelhança significativo entre si. No contexto deste projeto, o *cluster* foi aplicado para identificar padrões nos perfis metabólicos dos doentes, analisando as amostras de plasma e urina disponibilizadas.

Ao longo desta etapa, foram elaboradas duas tarefas distintas de *clustering*:

- **Clustering Geral**, aplicado a todas as amostras disponibilizadas (tanto de urina, como de plasma), independentemente do diagnóstico clínico. Esta etapa tinha como propósito a identificação de grupos e de possíveis subpopulações na amostra total.
- **Clustering de Crohn**, aplicado apenas às amostras pertencentes a doentes de Crohn, visando identificar potenciais sub-perfis metabólicos dentro deste grupo específico. Esta etapa poderá ser importante para analisar metabólito associados a possíveis agravamentos da doença, como as fístulas anais.

5.1 Técnicas de Clustering

Esta secção apresenta as técnicas de *clustering* utilizadas no estudo, apresentando os critérios adotados para definir a estrutura dos grupos e as estratégias aplicadas na sua implementação.

5.1.1 Determinação da Quantidade Ótima de Clusters

A primeira etapa de preparação para o *clustering* foi a seleção do número ideal de *clusters* (k). Este é um passo indispensável, sendo que influencia diretamente a qualidade e a interpretabilidade dos grupos gerados. Para o efeito, foram aplicados dois métodos distintos: o Método do Cotovelo e o Coeficiente da Silhueta.

Método do Cotovelo

O Método do Cotovelo (*Elbow Method*) consiste no cálculo da inércia (soma das distâncias quadráticas das amostras aos respetivos centróides), para diferentes valores de k (quantidade de *clusters*). Enquanto o valor de k aumenta, a inércia diminui, sendo que a melhoria deixa de ser significativa a partir de um determinado valor (o "cotovelo"da curva), que é então escolhido como o número de *clusters* a adotar [75].

Para o *Clustering* Geral, a quantidade sugerida de *clusters* pelo *Elbow Method* foi **k = 4** para as amostras de plasma e **k = 2** para as observações de urina, conforme o apresentado no gráfico da figura 5.1 .

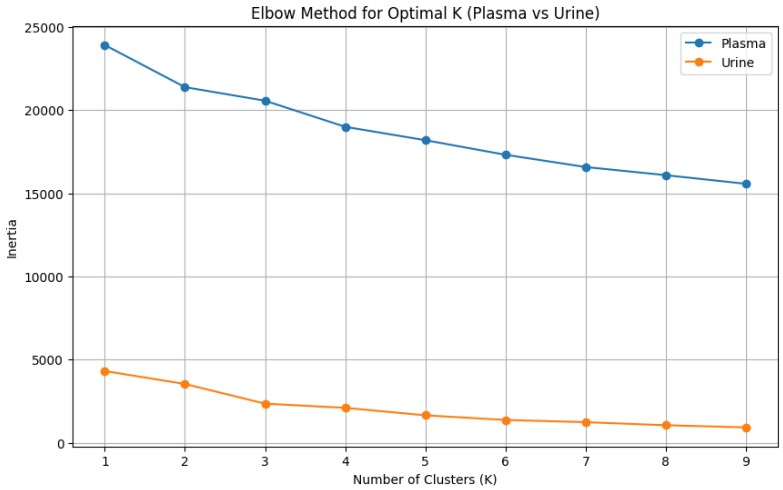

Figura 5.1: Método do Cotovelo (*Clustering* Geral)

No caso do *Clustering* específico em doentes de Crohn, o método do cotovelo indicou **k = 4** como a quantidade ideal de *clusters* para as amostras de plasma, e **k = 2** para as observações de urina (conforme apresentado na figura 5.2).

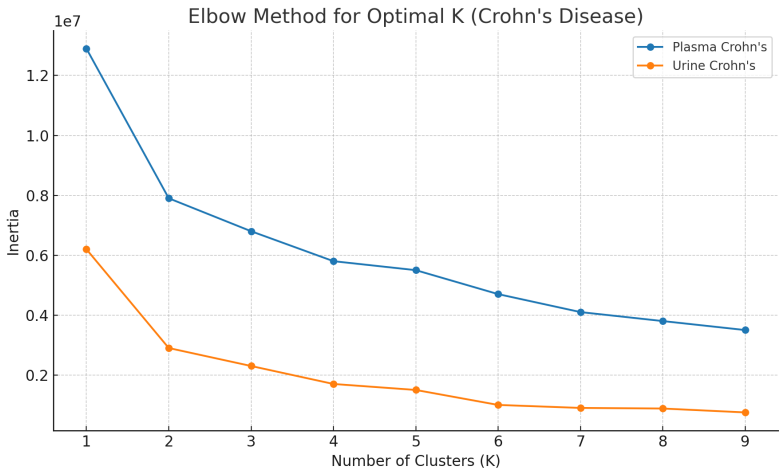

Figura 5.2: Método do Cotovelo (*Clustering* de Crohns)

Coeficiente de Silhueta

O Coeficiente de Silhueta permite encontrar o valor de grupos ideal, através da medição da coesão (distância média intra-*cluster*) e da separação (distância média para pontos do *cluster* mais próximo). O valor oscila entre -1 e 1, sendo que valores superiores indicam melhor separação entre os *clusters* [76].

Para o *Clustering* geral, os valores ideiais de *k* encontrados mantêm-se em concordância com o Método do Cotovelo, validando as escolhas de **k = 4** para as amostras de plasma e

de **k = 2** para as observações de urina. As tabelas 5.1 e 5.2 apresentam os valores obtidos para cada valor de k para as amostras de plasma e de urina, respetivamente.

Valor de k	Pontuação
$k = 2$	0.1242
$k = 3$	0.1239
$k = 4$	0.1321
$k = 5$	0.1294
$k = 6$	0.1134
$k = 7$	0.1214
$k = 8$	0.1186
$k = 9$	0.0985

Tabela 5.1: Pontuações do Coeficiente de Silhueta para cada valor de k (conjunto total de amostras de plasma).

Valor de k	Pontuação
$k = 2$	0.8126
$k = 3$	0.7790
$k = 4$	0.8117
$k = 5$	0.7050
$k = 6$	0.7071
$k = 7$	0.6643
$k = 8$	0.3092
$k = 9$	0.3091

Tabela 5.2: Pontuações do Coeficiente de Silhueta para cada valor de k (conjunto total de amostras de urina).

No *Clustering* de Crohns, os valores encontrados também se mantiveram razoavelmente consistentes, indicando **k = 3** para amostras de plasma e **k = 2** para as de urina. Nass tabelas 5.3 e 5.4 estão apresentados os valores obtidos para cada valor de k para as amostras de plasma e de urina, respetivamente.

Valor de k	Pontuação
$k = 2$	0.2309
$k = 3$	0.3895
$k = 4$	0.3675
$k = 5$	0.2366
$k = 6$	0.2097
$k = 7$	0.2057
$k = 8$	0.1927
$k = 9$	0.1868

Tabela 5.3: Pontuações do Coeficiente de Silhueta para cada valor de k (subconjunto de Crohn nas amostras de plasma).

Valor de k	Pontuação
$k = 2$	0.8547
$k = 3$	0.8429
$k = 4$	0.4676
$k = 5$	0.4062
$k = 6$	0.4107
$k = 7$	0.2150
$k = 8$	0.2186
$k = 9$	0.2148

Tabela 5.4: Pontuações do Coeficiente de Silhueta para cada valor de k (subconjunto de Crohn nas amostras de urina).

5.1.2 Estratégias aplicadas

A tarefa de *clustering* desenvolvida ao longo deste projeto, foi estruturada de forma a responder a duas necessidades distintas: por um lado, compreender a organização global das amostras, independentemente do diagnóstico associado, e por outro lado, analisar possíveis distinções intra-grupo no caso do diagnóstico positivo da Doença de Crohn. Assim, foram conduzidas duas análises de *clustering*: *clustering* geral e *clustering* específico para as amostras dos doentes de Crohn.

A seleção dos algoritmos foi orientada pela necessidade de explorar diferentes abordagens de modelação. No âmbito dos algoritmos particionais, elegeu-se o **K-Means**, capaz de segmentar os dados em k grupos mutuamente exclusivos. Este algoritmo é particularmente eficiente para grandes volumes de dados, uma vez que assume que os *clusters* são aproximadamente esféricos e possuem tamanhos idênticos [77]. Complementarmente, aplicou-se também, o algoritmo GMM, apto para modelar as amostras como se fossem provenientes de várias distribuições normais multivariadas. Em oposição ao **K-Means**, o **GMM** não impõe restrições rigorosas quanto à forma dos *clusters*, o que se pode tornar vantajoso em contextos onde a separação entre os grupos é gradual [78].

De forma a assegurar a consistência metodológica, implementou-se um modelo capaz de centralizar todo o fluxo de trabalho. Esta estrutura recebe como dados de entrada os *datasets* a analisar (dados completos para o *Clustering* Geral e as amostras de doentes de Crohn para o *Clustering* específico), o algoritmo a utilizar e a quantidade de *clusters* a ser considerada, de acordo com os resultados apresentados na secção 5.1.1. O modelo desenvolvido é responsável pela condução de todas as etapas essenciais: treino do modelo, atribuição dos rótulos de *cluster* para cada observação e cálculo da distribuição de diagnósticos por *cluster*.

Adicionalmente, os modelos integram funcionalidades de geração de representações visuais e execução de análises estatísticas, simplificando a leitura e validação dos grupos obtidos. De entre as ferramentas aplicadas, destaca-se a Análise de Componentes Principais (PCA), capaz de reduzir a dimensionalidade dos dados e projetá-la em duas dimensões, permitindo observar visualmente a separação dos grupos [79].

Além disso, foram gerados gráficos para facilitar a visualização dos metabolitos selecionados em cada *cluster*. Esta representação permite analisar o contributo de cada biomarcador para a caracterização dos grupos.

Finalmente, foi também incluída a técnica **Analysis of Variance (ANOVA)**, um procedimento estatístico capaz de comparar as médias de uma determinada variável em três ou mais grupos. O funcionamento deste método fundamenta-se na comparação de duas fontes de variação: variação entre grupos (inter-grupo) e a variação dentro dos grupos (intra-grupo). Embora a sua aplicação em estudos de metabolómica seja comum, a quantidade de comparações simultâneas representa um obstáculo: cada metabolito é testado individualmente e, sem correção, pode surgir uma quantidade considerável de falsos positivos. De forma a contornar esta barreira, recomenda-se a aplicação adicional do *False Discovery Rate* (FDR), que permite controlar a proporção esperada de resultados falsamente significativos [80]. Após a correção dos p-valores [1], os metabolitos são ordenados por q-*value*[2], focando na sua relevância. Relativamente à quantidade de compostos considerados, estudos recentes evidenciam vantagens da aplicação de listas mais curtas, como é o caso do *"Exploratory metabolomic analysis for characterizing the metabolic profile of the urinary bladder under estrogen deprivation"* [83], onde foram selecionados os dez metabolitos considerados mais relevantes para comparar a privação de estrogénios. Da mesma forma, os autores do estudo *"Metabolite signatures of metabolic risk factors and their longitudinal change"* [84], optaram pela avaliação dos dez metabolitos selecionados pelo ANOVA. Esta quantidade

[1] Um p-valor representa a probabilidade de obter resultados tão ou mais extremos quanto os que foram realmente obtidos num determinado estudo. Um p-valor reduzido indica que é improvável que os resultados observados tenham ocorrido por acaso e sustenta fortes evidências contra a hipótese nula [81].

