INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA INFORMÁTICA




Machine Learning na identificação de anomalias e clusters em casos de cancro do reto em função de alterações metabólicas

JOÃO CARLOS MARTINS RIBEIRO
Junho de 2023

POLITÉCNICO
DO PORTO

Instituto Superior de Engenharia do Porto

Machine Learning na identificação de anomalias e clusters em casos de cancro do reto em função de alterações metabólicas

João Carlos Martins Ribeiro

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

Orientador: José Reis Tavares

Co-orientador: Isabel Praça

Porto, junho 2023

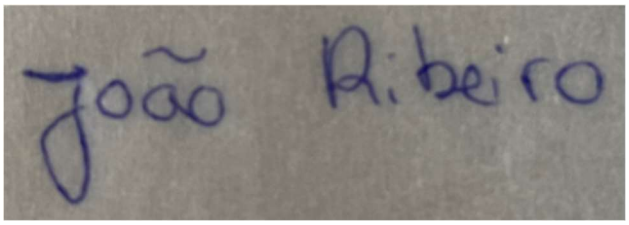
Declaração de Integridade

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, 29 de junho de 2023

Resumo

Atualmente, o setor da saúde é um dos setores que apresenta um crescimento mais rápido. Os dados crescem exponencialmente todos os dias e à medida que a quantidade de informação aumenta, novas formas de interagir e extrair significado vão surgindo.

Machine Learning é uma área da Inteligência Artificial que estuda como solucionar problemas complexos e intuitivos, apresentando potencial para ser a solução que permitirá reduzir o custo crescente dos cuidados de saúde e, auxiliar os médicos no tratamento dos seus pacientes, de forma mais rápida e eficaz.

A presente dissertação de mestrado tem como objetivo estudar os perfis metabólicos de aminoácidos e acilcarnitinas em doentes com cancro do reto localmente avançado e a forma como estes perfis influenciam a biologia tumoral e a resposta ao tratamento. Pretende-se analisar os respetivos perfis para a realização de duas tarefas de aprendizagem não supervisionada: Deteção de anomalias e *Clustering.*

No decorrer deste trabalho, serão testados diferentes algoritmos e técnicas de tratamento de dados para serem avaliados por métricas intrínsecas para determinar o nível de coesão, separação e semelhança nos resultados de grupos obtidos, seguido pelas respetivas conclusões.

Os melhores algoritmos serão utilizados num sistema de decisão que ajude de forma mais eficiente os profissionais de saúde na escolha de tratamento adequado para cada um dos futuros pacientes.

Palavras-chave: Machine Learning, aprendizagem não supervisionada, colorretal, aminoácidos, acilcarnitinas

Abstract

Currently, the healthcare sector is one of the fastest growing sectors. Data exponentially takes into account every day and as the amount of information increases, new ways of interacting and extracting meaning are emerging.

Machine Learning is an area of Artificial Intelligence that studies how to solve complex and intuitive problems, presenting the potential to be the solution that will reduce the growing cost of health care and help doctors in treating their patients more quickly and effectively.

This master's thesis aims to study the metabolic profiles of amino acids and acylcarnitines in patients with colorectal cancer and how these profiles influence tumor biology and response to treatment. It is intended to analyze the respective profiles for carrying out two unsupervised learning tasks: Anomaly detection and Clustering.

In the course of this work, different algorithms and data processing techniques will be tested to be evaluated by intrinsic metrics to determine the level of cohesion, separation and similarity in the obtained group results, followed by the respective understanding.

The best algorithms will be used in a decision system that will more efficiently help health professionals in choosing the right treatment for each of their future patients.

Keywords: Machine Learning, unsupervised learning, Coloretal, amino acids, acylcarnitines

Agradecimentos

Gostaria de começar por agradecer à minha família pelo apoio prestado ao longo dos anos, e por sempre me incentivarem e terem fornecido todas as condições para que conseguisse completar o meu percurso académico.

No final deste percurso académico, gostaria de agradecer a todos os meus amigos e colegas que me acompanharam ao longo dos anos letivos, em especial Pedro Marques, Tiago Barbosa, Alexandre Mendes e Rogério Alves

Por fim, ao professor José Reis Tavares e à professor Isabel Praça, por todo o apoio prestado durante o projeto, disponibilidade e transmissão de conhecimento ao longo deste ano.

Índice

Lista de Figuras

Lista de Tabelas

Acrónimos e Símbolos

Lista de Acrónimos

ML	*Machine Learning*
IA	Inteligência Artificial
CRISP-DM	*Cross Industry Standard Process For Data Mining*
DL	*Deep Learning*
API	*Application Programming Interface* (interface de programação de aplicações)
CPU	*Central Processing Unit*
GPU	*Graphics Processing Unit*
GMM	*Gaussian Mixture Model*
NCD	*New Concept development*
QFD	*Quality function deployment*
SC	*Silhouette Coefficient*
CH	*Calinski-Harabasz Index*
DB	*Davies-Bouldin Index*
PCA	*Principal Component Analysis*
VTFS	*Variance Threshold Feature Selection*

Lista de Símbolos

β	Largura de banda
π	Pi
μ	Micro

1 Introdução

Este capítulo representa a introdução ao tópico desta dissertação, realizada durante o mestrado de Engenharia Informática no Instituto Superior de Engenharia do Porto, e que vai incluir o contexto, o problema, o objetivo do trabalho, a metodologia aplicada e finalmente a estrutura do documento.

1.1 Contexto

Atualmente, o setor da saúde é um dos setores com um crescimento mais rápido e está no centro de uma revisão e transformação global completa, segundo [1]. *Russell Reynolds and Associates* afirma, em [2], que "os custos globais de saúde, atualmente estimados em US$ 6 triliões a US$ 7 triliões, devem atingir mais de US$ 12 triliões em apenas sete anos".

No mundo de hoje, os dados em todos os setores estão a crescer exponencialmente. À medida que a quantidade de dados aumenta, novas formas de interagir e extrair o significado destes vão aparecer. No passado, de acordo com [3], os dados eram analisados através de intervenção humana, no entanto, isso é ineficiente e um grande número de padrões ocultos nos dados muitas vezes não são encontrados, devido ao seu grande volume. Assim sendo, é um dos fatores de maior importância para o setor de saúde.

Machine learning (ML) pode ser a solução para reduzir o custo crescente dos cuidados de saúde e ajudar a estabelecer uma melhor relação médico-paciente. As soluções de ML e *Big data* podem ser usadas para uma infinidade de usos relacionados com a saúde, alguns incluem ajudar os médicos a determinar prescrições e tratamentos mais personalizados para os pacientes, segundo [4].

Como resultado, a saúde é uma das vertentes que mais pode beneficiar significativamente com as quantidades crescentes de dados e da sua disponibilidade [1].

1.2 Problema

Os pacientes com cancro do reto localmente avançado são tratados de forma multimodal com radioterapia, quimioterapia e cirurgia. A resposta terapêutica pode ser muito variável – alguns apresentando resposta completa, enquanto outros pouca ou nenhuma resposta – sendo que a resposta patológica se tornou um importante preditor de bom resultado oncológico. O conhecimento em torno dos mecanismos moleculares do cancro do reto tem vindo a aumentar, no entanto sem que isso se traduza em métodos mais precisos na seleção dos doentes e em novas estratégias para melhorar a resposta terapêutica. Estudos recentes mostram que os aminoácidos e acilcarnitinas desempenham um papel importante na biologia do cancro, podendo influenciar a resposta ao tratamento e a agressividade tumoral.

Este estudo procura analisar os perfis metabólicos de aminoácidos e acilcarnitinas em doentes com cancro do reto e de que forma este perfil influencia a biologia tumoral e a resposta ao tratamento. Este será um estudo exploratório, de não intervenção, com um componente retrospetivo e um componente prospetivo. Serão obtidos dados demográficos, clínicos e histológicos de forma a identificar fatores associados à resposta terapêutica e ao prognóstico.

O problema principal consiste em identificar uma associação entre os respetivos indicadores e a resposta patológica, na tentativa de identificação de anomalias e *clusters*, em função destes biomarcadores moleculares preditores de resposta.

1.3 Objetivo

ML é uma área da Inteligência Artificial (IA) que estuda como solucionar problemas complexos e intuitivos. As metodologias propostas permitem, com recurso a meios computacionais, que as máquinas aprendam e compreendam o mundo em determinados contextos a partir de experiências anteriores e que com base na hierarquia de conceitos possam compreender conceitos mais complexos, de modo a solucionarem eficientemente a mais variadíssima gama de problemas.

Este projeto insere-se no domínio da aprendizagem não supervisionada que consiste numa técnica de ML na qual os engenheiros de dados não precisam de supervisionar o processo de treino do modelo. Em vez disso, este tipo de aprendizagem permite que o modelo funcione de forma independente, sem qualquer supervisão por forma a descobrir padrões ocultos e informações que não foram detetados anteriormente.

O principal objetivo neste trabalho consiste em analisar um conjunto de dados com base no perfil de aminoácidos e acilcarnitinas para a realização de duas das tarefas de aprendizagem não supervisionadas:

- Deteção de anomalias - a tarefa de detetar instâncias que são muito diferentes da norma;
- *Clustering* - a tarefa de agrupar instâncias semelhantes em *clusters*.

1.4 Metodologia

A metodologia de trabalho utilizada para a construção da solução deste projeto vai ser baseada no modelo CRISP-DM(Cross Industry Standard Process For Data Mining). De acordo com N. Hotz [5], este prevê seis fases distintas:

- Compreensão do negócio – o objetivo principal é compreender o projeto e os requisitos necessários para o cliente;
- Compreensão dos dados – inclui a exploração dos dados fornecidos ou obtidos, permitindo assim uma melhor compreensão dos mesmos e do que é apresentado;
- Preparação dos dados – consiste em preparar os dados para serem consumidos pelos modelos, corrigindo-se inconsistências ou erros que possam existir;
- Modelação – corresponde ao processo de selecionar diferentes técnicas, algoritmos, definir parâmetros de ML, resultando em diferentes modelos de ML. Estes modelos vão consumir os dados anteriormente preparados;
- Avaliação – corresponde à avaliação feita dos modelos da fase anterior com base nos critérios de aceitação definidos;
- Implementação – consiste na apresentação dos modelos ao cliente, para que este possa usufruir dos mesmos.

É possível verificar as várias fases do modelo CRISP-DM na Figura 1.

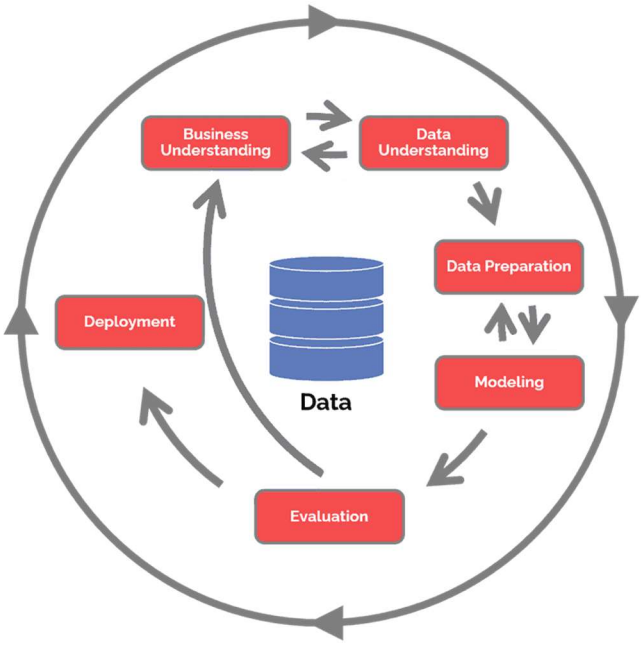


Figura 1 - Fases do modelo CRISP-DM, obtida de [5]

1.5 Estrutura do Documento

A estrutura prevista do documento começará com a introdução que descreve o tópico da dissertação, o contexto, o problema a ser tratado, os objetivos a alcançar e a metodologia de trabalho a seguir. De seguida, é apresentado o estado de arte onde se descrevem os principais conceitos, para esclarecer o leitor sobre o cancro colorretal e os perfis metabólicos a ter em conta neste trabalho. Também vão ser abordados os vários tipos de aprendizagem de ML, mais concretamente, a não supervisionada, trabalhos de literatura já existentes na área de saúde sobre a aplicação de ML para prognóstico, identificação de anomalias e agrupamento.

Existirá também um capítulo de análise de valor que será dedicado a analisar o valor desta dissertação através do uso de modelos que permitem analisar e avaliar o valor que a solução pode fornecer tanto para o cliente como para a área de saúde.

Posteriormente, começa-se com a análise e desenho da solução para descrever o domínio do problema, como são tratados os dados de forma a serem consumidos pelos algoritmos de ML e é apresentada a arquitetura da solução através de diferentes vistas.

A solução vai apresentar um capítulo de experimentação e avaliação onde procura descrever os experimentos a se realizar e a avaliação dos resultados obtidos.

Por fim, uma secção de conclusões para apresentar as respetivas conclusões da dissertação, apontando-se os objetivos alcançados e as possíveis melhorias para o futuro.

2 Estado de arte

Neste capítulo são descritos os conceitos fundamentais a serem compreendidos pelo leitor respetivamente ao trabalho a ser desenvolvido: Cancro Colorretal, Perfis metabólicos, ML, exemplos de aplicação de ML na área da saúde na ajuda ao tratamento do cancro e as várias bibliotecas a ser consideradas.

2.1 Cancro Colorretal

O cancro do reto e/ou cólon, cancro colorretal ou cancro do intestino grosso é uma das doenças oncológicas mais comuns nos países desenvolvidos. Em Portugal, por ano, são diagnosticados mais de 7 mil novos casos, sendo este o tipo de cancro mais fatal. Em termos mundiais, estamos a falar de 1,4 milhões de casos diagnosticados anualmente, sendo a terceira principal causa de morte por cancro. [6]

No que respeita aos fatores de risco do cancro colorretal, temos como principais: Alimentação; Idade e género; História familiar e Fatores genéticos.

Este tipo de cancro pode ter origem em alterações/mutações, hereditárias ou espontâneas, dos genes que controlam as células; e desenvolve-se devido à produção descontrolada de células da camada de revestimento interior do cólon ou do reto. Com a multiplicação celular desorganizada é formado um pequeno tumor benigno (não cancerígeno), designado como pólipo. Este último consiste numa saliência do tecido da parede abdominal, podendo tornar-se em tumor maligno (cancro) à medida que aumenta de dimensão.

No que diz respeito ao tratamento do cancro colorretal, o mesmo depende especialmente, entre outros fatores, do estado do cancro. É também importante saber qual a sua localização, o tipo de lesão, os possíveis efeitos do tratamento, bem como o estado clínico do paciente, para selecionar o tipo de tratamentos a administrar. [7]

No caso de doentes com cancro do reto localmente mais avançado, o tratamento é variável. Contudo, a resposta patológica tornou-se num importante preditor de bom resultado oncológico. Vários estudos recentes demonstram que os aminoácidos e acilcarnitinas desempenham um papel importante na biologia do cancro, podendo influenciar a resposta ao tratamento e a agressividade tumoral.

2.2 Perfis metabólicos

Nesta secção são apresentados os diferentes dados de perfis metabólicos que serão analisados pelos algoritmos de ML. No capítulo 4 é possível compreender com mais detalhe os elementos que constituem os respetivos dados.

2.2.1 Acilcarnitinas

As acilcarnitinas são metabólitos de ácidos gordos envolvidos na produção de energia e na manutenção da atividade celular dos organismos [8]. Esta substância é principalmente utilizada no estudo de diversas doenças, tal como, distúrbios metabólicos, neurológicos, doenças cardiovasculares e certos tipos de cancro. Estes são considerados marcadores de diagnósticos de: erros de oxidação dos ácidos gordos, ou de insuficiência na atividade de oxidação mitocondrial e peroxissomal, metabolismo energético, resistência à insulina e atividade física [9].

2.2.2 Aminoácidos

Os aminoácidos são partículas constituídas por proteínas utilizados como fonte de energia pelos organismos. As concentrações deste composto, normalmente, são estáveis e equilibrados, contudo elas também são afetadas pela dieta, metabolismo, estilo devida e fatores genéticos. Um número elevado na produção destas partículas permite sustentar o impulso do cancro para se multiplicar [10].

Atualmente, existem vários estudos a demonstrar as mudanças significativas nos perfis de aminoácidos em casos de pacientes com cancro sem caquexia, em pacientes com cancro em diferentes órgãos e para pacientes com cancro colorretal quando analisados os perfis de aminoácidos plasmáticos [11].

2.3 Tipos de aprendizagem

ML foi definido em 1959, por Arthur Samuel, como a habilidade do “computador de aprender sem ser explicitamente programado”. Simplificado, ML utiliza algoritmos programados que recebem e analisam dados de entrada e preveem o valor de saída dos resultados dentro dos parâmetros de aceitação. Desde que novos dados sejam constantemente adicionados ao

algoritmo, eles aprendem e otimizam as suas operações, obtendo-se resultados mais precisos e exatos, desenvolvendo-se assim "inteligência" ao longo do tempo, segundo K. Wakefield [12]. Este processo automático permite às máquinas aprender e melhorar sem ser preciso programá-las diretamente [13].

Atualmente, existem muitos tipos de ML, sendo que estes são classificados normalmente de acordo com a quantidade e tipo de supervisão utilizados. De seguida serão detalhadamente apresentados os diferentes tipos de ML que se consideram ser mais importantes.

2.3.1 Supervised learning

A aprendizagem supervisionada consiste em ensinar a máquina por exemplo. Os modelos são treinados através de exemplos de dados de entrada e dados de saída esperados.

Os operadores destas aprendizagens fornecem ao algoritmo dados que já incluem os valores desejados de entrada e saída esperados. Deste modo, o algoritmo tem de encontrar métodos para determinar como chegar ao valor de saída esperado. Se o algoritmo estiver correto, este vai identificar padrões nos dados, aprender através de observação e fazer previsões no final. Estas previsões são depois validadas pelo operador, até o algoritmo alcançar um nível elevado de precisão/desempenho, segundo o artigo em [12].

De acordo com o autor em [14], algoritmos de Regressão e Classificação são os dois ramos mais importantes em aprendizagem supervisionada. Nos algoritmos de classificação são obtidas previsões de "Sim" e "Não", enquanto os de Regressão permitem fazer previsões de quantidade, respondendo a questões, por exemplo, de "Quantos?".

2.3.2 Unsupervised learning

A aprendizagem não supervisionada vai ser o foco principal deste trabalho. Neste tipo de aprendizagem, segundo o autor em [15], os dados de entrada fornecidos ao algoritmo não são rotulados, sendo que o algoritmo tem de encontrar padrões, semelhanças, que possam existir entre os dados. Para este tipo de aprendizagem não é fornecida qualquer instrução por parte do operador, este não intervém, sendo que, o algoritmo tem de determinar correlações e relações enquanto analisa os dados disponíveis [12]. Compete à máquina compreender os vários dados, identificar qualquer semelhança e tentar organizar os dados de alguma forma ou estrutura. Isto significa agrupar os dados em *clusters* para parecerem mais organizados. Ao longo do tempo, quantos mais dados válidos forem analisados pelo algoritmo, melhor serão as decisões, tornando assim os dados mais refinados.

2.3.3 Semi-supervised learning

Segundo [12], a aprendizagem semi-supervisionada é muito similar à aprendizagem supervisionada. A diferença prende-se com a utilização de dados por classificar além dos dados

já identificados/esperados para o algoritmo aprender. Através de dados que já se encontram identificados, com respetivas *tags* e identificações, o algoritmo consegue reconhecer semelhanças entre as características dos restantes dados, tal como as técnicas de aprendizagem não supervisionada.

Esta técnica é principalmente aplicada para casos de imagens quando as mesmas não estão identificadas, segundo Brian A. [14]. Por exemplo, durante o processo de carregamento de várias imagens para o Facebook onde aparece a mesma pessoa, o algoritmo consegue perceber nas várias imagens que se trata da mesma pessoa. Depois de a identificar uma única vez, o algoritmo consegue reconhecê-la nas restantes imagens em que possa estar presente.

2.3.4 Reinforcement learning

De acordo com [15], em técnicas de aprendizagem por reforço, o algoritmo vai se desenvolvendo através de um Sistema de *Reward feedback*. Por norma, nestes algoritmos de ML é fornecido um conjunto de ações, parâmetros, valores finais esperados, de modo que o algoritmo compreenda quando o caminho que está a percorrer está errado, mas nunca é informado de quando está correto, segundo [16]. Este tipo de regras e aprendizagem permite explorar, testar e avaliar várias possibilidades até que encontre a melhor resposta possível. Aprendizagem por reforço ensina a máquina acerca de tentativa erro. Esta aprende através de experiências passadas e começa a adaptar os seus métodos em resposta à situação, permitindo alcançar o melhor resultado possível [12].

2.3.5 Deep learning

Deep learning (DL) é uma especialização de ML. Particularmente, esta técnica distingue-se dos métodos clássicos, e anteriormente falados, pelo tipo de dados e informação com que trabalha, tal como os algoritmos que utiliza. Esta técnica permite às máquinas trabalhar como os humanos, aprender através de exemplo e combinação de vários dados [17].

A maioria dos métodos DL usam arquiteturas de redes neuronais, sendo referidas normalmente como *Deep neural networks*. Estas, de acordo com o artigo em [18], consistem em múltiplas camadas conectadas através de nós. Cada camada aprende através da camada anterior, refinando-se e aperfeiçoando as previsões, categorização e descrição dos objetos dos dados.

Em DL os modelos aprendem a classificar tarefas diretamente de imagens, textos ou sons. Uma aplicação de DL com sucesso necessita de uma grande quantidade de dados para treinar o modelo, tal como *GPUs* (unidades de processamento gráfico) de forma a processar a vasta quantidade de dados mais rapidamente [17].

2.4 Algoritmos de aprendizagem não supervisionada

Esta dissertação vai abordar essencialmente ML não supervisionada. De acordo com o referido anteriormente, este tipo de aprendizagem não necessita de supervisão no modelo por parte dos utilizadores, sendo possível descobrir padrões e informações que não foram encontrados previamente, sobretudo porque funciona através de dados que não se encontram identificados, rotulados.

Em comparação aos outros tipos de aprendizagens, a aprendizagem não supervisionada permite efetuar tarefas bastante complexas, contudo pode apresentar resultados muito imprevisíveis.

Este tipo de aprendizagem é principalmente utilizado para resolver problemas de agrupar ou associar dados. Uma vez que o principal foco deste trabalho é agrupar dados e identificar anomalias em função dos perfis metabólicos dos pacientes de Colorretal, apenas serão apresentados os tipos de algoritmos de cluster da aprendizagem não supervisionada considerados mais relevantes de acordo com o artigo em [19]. No entanto, existem outros que poderão a vir a ser utilizados neste trabalho:

- *Exclusive Clustering* – é uma forma de agrupar os dados em que estes apenas podem pertencer a um único grupo (cluster). O *K-means clustering* é um dos algoritmos exemplo para agrupamento exclusivo.
    - *K-means clustering* é um método de agrupamento em que os dados são atribuídos a K grupos, onde K representa o número de grupos com base na distância aos centros de cada grupo. Quanto mais próximos do centro de cada grupo, os dados são considerados parte da mesma categoria.

      Resumidamente, quanto maior for a variável K, existirão grupos mais pequenos com maior granularidade; enquanto se o valor de K for menor, existirão maiores grupos de dados e menor granularidade. Este tipo de algoritmo é principalmente utilizado na segmentação de mercados, agrupamento de documentos, segmentação de imagens e compactação de imagens.
- *Overlapping Clustering* – é uma forma de agrupar os dados bastante semelhante ao agrupamento de exclusão. Contudo, neste caso, os dados podem pertencer a vários clusters com graus de associação diferentes para cada grupo. O *Fuzzy C-means clustering* é um exemplo para agrupamento por sobreposição.
    - *Fuzzy C-means*, de acordo com MATLAB em [20], é uma técnica de agrupamento, na qual o conjunto de dados agrupados em C clusters em que cada dado pertence a todos os grupos em um determinado grau. Caso um dado, por exemplo, se encontre mais perto do dado central de um grupo do que de outro centro, significa que tem um grau de associação maior a esse grupo do que ao que se encontra mais distante;
- *Agglomerative clustering* – é uma forma de agrupamento em que todos os dados são um grupo/cluster, sendo que se encontram inicialmente isolados como clusters separados e, em seguida, são unidos iterativamente com base nas semelhanças

encontradas até que apenas um único cluster seja alcançado. O principal exemplo deste tipo de agrupamento é a técnica *Hierarchical clustering*.

- *Hierarchical clustering*, também conhecida como, analise hierárquica de agrupamento (HCA), é uma forma de agrupar dados com o intuito de construir uma hierarquia de clusters. Todos os dados encontram-se num cluster próprio, sendo que ao longo das iterações, os clusters próximos tornam-se no mesmo grupo. O número de grupos, ao longo das iterações, vai diminuindo e o número de dados que os constitui aumenta até o algoritmo terminar e restar apenas um único cluster [21];

- *Probabilistic clustering* – é uma técnica não supervisionada que procura resolver principalmente a densidade dos grupos. No agrupamento probabilístico, os dados são agrupados com base na probabilidade de pertencerem a uma distribuição específica. Um dos métodos mais comuns para este tipo de agrupamento é o *Gaussian Mixture Model* (GMM).
  - *Gaussian Mixture Model* é classificado como um modelo de mistura, uma vez que é constituído por um número indeterminado de funções de distribuição de probabilidade. Estes modelos são usados para determinar a qual distribuição de probabilidade Gaussiana, ou normal, um determinado dado pertence. Caso se conheça a média ou a variância, é possível determinar qual a distribuição a que um dado pertence. No entanto, em GMMs, essas variáveis não são conhecidas, logo, assume-se a existência de uma variável oculta para agrupar os dados adequadamente.

Consideraram-se apenas estes algoritmos de *clustering* de forma a introduzir o leitor à diversidade de técnicas de aprendizagem não supervisionada. Estas técnicas de ML apresentam várias características/*features* que permitem aos utilizadores obter diferentes resultados em função dos dados que pretendem agrupar.

2.5 Aplicações de ML na medicina

Esta dissertação aborda a aprendizagem não-supervisionada, mais concretamente *Clustering*, sendo que de seguida serão revistas as diferentes aplicações da aprendizagem não supervisionada na Área da Saúde.

Nos últimos anos, a IA tem tido um crescimento sustentado, o que tem permitido solucionar problemas em vários setores: construção, educação, marketing e medicina. Sendo o setor de saúde o mais presente na vida de toda a população, a aplicação de IA no mesmo vai facilitar a análise de grandes quantidades de dados, a automatização de processos e a obtenção de recomendações de tratamentos para auxiliar o trabalho dos profissionais de saúde no seu dia-a-dia. Estas tecnologias atualmente estão em constante desenvolvimento, aprendem rapidamente, preveem resultados, analisam e tiram conclusões muitas vezes sem ser necessária a supervisão do ser humano.

O *Clustering* permite agrupar os dados dos pacientes que apresentem os mesmos indicadores ou combinação de indicadores e criar perfis para estes grupos. Considera-se que estes são mais fáceis de interpretar do que analisar todos os elementos individualmente, tornando-se cansativo para os médicos. Durante este processo é possível identificar anomalias que foram excluídas e que poderiam ter passado despercebidas ao olho humano, sendo consideradas casos especiais. Identificar estas anomalias antecipadamente permite estudá-las e investigá-las melhor para ser aplicado um melhor tratamento possível aos respetivos pacientes.

Seguidamente, na Tabela 1 são apresentadas algumas publicações onde se aplica a aprendizagem não supervisionada na área da saúde.

Tabela 1 - Diferentes aplicações de unsupervised learning in Healthcare

Publicação	**Área da Saúde**	**Algoritmo ML**
Chien-Hsing C [22]	Diagnóstico Cancro Mama	• K-Nearest-Neighbour (KNN) • Density Based Clustering • K-means • Hierarchical Clustering
Kemal P [23]	Classificação doença de Parkinson	• Fuzzy C-means Clustering
Nilashi M et al. [24]	Prever doença Parkinson	• K-means (Expectation Maximization)
Yanping W et al. [25]	Diagnóstico de dores de cabeça	• Multiple Fuzzy c-means Clustering
Liam T et al. [26]	Diagnóstico de saúde mental	• Mental Health Clustering Tool
Nihat Y et al. [27]	Diabetes e doenças de coração	• K-means
Jason N et al. [28]	Doença de Huntington	• K-means • Fuzzy C-means • Hierarchical • K-medoids
Hany A et al[29]	Doença de Alzheimers	• K-means • K-means-Mode • Multi-Layer Clustering • Hierarchal Agglomerative

Na tabela apresentada, sintetizam-se as publicações consideradas relevantes relativamente ao uso da técnica de *Clustering* na área da saúde. Atualmente, existe artigos na área da saúde que comprovam o uso de diferentes algoritmos e técnicas de agrupar os dados para um sistema de

diagnóstico de doenças mentais, como é o caso de Jason N et al. [28], em que o uso de *Fuzzy c-means Clustering* indica até que percentagem um paciente pertence a um grupo de pacientes. O algoritmo utilizado neste artigo é muito comum para diagnóstico de doenças mentais de pacientes, por exemplo, Parkinson. Este tipo de doença mental, de acordo com o artigo em [23], é possível de se classificar em doentes reais através da utilização do algoritmo *Fuzzy c-means* em função do peso das características. Um dos principais problemas no uso destas técnicas para diagnóstico de doenças, tal como, cancro da mama em [22], é a necessidade de especialistas na área para identificar os primeiros doentes e rotular as instâncias dos dados para o algoritmo conseguir aprender com alguma verdade absoluta. Este processo de rotular os dados torna-se difícil, demorado e dispendioso.

Atualmente, existem trabalhos que utilizam o algoritmo *K-means* modificado e *Fuzzy Hierarchical* para eliminar o ruido existente nos dados de forma a classificar doentes com problemas de coração e diabetes, [27], ou estudar perfis metabólicos de doença de Huntington em ratos. Existem artigos a demonstrar a utilização de diferentes técnicas de redução de dimensionalidade, para identificar melhor padrões existentes nos dados de pacientes com doença de Parkinson através de métodos de predição de agrupamento, segundo [24]. A diversidade de utilização de algoritmos de agrupamento ou através de aprendizagem não supervisionada é vasta, apenas é necessário experimentar, através de tentativa-erro muitas vezes, até encontrar os melhores resultados, melhores métodos e práticas e selecionar as combinações de técnicas mais eficientes e adequadas para cada problema no setor da saúde.

Na área da saúde, o *Clustering* pode ter diversas aplicações para diferentes utilidades. Pode ser aplicado a vários pacientes com a mesma doença, como é o caso deste trabalho, onde serão identificados vários grupos de pacientes com cancro do reto localmente mais avançado em função do perfil metabólico, e fornecida melhores recomendações quanto ao tratamento e resposta patológica. Para além disso, também se considera que este tipo de aprendizagem possibilita na área de saúde agrupar diferentes documentos de acordo com o texto que contêm. Agrupar documentos, prescrições e outro tipo de dados médicos permite aos técnicos de saúde reduzir o seu trabalho.

Concluindo e analisando apenas para a área de saúde, o *Clustering* pode ter diversas aplicações que possibilitam a resolução de diferentes problemas. Todos os anos são realizadas mais pesquisas e investigações, permitindo aplicar estes algoritmos de ML na medicina e ajudar os profissionais de saúde a avaliar e tratar mais eficientemente os seus pacientes.

2.6 Bibliotecas de Machine Learning

Antigamente, as tarefas de ML eram programadas manualmente, tal como, todos os algoritmos e fórmulas estatísticas, despendendo-se muito tempo. Atualmente, graças às diversas blibliotecas Python, *frameworks* e módulos, este processo tornou-se muito mais simples, rápido e eficiente. A vasta coleção de bibliotecas de ML foi umas das principais razões que conduziu à popularidade da linguagem *Python*.