[2] O q-*value* é uma métrica que permite ajustar os p-valores em contextos de testes múltiplos, indicando a taxa mínima de falsa descoberta (FDR) em que uma descoberta específica pode ser considerada estatisticamente significativa [82].

representa um compromisso entre a exaustividade e a clareza, reduzindo a complexidade e facilitando a interpretação da literatura. Assim, definiu-se que, para o presente estudo, seriam avaliados os dez aminoácidos considerados mais relevantes pela combinação do ANOVA com FDR para analisar e identificar os possíveis perfis metabólicos e biomarcadores relacionados com a Doença de Crohn. A análise de variância entre *clusters* foi implementada recorrendo a `scipy.stats.f_oneway` e à correção de múltiplas comparações pelo método FDR (`statsmodels.multipletests`). O excerto seguinte ilustra a identificação dos 10 metabolitos mais relevantes, classificados segundo o q-value.

```
import numpy as np
import pandas as pd
from scipy.stats import f_oneway
from statsmodels.stats.multitest import multipletests

cluster_col = "KMEANS_Cluster"  # or "GMM_Cluster"

# Select numeric features only (exclude IDs/labels)
features = [c for c in df.columns
            if c not in {cluster_col, "Diagnosis"}
            and np.issubdtype(df[c].dtype, np.number)]

pvals, Fvals, vars_ = [], [], []
for v in features:
    # Build groups
    groups = [g[v].dropna().values for _, g in df.groupby(
    cluster_col)]

    if len(groups) >= 2 and all(len(g) >= 2 for g in groups):
        # One-way ANOVA across clusters for variable v
        F, p = f_oneway(*groups)
        Fvals.append(F); pvals.append(p); vars_.append(v)

# Multiple testing correction (FDR)
qvals = multipletests(pvals, method="fdr_bh")[1]

# Rank variables by adjusted and keep the top-10
top10 = (pd.DataFrame({"variable": vars_, "F": Fvals, "p": pvals,
     "q": qvals})
          .sort_values("q", ascending=True)
          .head(10))
```

Listing 5.1: Excerto de código ilustrando o procedimento de ANOVA + FDR

5.2 Clustering Geral

A tarefa de *clustering* aplicada ao conjunto completo de amostras de plasma e urina, permitiu avaliar a existência de perfis metabólicos que pudessem ser associados à Doença de Crohn. Esta análise procurou avaliar a existência de padrões nos dados, de forma a perceber se a distribuição natural das amostras revela grupos distintos e biologicamente relevantes. O processo foi conduzido separadamente para plasma e urina, permitindo a comparação entre os dois tipos de matriz biológica.

5.2.1 Avaliação dos Clusters

Esta secção tem como objetivo verificar a qualidade da segmentação obtida e a sua capacidade de refletir padrões consistentes no perfil metabólico das amostras.

Amostras de Plasma

No *clustering* aplicado às observações de plasma, ambas as abordagens - **K-Means** e **GMM** - produziram soluções contendo quatro grupos de estrutura semelhante.

A tabela 5.5 representa a distribuição dos diagnósticos por *cluster* obtida com a aplicação do *K-Means*. Sublinha-se que os valores percentuais presentes na última coluna da tabela se referem à composição interna de cada *cluster*, quantificando a presença do diagnóstico positivo em cada grupo.

Por observação da tabela, verifica-se que o **Cluster C2** possui uma elevada quantidade de amostras de doentes de Crohn (cerca de 95% da constituição do grupo). Por outro lado, os restantes *clusters* apresentam proporções reduzidas deste diagnóstico, sendo que o *cluster* C3 não possui uma única observação da doença, e C0 e C1 contêm proporções semelhantes entre si, de $\approx 15\%$ e de $\approx 14\%$, respetivamente.

Cluster	Outros doentes com DHM	Doentes de Crohn	% de Crohns
C0	405	71	$\approx 15\%$
C1	380	64	$\approx 14\%$
C2	2	41	$\approx 95\%$
C3	33	0	0

Tabela 5.5: Distribuição dos diagnósticos nos *clusters* gerados pelo algoritmo *K-Means* nas amostras de plasma

Esta distribuição sustenta a teoria da existência de um grupo metabólico fortemente associado à Doença de Crohn, incentivando a análise das amostras constituintes neste grupo, bem como dos seus constituintes metabólicos.

As projeções de PCA obtidas com o *K-Means*, observáveis na figura 5.3, apresentam os quatro grupos resultantes, visivelmente separados. Evidenciam-se dois perfis extremos: o **Cluster C2**, maioritariamente populado com amostras de doentes de Crohn, e **C3**, sem presença de casos da doença.

Esta separação não é somente geométrica, mas proveniente de disparidades significativas na concentração de múltiplos compostos biológicos. De forma a quantificar os determinantes da separação observadas no gráfico PCA (gráfico 5.3), efetuou-se uma avaliação, por ANOVA, dos dez aminoácidos mais relevantes e comparou-se, para cada *cluster*, a sua média com a média global da totalidade de amostras (tabela 5.6).

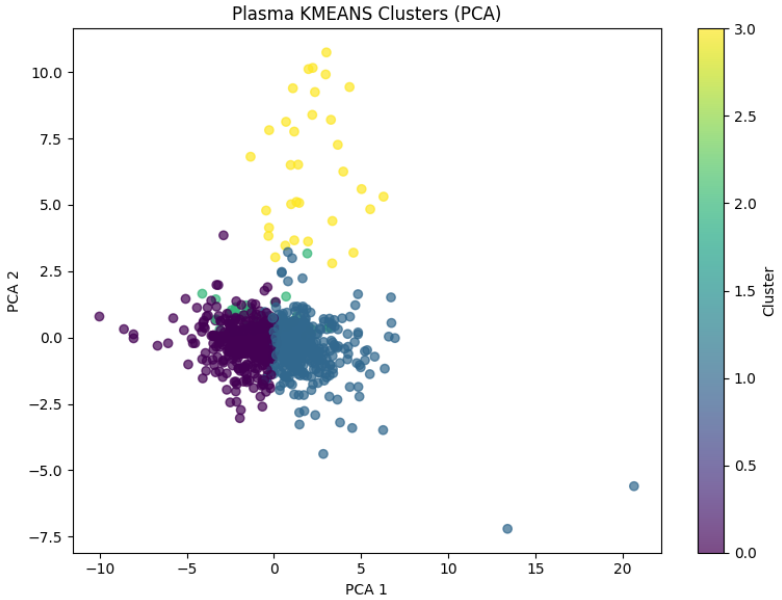

Figura 5.3: PCA do *clustering* geral com *K-Means* para amostras de plasma

Metabolito	Média global	C0 (média)	C1 (média)	C2 (média)	C3 (média)
Cysta	4.76	1.2	2.5	3.1	88.4
Glu	77.15	60.0	73.3	333.8	42.5
Leu	143.12	113.7	152.6	141.3	441.8
Ile	63.72	48.4	68.4	69.0	214.7
Thr	136.90	106.2	166.8	128.4	188.6
Lys	174.12	145.0	206.3	172.5	162.8
Gln	604.93	567.6	678.4	264.7	598.4
Asn	77.48	67.0	90.4	70.7	63.8
PRO	195.83	151.9	240.1	202.0	224.4
His	79.40	70.9	89.1	70.0	83.1

Tabela 5.6: Top-10 aminoácidos (ANOVA) — médias absolutas por *cluster* e média global para amostras de plasma com aplicação do *K-Means*

Os valores médios absolutos obtidos com o *K-Means* apontam para uma distribuição de componentes metabólicos que espelha a distribuição dos diagnósticos previamente apresentada na tabela 5.5.

Os grupos **C0** e **C1** representam *clusters* heterogéneos, possuindo maioritariamente amostras de controlos, embora também incluam observações de doentes de Crohn. Os perfis metabólicos destes *clusters* apresentam valores equilibrados, sem alterações extremas, sustentando a sua composição mista.

O *cluster* **C2** é particularmente relevante, sendo praticamente constituído por amostras de doentes de Crohn. O seu perfil é marcado por valores elevados de Glutamato (Glu) e por níveis reduzidos de Glutamina (Gln), refletindo uma forte desregulação do eixo Glu/Gln. Assim, este resultado sugere que este grupo poderá corresponder a uma assinatura metabólica característica da Doença de Crohn.

Por outro lado, o *cluster* C3, sem presença de amostras de doentes de Crohn, apresenta um perfil caracterizado por *Branched-Chain Amino Acids* (BCAA) [3] com concentrações marcadamente elevadas de Leucina (Leu) e Isoleucina (Ile). Também a Cistationina (Cysta) se encontra acentuadamente alta, enquanto que o Glutamato (Glu) apresenta concentrações baixas e a Glutamina (Gln) se mantém próxima da média global.

Assim, pode haver tendência para um eixo **Glu/Gln pró-doença de Crohn** (combinando elevadas concentrações de Glutamato com concentrações mais reduzidas de Glutamina) e um eixo BCAA/Cysta protetivo no contexto da doença (elevadas concentrações de BCAA e de Cistationina). Os *boxplots* das figuras 5.4 e 5.5 sustentam as disparidades encontradas.

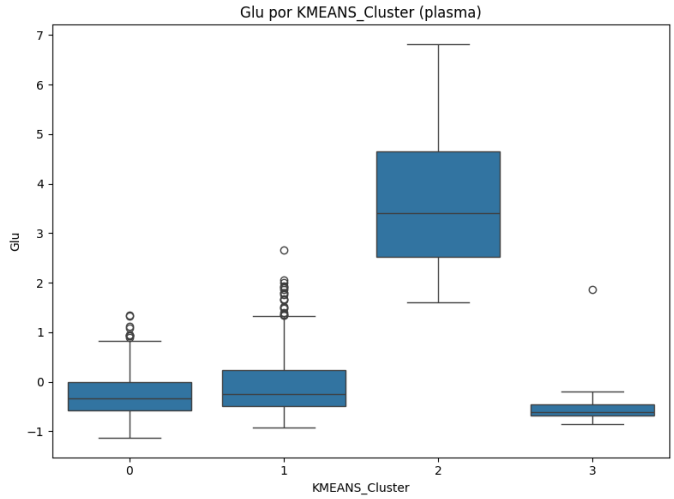

Figura 5.4: Variância da concentração de Glutamato por *cluster*

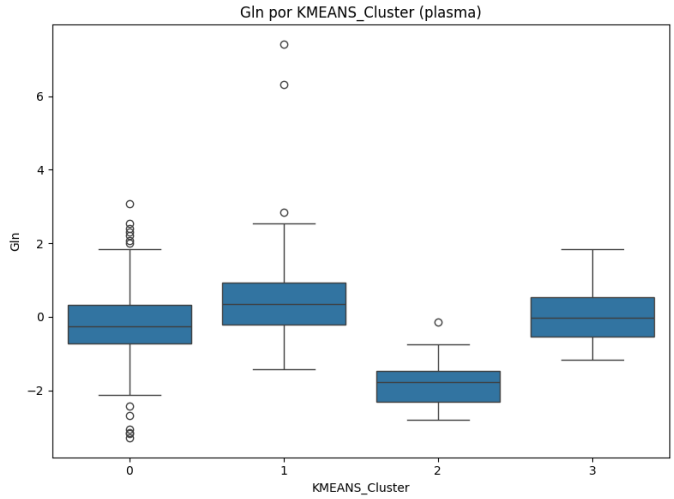

Figura 5.5: Variância da concentração de Glutamina por *cluster*

[3] Os **Branched-Chain Amino Acids (BCAA)** incluem os aminoácidos Leucina, Isoleucina e Valina.