De acordo com V. Prabhu [30], bibliotecas de ML ou *frameworks*, consistem num conjunto de rotinas e funções. Estas ferramentas permitem ajudar os programadores num desenvolvimento mais fácil e rápido de modelos de ML, ultrapassando os detalhes básicos nos algoritmos e efetuando tarefas mais complexas, sem ter de repetir muitas linhas de código.

De seguida serão apresentadas algumas bibliotecas relevantes para o tema.

2.6.1 TensorFlow

TensorFlow é uma biblioteca grátis e *open-source*, originalmente desenvolvida pela Google Brain Team, para investigação e produção. Esta biblioteca é principalmente utilizada para DL e fornece APIs estáveis para Python e C, através de processamento em paralelo. Esta ferramenta permite treinar com muita facilidade em CPU e GPU para computação distribuída.

Esta biblioteca é baseada no princípio de *dataflow*, em que o programa está organizado na forma de blocos computacionais associados uns aos outros na forma de *directed graph*, sendo assim chamado de *computational graph*. A estrutura dos dados usados é chamada de tensor, pois representa um vetor de elementos multidimensional [31].

Esta arquitetura do TensorFlow facilita o processamento de cálculos em paralelo em CPUs de múltiplos núcleos, sendo bastante adequada para construir redes neurais.

2.6.2 Keras

A biblioteca Keras, segundo [31], fornece uma interface de alto-nível de programação para construir redes neurais, tal como TensorFlow. A principal diferença destas duas ferramentas é que enquanto o TensorFlow é uma biblioteca *end-to-end*, o Keras é uma *interface* ou camada de abstração que normalmente opera "por cima" de outra biblioteca *back-end*. Esta *interface open-source* foi desenvolvida apenas em Python e segue os seguintes princípios: facilidade de uso, modularidade e extensibilidade. As principais vantagens desta ferramenta são as seguintes:

- A modularidade permite aos modelos serem entendidos como uma sequência ou um gráfico sozinho;
- O minimalismo para obter um resultado enquanto se fornece tão pouco;
- A maximização da legibilidade e extensibilidade que permite aos pesquisadores efetuarem testes.

2.6.3 Scikit-learn

SciKit Learn é uma biblioteca grátis para Python focada em processamento de dados. Contém diversos métodos de classificação, análise de regressão, *clustering* e outros algoritmos relacionados com ML, referido em [32].

Esta biblioteca é considerada uma das mais eficientes para *data mining* e *data analysis*, contudo, atualmente é frequentemente utilizada para seleção de modelos e *clustering*. É considerada uma das bibliotecas mais populares relativamente ao facto de possuir uma API intuitiva, fácil de usar, rápida, compreensível e com boa documentação para os programadores usufruírem.

2.6.4 Comparação das bibliotecas

Nesta secção, será efetuado um breve sumário das bibliotecas anteriormente apresentadas.

De acordo com as limitações apresentadas, tanto a biblioteca TensorFlow como a Keras são mais adequadas para os casos de construção de redes neurais. Dado não ser esse o tema principal desta dissertação e uma vez que existe um baixo volume de dados, estas não serão escolhidas.

Nesse sentido, a biblioteca selecionada será SciKit-Learn. Trata-se de uma biblioteca que apresenta várias ferramentas para algoritmos de ML. Na medida em que, existe bastante suporte de informação e documentação sobre a mesma, ao que acresce a sua fácil utilização, considera-se ser a escolha ideal.

2.7 Conclusões

Nesta secção são apresentadas as conclusões efetuadas ao longo do capítulo do estado de arte. No setor de saúde existem inúmeras utilidades de agrupamento de pacientes, assim como foi demonstrado. Atualmente há artigos sobre a utilização dos algoritmos *fuzzy c-means*, *k-means*, *Hierarchical Clustering*, entre outros, para diagnosticar, classificar ou identificar prognósticos para doentes com doenças mentais, diabetes, problemas de coração, cancro da mama, entre outros.

Assim como foi apresentado anteriormente, a biblioteca utilizada para este trabalho vai ser o Scikit-learn, pois apresenta melhor suporte de informação e foi recomendada pelo orientador. Os algoritmos de agrupamento a ser estudados são os melhores que a biblioteca tem a oferecer para o conjunto de dados em estudo, por exemplo, *K-means, Agglomerative e GMM*.

O cancro do reto e/ou colon é uma das doenças oncológicas mais comuns nos países desenvolvidos, sendo o tratamento variável e a resposta patológica um importante preditor de bom resultado. Este trabalho procura identificar a melhor combinação de técnicas de redução de dimensionalidade, tratamento de dados e algoritmos de aprendizagem que permitem melhor agrupar e identificar anomalias no conjunto de dados de perfis metabólicos de aminoácidos e acilcarnitinas de pacientes com cancro do reto localmente mais avançado.

3 Análise de Valor

Neste capítulo será analisado o valor do produto em desenvolvimento durante este trabalho. O principal objetivo passa pela redução de custos ou pelo aumento de desempenho, ou até mesmo, pela conjugação de ambos, através de uma avaliação e análise do trabalho.

A análise de valor foi baseada nos seguintes componentes:

- Processo de Inovação, usando o modelo *New Concept development* (NCD), tendo em consideração o valor da solução para o cliente e a definição da proposta de valor*;*
- Análise funcional - *Quality function deployment* (QFD).

Nas próximas secções, serão explorados os componentes referidos em detalhe.

3.1 Processo de inovação

A evolução constante da tecnologia leva ao aparecimento de novos produtos. Estes produtos alimentam assim a necessidade dos consumidores, que procuram por produtos de boa qualidade. De forma a assegurar a qualidade do produto e adicionar valor à organização foi desenvolvido um processo de inovação constituído por várias etapas. O típico processo de inovação está dividido em três partes principais: *Fuzzy Front End* (FFE), *New Product Development* (NPD) e Comercialização.

A parte do *Fuzzy Front End*, corresponde ao estado inicial do novo processo de desenvolvimento. Esta etapa é compreendida como uma fase experimental, incerta, e caótica, e uma vez que vai servir de base para as últimas partes é necessário adaptá-la. Para remover o incontrolável, foi desenvolvido um novo termo: *Front End of Inovation*, que permitiu resolver a falta de padrões

no FFE, permitindo assim criar o modelo NCD, de acordo com [33], tal como é possível verificar na Figura 2.

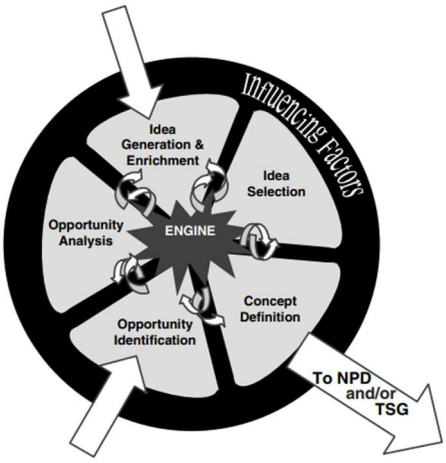


Figura 2 - New Concept Development. Retirado de [33]

O modelo NCD está dividido em três partes:

- O *Engine* é composto pelas características da organização que vão influenciar o processo, tal como, visão, estratégia de negócio, cultura e liderança;
- Fatores que influenciam o motor e moldam a roda, tal como, tecnologias, capacidades organizacionais, cliente, entre outros;
- Roda que consiste nos cinco elementos do *Front End*: Identificação de Oportunidade, Análise de oportunidade, Produção de Ideias, Seleção de ideia e Conceito & Tecnologia.

3.1.1 Influencing factores

Os fatores que influenciam o resto dos componentes do modelo NCD são, por exemplo, representados pela influência de competidores, pelo cliente, pelas tecnologias usadas e pela competência da organização.

Os fatores que influenciam este projeto são:

- A qualidade e quantidade dos dados obtidos dos pacientes, na medida em que devem ser obtidos frequentemente, sendo assim atualizados, uma vez que terão impacto no resultado dos algoritmos;
- A tecnologia utilizada, que deverá estar sempre atualizada, sendo que novas melhorias podem ser adicionadas aos algoritmos, tornando-os assim mais eficientes e precisos;
- A satisfação e comentários dos profissionais de saúde, uma vez que são essenciais para desenvolver uma solução que lhes permita facilitar e melhorar o seu trabalho no futuro.

3.1.2 Engine

Nesta secção são apresentados os fatores que permitem conduzir e influenciar o processo, tal como, a liderança e a cultura da organização. O principal objetivo deste projeto é dar uso à vasta quantidade de dados dos perfis metabólicos e combinar com as tecnologias disponíveis, permitindo melhorar o dia-a-dia dos profissionais de saúde no desempenho do seu trabalho, o que, por sua vez, permite melhorar a qualidade de vida dos pacientes.

Este projeto vai ser realizado em parceria com os profissionais de saúde do Centro Hospitalar Universitário do Porto (CHUPorto), que, de acordo com [34], tem como missão "prestar cuidados de saúde humanizados, competitivos e de referência, promovendo a articulação com os outros parceiros do sistema, a valorização do ensino pré e pós-graduado e da formação profissional, a dinamização e incentivo à investigação e desenvolvimento científico na área da saúde".

3.1.3 Identificação de Oportunidade

Este processo permite identificar as oportunidades que são tipicamente relevantes para a organização explorar.

A abundância de dados biomédicos de pacientes com diversos tipos de cancro é o que permite aos profissionais de saúde compreender melhor as características moleculares destes cancros. Na medida em que, o cancro do colorretal, tal como supramencionado, é um dos mais comuns nos países desenvolvidos e que está em crescimento, a quantidade de dados gerada diariamente é imensurável. Nesse sentido, existe oportunidade para estes dados serem explorados, de modo a facilitarem o trabalho dos profissionais de saúde, quer seja no processo de agrupar os casos semelhantes ou de identificar possíveis anomalias, o que permitirá um melhor tratamento dos pacientes.

Para uma melhor compreensão do valor que este projeto trará para o cliente, serão explicitados os seguintes concentos: Valor para o cliente (VC), *Perceived Value* e Proposta de Valor.

Valor

De acordo com Zeithaml em [35], o valor percebido consiste na avaliação geral efetuada pelo cliente sobre a utilidade de um produto, com base nas suas perceções do que é recebido e do que é dado. Embora o que seja recebido e dado varie entre os clientes, o valor representa uma compensação dos principais componentes de dar e receber.

O valor percebido pode ser considerado como a relação entre benefícios e sacrifícios, sendo principalmente o resultado da análise da utilidade do produto com base nas perceções. Diferentes clientes podem ter diferentes perceções do valor sobre o mesmo produto/serviço.

De modo a avaliar o valor para o cliente deste projeto, foi criada a Tabela 2 que apresenta os benefícios e os sacrifícios associados.

Tabela 2 - Benefícios e Sacrifícios associados ao projeto

Benefícios	Sacrifícios
• Agrupar pacientes entre casos similares; • Identificar casos irregulares; • Melhor decisão de tratamento dos pacientes; • Facilitar trabalho dos profissionais de saúde; • Melhorar qualidade de vida do paciente.	• Esforço em obter os dados; • O primeiro momento da aplicação não vai apresentar resultados 100% exatos; • Manutenção e custos do software.

A tabela anterior apresenta os benefícios e os sacrifícios no âmbito deste projeto. Os principais benefícios para o cliente são o de poder agrupar os pacientes em casos similares, uma vez que permitirá um tratamento mais rápido e identificar casos irregulares. A nível dos sacrifícios, apenas se consideram os primeiros momentos da aplicação, pelo esforço para obter os dados dos pacientes e uma vez que o algoritmo tem uma curva de aprendizagem no início.

Proposta de Valor

Uma proposta de valor é uma prática originária do *Marketing* e resume, através de ideias claras e transparentes, as vantagens para um cliente pela utilização de um determinado produto, definindo o valor que este vai receber, segundo o autor em [36]. Esta representa um conjunto de benefícios ou valores que se promete oferecer aos consumidores para satisfazerem as suas necessidades.

Através de uma proposta de valor é possível diferenciar a sua marca ou produto da concorrência, despertando o interesse do público, sendo este o primeiro passo para demonstrar como uma empresa pode atender às necessidades dos clientes de forma satisfatória e certeira.

A proposta de valor é uma vista geral do conjunto de produtos e serviços de uma empresa que são de valor para o cliente. De modo a descrever a proposta de valor para o cliente deste trabalho, vai ser utilizada a ferramenta *Value Proposition Canvas*[36].

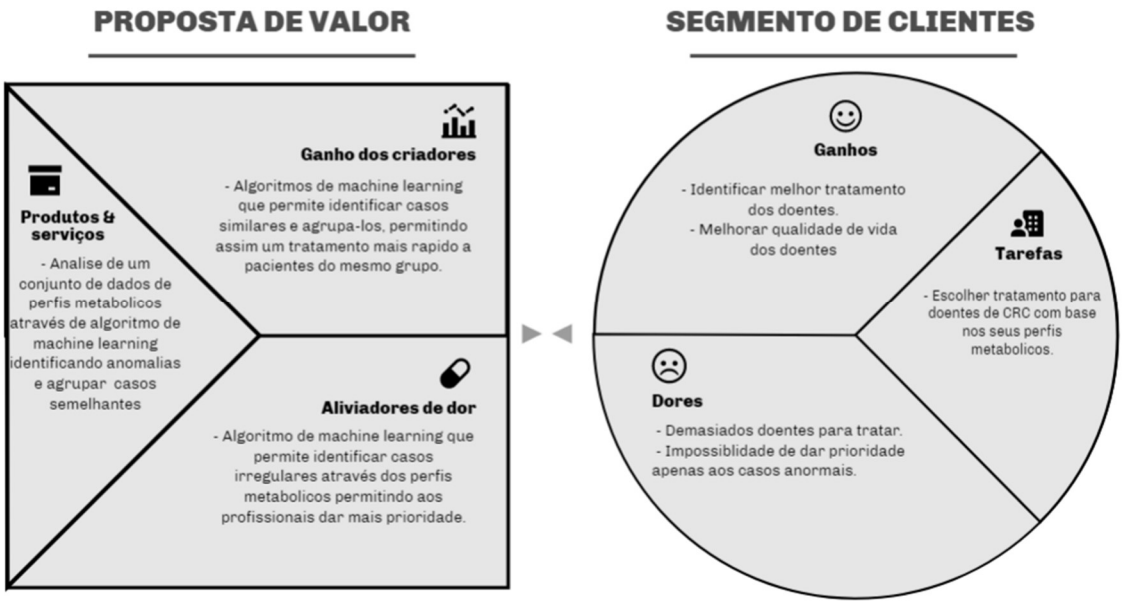


Figura 3 - Proposta de valor baseado no modelo Osterwalder

3.1.4 Análise de oportunidade

A análise de oportunidade consiste numa avaliação do ambiente externo, sendo um processo de estudo de oportunidades do mercado, de modo a entender se existe um mercado e uma necessidade para a solução. É necessário considerar o esforço investido no seu desenvolvimento, vantagens para o futuro, existência de competitividade, análise das tendências da tecnologia e tema do projeto para validar a sua rentabilidade.

Com o já referido anteriormente, ML apresenta uma vasta variedade de algoritmos que possibilitam simplificar o trabalho dos profissionais de saúde, através da análise de grandes quantidades de dados, evitando o erro humano no processamento dos mesmos.

De acordo com o mencionado na secção de Aplicações de ML na medicina, no capítulo 2, existem atualmente várias publicações relativas às aplicações de aprendizagem não supervisionada na medicina, muitas das quais aplicam ML para diagnóstico e previsão de doenças/cancro.

O principal objetivo deste trabalho consiste no agrupamento e identificação de anomalias de pacientes com cancro do reto localmente mais avançado, em função dos seus perfis metabólicos. Não obstante, existirem trabalhos sobre aplicações de ML supervisionada e DL, não foi possível encontrar aplicações de aprendizagem não supervisionada para os perfis metabólicos em estudo.

Conclui-se que existe oportunidade de mercado ao combinar o poder de ML, mais concretamente aprendizagem não supervisionada, com as respetivas tecnologias de modo a

identificar uma associação entre os respetivos indicadores e a resposta patológica, na tentativa de identificação de anomalias e *clusters*, em função destes biomarcadores moleculares preditores de resposta.