A aplicação do algoritmo **GMM** nas amostras de plasma obteve uma solução de quatro grupos com as distribuições de diagnóstico apresentadas na tabela 5.7. À semelhança do *K-Means*, o **cluster C2** apresenta a maior proporção de casos de Doença de Crohn ($\approx 26\%$), enquanto o **cluster C3** não contém casos positivos do diagnóstico. Uma vez mais, os restantes *clusters* apresentam quantidades residuais de observações pertencentes a doentes de Crohn ($\approx 1\%$ e $\approx 3\%$). Enquanto o *K-Means* fragmenta regiões alongadas (C0 e C1), por impor fronteiras esféricas, o GMM modela a covariância e, por isso, capta a mesma região como um único núcleo elíptico (C2). Por outro lado, em ambos os modelos, o grupo sem casos de Crohn (C3) mantém-se estável.

Cluster	Outros doentes com DHM	Doentes de Crohn	Proporção de casos de Crohn
C0	176	6	$\approx 3\%$
C1	128	1	$\approx 1\%$
C2	476	169	$\approx 26\%$
C3	40	0	0

Tabela 5.7: Distribuição dos diagnósticos nos *clusters* gerados pelo algoritmo *GMM* nas amostras de plasma

Na figura 5.6 é apresentada a projeção a duas dimensões (PCA) dos grupos obtidos com o algoritmo GMM. Novamente, os grupos mais destacados são o C2 e C3.

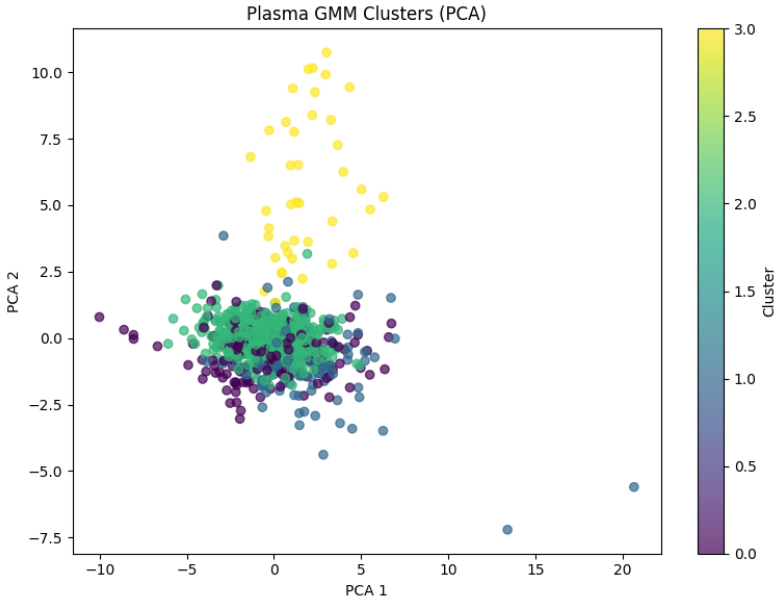

Figura 5.6: PCA do *clustering* geral com *GMM* para amostras de plasma

Após a obtenção dos resultados da execução do modelo implementado com recurso ao GMM, foi também, aplicado o ANOVA, de forma a comparar as concentrações dos aminoácidos mais relevantes na tarefa. A tabela 5.8 espelha as concentrações médias globais e individuais, para cada *cluster* identificado, dos componentes selecionados por esta técnica.

Em semelhança ao identificado com o *K-Means*, os grupos **C0** e **C1**, apresentam valores equilibrados para a maior parte dos aminoácidos. Assim, e tendo em conta a baixíssima representação de amostras de doentes de Crohn, estes perfis poderão ser menos relevantes para a caracterização da doença.

Metabolito	Média global	C0 (média)	C1 (média)	C2 (média)	C3 (média)
Cysta	4.76	1.3	3.8	1.5	76.7
Leu	143.12	126.5	119.8	135.9	410.9
Ile	63.72	52.7	53.5	60.3	201.8
Met	31.09	30.4	73.1	23.2	26.8
Sarc	54.35	58.4	68.3	50.2	57.7
Orn	112.51	92.2	170.9	105.3	133.3
Gln	604.93	674.6	690.9	566.3	633.9
Ala	365.12	378.8	448.8	349.9	278.8
Thr	136.90	129.6	163.1	131.0	180.0
Glu	77.15	63.93	75.60	83.11	46.25

Tabela 5.8: Top-10 aminoácidos (ANOVA) — médias absolutas por *cluster* e média global para amostras de plasma com aplicação do *GMM*

O *cluster* **C2** pode ser considerado como o mais relevante, por reunir a maior parte das observações de Crohn. Este grupo apresenta valores elevados de Glutamato e redução de Glutamina, reforçando a importância do eixo **Glu/Gln** na caracterização da Doença de Crohn.

Finalmente, o grupo **C3**, que não possui uma única amostra de Crohn, destaca-se por possuir concentrações de BCAA elevadas, podendo refletir um perfil mais anabólico, não associado à Doença de Crohn.

Amostras de Urina

Nas amostras de urina com aplicação do **K-Means**, obtiveram-se dois grupos extremamente desequilibrados, conforme apresentado na tabela 5.9. O grupo **C0** contém praticamente a totalidade de amostras de urina disponíveis, sendo constituído por 254 observações e agrupando a totalidade das amostras de Crohn. Por outro lado, o *cluster* **C1** apenas possui uma amostra isolada, sem diagnóstico positivo. A existência de um *cluster* com uma única amostra impossibilita a aplicação de ANOVA e indica que a estrutura é dominada por um *outlier* e não por grupos estáveis.

Cluster	Outros doentes com DHM	Doentes de Crohn	Proporção de casos de Crohn
C0	79	175	$\approx 69\%$
C1	1	0	0

Tabela 5.9: Distribuição dos diagnósticos nos *clusters* gerados pelos dois algoritmos (*K-Means* e GMM) nas amostras de urina

A utilização do **GMM** no mesmo conjunto de dados resultou na repetição do cenário, possuindo exatamente a mesma distribuição observada com o algoritmo alternativo.

Face a esta distribuição, realizaram-se duas análises adicionais nas amostras de urina. Primeiro, manteve-se o $k = 2$ (valor indicado pelos métodos do cotovelo e da silhueta) e removeu-se a amostra isolada, tendo-se obtido os resultados apontados na tabela 5.10.

Cluster	Outros doentes com DHM	Doentes de Crohn	Proporção de casos de Crohn
C0	76	175	$\approx 70\%$
C1	3	0	0

Tabela 5.10: Distribuição dos diagnósticos nos *clusters* gerados pelos dois algoritmos (*K-Means* e GMM) nas amostras de urina, excluindo a amostra isolada

Estes resultados são também desproporcionais, com C0 a concentrar a grande maioria dos casos (76 casos de outros doentes com DHM e 175 com Crohn) e um C1 residual com somente 3 observações relativas a outros doentes com DHM. Assim, esta análise é insuficiente para constituir uma inferência estática robusta.

Adicionalmente, realizou-se outra experiência, aplicando $k = 3$ (apesar de não ter sido o valor indicado pelos métodos) à totalidade de amostras de urina (tabela 5.11).

Cluster	Outros doentes com DHM	Doentes de Crohn	Proporção de casos de Crohn
C0	77	171	$\approx 69\%$
C1	1	0	0
C2	2	4	$\approx 66\%$

Tabela 5.11: Distribuição dos diagnósticos nos *clusters* gerados pelos dois algoritmos (*K-Means* e GMM) nas amostras de urina, com $k=3$

Ainda assim, obteve-se novamente um *cluster* C0 dominante (contendo 77 amostras de outros doentes com DHM e 171 observações de Crohn), um *cluster* C1 quase vazio (contendo apenas uma amostra não-Crohn) e um micro-*cluster* C2 (com somente 2 observações de outras patologias e 4 amostras de Crohn). Desta forma, a disparidade extrema entre os *clusters* obtidos não permitiu a obtenção de análises comparativas estáveis. Em semelhança à primeira experiência realizada, estas duas análises adicionais demonstraram os mesmos resultados independentemente do algoritmo aplicado.

A análise conjunta destes resultados sugere que o plasma exibe uma estrutura de *clusters* interpretável e clinicamente mais relevante, com um grupo que consistentemente não apresenta casos (C3) e outro bastante rico em observações referentes a diagnósticos de Crohn (C2). Em contrapartida, a urina, no conjunto de dados geral, não evidencia segmentação informativa e é fortemente condicionada por uma observação atípica. Esta assimetria comportamental orienta a matriz de plasma como fonte primária para a identificação de perfis no *clustering* geral.

5.2.2 Interpretação dos Grupos

A análise dos perfis metabólicos no plasma permitiu a identificação de dois grupos com características distintas. No caso dos **clusters C2**, correspondente ao subconjunto de indivíduos com maior prevalência de diagnósticos de Crohn, verificou-se um conjunto de características que podem remeter para um estado de catabolismo e de intensa atividade inflamatória intestinal. Entre as diferenças mais significativas, destacam-se a elevada concentração do Glutamato (Glu), acompanhada por uma redução acentuada da quantidade de Glutamina

(Gln). Este desequilíbrio pode sugerir maior catabolismo[4] de proteínas e de aminoácidos e uma maior utilização de compostos azotados como fontes de energia [86]. Estes processos estão, muitas vezes, associados a respostas inflamatórias, *stress* metabólico ou à crescente necessidade de reparação de tecidos, que em conjunto, representam sintomas recorrentes da Doença de Crohn [87]. Em simultâneo, verificou-se que os aminoácidos de cadeia ramificada (BCAA), nomeadamente a Leucina (Leu) e a Isoleucina (Ile), apresentaram valores significativamente inferiores em comparação com as concentrações médias presentes no *cluster* C3. Estes componentes são essenciais para a produção proteica e para a preservação da massa muscular. A sua quantidade ínfima no *cluster* C2 é compatível com lesões musculares ou estados de défice de nutrientes, frequentemente associados à atividade inflamatória crónica [88]. Adicionalmente, os níveis reduzidos de **Cistationina (Cysta)**, sugerem perturbações na via intestinal e aumento do *stress* oxidativo intestinal. Em suma, a redução simultânea de Ácido β-aminoisobutírico (BAIBA) e de Cistationina, combinada com o aumento de Glutamato e a diminuição de Glutamina, Leucina e Isoleucina, define um perfil metabólico pró-inflamatório, observado nos grupos com maior proporção de análises doentes de Crohn.

No *cluster* C3, por outro lado, sublinha-se o perfil oposto: os valores de Leucina (Leu), Isoleucina (Ile) e de Cistationina (Cysta) estão expressivamente mais elevados, enquanto o Glutamato (Glu) se mantém reduzido e a Glutamina (Gln) elevada. Este é um padrão consistente com um estado metabólico mais equilibrados e sem sinal aparente de inflamação ativa. Adicionalmente, a presença abundante dos aminoácidos de cadeia ramificada e de Cistationina sugere uma boa integridade muscular.

A coerência verificada entre os resultados obtidos em ambas as abordagens de *clustering*, reforça a robustez das conclusões alcançadas. Os dois modelos encontraram um grupo minoritário (C3) metabolicamente coeso e sem presença de casos de Crohn, e um grupo fortemente associado à doença (C2) que contém as alterações anteriormente descritas. As observações apresentadas corroboram a hipótese de que determinadas alterações na concentração dos compostos metabólicos plasmáticos, nomeadamente o aumento do Glutamato combinado com a redução da Glutamina, a supressão dos amniácidos de cadeia ramificada e dos derivados da BAIBA e Cistationina, refletem a presença de atividade intestinal inflamatória, bem como a possibilidade de desenvolvimento de lesões musculares associadas à Doença de Crohn.

5.2.3 Análise de Outliers

De forma a aumentar a robustez da análise com *clustering*, foi, também, realizado um estudo dos *outliers*, ou seja, as amostras cujo perfil metabólico se afasta de forma significativa do padrão dominante. A deteção destas observações, foi realizada com recurso ao algoritmo *Isolation Forest*, capaz de identificar as amostras que mais se afastam da maior concentração de observações.

Na tarefa de *clustering* incluindo o conjunto de amostras na sua totalidade, foram encontrados 50 *outliers* no plasma (cerca de 3%), distribuídos de forma equilibrada entre dois grupos internos definidos pelo algoritmo: 26 observações no *cluster* interno 0 e 24 no *cluster* interno 1. A partir da análise dos perfis destas observações, verificaram-se alguns casos com valores extremamente elevados de Glutamato, Alanina e de BCAAs , bem como situações com valores de Glutamina bastante reduzida. Assim, os *outliers* encontrados podem ser

[4] O catabolismo é um processo metabólico segundo o qual são produzidas moléculas simples a partir de outras mais complexas. Um exemplo deste processo é a síntese de aminoácidos a partir de proteínas [85].

representativos de estados de *stress* metabólico severo ou, em alternativa, erros de medição. A figura 5.7 é apresentada a projeção a duas dimensões do *outliers* encontrados nas amostras de plasma.

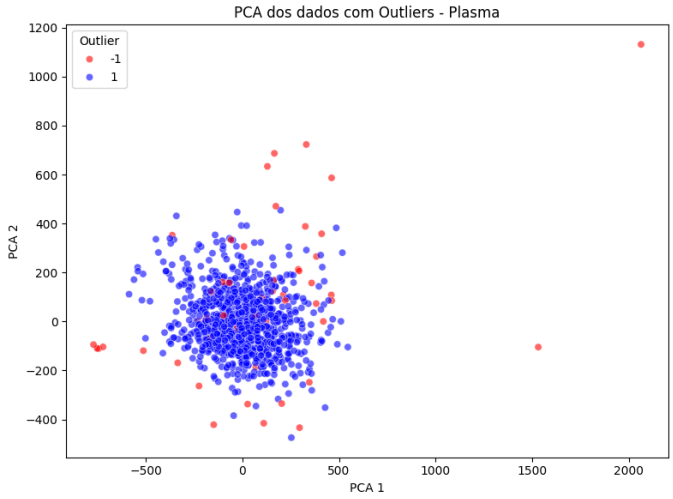

Figura 5.7: PCA dos *outliers* encontrados no *clustering* geral de amostras de plasma

Nas amostras de urina, identificaram-se 13 *outliers* (aproximadamente 5%), agrupados maioritariamente num micro-*cluster* com valores execionalmente elevados para vários componentes. Na figura 5.8, onde está representada a distribuição das amostras, observa-se que os *outliers* estão situados nos extremos dos eixos principais, afastando-se do conjunto principal de amostras. Os *outliers* encontrados possuem concentrações anormalmente elevadas de determinados aminoácidos e, noutros casos, valores extremamente reduzidos de Glutamina. À semelhança do plasma, estes perfis extremos podem indicar problemas analíticos ou perfis em *stress* fisiológico intensivo.

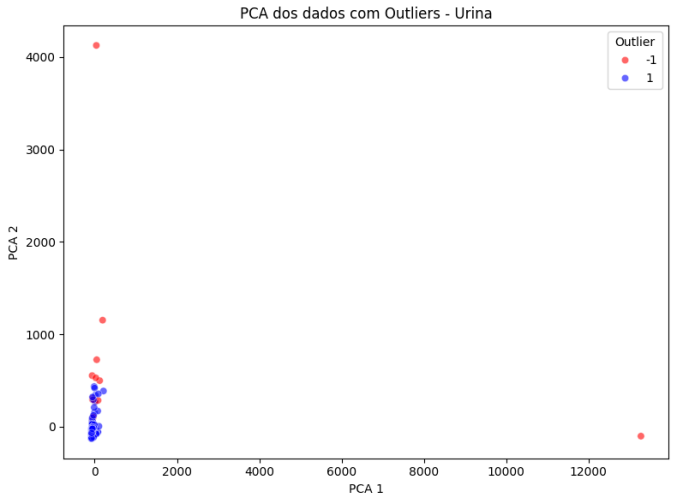

Figura 5.8: PCA dos *outliers* encontrados no *clustering* geral de amostras de urina

5.3 Clustering de Crohn

Com o objetivo de identificar sub-perfis metabólicos dentro do conjunto de doentes de Crohn, realizou-se uma tarefa de *clustering* somente aplicada ao conjunto de observações com o diagnóstico positivo. Desta forma, é possível analisar se alguns dos grupos encontrados exibem uma assinatura bioquímica compatível com maior atividade intestinal e/ou lesões musculares, características frequentemente associadas a complicações, por exemplo fístulas anais. Para tal, foram seguidos os mesmos passos usados no *Clustering* Geral, tendo-se iniciado pela determinação da quantidade ideal de grupos. A análise foi também conduzida de forma independente para plasma e urina.

5.3.1 Avaliação dos Clusters

A presente secção descreve a etapa de avaliação dos *clusters* obtidos, processo fundamental para aferir a robustez dos agrupamentos e estimar diferenças relevantes entre os indivíduos analisados.

Amostras de Plasma

Nas amostras de plasma, o *clustering* com **K-Means** apresentou uma distribuição relativamente equilibrada, conforme apresentado na tabela 5.12.

Cluster	Quantidade de Amostras
C0	66
C1	45
C2	5
C3	60

Tabela 5.12: Distribuição das amostras nos *clusters* gerados pelo *K-Means* nas amostras de plasma (*clustering* de Crohns)

Por outro lado, o **GMM** também encontrou quatro grupos, mas detetou um *cluster* maioritário e dois grupos intermédios: C0 com 116 amostras, C1 com 26, C2 com 5 observações e C3 com 29.

Cluster	Quantidade de Amostras
C0	116
C1	26
C2	5
C3	29

Tabela 5.13: Distribuição das amostras nos *clusters* gerados pelo *GMM* nas amostras de plasma (*clustering* de Crohns)

Embora os tamanhos absolutos nas duas abordagens não coincidam, foi detetado um micro-grupo constituído por apenas cinco observações clínicas, com perfis metabólicos extremos. Por ser um nicho tão distinto, este grupo merece uma interpretação cuidadosa, pois pode corresponder a um subtipo raro, a complicações da doença ou a erros de medição.

As projeções apresentadas nas figuras 5.9 e 5.10 demonstram claramente que existem diferenças entre a forma de distribuição dos doentes consoante a abordagem aplicada: enquanto

que no *K-Means* há limites mais bem definidos entre os núcleos das amostras, no GMM as mudanças entre os núcleos são graduais. Ainda assim, os dois algoritmos apresentam distribuições relativamente semelhantes, contendo um grupo central compacto, dois grupos mais dispersos e um pequeno grupo isolado. Apesar da variação do ponto exato de separação entre os grupos C0 e C3, a organização das amostras e as diferenças de perfil entre esses grupos permanecem claras.

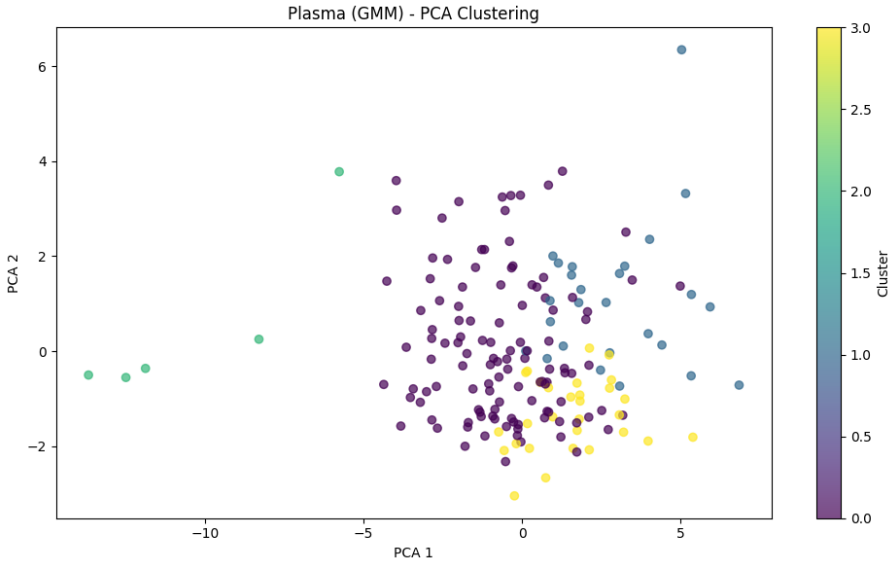

Figura 5.9: PCA do *clustering* geral com *K-Means* para amostras de plasma

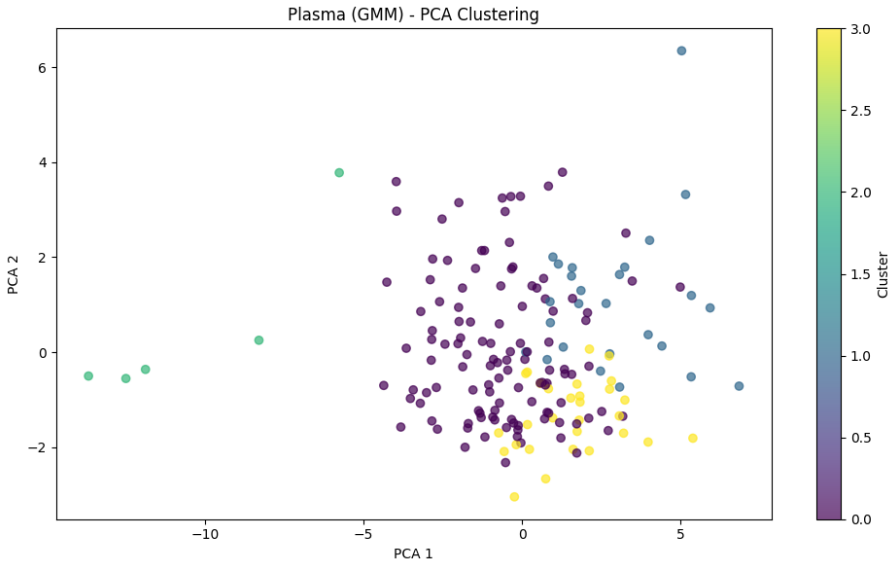

Figura 5.10: PCA do *clustering* geral com *GMM* para amostras de plasma

Amostras de Urina

Tal como se verificou no *Clustering* Geral, os resultados do *clustering* apenas em amostras de urina de doentes de Crohn, não revelou uma estrutura interna estável, conforme apresentado na tabela 5.14. Tanto o *K-Means* como o GMM, geraram um *cluster* C0 contendo a maior parte das amostras (175) e um grupo com apenas 5 amostras. Assim, as amostras de urina não contém informação suficiente para identificar diferentes perfis dentro dos Doentes de Crohn. Desta forma a análise dos perfis centrar-se -á somente nos resultados de plasma.