3.1.5 Geração de Ideia

O principal objetivo desta fase é a geração de ideias que, posteriormente, vão permitir desenvolver e gerar uma ideia concreta. De acordo com os autores em [37], as ideias podem passar por muitas iterações e mudanças à medida que são examinadas, estudadas, discutidas e desenvolvidas ao longo do processo de NCD até estarem completamente definidas e bem identificadas. O contacto direto entre os clientes/utilizadores normalmente potencializam esta fase.

Nesse sentido, o projeto foi discutido com profissionais de saúde, permitindo elencar os vários requisitos cruciais a este trabalho:

- Identificação das características principais que permitem agrupar os vários pacientes;
- Observação de padrões ocultos nos dados;
- Divisão dos pacientes em diferentes grupos, de acordo com os seus perfis metabólicos;
- Identificação de qualquer tipo de anomalia;
- Desenvolvimento de uma interface de fácil uso e compreensão para os utilizadores da aplicação.

3.1.6 Seleção de Ideias

Nesta fase são selecionadas as ideias mais importantes para se atingir um maior valor de negócio do projeto, considerando o que é necessário, custos e benefícios para as partes interessadas.

Conclui-se que os requisitos mais importantes no contexto desta dissertação são a identificação de diferentes grupos de acordo com os seus perfis metabólicos e respetivas anomalias, na medida em que são estes os que realmente permitem aos profissionais de saúde um melhor tratamento do doente, facilitando as dificuldades do dia-a-dia.

3.1.7 Conceito e desenvolvimento tecnológico

O último elemento do modelo NCD envolve o desenvolvimento de um caso de negócio com base em estimativas: de um potencial mercado, das necessidades do cliente, do risco geral do projeto e de incógnitas tecnológicas.

Considerando o exposto anteriormente, o principal conceito associado ao produto desta dissertação é o desenvolvimento de um sistema que permita analisar e processar dados de perfis metabólicos de pacientes com Cancro do Colorretal, utilizando algoritmos de ML. O objetivo prende-se com ajudar os profissionais de saúde no seu trabalho diário na identificação de casos semelhantes entre pacientes e possíveis anomalias, permitindo assim um melhor controlo da doença nos pacientes, de modo que estes tenham uma melhor qualidade de vida.

Na medida em que, o valor deste negócio para o cliente depende muito da eficácia dos algoritmos em agrupar e identificar anomalias, estes são considerados a maior incógnita do projeto.

3.2 Análise funcional

Nesta secção vão ser identificados os requisitos funcionais mais críticos para o produto, através do QFD. A sigla QFD significa, em português, qualidade, funcionalidade e implantação. Este é um sistema utilizado tanto para projetar um produto ou serviço, baseado nas necessidades do cliente, bem como para demonstrar as funcionalidades em função dos requisitos. Através da "House of Quality" é possível verificar relação entre ambos [38].

A "House of Quality" é um dos elementos do QFD, que, de acordo com o contexto desta dissertação, foi desenvolvido para assegurar qualidade e potencial para o cliente. Visto que, este produto é para um cliente específico, não vai ser considerada a análise competitiva.

Na medida em que, já foram identificados os requisitos necessários do software na fase de Geração de Ideias, deve agora ser definido como identificar as funcionalidades. No contexto deste trabalho foi possível identificar as seguintes tecnologias necessárias: Algoritmos de ML, Desenvolvimento de UI, Tratamento de Dados, Carregamento de dados e *Clustering*.

Na Figura 4 é apresentada a "*House of Quality*" para o trabalho a ser desenvolvido nesta dissertação. A coluna WHATs contém as funcionalidades anteriormente definidas das necessidades do cliente e no centro é efetuada a relação com as tecnologias anteriormente identificadas. A escala usada para representar a relação começa em "Fraco", o qual significa que "não existe relação ou é muito fraca", e termina em "Forte" que representa uma "relação forte" entre requisito e tecnologia, representado na Tabela 4. No topo da "casa" é possível verificar a relação entre as várias tecnologias, de acordo com Tabela 3.

Tabela 3 - Escala de avaliação utilizada entre tecnologias

+	Positivo
.	Nenhum
-	Fraco

Tabela 4 - Escala de relação entre tecnologias e requisitos do cliente

Relações	
Forte	●
Moderado	○
Fraco	▽

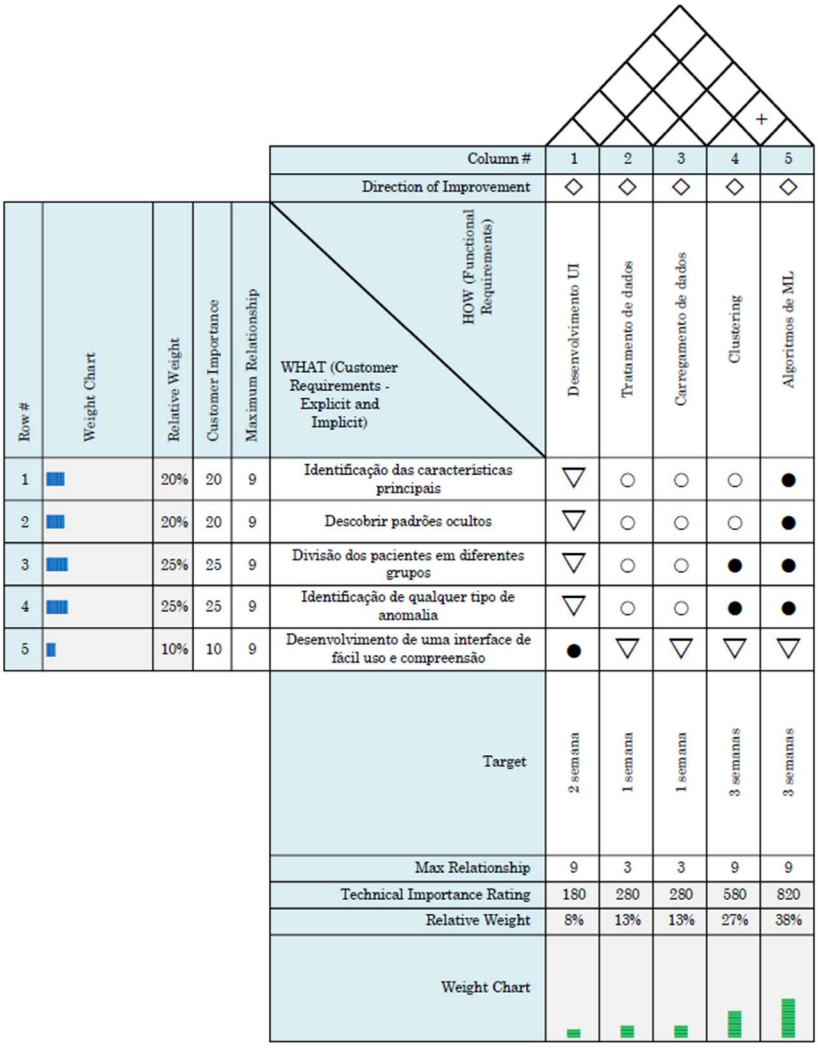


Figura 4 - House of Quality

4 Análise e desenho da solução

Neste capítulo serão apresentados a análise e desenho da solução. A análise corresponde à compreensão e preparação dos dados. Na subseção do desenho é apresentada a proposta de arquitetura esperada da solução através de diferentes vistas da mesma.

4.1 Análise de dados

Este estudo procura analisar diferentes conjuntos de dados de perfis metabólicos de aminoácidos e acilcarnitinas de pacientes de cancro colorretal. Os respetivos dados são fornecidos pelo Centro Hospitalar Universitário do Porto (CHUP). Os vários conjuntos de dados encontram-se divididos em múltiplos ficheiros e contém resultados de 902 perfis de acilcarnitinas e 4052 de aminoácidos de pacientes reais com diversos diagnósticos desde outubro de 2019 até junho de 2023.

Aminoácidos

Na Figura 5 é demonstrado um exemplo parcial do conjunto de dados de aminoácidos. Este é constituído por 32 indicadores, sendo o DIAG (se existe diagnóstico) e o sexo da pessoa representado por valores binários, 0 e 1; e os restantes indicadores por valores inteiros.

MET	CYSTA	ILE	LEU	TYR	PHE	ORN	LYS	1MHIS	HIS	3MHIS	ARG	SEXO	IDADE_NA_COLHEITA_ANOS	DIAG	DIAG_ID
15	0	34	83	30	530	34	114	0	51	0	37	0	12	1	19
33	0	52	112	54	223	51	166	0	86	0	55	1	13	1	19
29	0	58	89	42	877	36	125	0	66	0	65	0	19	1	19
23	0	55	114	52	601	42	149	0	55	0	41	0	17	1	19
19	0	47	78	45	203	52	135	0	57	0	47	0	12	1	19
37	0	45	91	55	491	34	143	0	49	0	62	0	11	1	19
17	0	34	69	29	431	57	150	0	45	0	38	1	13	1	19
33	0	64	129	43	799	69	169	0	66	0	30	0	17	1	19
31	0	46	96	46	275	75	146	0	51	0	44	0	20	1	19
25	0	54	108	47	139	82	184	0	71	0	54	0	21	1	19
27	0	52	83	39	768	38	164	0	46	0	46	0	16	1	19
18	0	58	114	39	397	66	155	0	68	0	34	0	15	1	19
25	0	37	73	43	211	31	112	0	47	0	42	1	18	1	19
19	0	58	111	72	210	77	127	0	64	0	57	0	13	1	19
17	0	49	103	66	164	56	125	0	66	0	62	0	12	1	19
24	0	68	127	64	549	61	175	0	69	0	65	0	16	1	19
23	0	79	167	73	520	171	231	0	78	0	1	0	17	1	19
30	0	48	94	52	1114	38	151	0	53	0	56	1	17	1	19
17	0	47	86	40	630	51	129	0	58	0	33	1	13	1	19
19	0	53	98	48	399	66	122	0	71	0	29	0	16	1	19
24	3	44	91	44	498	33	147	0	80	0	32	1	22	1	19
35	0	45	84	47	223	43	122	0	53	0	61	0	13	1	19
21	0	39	89	37	955	66	153	0	63	0	40	0	15	1	19
19	0	50	110	54	168	56	162	0	57	0	57	0	10	1	19
29	0	49	107	40	333	49	139	0	53	0	49	1	10	1	19
26	0	46	95	43	405	47	121	0	58	0	43	1	9	1	19
25	0	55	126	56	272	83	155	0	75	0	47	0	18	1	19
22	1	55	108	39	233	53	127	0	62	0	54	0	17	1	19
25	0	37	68	33	675	44	135	0	40	0	34	1	18	1	19
14	0	39	74	32	505	46	121	0	56	0	21	1	19	1	19
19	0	42	94	38	677	50	131	0	43	0	37	0	16	1	19

Figura 5 – Exemplo parcial dos dados utilizados

Nos dados fornecidos existem 28 indicadores referentes às concentrações de plasma de aminoácidos, sendo que os restantes dizem respeito à idade, sexo, se existe diagnóstico e respetivo diagnostico do paciente. Na Tabela 5 são demonstrados os vários indicadores de aminoácidos considerados, as respetivas abreviações e unidades de medida.

Tabela 5 - Indicadores existentes nos dados de perfis de aminoácidos

Nome Indicador	Abreviação	Unidade de medida
Taurine	Tau	Micromoles/Litro (µmol/L)
Aspartic Acid	Asp	
Hydroxyproline	Hyp	
Threonine	Thr	
Serine	Ser	
Asparagine	Asn	
Glutamic acid	Glu	
Glutamine	Gln	
Alpha-Aminoadipic acid	AAA	
Proline	Pro	
Glycine	Gly	
Alanine	Ala	
Citrulline	Cit	
Aminobutyric acid	ABU	
Valine	Val	
Cystine	Cys2	
Methionine	Met	
Cystathionine	Cysta	
Isoleucine	Ile	
Leucine	Leu	
Tyrosine	Tyr	
Phenylalanine	Phe	
Ornithine	Orn	
Lysine	Lys	
1-Methylhistidine	1Mhis	
Histidine	His	
3-Methylhistidine	3Mhis	
Arginine	Arg	
Idade no momento da colheita	Age	Inteiro
Sexo	Gender	Binário: 0 Feminino, 1 Masculino

O conjunto de dados é composto por 4052 exemplos, classificados em diferentes classes de diagnóstico, nomeadamente a diferentes doenças metabólicas, cancro colorretal ou sem diagnóstico. Dado que, o foco deste trabalho é o estudo das alterações metabólicas em pacientes com cancro do reto localmente avançado, dos 4052 exemplos fornecidos pelo CHUP, apenas 273 instâncias dos dados correspondem a diagnósticos do cancro em estudo, sendo que apenas 6,78% dos dados vão ser utilizados nos casos de estudos dos diferentes algoritmos não supervisionados para esta dissertação, os restantes dados não vão ser considerados.

Acilcarnitinas

Na medida em que, os dados fornecidos pelos profissionais de saúde estão distribuídos por múltiplos ficheiros, obtidos por máquinas especializadas para fornecer os resultados dos exames de perfis de acilcarnitinas, vai ser necessário o tratamento dos dados. Na Tabela 6 e 7 estão demonstrados dois exemplos dos resultados obtidos dos exames de uma mesma amostra de sangue, com as respostas às diferentes carnitinas.

Tabela 6 - Exemplo de resultados de perfil de acilcarnitinas avançado

#	Name	Trace	RT	Area	IS Area	Response	Detection Flags
1	C3DC	248 > 85	0.43	16.629	2196.167	0.035	MM
2	C4OH	248 > 85	0.44	6.147	2917.427	0.005	MM
3	C4DC	262 > 85	0.68	586.968	2917.427	0.479	bb
4	C5:1	244 > 85	0.63	13.447	3059.292	0.010	MM
5	C5OH	262 > 85	0.69	398.293	3059.292	0.310	bb
6	C6DC	289.1 > 85	0.76	0.609	1420.675	0.001	bb
7	C8:1	286 > 85	0.64	861.224	5969.741	0.329	bb
8	C10:1	314 > 85	0.83	250.373	8856.615	0.061	bb
9	C10:2	311.21 > 85	0.77	0.313	8856.615	0.000	MM
10	C12:1	341.26 > 85	0.59	1836.286	13098.196	0.349	bb
11	C14:1	370 > 85	0.65	705.236	14242.084	0.118	bb
12	C14:1OH	385.5 > 85	0.71	41.455	14242.084	0.007	bb
13	C14:2	368 > 85	0.47	29.347	14242.084	0.005	MM
14	C14OH	388 > 85	0.16	0.579	14242.084	0.000	bb
15	C16:1	398 > 85	0.60	257.577	23812.861	0.054	bb
16	C16OH	416 > 85	0.66	5.129	23812.861	0.001	bb
17	C18:1	426 > 85	0.61	5834.324	38039.688	0.653	bb
18	C18:1OH	442 > 85	0.67	64.989	38039.688	0.007	MM
19	C18:2	424.33 > 85	0.61	1761.418	38039.688	0.197	bb
20	C18OH	444.3 > 85	0.85	7.591	38039.688	0.001	MM
21	C3 d3	221.2 > 85	0.62	2196.167		2196.167	bb
22	C4 d3	235 > 85	0.62	2917.427		2917.427	bb
23	C5DC d6	282 > 85	0.32	33.313		33.313	bb
24	C5 d9	255 > 85	0.63	3059.292		3059.292	bb
25	C6 d3	263 > 85	0.64	1420.675		1420.675	bb
26	C8 d3	291.1 > 85	0.64	5969.741		5969.741	bb
27	C10 d3	319.3 > 85	0.65	8856.615		8856.615	bb
28	C14 d3	375.2 > 85	0.60	14242.084		14242.084	bb
29	C12 d3	347.3 > 85	0.65	13098.196		13098.196	bb
30	C16 d3	403.5 > 85	0.60	23812.861		23812.861	bb
31	C18 d3	431.6 > 85	0.62	38039.688		38039.688	bb

Tabela 7 - Exemplo de resultados de perfil de acilcarnitinas básico

#	Name	Trace	RT	Area	IS Area	Response	Detection Flags
1	Carnitina	162.1 > 85	0.61	2675.432	6805.857	16.039	bb
2	C2	204.1 > 85	0.61	8528.329	15748.846	9.487	bb
3	C3	218.2 > 85	0.62	1107.154	4212.312	1.209	bb
4	C4	232 > 85	0.57	92.063	3065.351	0.071	MM
5	C5	246 > 85	0.57	360.178	6063.218	0.141	bb
6	C5DC	276 > 85	0.77	2.812	56.870	0.082	MM
7	C6	260 > 85	0.58	52.427	4442.634	0.027	bb
8	C8	288.1 > 85	0.58	119.264	8898.764	0.031	bb
9	C10	316.3 > 85	0.64	171.397	7768.977	0.048	bb
10	C12	344.3 > 85	0.64	142.704	14525.496	0.024	MM
11	C14	372.2 > 85	0.65	265.618	16256.237	0.039	bb
12	C16	400 > 85	0.65	6345.880	40523.078	0.780	bb
13	C18	428.6 > 85	0.60	4185.330	39662.383	0.450	bb
14	Carnitina d9	171.1 > 85	0.61	6805.857		6805.857	bb
15	C2 d3	207.1 > 85	0.61	15748.846		15748.846	bb
16	C3 d3	221.2 > 85	0.62	4212.312		4212.312	bb
17	C4 d3	235 > 85	0.62	3065.351		3065.351	bb
18	C5 d9	255 > 85	0.62	6063.218		6063.218	bb
19	C5DC d6	282 > 85	0.55	56.870		56.870	MM
20	C6 d3	263 > 85	0.63	4442.634		4442.634	bb
21	C8 d3	291.1 > 85	0.63	8898.764		8898.764	bb
22	C10 d3	319.3 > 85	0.64	7768.977		7768.977	bb
23	C12 d3	347.3 > 85	0.64	14525.496		14525.496	bb
24	C14 d3	375.2 > 85	0.64	16256.237		16256.237	bb
25	C16 d3	403.5 > 85	0.65	40523.078		40523.078	bb
26	C18 d3	431.6 > 85	0.65	39662.383		39662.383	bb

Porquanto, este trabalho é a continuação de estudos anteriores ao mesmo conjunto de dados de acilcarnitinas, os vários ficheiros foram tratados previamente para um único ficheiro Excel, facilitando assim o tratamento e análise do mesmo.