Cluster	Quantidade de Amostras
C0	170
C1	5

Tabela 5.14: Distribuição das amostras nos *clusters* obtidos nas amostras de urina

5.3.2 Interpretação dos Grupos

Após a identificação da existência de quatro grupos distintos nas amostras de plasma, a exploração dos perfis metabólicos centra-se na análise das médias dos aminoácidos que distinguem cada *cluster*. Para tal, recorreu-se ao ANOVA para identificar, em cada abordagem, os dez compostos mais relevantes na distinção dos sub-grupos dentro do conjunto completo de amostras de Crohn. De seguida, foram calculadas as médias das concentrações de cada um desses componentes em cada *cluster* e no conjunto global, facilitando a análise dos diferentes perfis encontrados. A tabela 5.15 apresenta os valores resultantes da aplicação desta técnica para o modelo implementado com recurso ao *K-Means*.

Metabolito	Média global	C0 (média)	C1 (média)	C2 (média)	C3 (média)
Lys	180.81	154.3	207.2	51.0	201.0
Met	22.91	18.7	27.5	7.0	25.4
Ala	360.46	311.8	451.2	63.6	370.6
Thr	137.24	117.5	147.5	34.2	159.8
Leu	141.04	122.3	168.5	40.4	149.5
Tyr	60.11	50.9	76.2	17.6	61.7
Baiba	0.66	0.3	0.5	8.6	0.5
Orn	116.12	95.2	132.1	18.8	135.3
His	74.69	64.6	77.4	51.6	85.7
Ile	65.93	56.8	85.1	21.2	65.4

Tabela 5.15: Top-10 aminoácidos (ANOVA) — médias absolutas por *cluster* e média global no conjunto de doentes de Crohn com aplicação do *K-Means*

Conforme referido anteriormente, o ANOVA testa as diferenças entre os *clusters* formados. Desta forma, os marcadores mais relevantes na distinção entre doentes de Crohn e controlos (como a Glutamina, o Glutamato e a Cistationina) podem não ser os mais discriminativos entre os sub-*clusters* de Crohn. Ainda assim, apesar de a seleção automática não os ter apontado, estes aminoácidos foram explicitamente analisados na tabela 5.16 devido à importância que demonstraram durante o *clustering* Geral (analisado na secção 5.2), tendo sido apontados como potenciais biomarcadores da Doença de Crohn, podendo ser também relevantes na identificação de grupos intra-Crohn.

Metabolito	Média global	C0 (média)	C1 (média)	C2 (média)	C3 (média)
Glu	77.15	59.96	73.31	333.84	42.45
Gln	604.93	567.55	678.43	264.70	598.39
Cysta	4.76	1.25	2.47	3.07	88.42

Tabela 5.16: Comparação das médias absolutas por *cluster* e média global da Glutamina e Glutamato para amostras de plasma com aplicação do *K-Means* (*Clustering* de Crohns)

O **cluster C0** encontra-se numa posição intermédia, possuindo, de modo geral, valores próximos da média global. No entanto, apresenta valores de Lisina (Lys), Leucina (Leu) e Treonina (Thr) ligeiramente abaixo da média. Simultaneamente, a Cistationina (Cys) e a BAIBA também se revelam em concentrações reduzidas. Estes padrões podem estar associados a um estado inflamatório moderado, compatível com doentes em fase de estabilização ou em recuperação parcial.

O **cluster C1** distingue-se por exibir valores elevados para a maioria dos aminoácidos, com destaque para os aminoácidos de cadeia ramificada (Leunina e Isoleunina), para a Alanina (Ala) e para o Glutamato (Glu), que embora próximo da média global, apresenta valores mais elevados do que C0 e C3. Neste *cluster*, a Glutamina (Gln) atinge o valor mais elevado entre os grupos, refletindo a possibilidade de maior disponibilidade anabólica. Este perfil pode ser indicativo de um estado pró-inflamatório e catabólico, refletindo uma atividade mais intensa do sistema imunitário [88]. A concentração extrema de Alanina (Ala) pode sugerir a sua libertação como essência para a glucogénese e os valores altos de Cistationina (Cys) podem indicar a ativação da via da transulfuração, processo metabólico no qual são sintetizadas a Cisteína e a Glutationina [89] [90]. Assim, este grupo poderá ser ser correspondente a um nicho de doentes com maior atividade da Doença de Crohn ou com complicações da mesma, como fístulas anais, onde a sobrecarga inflamatória e a perda de massa magra são recorrentes.

Por outro lado, o **cluster C2** agrupa somente cinco observações e revela um perfil metabólico extremo: as concentrações de Glutamina (Gln) e Alanina (Ala) estão muito reduzidas e BAIBA e Glutamato (Glu) extremamente elevados. Os BCAA apresentam-se, também, em concentrações ínfimas. A combinação Glutamina/Alanina reduzida pode sugerir o esgotamento de substratos, enquanto que a quantidade elevada de BAIBA é sugestão de atividade física intensiva ou catabolismo intensivo de Valina [91]. O conjunto de valores elevados de BAIBA com as concentrações reduzidas de BCAA pode corresponder a um fenótipo raro que pode estar associado a exercício intensivo ou a perturbações musculares específicas. No entanto, os dados fornecidos não fornecem informações suficientes para analisar este microperfil com rigor. Ainda assim, a persistência deste grupo restrito em ambas abordagens, aponta para um subtipo consistente merecedor de investigação futura.

Por fim, o **cluster C3** diferencia-se por apresentar valores baixos de Glutamato (Glu) e uma concentração de Glutamina (Gln) bastante elevada. Simultaneamente, este grupo apresenta concentrações relativamente elevadas dos aminoácidos de cadeia ramificada (Leucina e Isoleucina) e de Cistationina (Cysta). Os níveis elevados de Leucina (Leu) e de Isoleucina (Ile) apontam para um estado de integridade muscular e balanço anabólico. A Cistationina (Cysta) exagerada pode ser reflexo do funcionamento regular da via de transulfuração. Este padrão metabólico poderá correspondem a um conjunto específico de doentes, onde a transulfuração desepenha um papel importante na regulação da resposta inflamatória. Por outro

lado, os valores díspares de Cistationina (Cysta) poderão, também, corresponder a possíveis erros clínicos.

A mesma análise foi conduzida para os resultados obtidos com recurso ao GMM, tendo-se obtido os valores apresentados nas tabelas 5.17 e 5.18.

Metabolito	Média global	C0 (média)	C1 (média)	C2 (média)	C3 (média)
Leu	141.04	129.9	112.2	131.7	252.3
Lys	180.81	187.4	141.1	184.6	206.7
Baiba	0.66	0.5	0.1	0.2	3.7
Ile	65.93	66.9	50.7	60.8	105.6
Met	22.91	24.0	19.8	24.6	25.7
Orn	116.12	119.3	90.1	112.3	138.7
Thr	137.24	140.1	109.6	135.0	160.5
Tyr	60.11	61.5	52.2	60.5	67.0
Asn	74.54	76.8	60.6	74.6	83.1
PRO	191.73	190.7	180.3	196.5	197.1

Tabela 5.17: Top-10 aminoácidos (ANOVA) — médias absolutas por *cluster* e média global no conjunto de doentes de Crohn com aplicação do *GMM*

Metabolito	Média global	C0 (média)	C1 (média)	C2 (média)	C3 (média)
Glu	77.15	93.93	75.60	83.11	46.25
Gln	604.94	674.61	690.95	566.27	633.88
Cysta	4.76	1.34	3.82	1.45	76.70

Tabela 5.18: Comparação das médias absolutas por *cluster* e média global da Glutamina, Glutamato e Cistationina para amostras de plasma com aplicação do *GMM* (*Clustering* de Crohns)

Nesta abordagem, em semelhança ao observado na avaliação com *K-Means*, o **cluster C0** apresenta concentrações com valores próximos às médias globais. Da mesma forma, o **cluster C2** também se encontra relativamente próximo da média global. Assim, ambos os perfis podem ser reflexo de um estado intermédio entre equilíbrio e atividade inflamatória.

O **cluster C1** apresenta valores ligeiramente inferiores às médias globais, indicando um perfil com menor pronunciação catabólica que o observado na abordagem com recurso ao *K-Means*.

Finalmente, o **cluster C3**, representa um perfil mais disitinto, possuindo concentrações acima da média para a maioria dos aminoácidos, destacando-se os níveis elevados de BCAA, Lisina (Lys), Treonina (Thr) e Prolina (Pro). Como referido anteriormente, os BCAA são compostos essenciais para a síntese proteica e manutenção da massa muscular [88], enquanto que a Lisina (Lys) e a Treonina (Thr) são importantes para a reparação tecidular e para a síntese de mucinas intestinais, respetivamente [92]. Por outro lado, a Prolina (Pro) é um aminoácido indispensável para a síntese de colagénio e para a regeneração da mucosa [93]. Desta forma, a presença destes compostos em concentrações elevadas, indica que este perfil poderá indicar menor gravidade clínica, preservando aminoácidos essenciais (BCAA, Lisina, Treonina) e destacando metabolitos de reparação tecidual (Prolina). Assim, este conjunto

poderá sugerir uma maior capacidade regenerativa, implicando uma menor probabilidade de complicações severas, como as fístulas anais.

A leitura das tabelas anteriores evidencia a existência de diferentes perfis metabólicos dentro do conjunto de doentes de Crohn, tendo-se verificado uma concordância entre os resultados obtidos pelo *K-Means* e pelo GMM: ambos identificaram a relevância dos BCAA e da Lisina como principais biomarcadores. No entanto, existem diferenças entre os dois modelos: enquanto que o modelo baseado em *K-Means* enfantizou a existência de um perfil extremo, marcado por excesso de atividade metabólica e elevação da concentração de BAIBA, o modelo com aplicação de GMM evidenciou grupos com perfis mais equilibrados. Estas diferenças sugerem que a aplicação conjunta de ambas as abordagens poderá ser benéfica para auxiliar a interpretação dos fenótipos metabólicos da Doença de Crohn.

Em síntese, a estratificação metabólica identificou os perfis **C1** e **C2** originados pela aplicação do *K-Means* como os perfis com maior potencial de complicações severas como fístulas perianais: o grupo C1 por possuir um padrão catabólico e pró-inflamação marcado por níveis elevados de Glutamato (Glu), Alanina (Ala) e Cistationia (Cysta), e o grupo C2 por apresentar sinais de esgotamento metabólico, com níveis reduzidos de BCAA e Glutamina (Gln), acompanhados por concentrações de BAIBA aumentadas. Em oposição, o perfil **C3** identificado pelo modelo implementado com recurso ao GMM, caracterizado por quantidades elevadas de BCAA, Treonina (Thr) e Prolina (Pro), sugere uma maior capacidade regenerativa da mucosa intestinal, podendo estar associado a um estado de menor gravidade clínica.

5.3.3 Análise de Outliers

Conforme realizado para a tarefa de *clustering* geral, também foi feita uma análise dos *outliers* encontrados no *clustering* de Crohns, tendo-se aplicado o algoritmo **Isolation Forest**. Nas amostras de plasma, foram sinalizados 9 casos: 2 em C0, 4 em C1 e 3 em C3. Além disso, e conforme o analisado na secção 5.3.2, os resultados do *clustering* identificaram um micro-*cluster* C2, composto por apenas 5 amostras. No entanto, os *outliers* encontrados não coincidem com as observações presentes em C2, representando observações dispersas com valores metabólicos extremos inseridas nos *clusters* de maiores dimensões. Por observação das projeções no espaço PCA, representado na figura 5.11, estas amostras ocupam posições em coordenadas bastante afastadas do conjunto de observações central e os seus perfis são sugestivos de um estado de catabolismo excessivo, possuindo níveis de Glutamina ínfimos, Glutamato e Alanina elevados e um desequilíbrio nos BCAA. Desta forma, este *outliers* podem representar amostras de doentes num estado grave da doença, recetivo a complicações como fístulas anais. Não obstante, não se pode descartar a possibilidade de ter ocorrido interferência analítica, sendo imperiosa uma investigação mais aprofundada para se alcançarem conclusões mais robustas.

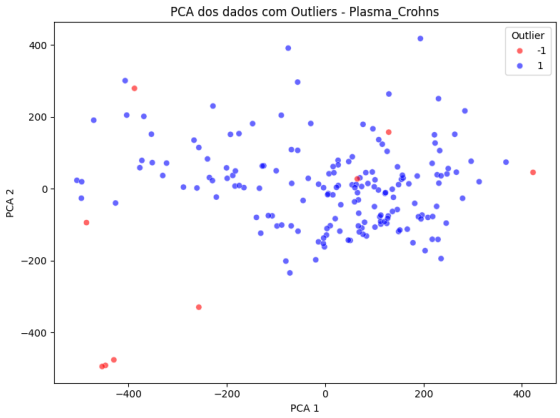

Figura 5.11: PCA dos *outliers* encontrados no *clustering* de Crohns nas amostras de plasma

Quanto à análise das amostras de urina, também se identificaram 9 *outliers* distribuídos de forma desigual: 4 amostras no *cluster* 0 e 5 no *cluster* 1. Tal como se verificou no plasma, estes casos formam um pequeno grupo com perfis muito diferentes da maioria das observações, apresentando desregulações severas, com concentrações extremas. Assim, os perfis destes *outliers* apontam para situações clínicas críticas ou para erros de medição. A figura 5.12 representa a projeção a duas dimensões dos *outliers* encontrados no subconjunto de amostras de Crohn na urina.

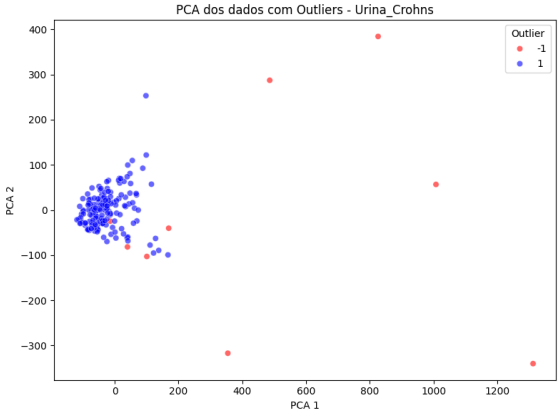

Figura 5.12: PCA dos *outliers* encontrados no *clustering* geral de amostras de urina

Capítulo 6

Conclusão

O presente trabalho teve como objetivo primordial explorar o potencial da aplicação de metodologias de ML a dados metabólicos provenientes de análises clínicas para caracterizar a Doença de Crohn, com o intuito de identificar potenciais biomarcadores e perfis associados à patologia. Para tal, foram realizadas diferentes etapas metodológicas: pré-processamento dos dados, incluindo a normalização e a seleção de variáveis; implementação e aplicação de modelos de classificação supervisionada, avaliando a capacidade das diferentes abordagens utilizadas em distinguir controlos de doentes de Crohn; análise de *clustering* no conjunto global de amostras, permitindo detetar perfis metabólicos distintos e potenciais biomarcadores da doença; e *clustering* restrito a doentes de Crohn, com vista à exploração da heterogeneidade interna desta população. Com o objetivo de orientar a pesquisa, foram estabelecidas questões de investigação, às quais as metodologias adotadas procuraram dar resposta.

- **RQ1 — Que algoritmos de classificação se demonstram mais eficientes a realizar a distinção entre doentes de Crohn e controlos?**

  Em termos globais, observaram-se desempenhos satisfatórios, com destaque para modelos *ensemble* como o *Random Forest* e, em particular nos dados de urina, também para o kNN. No plasma, o melhor cenário correspondeu ao *Random Forest* com *oversampling* (resultando num *f1-score* médio de 0.72). Na urina, o *Random Forest* e o kNN atingiram *f1-score* de 0.85–0.86 sem balanceamento, superando as restantes alternativas.

- **RQ2 — De que forma as técnicas de ML podem ser aplicadas a dados clínicos e metabólicos de doentes, de modo a prever a presença da Doença de Crohn?**

  Os resultados confirmam a viabilidade da classificação supervisionada para prever a presença da Doença de Crohn, a partir de dados clínicos/metabólicos, desde que se combinem pré-processamento adequado e técnicas de balanceamento. No plasma, sem balanceamento, o *Random Forest* apresentou *accuracy* elevada (cerca de 0.90) e *precision* alta (0.94–0.95), mas *recall* mais baixo (0.43–0.46), refletindo maior conservadorismo na classe Crohn. Já o kNN exibiu um perfil mais equilibrado (*accuracy* 0.88–0.89; *recall* médio de 0.59–0.61; *f1-score* médio de 0.65). Com *oversampling*, o *Random Forest* atingiu o melhor *f1-score* médio do plasma (0.72), seguindo-se a *Regressão Logística* (0.66). Assim, conclui-se que alguns dos modelos são adequados para esta tarefa, tendo apresentado maior robustez na urina.

- **RQ3 – De que forma se podem aplicar algoritmos de clustering para identificar subgrupos de doentes de Crohn com perfis metabólicos diferentes?**

  A utilização de técnica de *clustering*, tanto no conjunto global de amostras como especificamente nos doentes com Crohn, permitiu identificar subgrupos com perfis metabólicos distintos, confirmando a heterogeneidade metabólica associada à doença. No caso do *clustering* com o conjunto total de amostras, observaram-se *clusters* com predominância de casos positivos caracterizados por concentrações elevadas de Glutamato e níveis reduzidos de Glutamina, em contraste com os grupos de controlos onde esta relação se inverteu. Na análise restrita a doentes de Crohn, identificaram-se grupos com perfis distintos, nomeadamente na concentração dos aminoácidos de cadeia ramificada (BCAA) , sendo que alguns *clusters* apresentaram concentrações significativamente superiores em relação aos restantes grupos. Estes resultados confirmam a heterogeneidade da Doença de Crohn e sustentam a hipótese de que o *clustering* pode ser uma ferramenta útil na identificação de potenciais biomarcadores da doença.

- **RQ4 – Que perfis metabólicos poderão estar associados a complicações da Doença de Crohn, como o desenvolvimento de fístulas perianais?**

  Embora não tenha sido possível estabelecer associações definitivas, foram identificados perfis metabólicos que se poderão relacionar com uma maior probabilidade de desenvolver complicações. Os resultados apontaram para um perfil pró-inflamação constituído por Glutamato (Glu) elevado e Glutamina (Gln) reduzida, acompanhado de Alanina (Ala) e Ornitina (Orn) altas e BCAA reduzidos. Por outro lado, encontrou-se um potencial perfil protetivo que combina BCAA e Cistationina (Cysta) elevados e Glutamato baixo. Assim, propõe-se a hipótese da existência dos seguintes potenciais biomarcadores associados ao aparecimento de complicações: Glutamato/Glutamina, Alanina, Ornitina, BCAA e Cistationina. Sublinha-se que estes resultados poderão não ser definitivos, sendo essencial uma posterior validação clínica.

- **RQ5 – Qual a utilidade do desenvolvimento de modelos de ML aplicados em tarefas de classificação e clustering enquanto ferramentas de apoio ao diagnóstico e de monitorização da Doença de Crohn?**

  Os resultados obtidos demonstraram que as abordagens aplicadas poderão constituir ferramentas exploratórias úteis, com potencial de apoio à decisão clínica. Na tarefa de classificação, foi possível distinguir com precisão doentes de Crohn de controlos, alcançando *accuracies* entre 0.79 e 0.90 e *F1-scores* até 0.86. No *clustering* foi possível identificar subgrupos metabólicos distintos, apontando diferenças entre os grupos com grande proposção de diagnósticos positivos e os grupos de controlo, tendo-se identificado, também disparidades de perfis entre os doentes de Crohn, podendo indicar a presença de amostras positivas em diferentes estados da doença. Embora ainda em fase exploratória, estes resultados demonstram que a aplicação de modelos de ML pode auxiliar o diagnóstico precoce da doença, facilitar a estratificação dos doentes e sustentar a sua monitorização.

6.1 Contributos

O desenvolvimento deste trabalho contribuiu para demonstrar a aplicabilidade de técnica de ML na análise de dados metabólicos associados à Doença de Crohn. A avaliação comparativa de diferentes algoritmos de classificação e a sua combinação com estratégias de *feature*

selection e de balanceamento de dados, evidenciou que é viável alcançar previsões relevantes a partir do processamento de dados metabólicos.

O estudo também colaborou para identificar alterações metabólicas interessantes no contexto da doença, destacando-se o eixo Glutamina/Glutamato (Glu/Gln) como um dos principais discriminadores entre doentes e controlos. A Cistationina (Cysta) surgiu com concentrações elevadas em determinados grupos, apontando para potenciais alterações no metabolismo. Por sua vez, os aminácidos de cadeia ramificada (BCAA) revelaram concentrações superiores em *clusters* menos associados a doentes de Crohn, sugerindo uma possível função protetiva. Estes resultados constituem evidência do contributo da metabolómica clínica e da gastroenterologia na compreensão da heterogeneidade metabólica da Doença de Crohn.

Por fim, os resultados obtidos apoiam a combinação da metabolómica e ML como ferramenta de apoio ao diagnóstico e monitorização da Doença de Crohn. A capacidade de identificar diferentes perfis metabólicos abre caminho para a o estudo mais aprofundado da aplicação destas metodologias em medicina de precisão, contribuindo para estratégias terapêuticas mais personalizadas.

6.2 Limitações

Ao longo do desenvolvimento do trabalho, foram encontrados alguns obstáculos, sobretudo na fase de pré-processamento dos dados. A heterogeneidade dos formatos iniciais, a presença de registos incompletos ou inconsistentes e a necessidade de efetuar um mapeamento de variáveis com diferentes designações, implicaram um esforço acrescido para harmonizar e limpar os dados.