A Tabela 6 e 7 apresentam resultados de diferentes indicadores de acilcarnitinas obtidos de uma mesma amostra de sangue. A construção do conjunto de dados dos vários ficheiros exemplos demonstrados nas tabelas é efetuada do seguinte modo: os valores de RT (tempo de retenção) e "Response" são obtidos nas respetivas colunas. Os valores do tempo de retenção são comparados com os valores padrões definidos anteriormente com os profissionais de saúde. Quando este valor desvia demasiado do padrão é considerado 0; ou no caso de 0 não ser possível para o respetivo indicador, o resultado da amostra é descartado. Se o valor se encontrar dentro do padrão, o valor de resposta é aceite e considerado.

Depois de reunir e juntar os dados fornecidos dos ficheiros de textos, através de processos automáticos desenvolvidos em Python, validando os valores de acordo com os respetivos RT, os dados encontram-se disponíveis para ser adicionados aos dados já processados anteriormente e para ser consumidos por algoritmos de ML. Na Figura 6 é possível verificar um

exemplo parcial dos dados de perfis de acilcarnitinas depois de tratados os múltiplos ficheiros para serem utilizados pelos algoritmos.

C50H	C6DC	C8:1	C10:1	C10:2	C12:1	C14:1	C14:10H	C14:2	C140H	C16:1	C160H	C18:1	C18:10H	C18:2	C180H	DIAG_ID
0.239	0	0.146	0.048	0	0.394	0.14	0.005	0.01	0	0.047	0.002	0.99	0.008	0.197	0.002	CRC
0	0	0.158	0.046	0	0.405	0.118	0.004	0	0	0.06	0.001	0.587	0.006	0.108	0.005	CRC
0	0	0.144	0.053	0	0.392	0.141	0.008	0	0.006	0.061	0	0	0	0	0	CRC
0.216	0	0	0.073	0	0.395	0.126	0.007	0.080	0	0.030	0.005	0.843	0.003	0	0.005	CRC
0	0	0	0	0	0.182	0	0	0	0	0	0	0	0	0.180	0	CRC
0.155	0	0.058	0.066	0	0.335	0.082	0.003	0	0	0	0.014	1.088	0	0.330	0	CRC
0	0	0.037	0.138	0	0.353	0.098	0	0	0	0.042	0.007	0	0	0.223	0	CRC
0	0	0.037	0.138	0	0.353	0.098	0	0	0	0.042	0.007	0	0	0.223	0	CRC
0	0	0.164	0.126	0.011	0	0.189	0.017	0.033	0	0	0	2.034	0.015	0	0	CRC
0.247	0	0	0.288	0	3.184	0	0	0	0	0	0	0.607	0	0	0	CRC
0.088	0	0	0.266	0	3.151	0.242	0	0.057	0	0	0	0.765	0	0.290	0	CRC
0	0	0	0	0.038	0	0	0	0	0	0.106	0	0	0	0.368	0.003	CRC
0.143	0	0	0.334	0	1.116	0.200	0	0.019	0	0.034	0.012	0.769	0.009	0.129	0	CRC
0.482	0	0	0.107	0	0.853	0.182	0	0.013	0	0.064	0	0.871	0	0.334	0	CRC
0.078	0	0.168	0.102	0	0.921	0.182	0	0	0	0.085	0	0.983	0	0.255	0	CRC
0.239	0	0.195	0.221	0.01	1.190	0.328	0.017	0.054	0	0.146	0	2.639	0.02	0.448	0	CRC
0	0	0.203	0.105	0	0.907	0.203	0	0	0	0.050	0.019	0.805	0.008	0.150	0	CRC
0	0	0.211	0.337	0	0.850	0	0.012	0	0.008	0	0.010	1.098	0	0.358	0	CRC
0	0	0.180	0.077	0	0.845	0.222	0	0	0	0.084	0.014	1.086	0.007	0.208	0	CRC
0	0	0.381	0.271	0	1.573	0.244	0	0	0	0	0	0.735	0.014	0.351	0	CRC
0.181	0	0.031	0.132	0	0.152	0.068	0.008	0.029	0.006	0.115	0.018	1.701	0.018	0.611	0	CRC
0.092	0	0	0.064	0	0	0.040	0	0	0	0.045	0.014	0	0.011	0.257	0.004	CRC
0.095	0	0	0.083	0	0.159	0	0.010	0	0.004	0.105	0.006	1.351	0.016	0.236	0	CRC
0.277	0	0.030	0.049	0	0	0.050	0	0	0	0.070	0	0.875	0.008	0.175	0.003	CRC
0.156	0	0	0	0	0.142	0.170	0	0.039	0	0.142	0.012	2.087	0.019	0.353	0	CRC
0.188	0	0.043	0	0	0	0.160	0.018	0.045	0.007	0.188	0.016	3.213	0.036	0.423	0.011	CRC
0.282	0	0.027	0	0	0	0.100	0.011	0.022	0	0.151	0	1.549	0.022	0.278	0.004	CRC
0.155	0.010	0.147	0	0	0	0.051	0.006	0.017	0	0.054	0	1.700	0.009	0.368	0.005	CRC
0.144	0	0.050	0	0.003	0	0.075	0	0	0.006	0	0	0	0.010	0.170	0	CRC
0.077	0.012	0	0	0	0	0.083	0.007	0	0	0.066	0	1.130	0	0.139	0.005	CRC
0.202	0.016	0	0.085	0	0.068	0.106	0	0	0	0.131	0.012	0	0.026	0.289	0	CRC
0	0	0	0.017	0	0.047	0.046	0	0	0	0.077	0	0	0	0	0	CRC
0.178	0	0.011	0.021	0	0.053	0.051	0.009	0.007	0	0	0.017	1.622	0	0.300	0	CRC

Figura 6 - Exemplo parcial de perfis de acilcarnitina

Os dados apresentados na Figura 6 dos perfis de acilcarnitinas são constituídos por 36 indicadores, dos quais, 33 estão relacionados com a resposta de acilcarnitinas e são medidos em micromoles/Litro (µmol/L) e os restantes dizem respeito ao sexo, idade do paciente e do diagnostico. Na Tabela 8 é possível verificar os indicadores existentes e as respetivas abreviações.

Tabela 8 - Indicadores existentes nos dados de acilcarnitinas

Nome Indicador	Abreviação
Free Carnitine	CARNITINE
Acetylcarnitine	C2
Propionylcarnitine	C3
Malonylcarnitine	C3DC
Butyrylcarnitine Isobutyrylcarnitine	C4
3-Hydroxybutyrylcarnitine/3- Hydroxyisobutyrylcarnitine	C4OH
Methylmalonylcarnitine/Succinylcarnitine	C4DC
Tiglylcarnitine/3-Methylcrotonylcarnitine	C5
Glutarylcarnitine	C5DC
Tiglylcarnitine/3-Methylcrotonylcarnitine	C5:1
3-Hydroxyisovalerylcarnitine/3-Hydroxy-2- methylbutyrylcarnitine	C5OH
Hexanoylcarnitine	C6
3-Methylglutarylcarnitine	C6DC
Octanoylcarnitine	C8
Octenoycarnitine	C8:1
Decanoylcarnitine	C10
Decenoylcarnitine	C10:1
Decadienoylcarnitine	C10:2
Dodecanoylcarnitine	C12
Dodecenoylcarnitine	C12:1
Tetradecanoylcarnitine	C14
Tetradecenoylcarnitine	C14:1
3-Hydroxytetradecenoylcarnitine	C14:1OH
Tetradecadienoylcarnitine	C14:2
3-Hydroxytetradecenoylcarnitine	C14OH
Hexadecanoylcarnitine	C16
Hexadecenoylcarnitine	C16:1
3-Hydroxyhexadecanoylcarnitine	C16OH
Octadecanoylcarnitine	C18
Octadecenoylcarnitine	C18:1
3-Hydroxyoctadecenoylcarnitine	C18:1OH
Octadecadienoylcarnitine	C18:2
3-Hydroxyoctadecanoylcarnitine	C18OH
Sexo do Paciente	SEX
Idade quando for feita a colheita	AGE
Diagnóstico	DIAG_ID

4.2 Desenho

Neste subcapítulo será realizada uma análise do desenho da proposta de solução e serão apresentados diagramas de componentes, segundo a anotação UML, com o intuito de demonstrar a arquitetura da solução a ser desenvolvida.

4.2.1 Diagrama de casos de uso

De acordo com o referido anteriormente, é necessário tratar e validar os dados fornecidos pelos utilizadores na solução, para criar um conjunto de dados prontos a ser consumidos pelos algoritmos de ML. Assim sendo, a figura seguinte apresenta os diferentes casos de uso considerados durante o desenvolvimento da aplicação.

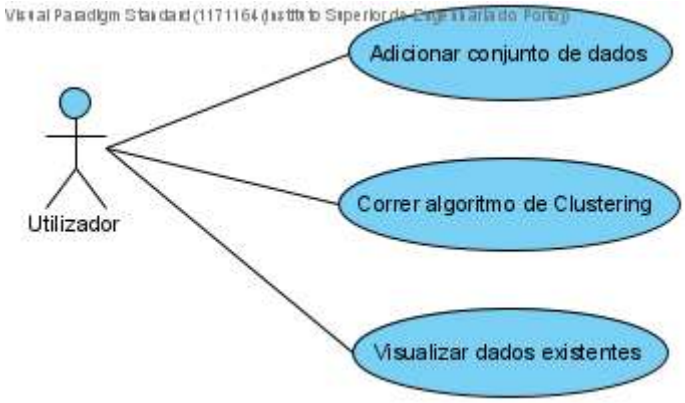


Figura 7 - Diagrama de casos de uso

4.2.2 Proposta da arquitetura do sistema

Na medida em que, o objetivo desta dissertação apenas se foca no estudo dos algoritmos de aprendizagem não supervisionada para agrupar e identificar anomalias existentes nos perfis metabólicos de pacientes com cancro do reto localmente avançado, é, de seguida, apresentada na Figura 8 uma proposta de arquitetura de um sistema a desenvolver para a próxima etapa deste projeto que permita incorporar os melhores algoritmos identificados neste trabalho e facilitar o trabalho dos profissionais de saúde do Centro Hospital.

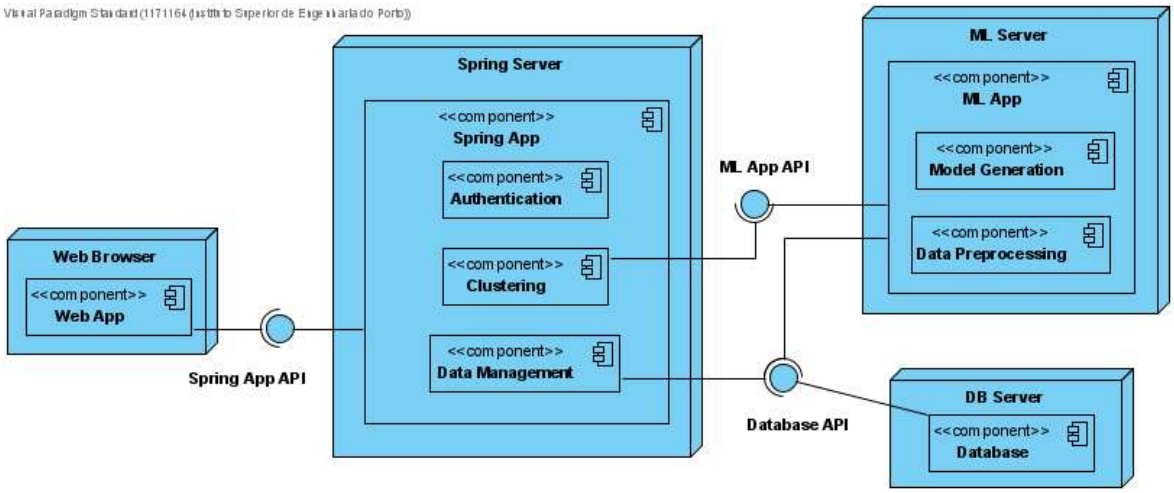


Figura 8 - Proposta de arquitetura de sistema

A parte principal da proposta de arquitetura apresentada é “ML Server”, que será o servidor que vai conter os diferentes componentes relacionados com tarefas de ML. Este contém componentes de geração de modelos, como pré-processamento e tratamento dos dados. Dado que, o servidor se encontra isolado e desacoplado dos restantes servidores e componentes, possibilita no futuro a comunicações externas com o serviço através de uma API, como se encontra representado na figura como “ML App API”. É da responsabilidade deste serviço receber pedidos externos de outros sistemas, carregar dados, tratar os dados, processá-los e gerar os modelos de ML.

Os utilizadores do sistema representado, profissionais de saúde, conseguem aceder através de uma aplicação web que comunica com um servidor intermediário desenvolvido em Spring. Este é responsável por tratar diferentes pedidos: autenticação, gestão de dados, aplicação de algoritmos de ML. No caso de algoritmos de ML é feita comunicação ao “ML App API”.

O servidor e o componente de base de dados são utilizados pelo sistema para guardar e armazenar os vários dados dos utilizadores, por exemplo, credenciais e conjunto de dados utilizados pelos algoritmos de ML.

5 Experimentação e Avaliação

Atualmente, existem bastantes pacientes com cancro do reto localmente avançado, sendo que uns respondem a tratamento e outros não. O intuito desta investigação prende-se em identificar com base nos perfis metabólicos, se estes podem fornecer informação sobre os respetivos grupos em que se enquadrem.

Os dados de perfis metabólicos permitem agrupar por grupos os vários pacientes, de acordo com a resposta patológica, usando técnicas de ML. O principal objetivo consiste em provar, através do uso, destas técnicas se realmente é possível tirar informação sobre estes grupos, conclusões, ou até mesmo, identificar anomalias.