Outra limitação encontrada relaciona-se com a dimensão e a composição da amostra. No total, foram analisadas 996 amostras de plasma (sendo 820 relativas a doentes com outras patologias e 176 pertencentes a doentes de Crohn) e 225 amostras de urina (80 de outras patologias e 175 de doentes de Crohn). A diferença na quantidade de diagnósticos positivos e negativos da doença não assegura representatividade suficiente da variabilidade clínica e metabólica da Doença de Crohn. Desta forma, existe a possibilidade de não se verificarem os padrões identificados em populações de maior dimensão e heterogeneidade. Este obstáculo poderá traduzir ameaças à validade dos resultados, restringindo a sua generalização.

6.3 Trabalho Futuro

Apesar dos resultados alcançados neste trabalho, existe ainda um enorme potencial exploratório que poderá ser investigado em estudos futuros. Primeiramente, poderá ser relevante estender a análise realizada a conjuntos de dados maiores e mais variados, de modo a validar a robustez dos modelos desenvolvidos e a aumentar a sua capacidade de generalização. A aplicação da abordagem desenvolvida a outras fontes de informação, como dados genéticos ou imagens médicas, poderá igualmente enriquecer a caracterização dos perfis metabólicos e melhorar a previsão de complicações associadas à Doença de Crohn.

Outro caminho promissor poderá, também, passar pela aplicação de outros modelos de classificação mais adequados ou de técnicas mais avançadas, nomeadamente métodos de *deep learning*, que poderão contribuir para aumentar a precisão das previsões e oferecer uma interpretação mais clara dos fatores metabólicos subjacentes.

Os dados e as conclusões obtidas constituem uma base para a análise e o processamento futuro por parte dos profissionais de saúde do Hospital de Santo António do Porto. Aos profissionais compete a tarefa de estudar e integrar as informações alcançadas no contexto clínico da Doença de Crohn. Esta é uma etapa fundamental para a validação clínica dos resultados.

6.4 Considerações Finais

Em síntese, esta dissertação cumpriu os objetivos delineados, fornecendo uma base para futuras investigações e sublinhando o potencial da metabolómica e da Inteligência Artificial como ferramentas de apoio para os processos clínicos adjacentes à Doença de Crohn. Ainda assim, importa salientar que este trabalho deixa espaço para investigações futuras, incluindo a validação dos resultados atingidos numa maior quantidade de amostras, de forma a avaliar a robustez das conclusões obtidas. Adicionalmente, a aplicação de métodos mais avançados de aprendizagem automática e de *clustering* poderá contribuir para suportar a previsão da evolução clínica individual.

Bibliografia

[1] *The Importance of Artificial Intelligence in Everyday Life*. url: https://www.aeologic.com/blog/the-importance-of-artificial-intelligence-in-everyday-life/ (acedido em 07/10/2024).

[2] Alvin Rajkomar et al. «Ensuring fairness in machine learning to advance health equity». Em: *Annals of Internal Medicine* 169 (12 dez. de 2018), pp. 866–872. issn: 15393704. doi: 10.7326/M18-1990.

[3] *Crohn's disease and ulcerative colitis - Better Health Channel*. url: https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/crohns-disease-and-ulcerative-colitis (acedido em 14/10/2024).

[4] Ralf Bender e Stefan Lange. «Adjusting for multiple testing—when and how?» Em: *JAMA* 325.14 (2021), pp. 1465–1466. doi: 10.1001/jama.2021.3077. url: https://jamanetwork.com/journals/jama/fullarticle/2779985.

[5] Fernando Magro et al. «Burden of Disease and Cost of Illness of Inflammatory Bowel Diseases in Portugal». Em: *GE Portuguese Journal of Gastroenterology* 30 (4 ago. de 2023), p. 283. issn: 23414545. doi: 10.1159/000525206. url: /pmc/articles/PMC10521318/%20/pmc/articles/PMC10521318/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10521318/.

[6] *Fístula anal | Hospital da Luz*. url: https://www.hospitaldaluz.pt/pt/dicionario-de-saude/fistula-anal-sintomas-tratamentos (acedido em 12/10/2024).

[7] Cheng Mei Tian et al. «Stem Cell Therapy in Inflammatory Bowel Disease: A Review of Achievements and Challenges». Em: *Journal of Inflammation Research* 16 (2023), p. 2089. issn: 11787031. doi: 10.2147/JIR.S400447. url: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10199681/.

[8] Mark Haakman et al. «AI lifecycle models need to be revised: An exploratory study in Fintech». Em: *Empirical Software Engineering* 26 (set. de 2021). issn: 15737616. doi: 10.1007/s10664-021-09993-1.

[9] Jeffrey S. Saltz e Iva Krasteva. «Current approaches for executing big data science projects—a systematic literature review». Em: *PeerJ Computer Science* 8 (2022). issn: 23765992. doi: 10.7717/PEERJ-CS.862.

[10] *CRISP-DM surface technology*. url: https://www.ist.fraunhofer.de/en/expertise/simulation-digital-services/data-acquisition-model-based-process-optimization/crisp-dm-surface-technology.html (acedido em 07/12/2024).

[11] *Inflammatory bowel disease (IBD) - Symptoms and causes - Mayo Clinic*. url: https://www.mayoclinic.org/diseases-conditions/inflammatory-bowel-disease/symptoms-causes/syc-20353315 (acedido em 24/11/2024).

[12] Gautam Maddineni et al. «izae020.082». Em: *Inflammatory Bowel Diseases* 30 (jan. de 2024).

[13] Daniel C. Baumgart e William J. Sandborn. «Crohns disease». Em: *Elsevier B.V.* (2012).

[14] *Genome-Wide Association Studies (GWAS)*. url: https://www.genome.gov/genetics-glossary/Genome-Wide-Association-Studies-GWAS.

[15] Keyu Jia e Jun Shen. «Transcriptome-wide association studies associated with Crohn's disease: challenges and perspectives». Em: *Cell and Bioscience 2024 14:1* 14 (1 fev. de 2024), pp. 1–19. issn: 2045-3701. doi: 10.1186/S13578-024-01204-W. url: https://cellandbioscience.biomedcentral.com/articles/10.1186/s13578-024-01204-w.

[16] H. Guadalajara et al. *New Perspectives in the Treatment of Anal Fistulas*. 2022. doi: 10.1007/978-3-030-76670-2_40. (Acedido em 15/11/2024).

[17] *What Is An Anal Fistula?* url: https://my.clevelandclinic.org/health/diseases/14466-anal-fistula (acedido em 30/11/2024).

[18] Zhou Zhou et al. «Crohn's Disease-Associated and Cryptoglandular Fistulas: Differences and Similarities». Em: *Journal of Clinical Medicine* 12 (2 jan. de 2023). issn: 20770383. doi: 10.3390/jcm12020466.

[19] *What Is Gut Dysbiosis?* url: https://my.clevelandclinic.org/health/diseases/dysbiosis (acedido em 07/12/2024).

[20] Antonio Blanco e Gustavo Blanco. «Carbohydrates». Em: *Medical Biochemistry* (2017), pp. 73–97. doi: 10.1016/B978-0-12-803550-4.00004-5.

[21] Horatio C. Brenchley. «NECROSIS.» Em: *The Lancet* 80 (nov. de 1862), pp. 547–548. issn: 01406736. doi: 10.1016/S0140-6736(02)41872-7.

[22] Dan In Jang et al. «The role of tumor necrosis factor alpha (Tnf-alfa) in autoimmune disease and current tnf-alfa inhibitors in therapeutics». Em: *International Journal of Molecular Sciences* 22 (mar. de 2021), pp. 1–16. issn: 14220067. doi: 10.3390/ijms22052719.

[23] Masato Morikawa, Rik Derynck e Kohei Miyazono. «TGF- beta and the TGF-beta family: Context-dependent roles in cell and tissue physiology». Em: *Cold Spring Harbor Perspectives in Biology* 8 (mai. de 2016). issn: 19430264. doi: 10.1101/cshperspect.a021873.

[24] Sozaburo Ihara, Yoshihiro Hirata e Kazuhiko Koike. *TGF-beta in inflammatory bowel disease: a key regulator of immune cells, epithelium, and the intestinal microbiota*. Jul. de 2017. doi: 10.1007/s00535-017-1350-1.

[25] Laurence J. Egan. «JOURNAL OF CROHN'S AND COLITIS International Journal Devoted to Inflammatory Bowel Diseases Official Journal of the European Crohn's and Colitis Organisation». Em: *Journal of Crohn's and Colitis* 18/S1 (2024). url: www.oxfordjournals.org/permissions.

[26] Weijie Zhou et al. «Exploration of the molecular linkage between endometriosis and Crohn disease by bioinformatics methods». Em: *Medicine (United States)* 103 (mai. de 2024), E38097. issn: 15365964. doi: 10.1097/MD.0000000000038097.

[27] Katherine A. Abrahams e Gurdyal S. Besra. *Synthesis and recycling of the mycobacterial cell envelope*. Abr. de 2021. doi: 10.1016/j.mib.2021.01.012.

[28] G. R. Lichtenstein. «Treatment of fistulizing Crohn's disease». Em: *Gastroenterology* 119 (2000), pp. 1132–1147. issn: 00165085. doi: 10.1053/gast.2000.18165.

[29] Vinod K. Dhawan. *Gram-negative bacteria*. Jan. de 2006. doi: 10.1385/1-59259-036-5:43.

[30] *What are immunomodulators?* url: https://my.clevelandclinic.org/health/drugs/24987-immunomodulators (acedido em 12/11/2024).

[31] Juliette Cooke et al. «Genome scale metabolic network modelling for metabolic profile predictions». Em: *PLoS Computational Biology* 20 (fev. de 2024). issn: 15537358. doi: 10.1371/journal.pcbi.1011381.

[32] Paula Trumbo et al. «Dietary reference intakes for energy, carbohydrate, fiber, fat, fatty acids, cholesterol, protein and amino acids». Em: *Journal of the American Dietetic Association* 102 (2002), pp. 1621–1630. issn: 00028223. doi: 10.1016/S0002-8223(02)90346-9.

[33] Samuel O. Adegbola et al. «Differences in amino acid and lipid metabolism distinguish Crohn's from idiopathic/cryptoglandular perianal fistulas by tissue metabonomic profiling and may offer clues to underlying pathogenesis». Em: *European Journal of Gastroenterology and Hepatology* 33 (dez. de 2021), pp. 1469–1479. issn: 14735687. doi: 10.1097/MEG.0000000000001976.

[34] *Types of Machine Learning | IBM*. url: https://www.ibm.com/think/topics/machine-learning-types (acedido em 02/12/2024).

[35] *New Advances in Machine Learning*. url: https://books.google.pt/books?hl=pt-PT%5C&lr=%5C&id=XAqhDwAAQBAJ%5C&oi=fnd&pg=PA19%5C&dq=machine+learning+algorithms+types%5C&ots=r3Im7TEjLo%5C&sig=_mcNYQMzkozNmvqWdbBPMcwqPu0&redir_esc=y%5C#v=onepage%5C&q=machine%5C%20learning%5C%20algorithms%5C%20types%5C&f=false (acedido em 12/11/2024).

[36] Emilian R. Vankov e Kais Gadhoumi. «Supervised Learning». Em: *Statistical Methods in Epilepsy* (jan. de 2024), pp. 272–301. doi: 10.1201/9781003254515-12.

[37] Frank Acito. «Logistic Regression». Em: *Predictive Analytics with KNIME* (2023), pp. 125–167. doi: 10.1007/978-3-031-45630-5_7.

[38] Tony Thomas, Athira P. Vijayaraghavan e Sabu Emmanuel. «Applications of Decision Trees». Em: *Machine Learning Approaches in Cyber Security Analytics* (2020), pp. 157–184. doi: 10.1007/978-981-15-1706-8_9.

[39] *Decision Tree*. url: https://machinelearningtheory.org/docs/Random-Forest/tree/ (acedido em 12/05/2024).

[40] Rajib Kumar Halder et al. «Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications». Em: *Journal of Big Data* 11 (1 dez. de 2024). issn: 21961115. doi: 10.1186/s40537-024-00973-y.

[41] Nour Al-Rahman Al-Serw. *K-nearest Neighbor: The maths behind it, how it works and an example*. (Acedido em 05/12/2025).

[42] Andy Liaw e Matthew Wiener. «Classification and Regression by randomForest». Em: *R News* 2.3 (2002), pp. 18–22. url: https://cran.r-project.org/doc/Rnews/.

[43] Leo Breiman. «Random forests». Em: *Machine learning* 45.1 (2001), pp. 5–32. doi: 10.1023/A:1010933404324.

[44] Trevor Hastie, Robert Tibshirani e Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd. Springer Science & Business Media, 2009. isbn: 9780387848570.

[45] D. Richard Cutler et al. «Random Forests for Classification in Ecology». Em: *Ecology* 88.11 (2007), pp. 2783–2792. doi: 10.1890/07-0539.1.

[46] Derek Greene, Padraig Cunningham e Rudolf Mayer. *Unsupervised Learning and Clustering*. doi: 10.1007/978-3-540-75171-7-3. url: https://www.researchgate.net/publication/235328198 (acedido em 28/08/2025).

[47] Hong Wang et al. *Phenotype clustering in health care: A narrative review for clinicians*.

[48] Douglas A. Reynolds. «Gaussian Mixture Models». Em: *Encyclopedia of Biometrics*. Springer, 2009, pp. 659–663. doi: 10.1007/978-0-387-73003-5_196.

[49] Geoffrey J. McLachlan e David Peel. *Finite Mixture Models*. John Wiley & Sons, 2000. doi: 10.1002/0471721182.

[50] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. isbn: 9780387310732.

[51] Rizgar Zebari et al. «A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction». Em: *Journal of Applied Science and Technology Trends* 1 (1 mai. de 2020), pp. 56–70. doi: 10.38094/jastt1224.

[52] Guangzhou et al. «2024 International Conference on Smart Healthcare and Wearable Intelligent Devices (SHWID 2024)». Em: *The Association for Computing Machinery* (out. de 2024).

[53] Young Joon Ko et al. «Identification of Functional Microbial Modules Through Network-Based Analysis of Meta-Microbial Features Using Matrix Factorization». Em: *IEEE-E/ACM Transactions on Computational Biology and Bioinformatics* 19 (5 2022), pp. 2851–2862. issn: 15579964. doi: 10.1109/TCBB.2021.3100893.

[54] T. G. Clark et al. *Survival Analysis Part I: Basic concepts and first analyses*. Jul. de 2003. doi: 10.1038/sj.bjc.6601118.

[55] Carlos Ramos-Carreño et al. «scikit-fda: A Python Package for Functional Data Analysis». Em: (nov. de 2022). doi: 10.18637/jss.v109.i02. url: http://arxiv.org/abs/2211.02566%20http://dx.doi.org/10.18637/jss.v109.i02.

[56] Kenneth Fricklas. *Nishant Shukla*. Manning Publications, 2018.

[57] *What is PyTorch?* url: https://www.techtarget.com/searchenterpriseai/definition/PyTorch (acedido em 10/12/2024).

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

[59] Wuxing Chen et al. «A survey on imbalanced learning: latest research, applications and future directions». Em: *Artificial Intelligence Review* 57 (2024). doi: 10.1007/s10462-024-10759-6. url: https://doi.org/10.1007/s10462-024-10759-6.

[60] Enzo Battistella, Dina Ghiassian e Albert-László Barabási. «Improving the performance and interpretability on medical datasets using graphical ensemble feature selection». Em: *Bioinformatics* 40.6 (2024), btae341. doi: 10.1093/bioinformatics/btae341. url: https://doi.org/10.1093/bioinformatics/btae341.

[61] *Omics Technologies*. 2023. url: https://humanspecificresearch.org/omics-technologies/ (acedido em 05/09/2025).

[62] T. Ghosh et al. «Predictive Modeling for Metabolomics Data». Em: *Metabolites* 10.8 (2020), p. 318. doi: 10.3390/metabo10080318. url: https://www.mdpi.com/2218-1989/10/8/318.

[63] scikit-learn developers. *Decision Trees — User Guide*. 2025. url: https://scikit-learn.org/stable/modules/tree.html (acedido em 20/08/2025).

[64] scikit-learn developers. *sklearn.tree.DecisionTreeClassifier — scikit-learn documentation*. 2025. url: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html (acedido em 20/08/2025).

[65] Fabian Pedregosa et al. *Scikit-learn: Machine learning in Python*. 2011.

[66] Szilvia Szeghalmy e Attila Fazekas. «A Comparative Study of the Use of Stratified Cross-Validation and Distribution-Balanced Stratified Cross-Validation in Imbalanced Learning». Em: *Sensors* 23.4 (2023), p. 2333. doi: 10.3390/s23042333. url: https://pmc.ncbi.nlm.nih.gov/articles/PMC9967638/.

[67] scikit-learn developers. $sklearn.linear_model.LogisticRegression$ *scikit – learndocumentation*. 2025. url: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html (acedido em 20/08/2025).

[68] Gareth James et al. *An Introduction to Statistical Learning: With Applications in R*. 2ª ed. Springer, 2021. doi: 10.1007/978-1-0716-1418-1. url: https://www.statlearning.com/.

[69] Thomas G Dietterich. «Ensemble methods in machine learning». Em: *International workshop on multiple classifier systems* (2000), pp. 1–15.

[70] Jason Brownlee. *Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning*. Machine Learning Mastery, 2020.

[71] Trevor Hastie, Robert Tibshirani e Jerome Friedman. *Springer Series in Statistics The Elements of Statistical Learning Data Mining, Inference, and Prediction*. 2009.

[72] David M W Powers. *Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation*. Vol. 2. 1. 2011, pp. 37–63.

[73] Tom Fawcett. «An introduction to ROC analysis». Em: *Pattern recognition letters* 27.8 (2006), pp. 861–874.

[74] Nathalie Japkowicz e Mohak Shah. *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, 2011.

[75] Pawan Bholowalia e Arvind Kumar. «EBK-means: A clustering technique based on elbow method and k-means in WSN». Em: *International Journal of Computer Applications* 105.9 (2014), pp. 17–24. doi: 10.5120/18405-9674.

[76] Peter J Rousseeuw. «Silhouettes: a graphical aid to the interpretation and validation of cluster analysis». Em: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65. doi: 10.1016/0377-0427(87)90125-7.

[77] Anil K. Jain. «Data clustering: 50 years beyond K-means». Em: *Pattern Recognition Letters* 31.8 (2010), pp. 651–666. doi: 10.1016/j.patrec.2009.09.011.

[78] Douglas A. Reynolds. «Gaussian Mixture Models». Em: *Encyclopedia of Biometrics* (2009), pp. 659–663. doi: 10.1007/978-0-387-73003-5_196.

[79] Ian T. Jolliffe e Jorge Cadima. *Principal Component Analysis*. Springer, 2016. doi: 10.1007/978-1-4899-7514-3.

[80] Tae Kyun Kim. «Understanding one-way ANOVA using conceptual figures». Em: *Korean Journal of Anesthesiology* 70.1 (2017), pp. 22–26. doi: 10.4097/kjae.2017.70.1.22. url: https://pubmed.ncbi.nlm.nih.gov/28184262/.

[81] S. McLeod et al. *Biostatistics, P Value*. Treasure Island (FL): StatPearls Publishing, 2023. url: https://www.ncbi.nlm.nih.gov/books/NBK557421/.

[82] John D. Storey e Robert Tibshirani. «Statistical significance for genomewide studies». Em: *Proceedings of the National Academy of Sciences* 100.16 (2003), pp. 9440–9445. doi: 10.1073/pnas.1530509100. url: https://www.pnas.org/doi/10.1073/pnas.1530509100.

[83] Wei Zhang et al. «Exploratory metabolomic analysis for characterizing the metabolic profile of the urinary bladder under estrogen deprivation». Em: *Frontiers in Endocrinology* 15 (2024), p. 1384115. doi: 10.3389/fendo.2024.1384115. url: https://www.frontiersin.org/articles/10.3389/fendo.2024.1384115.

[84] Xiaoyan Yin, Subha Subramanian, Christine M. Willinger et al. «Metabolite signatures of metabolic risk factors and their longitudinal changes». Em: *Journal of Clinical Endocrinology Metabolism* 101.4 (2016), pp. 1779–1789. doi: 10.1210/jc.2015-2555. url: https://pubmed.ncbi.nlm.nih.gov/26894262/.

[85] Tatyana S. Gurina e Shamim S. Mohiuddin. «Biochemistry, Protein Catabolism». Em: *StatPearls* (2022). url: https://www.ncbi.nlm.nih.gov/books/NBK556047/.

[86] Vinicius Cruzat et al. «Glutamine: Metabolism and Immune Function, Supplementation and Clinical Translation». Em: *Nutrients* 10.11 (2018), p. 1564. doi: 10.3390/nu10111564.

[87] Chih-Chieh Hsu et al. «Metabolism of Proteins and Amino Acids in Critical Illness: From Physiological Alterations to Relevant Clinical Practice». Em: *Journal of Multidisciplinary Healthcare* 14 (2021), pp. 1107–1117. doi: 10.2147/JMDH.S306350.

[88] Iolanda Cioffi et al. «Amino Acid Profiles, Disease Activity, and Protein Intake in Adult Patients with Crohn's Disease». Em: *Frontiers in Nutrition* 10 (2023), p. 1245574. doi: 10.3389/fnut.2023.1245574.

[89] Gagandeep Mann et al. «Branched-chain amino acids: catabolism in skeletal muscle and implications for muscle and whole-body metabolism». Em: *Frontiers in Physiology* 12 (2021), p. 702826. doi: 10.3389/fphys.2021.702826.

[90] Juan I. Sbodio, Solomon H. Snyder e Bindu D. Paul. «Regulators of the transsulfuration pathway». Em: *British Journal of Pharmacology* 176.4 (2019), pp. 583–593. doi: 10.1111/bph.14446.

[91] Dmitrii A. Tanianskii et al. «Beta-Aminoisobutyric Acid as a Novel Regulator of Carbohydrate and Lipid Metabolism». Em: *Nutrients* 11.3 (2019), p. 524. doi: 10.3390/nu11030524.

[92] Yulan Liu. «The role of amino acids in inflammatory bowel disease: implications for clinical practice». Em: *Nutrients* 9.9 (2017), p. 920. doi: 10.3390/nu9090920. url: https://www.mdpi.com/2072-6643/9/9/920.

[93] Guoyao Wu et al. «Important roles of dietary taurine, creatine, carnosine, anserine and 4-hydroxyproline in human nutrition and health». Em: *Amino Acids* 40.5 (2011), pp. 1161–1175. doi: 10.1007/s00726-010-0740-7.