Nesta experimentação, os dados de perfis metabólicos serão tratados em função de deteção de possível ruído e inconsistências. Pretende-se equilibrar o conteúdo dos dados, de modo a não influenciar negativamente a qualidade destes, e aplicar métodos de redução de dimensionalidade, os quais serão utilizados por diferentes algoritmos não supervisionados.

Para cada um dos testes realizados, é efetuada a avaliação dos resultados obtidos dos grupos, de acordo com medidas intrínsecas, em que são avaliados internamente, com base nos níveis de coesão e separação, avaliando-se assim a semelhança entre dados pertencentes a um mesmo grupo com os restantes grupos.

Uma vez que este trabalho se debruça na aprendizagem não-supervisionada e não existem dados identificados, não é possível efetuar validações através de informação externa, sendo a única possibilidade de validação externa efetuada pelos profissionais de saúde.

Por fim, mal sejam identificados os melhores algoritmos, estes serão otimizados de acordo com os seus *hyperparameters* (parâmetros especificados pelo utilizador antes da aprendizagem).

5.1 Pré-processamento

No estado atual, é complicado extrair informação importante de grandes volumes de dados. Nesta secção será abordada a terceira etapa do modelo de CRISP, modelo previamente referido.

O pré-processamento de dados é um conjunto de atividades que corresponde à preparação, organização, tratamento e estruturação de dados, para estes serem consumidos pelos algoritmos de ML. É uma etapa fundamental que vai permitir tornar a informação obtida dos algoritmos, confiável e rentável.

Com base nos artigos de Pedro Gomes em [39] e Rajaratne em [40] consideraram-se as seguintes técnicas de pré-processamento de dados:

- Tratamento/limpeza de dados
- Transformação de dados
- Redução de dimensões

5.1.1 Tratamento e limpeza de dados

Este processo é utilizado para melhorar a qualidade dos dados. É efetuada a remoção de ruído, dados incompletos e inconsistências (no caso de existirem valores que não vão de encontro com os esperados). Através deste processo de limpeza, os dados tornam-se mais confiáveis, visto que, já não existem dados fora do esperado, que poderiam ter um impacto negativo no processo de aprendizagem do algoritmo [41].

Na medida em que, os dados de acilcarnitinas se encontravam divididos em múltiplos ficheiros, o processo de agrupamento dos mesmos em um único ficheiro, permitiu efetuar uma limpeza dos dados, remover valores fora do padrão esperado, pois estavam incompletos, e remover qualquer conjunto de valores de pacientes que estivesse repetido.

De acordo com mencionado na análise de dados da secção anterior, das 36 características disponíveis em cada um dos dados de acilcarnitinas, 33 correspondem a concentrações de carnitinas. Nos dados de aminoácidos 28 das 32 características disponíveis são concentrações de aminoácidos sendo que as restantes correspondem à idade do paciente, sexo e o respetivo diagnóstico. Apenas as concentrações de acilcarnitinas e aminoácidos vão ser estudadas e utilizadas no processo de aprendizagem dos algoritmos de ML.

Durante o pré-processamento dos dados, os valores das diferentes concentrações dos perfis de aminoácidos e acilcarnitinas mantiveram-se, uma vez que, as medidas já se encontravam em micromoles/Litro. Para a idade do paciente apenas foi considerada a idade do paciente na altura da colheita de sangue (por exemplo no caso de 46 anos e 7 meses apenas se considerou a idade de 46 anos). O sexo do paciente foi convertido em valores numéricos e binários, sendo atribuído “0” ao sexo masculino e “1” ao sexo feminino.

5.1.2 Transformação de dados

O processo de transformação de dados deste trabalho diz respeito à mudança de valores dos dados de perfis metabólicos de pacientes.

Os dados fornecidos pelo CHUPorto são apresentados com diversas características/diferentes concentrações. As diferentes características dos dados com escalas maiores podem dominar as restantes, uma vez que podem influenciar negativamente os algoritmos de aprendizagem, induzindo a erros durante a aprendizagem e criação dos modelos. Igualmente, valores extremos de certos dados (*outliers*) que se apresentam como anomalias, podem ofuscar certos pontos importantes, influenciando assim negativamente os restantes métodos de transformação.

Com este problema em mente serão aplicados quatro tipo de transformações de dados:

- **Normalização e Uniformização**

Para transformação de dados, dois processos mais utilizados, são: Normalização (*MinMax*) e Uniformização (*Standardization*).

A Normalização é um dos métodos mais populares para preparar os dados, uma vez que, permite transformar os resultados dos diferentes indicadores para a mesma escala. Este é um método de escalabilidade que utiliza os valores mínimos e máximos, por isso, esta designação, serve para fazer escalonamentos. A escala de valores de normalização compreende o intervalo é entre [0,1] e/ou [-1,1], permitindo aos algoritmos de ML considerar todos os atributos como iguais, dado que se encontram na mesma escala.

A segunda técnica de transformação de dados considerada bastante similar à anterior é a *Standardization*, também referido como *Z-score Normalization*. Este processo difere da Normalização, na medida em que, utiliza a média e os desvios padrões para fazer o escalonamento dos dados. A escala de valores varia de acordo com os resultados obtidos. O principal defeito deste método é o facto de normalizar o desvio padrão/variância das características de dados para o mesmo valor, não permitindo mais tarde aplicar certas técnicas de redução de características.

- **Escala Robusta**

O método de transformação Robusta permite escalonar o conjunto de dados das diversas características de dados, de modo que estes se tornem robustos aos valores extremos existentes. Este método consiste em remover a mediana (considerar como 0) e escalonar os dados entre o primeiro e terceiro quartil. Apesar de ser parecido com a Uniformização, Escala Robusta (*Robust Scaling)* utiliza a mediana e a distância interquartil ao invés da média e do desvio padrão para escalonar os dados. A utilização da distância interquartil, permite a este método de transformação, tornar-se imune ao ruído existente, principalmente porque os valores extremos não afetam a mediana [42].

❖ **Escala de Máximo Absoluto**

No método de transformação de Máximo Absoluto, cada valor numérico de uma coluna é dividido pelo valor maior existente dessa característica. Os resultados obtidos de cada umas das transformações efetuadas estarão compreendidas numa escala de -1 a 1.

5.1.3 Redução de dimensões

Em ML, as diferentes características de uma instância (amostra da análise de um paciente) podem corresponder a múltiplos atributos que o caracterizam e definem. Cada característica representa uma dimensão, sendo que um grupo de dimensões permite caracterizar os vários dados.

Quando se quer aumentar a dimensionalidade, pretende-se aumentar o número de características/indicadores utilizados para descrever o conjunto de dados. À medida que a dimensionalidade aumenta, o esforço necessário para alcançar resultados rentáveis, por parte de qualquer algoritmo de ML, cresce exponencialmente, dado que a quantidade de dados é maior, o que poderá induzir a prevalência de ruído, de dados irrelevantes e redundantes, aumentando, posteriormente, o erro de aprendizagem por parte do algoritmo, segundo [43].

O processo de agrupar os dados depende essencialmente de identificar conjuntos de dados similares, que possibilitem criar grupos. Contudo, quando existe uma elevada dimensionalidade de dados, todos as instâncias parecem dispersas de múltiplas formas.

De acordo com os autores em [44], o objetivo principal de redução de dimensões em aprendizagem não-supervisionada, consiste na seleção de conjuntos de indicadores que permitam explicar melhor o conjunto de dados, escolhendo apenas os indicadores que são mais relevantes e menos redundantes. Esta técnica permite identificar de “forma natural” os vários grupos formados nos dados, tal como, aumentar o desempenho e reduzir o tempo de execução por parte dos algoritmos.

Uma das vantagens quando utilizado um baixo nível de dimensionalidade é conseguir visualizar os dados, algo impossível quando existem mais de dois ou três indicadores. Neste trabalho, visto que não é possível visualizar os dados, devido ao elevado número de características para os diferentes perfis metabólicos, considerou-se prioritário não perder parte da informação, reduzindo assim a dimensão. Segundo o artigo em [45], o ideal é escolher um número de componentes que permita explicar a variância dos dados de até 80% para evitar *overfitting* (os dados a analisar exatamente iguais aos dados reais).

❖ **Variance Thresholding Feature Selecion (VTFS)**

O conceito Variância é um termo estatístico que diz respeito à distância que um valor médio apresenta dos restantes valores de um conjunto de dados.

Quando um indicador/característica apresenta um alto valor de variância, significa que os valores contidos nele variam ou que têm alta cardinalidade, ou seja, que a coluna contém dados com valores totalmente exclusivos ou extremos. Em contrapartida, indicadores com baixa variância demonstram que os valores dessa coluna são similares ou até mesmo iguais [46].

De acordo com o mencionado anteriormente, quanto maior a dimensão dos dados, maior a informação que se obtém de cada amostra. No entanto, possuírem características a mais, não significa que fornecem informação extra, especialmente se tiverem um nível de variância baixo.

Por essa razão, foi considerada a técnica de *Variance Threshold Feature Selection (VTFS)* fornecida pelo Scikit-learn. Segundo Bex T. em [47], esta técnica é bastante útil para reduzir a dimensionalidade dos dados especialmente para modelação não supervisionada. Este método de redução é de rápida aplicação, não envolve muita complexidade, o que permite facilitar a redução de dimensionalidade. São removidas as colunas que apresentam baixo nível de variância e sem informação útil. *VTFS* simplesmente necessita que seja definido um valor de variância mínimo (*threshold*) e que as restantes colunas, que não apresentem variância superior, sejam excluídas.

Na medida em que, as variâncias são afetadas pela escala numérica, é necessário normalizar os dados, para que as variâncias tenham os mesmos significados e para não haver colunas de dados a influenciar os resultados.

❖ **Principal Components Analysis (PCA)**

O PCA é uma técnica de redução de dimensionalidade não supervisionada que permite construir novas características/variáveis, mais relevantes, através de combinações das características originais. De acordo com Loukas em [48], esta técnica é normalmente útil no processamento de dados, para tirar o ruído ou até mesmo para comprimir os dados. É utilizada em situações onde existe multicolinearidade (variáveis independentes altamente correlacionadas) entre os vários atributos dos dados, pois as dimensões dos mesmos são elevadas. O resultado de dados obtidos do PCA são essencialmente combinações lineares dos dados originais.

❖ **T-distributed Stochastic Neighbor Embedding (t-SNE)**

No final da experiência/investigação será utilizado este método de redução para melhor visualizar os resultados obtidos. No entanto, é de salientar que ao reduzir um grande número de características apenas para duas ou três dimensões, pode induzir o utilizador em erro, no momento de visualização dos dados.

O t-SNE é um algoritmo de redução de dimensionalidade não linear que encontra padrões nos dados baseados nas suas semelhanças entre as várias características e permite principalmente visualizar características de dimensão superior em planos de duas ou três dimensões. Segundo o artigo em [49], as semelhanças entre os vários indicadores dos dados são calculadas, tendo em conta as probabilidades condicionais de um ponto A escolher o ponto B como seu vizinho, tentando minimizar as diferenças entre as probabilidades num plano com dimensão maior ou menor para uma representação perfeita dos pontos num plano com baixa dimensionalidade.

Em suma, t-SNE tenta preservar posições relativas dos pontos num mapeamento de baixa dimensionalidade, mantendo pontos idênticos mais próximos e pontos diferentes mais afastados.

5.1.4 Experimentação

Nesta secção são apresentadas as diferentes métricas de avaliação utilizadas para avaliar os modelos, nas diferentes experiências efetuadas, com vista à redução dos dados à mesma escala e redução de dimensionalidade.

5.1.4.1 Técnicas de avaliação de clusters

Após a conclusão do tratamento de dados e dos modelos gerados com base nos mesmos, começa uma das etapas finais dos sistemas de ML, que corresponde à análise dos clusters. Nesta etapa, pretende-se identificar qual dos diferentes resultados de cluster obtidos, pelos diferentes processos e algoritmos, é considerado o melhor.

Primeiramente, as principais questões de avaliação de resultados dizem respeito a como identificar qual dos diferentes resultados é considerado o melhor, originando assim os seguintes índices de avaliação [50]:

- Comparar algoritmos de *clustering* entre si;
- Comparar dois conjuntos de clusters
- Comparar dois clusters e verificar qual dos dois é mais compacto/denso (os dados que os compõem não se encontram dispersos).
- Determinar estruturas criadas devido ao ruído nos dados.

A avaliação de resultados é das etapas mais cruciais para o processo de seleção dos modelos. Nesta secção será identificado qual dos algoritmos é o mais apropriado para a tarefa em questão, de agrupar os diferentes dados de pacientes com cancro do reta localmente avançado. Apenas foi possível alcançar os resultados mais precisos através de diversas tentativas-erro, de modo a treinar e aperfeiçoar os algoritmos com diferentes parâmetros.

Métricas de avaliação para Clustering

Em aprendizagem não-supervisionada, o resultado esperado, verdade absoluta (*grounth truth*), não é conhecida. Consequentemente, torna-se difícil identificar até que ponto os resultados dos algoritmos estão corretos. Contudo, existem critérios de avaliação recorrentes que permitem identificar se um algoritmo de *clustering* é considerado bom, por exemplo, se este tem baixo nível de variância dentro dos clusters (os dados dos mesmos clusters são iguais ou idênticos entre si) e se a variância entre clusters é elevada (dados de diferentes clusters são muito diferentes).

Atualmente existem dois tipos de métricas de avaliação para *clustering*, segundo [50], a saber:

- *Extrinsic Measures* – Estas medidas necessitam de verdade absoluta, resultado esperado, não sendo disponível nesta questão prática;
- *Intrinsic Measures* – Estas medidas não necessitam de verdade absoluta, sendo aplicável para todos os resultados de aprendizagem não supervisionada.

Neste trabalho, visto que não existe uma verdade absoluta sobre os possíveis resultados de grupos, através da combinação dos dados dos pacientes, serão utilizadas as medidas intrínsecas para a avaliação dos resultados. As medidas de validação de *cluster* encontram-se categorizadas em três classes [51], a saber:

- Validação Interna de Cluster – o resultado dos clusters é avaliado com base nos dados que os compõem, utilizando unicamente informação interna, sem recorrer a qualquer referência externa;
- Validação Externa de Cluster – os resultados são avaliados com base em informação externa. Existe uma verdade absoluta conhecida previamente e espera-se que os resultados a validem (*ground-truth*);
- Validação Relativa de Cluster – os resultados são avaliados através de tentativas de diferentes parâmetros dos algoritmos no processo de aprendizagem.

A avaliação vai centrar-se em métricas de validação interna, que permitam indicar e avaliar o quão bem os grupos foram definidos, sem ser necessário utilizar informação externa sobre resultados esperados. No final da experimentação, vai ser efetuada uma validação relativa de *Cluster,* no momento da otimização dos *hyperparameters* dos algoritmos.

De acordo com [52], para identificar se o agrupamento de dados foi efetuado corretamente, os *clusters* criados têm de ser de boa qualidade, nos quais, a semelhança *intra-cluster* (dentro do grupo) tem de ser elevada e, entre clusters, baixa. Apresentar um baixo valor de variância, dentro do cluster é um sinal de robustez. O principal objetivo é criar grupos densos, bem definidos, e que apresentem dados do mesmo grupo semelhantes e diferenciados dos restantes grupos.

As métricas de avaliação mais utilizadas, segundo [53], que não necessitam de verdade absoluta para calcular a eficiência dos algoritmos de *clustering*, serão a seguir identificadas.

❖ **Silhouette Coefficient (SC)**

O coeficiente de Silhouette, de acordo com [50], mede a distância média de uma amostra aos restantes existentes no *cluster* mais próximo (*inter-cluster*), ou seja, a distância de separação entre *clusters* (pode ser calculado através de Euclidean, Manhattan, Minkowski, Hamming). Na Figura 9 está demonstrada a fórmula do coeficiente.

$$\begin{aligned} \text{a} &= \text{Average distance between sample and all other points in same cluster} \\ \text{b} &= \text{Average distance between sample and all other points in next nearest cluster} \\ \text{Silhouette Coefficient } (s) &= \frac{b-a}{\max(a,b)} \end{aligned}$$

Figura 9 - Fórmula para coeficiente de Silhouette, obtido em [50]

Os valores deste coeficiente estão compreendidos entre -1 e 1. Quanto maior o valor, significa que é maior a separação entre clusters e, por essa razão, o *cluster* se encontra mais distante do *cluster* mais próximo, estando este denso, bem definido e separado dos dados vizinhos. Caso o resultado seja nulo, significa que há *clusters* vizinhos que estão demasiado próximos entre si. No caso de ser negativo, significa que há amostras nos *clusters* errados [54].

❖ **Calinski-Harabasz Index (CH)**

O Índice de Calinski-Harabasz, ou Critério de Razão de Variação, resulta da multiplicação entre a dispersão de *clusters* e a soma de dispersão dentro dos *clusters*, segundo Wong em [50].

$$\begin{aligned} k &= \text{Number of clusters} \\ n_q &= \text{Number of points in cluster } q \\ c_q &= \text{Cluster center of cluster } q \\ n_E &= \text{Number of data points} \\ c_E &= \text{Cluster center of all points} \\ \text{Between-cluster dispersion, } B &= \sum_{q \in k} n_q (c_q - c_E)(c_q - c_E)^T \\ \text{Within-cluster dispersion, } W &= \sum_{q \in k} \sum_{x \in \text{cluster } q} (x - c_q)(x - c_q)^T \\ \text{Calinski-Harabasz score } (s) &= \frac{B}{W} \times \frac{n_E - k}{k - 1} \end{aligned}$$

Figura 10 - Fórmula do Índice de Calinski-Harabasz, obtido em [50]

Esta métrica de avaliação é eficiente e rápida de processar, permitindo assim identificar bons *clusters*, que se encontrem bem definidos, densos e separados dos restantes. Quanto maior for o índice, quer dizer que o *cluster* se encontra afastado dos restantes e que se encontra bem definido.

❖ **Davies-Bouldin Index (DB)**

O Índice de Davis-Bouldin é definido como a média de semelhança medida de cada *cluster*, com os restantes semelhantes. A semelhança corresponde ao rácio entre a distância dentro do *cluster* e a distância entre *clusters*, sendo o valor mínimo de 0.

$$
\begin{aligned}
k &= \text{Number of clusters} \\
s_i &= \text{Average distance between each point in cluster } i \text{ to cluster center } c_i \\
d_{ij} &= \text{Distance between cluster centers } c_i \text{ and } c_j \\
\text{Difference measure, } R_{ij} &= \frac{\text{Average within-cluster distance}}{\text{Between-cluster distance}} \\
&= \frac{s_i + s_j}{d_{ij}} \\
\text{Davies-Bouldin Index } (DB) &= \text{Average of maximum difference measure} \\
&= \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} R_{ij}
\end{aligned}
$$

Figura 11 - Fórmula de Índice de Davies-Bouldin, obtido em [50]

Resumindo, *clusters* que se encontrem mais afastados dos restantes e menos dispersos, apresentam melhores resultados, sendo que, valores menores correspondem a *clusters* bem definidos. Visto que, o resultado do *cluster* é menor, comparado com a distância para os restantes, considera-se como bem definido.

5.2 Avaliação

Nesta secção, serão apresentados os resultados obtidos das experiências efetuadas. As experiências foram realizadas, utilizando um processador Intel core i5-9600K com 16GB RAM, num Windows 10 Education 64 bits.

Com o intuito de tornar os resultados obtidos de cada algoritmo confiáveis, os valores demonstrados das avaliações, nas tabelas seguintes, correspondem à média das diferentes experiências, realizadas em múltiplos testes com diferentes sementes (*seeds)*. Alguns algoritmos permitem que se defina o valor da semente para replicar os resultados obtidos. As diferentes sementes utilizadas, permitiram replicar os resultados e ter em consideração diferentes pontos de início, utilizados pelos vários algoritmos que originaram diferentes resultados.

Nas primeiras experiências será demonstrado o quão efetivo são as técnicas de pré-processamento referente à normalização e redução de dimensionalidade dos dados nos diferentes algoritmos de ML utilizados. Seguidamente, aplicar-se-ão técnicas que permitem otimizar os resultados dos algoritmos, identificando os melhores parâmetros, tal como identificando situações que proporcionem melhores resultados. Por fim, os melhores modelos serão analisados e visualizados para retirar as respetivas conclusões.

5.2.1 Algoritmos de aprendizagem utilizados

Nesta subsecção, serão apresentados os vários algoritmos utilizados para efetuar as diferentes experiências nos dados de perfis metabólicos de acilcarnitinas e aminoácidos.

- ***K-means Clustering***

O algoritmo de agrupamento K-means, de acordo com [55], é um dos algoritmos de aprendizagem mais simples e populares. O objetivo deste algoritmo é simples, agrupar os pontos de dados que sejam semelhantes entre si e descobrir padrões subjacentes. Começa-se por definir um número destino de *clusters* no conjunto de dados, introduzido pelo utilizador como parâmetro. O algoritmo começa por identificar um número **k** de centroides, estes são escolhidos aleatoriamente, o que significa que pode levar a erro, por parte do algoritmo, sem que se encontrem os centroides mais apropriados. Por fim, o algoritmo aloca todos os pontos para o *cluster* mais próximo.

- ***Agglomerative Hierarquical Clustering***

O algoritmo *Agglomerative Clustering* é um dos tipos mais comuns de agrupamento hierárquico, utilizado para agrupar objetos em *clusters* com base nas suas semelhanças. Este algoritmo de agrupamento, começa com grupos de dados com um único elemento (folha). A cada iteração que passa, *clusters* mais semelhantes são combinados em um novo *cluster* maior (nós). Este procedimento é iterativo até que todos os pontos sejam membros de um único grande *cluster* ou o número pretendido de *clusters* seja alcançado, chama-se a este tipo de agrupamento "*bottom-up*", segundo os autores em [56].

O agrupamento hierárquico aglomerativo é baseado em conectividade. Os pontos de dados mais próximos são agrupados, em função da distância ou semelhanças destes, supondo que pontos de dados mais próximos são mais semelhantes ou estão mais relacionados entre si do que com os pontos mais distantes.

- ***Birch Clustering***

Normalmente os algoritmos de agrupamento utilizados não escalonam bem relativamente aos tempos de execução em função do tamanho do conjunto de dados. Birch (*Balanced Iterative Reducing and Clustering using Hierarchies*) é um algoritmo de agrupamento que pode agrupar grandes conjuntos de dados, pois, de acordo com [57], começa por gerar um resumo mais compacto, que retém o máximo possível de informação de distribuição, agrupando o resumo de dados em vez do conjunto original. Esta técnica de resumir o conjunto de dados original, é a que o permite diferenciar dos restantes algoritmos, pois, consegue, à medida que o tamanho do conjunto de dados aumenta, produzir bons resultados de execução e qualidade.

Normalmente este algoritmo é utilizado para complementar outros, mas nesta experimentação vai ser utilizado sozinho.

A principal desvantagem deste algoritmo de agrupamento é o facto de só conseguir lidar com atributos métricos (valores que podem ser representados por coordenadas explicitas em um espaço euclidiano) [58].

- ***Gaussian Mixture Model Clustering***

Este algoritmo de agrupamento é muitas vezes comparado com o K-means. No entanto, há bastantes vantagens em usar modelos de mistura Gaussian. O algoritmo GMM tem em consideração a variância dos dados durante o agrupamento e consegue lidar com *clusters* que possuem formas peculiares que, normalmente, são identificadas, por outros algoritmos, como mais que um *cluster*.

O algoritmo GMM agrupa os pontos de dados pertencentes a uma única distribuição. É, um modelo probabilístico e que usa abordagem de agrupamento suave para distribuir os pontos em diferentes grupos, fornecendo informação sobre as probabilidades de um determinado ponto de dados pertencer a cada um dos *clusters* possíveis. OS GMMs assumem que existe um certo número de distribuições Gaussianas e que cada uma destas distribuições representam um *cluster* [59].

- ***Spectral Clustering***

No algoritmo de agrupamento Spectral, segundo o autor em [60], os pontos de dados são tratados como nós de um gráfico, sendo os nós mapeados para um espaço de baixa dimensão, onde é facilitada a formação de *clusters*. O *Spectral Clustering* é uma técnica de análise de dados exploratória, que reduz a complexidade da multidimensionalidade do conjunto de dados, em grupo de dados que apresentem semelhanças. O principal objetivo é agrupar todo o espectro de pontos de dados desorganizados em vários grupos, com base nas suas características, e que estão conectados ou próximos uns dos outros no gráfico.

- ***MeanShift***

O último algoritmo de agrupamento de dados utilizado nesta experimentação é o MeanShift. Este algoritmo é utilizado na análise de dados do mundo real em diversas situações, uma vez que não é muito paramétrico e não requer nenhuma forma predefinida dos *clusters*. O MeanShift consiste em mudar (*Shift*) para a média (*Mean*) de forma iterativa. Cada ponto de dados é alterado para a média das características nas suas áreas, passo a passo, e a localização do destino final de cada ponto vai representar o *cluster* a que pertence, de acordo com Yufeng em [61].

Este algoritmo de deslocamento de pontos, procura mover os pontos até estes encontrarem o seu grupo ao longo das iterações, em que são movidos em direção à média ponderada nas suas respetivas áreas locais. O destino de cada ponto será o centroide do *cluster* de dados ao qual o ponto pertence. Todos os pontos com o mesmo ponto de destino podem ser identificados para o mesmo *cluster*. De todos os algoritmos anteriormente referidos, este é o único que não necessita de parâmetros de entrada, por exemplo, o número de *clusters* a procurar.

5.2.2 Método de seleção de resultados

Em seguida, será apresentada a avaliação de diferentes experiências efetuadas para cada um dos algoritmos anteriormente mencionados. As experiências realizaram-se em função de diferentes tipos de tratamento de dados e redução de dimensionalidade. Antes de se efetuar

as experiências é necessário apresentar as decisões que foram tomadas durante a decisão dos melhores resultados apresentados.

Qualquer algoritmo utilizado nas diferentes experiências, quer seja no tratamento de dados ou na redução de dimensionalidade, em cada um dos dados de perfis metabólicos, corre múltiplas vezes e a média dos resultados obtidos de todas das avaliações é apresentada. Visto que, é necessário parametrizar os algoritmos sobre o número de *clusters* que se pretende encontrar, são efetuadas várias experiências. Depois de uma análise dos dados e experiências prévias, estimou-se que o melhor número de *clusters*, para as acilcarnitinas, encontra-se entre 4 e 15 *clusters* e para os aminoácidos de 4 a 20.

Quanto menor for o número de clusters, maior será a separação destes, logo maior será o valor de Silhouette Coefficient. Contudo, os elementos que os constituem não são idênticos, prejudicando os valores de Davies-Bouldin. Contrariamente, um número elevado de *clusters* significa que poderão existir *clusters* com um único elemento e até mesmo pontos semelhantes em diferentes grupos. É necessário identificar, para cada uma das experiências, o número ideal de *clusters* que:

- Não desvalorize os valores de SC, tornando-os demasiado próximos de 0, significando que há clusters vizinhos demasiado próximos entre si;
- Identifique valores de DB baixos o suficiente para não existir semelhança entre os elementos de diferentes clusters;
- Um valor alto de CH, que demonstre que os elementos de um grupo não estão dispersos, mas se encontram dispersos de elementos pertencentes a outros grupos.

Na Tabela 9 apresenta-se um exemplo com diferentes resultados obtidos experimentando o algoritmo BIRCH para os dados de acilcarnitinas, com tratamento de Escala Robusta, com redução de dimensionalidade de PCA até 5 componentes, para demonstrar como é identificado a melhor avaliação para os diferentes números de *clusters*.

Tabela 9 - Exemplo de escolha de resultados durante experimentação

Número Clusters	Silhouette Coef. (SC)	Calinski-Harabasz Index (CH)	Davies-Bouldin Index (DB)
4	0.508	180.943	0.938
5	0.510	169.699	0.786
6	**0.534**	**174.585**	**0.766**
7	0.497	167.166	0.764
8	0.491	162.476	0.642
9	0.373	163.136	0.675
10	0.369	168.099	0.734
11	0.355	178.019	0.768
12	0.345	185.838	0.882
13	0.344	183.012	0.836
14	0.335	187.322	0.808

De acordo com a tabela anterior, é possível tirar as seguintes conclusões:

- O valor de SC diminui à medida que se aumenta o número de *clusters*. Logo, há grupos que ficam cada vez mais próximos entre si. O melhor valor de Silhouette foi obtido com 6 *clusters*;
- Os valores de DB, por instantes, desceram à medida que se aumentavam os *clusters*, sendo o que se pretende. Contudo, no final acabaram por subir. O melhor valor foi alcançado com 8 clusters.
- Ao contrário das restantes avaliações, o valor de CH aumentou à medida que o número de *clusters* foi aumentando. Existe, assim, uma menor dispersão dentro dos grupos.

Nestas experiências efetuadas, apesar de um elevado número de *clusters* produzir valores elevados de CH, os restantes valores de SC eram baixos e os de DB altos, sendo exatamente o oposto do que se pretendia.

Os principais resultados considerados os melhores do exemplo foram de 6 e 8 grupos, dado que apresentam bons valores de SC e DB, assim como valores altos de CH. Posto isto, considerou-se como melhor resultado a apresentar, para esta situação, o de 6 grupos, na medida em que, possui o melhor valor de SC e, em comparação com o de 8, um valor de CH mais alto. Nesta experiência, com 6 *clusters* foi obtido o melhor valor do coeficiente de Silhouette, com 0.534, com baixo índice de Davies-Bouldin de 0.766, estando este próximo do melhor valor (0.642).

5.2.3 Experimentação e avaliação da normalização

A primeiro experiência apresentada diz respeito à transformação das várias características dos dados de perfis metabólicos para a mesma escala. É necessário ter em consideração que, a existência de dados que, apesar das características se encontrarem nas mesmas medidas (µmol/L), podem influenciar os resultados dos algoritmos e ter um impacto negativo nos resultados devido ao facto de se encontrarem em escalas diferentes.

Normalmente em ML os algoritmos consideram que os dados estão escalonados e balanceados, sendo que consideram que as características apresentam uma distribuição padrão normal. Se uma característica tem uma variância muito maior que as restantes, sem ser normalizada, esta pode induzir a erro no momento da aprendizagem do algoritmo, fazendo com que este acredite que é a característica mais importante de todas, influenciando assim os resultados de forma negativa. Deste modo, uma vez que os limites de certos indicadores podem ser maiores do que outros, é necessário aplicar o tratamento de dados, para todos os indicadores terem a mesma importância durante o treino dos modelos dos algoritmos.

De seguida serão apresentados os diferentes resultados, para cada uma das experiências referentes à normalização dos dados de acilcarnitinas e aminoácidos. O "Grupo Controlo" apresenta os resultados de cada algoritmo, sem aplicação de qualquer tipo de pré processamento de dados, sem normalização. Os resultados obtidos para cada uma das experiências, dizem respeito à melhor avaliação dos diferentes numero de *clusters*, o tempo de

execução (ET) e avaliação do melhor número de grupos de acordo com coeficiente de Silhouette (SC), o índice de Calinski-Harabasz (CH) e o índice de Davies-Bouldin(DB).

5.2.3.1 Acilcarnitinas

Na Tabela 10 são apresentadas as quantidades ótimas de *clusters,* que apresentem os melhores resultados das avaliações efetuadas, para cada algoritmo, quando aplicados aos dados de acilcarnitinas, sem qualquer tipo de pré-processamento de dados (Controlo), com Uniformização (*Standard*), Normalização (*MinMax*), Escala Robusta (*Robust*) e Escala de Máximo Absoluto (*MaxAbs*).

Tabela 10 - Resultados da avaliação da experiência de múltiplos algoritmos com tratamentos de dados de acilcarnitinas

	CONTROLO	STANDARD	MINMAX	ROBUST	MAXABS
	Clusters: 5	Clusters: 4	Clusters: 4	**Clusters: 4**	Clusters: 4
	ET: 0.049s	ET: 0.044s	ET: 0.044s	**ET: 0.041s**	ET: 0.047s
KMEANS	**SC: 0.347**	SC: 0.211	SC: 0.224	**SC: 0.552**	SC: 0.224
	CH: 169.287	CH: 34.35	CH: 58.903	**CH: 127.62**	CH: 59.237
	DB: 0.762	DB: 1.823	DB: 2.043	**DB: 0.644**	DB: 2.054
	Clusters: 8	Clusters: 5	Clusters: 5	**Clusters: 4**	Clusters: 4
	ET: 0.002 s	ET: 0.003 s	ET: 0.003 s	**ET: 0.003 s**	ET: 0.003 s
AGGLOMERATIVE	**SC: 0.317**	SC: 0.193	SC: 0.178	**SC: 0.536**	SC: 0.297
	CH: 150.454	CH: 27.874	CH: 46.285	**CH: 116.897**	CH: 53.677
	DB: 0.806	DB: 1.568	DB: 1.972	**DB: 0.55**	DB: 1.398
	Clusters: 5	Clusters: 5	Clusters: 5	**Clusters: 4**	Clusters: 6
	ET: 0.010 s	ET: 0.014 s	ET: 0.006 s	**ET: 0.012 s**	ET: 0.006 s
BIRCH	**SC: 0.269**	SC: 0.195	SC: 0.245	**SC: 0.536**	SC: 0.3
	CH: 127.643	CH: 28.074	CH: 43.396	**CH: 116.897**	CH: 37.304
	DB: 0.98	DB: 1.558	DB: 1.599	**DB: 0.55**	DB: 1.424
	Clusters: 7	Clusters: 4	Clusters: 4	Clusters: 4	Clusters: 4
	ET: 0.021 s	ET: 0.016 s	ET: 0.019 s	ET: 0.013 s	ET: 0.013 s
GMM	**SC: 0.298**	SC: 0.169	SC: 0.212	SC: 0.379	SC: 0.215
	CH: 162.219	CH: 30.837	CH: 54.581	CH: 105.36	CH: 56.031
	DB: 0.873	DB: 2.092	DB: 2.029	DB: 1.155	DB: 2.021
	Clusters: 5	Clusters: 4	Clusters: 4	Clusters: 5	Clusters: 4
	ET: 0.031 s	ET: 0.029 s	ET: 0.029 s	ET: 0.028 s	ET: 0.028 s
SPECTRAL	**SC: 0.286**	SC: 0.084	SC: 0.119	SC: 0.181	SC: 0.129
	CH: 143.014	CH: 24.552	CH: 50.9	CH: 57.153	CH: 51.97
	DB: 1.004	DB: 2.472	DB: 2.01	DB: 1.62	DB: 1.987
	Clusters: 4	Clusters: 20	Clusters: 14	**Clusters: 18**	Clusters: 15
	ET: 2.491 s	ET: 1.788 s	ET: 1.755 s	**ET: 1.350**	ET: 1.737 s
MEANSHIFT	**SC: 0.338**	SC: 0.244	SC: 0.204	**SC: 0.414**	SC: 0.203
	CH: 51.733	CH: 10.027	CH: 9.652	**CH: 50.033**	CH: 9.459
	DB: 0.59	DB: 0.788	DB: 0.978	**DB: 0.518**	DB: 1.009

- **Discussão de resultados**

É possível verificar, na Tabela 10, em todas as experiências de controlo, que, muitas das vezes, os resultados em cada um dos algoritmos apresentam as melhores avaliações de clusters criados. Os valores de SC encontram-se perto de 0.3 ou às vezes acima. Isto é muito bom, mas também já seria de esperar, uma vez que, os valores dos dados de controlo não estão todos na mesma escala.

Existem indicadores que apresentam valores de variância muito maiores, sendo estes considerados pelos algoritmos como mais importantes em relação aos restantes, facilitando assim o processo de criação de grupos apenas com base nessas características.

É possível verificar que, assim que os dados são normalizados para as restantes experiências, as avaliações dos clusters baixam bastante a nível de SC e de CH e aumentam para DB. Isto ocorre porque se torna complicado para o algoritmo encontrar grupos que se destaquem entre tantos dados similares. Existe, assim semelhança entre grupos ou até mesmo grupos vizinhos demasiado próximos. Utilizando a transformação dos dados, para que todos fiquem à mesma escala, torna-se complicado, para o algoritmo, identificar quais as características mais importantes em grande dimensão de dados.

No caso das experiências com escala robusta, é possível verificar que, com a utilização dos algoritmos de K-means, de Agglomerative, de Birch e de Meanshift Clustering, os resultados superavam as avaliações de controlo, pois esta escala procura resistir aos valores extremos que possam existir nos dados. Esta abordagem permite escalonar os dados, mas continua a ser possível que valores demasiado extremos, em certas características, possam ter influenciado o algoritmo na sua decisão. Apesar de a escala robusta ser imune aos *outliers*, tornou-se um método de transformação de dados importante, uma vez que esta dissertação procura identificar anomalias que possam existir nos dados.

5.2.3.2 Aminoácidos

Na Tabela 11 são apresentadas as quantidades ótimas de *clusters,* que apresentem os melhores resultados das avaliações efetuados, para cada algoritmo, quando aplicados aos dados de aminoácidos sem qualquer tipo de pré-processamento de dados (Controlo), com Uniformização (*Standard*), Normalização (*MinMax*), Escala Robusta (*Robust*) e Escala de Máximo Absoluto (*MaxAbs*).

Tabela 11 - Resultados da avaliação da experiência de múltiplos tratamento de dados de aminoácidos

	CONTROLO	STANDARD	MINMAX	ROBUST	MAXABS
	Clusters: 5	Clusters: 4	Clusters: 5	Clusters: 6	Clusters: 4
	ET: 0.046s	ET: 0.038s	ET: 0.041s	ET: 0.046s	ET: 0.039s
KMEANS	**SC: 0.161**	SC: 0.09	SC: 0.096	SC: 0.093	SC: 0.1
	CH: 51.252	CH: 25.539	CH: 23.176	CH: 20.214	CH: 24.584
	DB: 1.435	DB: 2.479	DB: 2.326	DB: 2.186	DB: 2.346
	Clusters: 4	Clusters: 13	Clusters: 4	Clusters: 7	Clusters: 7
	ET: 0.002s	ET: 0.002s	ET: 0.002s	ET: 0.002s	ET: 0.002s
AGGLOMERATIVE	**SC: 0.176**	SC: 0.071	SC: 0.092	SC: 0.082	SC: 0.087
	CH: 50.884	CH: 12.344	CH: 23.368	CH: 16.909	CH: 16.548
	DB: 1.393	DB: 2.015	DB: 2.465	DB: 2.263	DB: 2.129
	Clusters: 4	Clusters: 4	Clusters: 5	Clusters: 11	Clusters: 6
	ET: 0.009s	ET: 0.009s	ET: 0.007s	ET: 0.010s	ET: 0.006s
BIRCH	**SC: 0.176**	SC: 0.087	SC: 0.089	SC: 0.078	SC: 0.105
	CH: 50.884	CH: 20.762	CH: 17.456	CH: 13.873	CH: 14.454
	DB: 1.393	DB: 2.673	DB: 2.527	DB: 2.059	DB: 2.29
	Clusters: 4	Clusters: 4	Clusters: 4	Clusters: 4	Clusters: 4
	ET: 0.011s	ET: 0.011s	ET: 0.010s	ET: 0.011s	ET: 0.011s
GMM	**SC: 0.154**	SC: 0.086	SC: 0.091	SC: 0.09	SC: 0.086
	CH: 51.054	CH: 24.007	CH: 25.285	CH: 22.664	CH: 23.005
	DB: 1.62	DB: 2.601	DB: 2.478	DB: 2.471	DB: 2.475
	Clusters: 5	Clusters: 11	Clusters: 6	Clusters: 5	Clusters: 16
	ET: 0.024s	ET: 0.033s	ET: 0.025s	ET: 0.026s	ET: 0.047s
SPECTRAL	**SC: 0.134**	SC: 0.064	SC: 0.078	SC: 0.075	SC: 0.076
	CH: 41.533	CH: 13.054	CH: 18.902	CH: 19.747	CH: 11.487
	DB: 1.704	DB: 2.161	DB: 2.34	DB: 2.3	DB: 1.981
	Clusters: 4	Clusters: 9	Clusters: 4	**Clusters: 12**	Clusters: 6
	ET: 0.454s	ET: 1.173s	ET: 1.100s	**ET: 1.109s**	ET: 1.093s
MEANSHIFT	**SC: 0.283**	SC: 0.189	SC: 0.231	**SC: 0.186**	SC: 0.197
	CH: 13.597	CH: 6.195	CH: 5.183	**CH: 6.134**	CH: 6.834
	DB: 0.893	DB: 1.13	DB: 1	**DB: 0.913**	DB: 1.364

- **Discussão de resultados**

Contrariamente à experiência efetuada com os dados de acilcarnitinas, relativamente aos dados de aminoácidos, é evidente que os resultados obtidos das avaliações efetuadas aos *clusters* formados com os dados de controlo foram os melhores.

Para todos os algoritmos, os dados de controlo apresentaram os melhores valores de SC e de CH, e os valores mais baixos de DB. O algoritmo MeanShift conseguiu encontrar bons resultados para a escala robusta, apesar de não serem os ideais, uma vez que os valores de SC e de CH são

relativamente baixos em comparação com os de controlo. É possível concluir que os dados de acilcarnitinas apresentam mais valores extremos que os dados de aminoácidos.

5.2.4 Avaliação da redução de dimensionalidade

Depois de ser demonstrado o impacto que a diferença de transformação de dados tem na aprendizagem dos algoritmos durante o agrupamento, foi analisada a redução de dimensionalidade, de modo a ter as melhores avaliações para os grupos formados.

A redução de dimensionalidade será efetuada através da utilização de PCA e VTFS. Estas técnicas apenas podem ser aplicadas a dados normalizados e, por essa razão, vão ser aplicadas a cada uma das técnicas de transformação anteriormente referidas e avaliadas. Para cada um dos dados de perfis metabólicos, começar-se-á por aplicar o PCA e, de seguida, a VTFS.

Durante o uso de PCA, vai ser utilizada a Taxa de Variância Explicada demonstrada por esta técnica, como métrica para avaliar a utilidade dos novos componentes principais formados, e conseguir identificar quantos se vão utilizar no modelo. Esta decisão é efetuada através da análise de gráficos. A taxa de variância explicada é a percentagem de variação atribuída a cada um dos componentes.

Neste tipo de técnica, o ideal é escolher um número de componentes que permita explicar um total de 80% dos dados para evitar *overfitting*. O PCA prioriza variáveis que têm variâncias mais altas do que baixas, por isso, é importante normalizar os dados na mesma escala para obter uma covariância razoável.

O método de VTFS somente exclui características que têm uma variância abaixo do valor determinado (*Threshold*). Por isso, durante a decisão deste valor priorizar-se-á manter as melhores características, com uma variância mais alta, definindo um valor que não diferencie da média e mediana dos indicadores.

5.2.4.1 Acilcarnitinas

A técnica PCA vai ser a primeira técnica de redução de dimensionalidade a ser estudada. De modo a identificar um número de componentes que permita explicar mais de 80% dos dados, sendo o ideal para evitar *overfitting*¸ segundo [45]. Seguidamente, são apresentados vários gráficos com o somatório das Taxas de Variancia Explicada de ordem crescente dos componentes, para cada uma das normalizações utilizadas, nos dados de acilcarnitinas. Cada figura apresenta, no primeiro gráfico o somatório das taxas de variação explicada de todos os componentes, logo o valor total é 100%, e no segundo gráfico até alcançar um somatório de 80% para o número mínimo de componentes.

- **PCA com Standard**

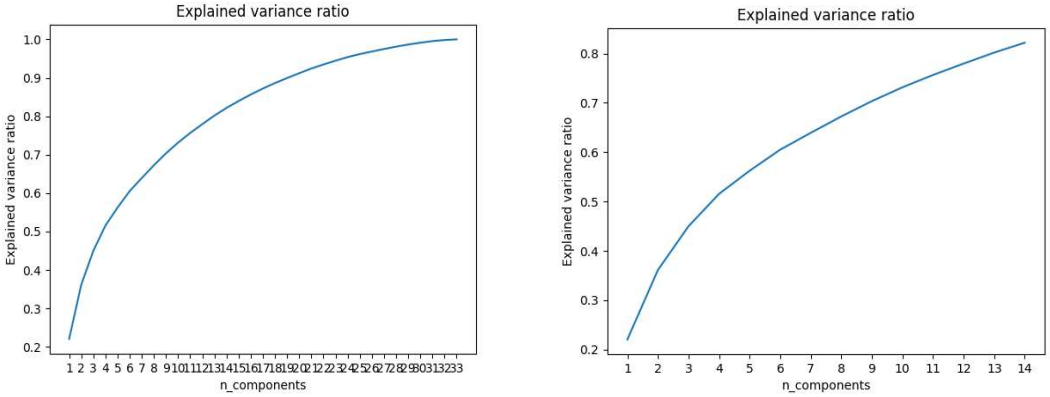


Figura 12 - Somatório das taxa de variância explicada ao longo dos componentes para os dados de acilcarnitinas com tratamento *Standard*

Para a transformação de dados de Uniformização, *Standard*, 14 componentes foram satisfatórios para explicar 82% dos dados, reduzindo dimensionalidade de 33 características para 14 componentes.

- **PCA com Normalização (MinMax)**

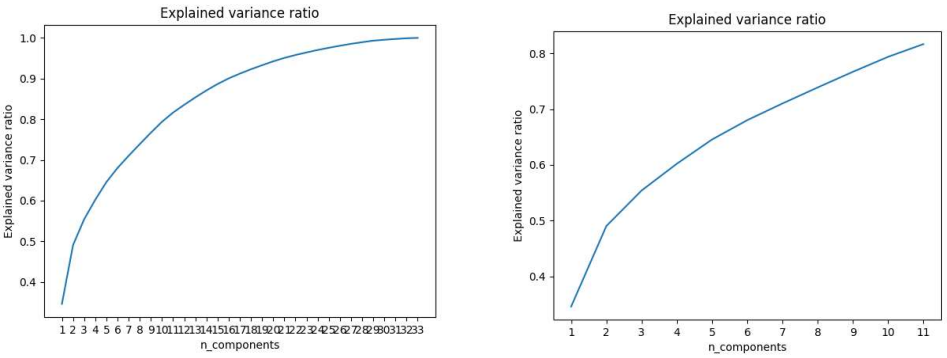


Figura 13 - Somatório das taxa de variância explicada ao longo dos componentes para os dados de acilcarnitinas com tratamento *MinMax*

Para a transformação de dados de Normalização, *MinMax*, o número de componentes considerados foi de 11, permitindo assim explicar 81,6% dos dados.

- **PCA com Escala Robusta**

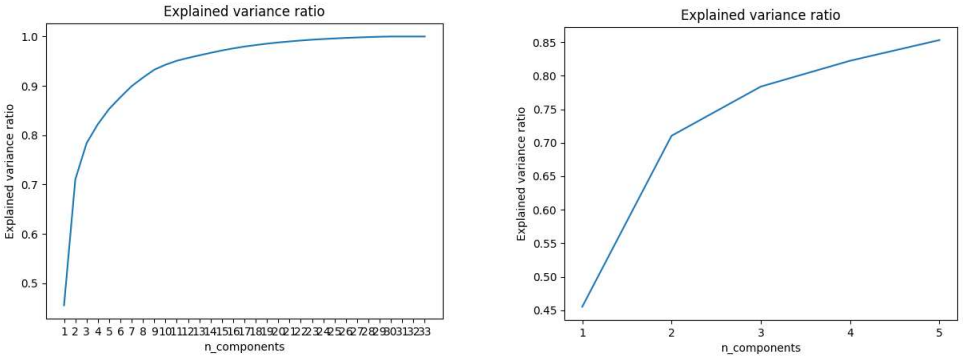


Figura 14 - Somatório das taxa de variância explicada ao longo dos componentes para os dados de acilcarnitinas com tratamento *Robust*

Para a transformação de Escala Robusta, *Robust*, simplesmente foram necessárias 5 componentes para explicar 85,3% dos dados.

- **PCA com Máximo Absoluto**

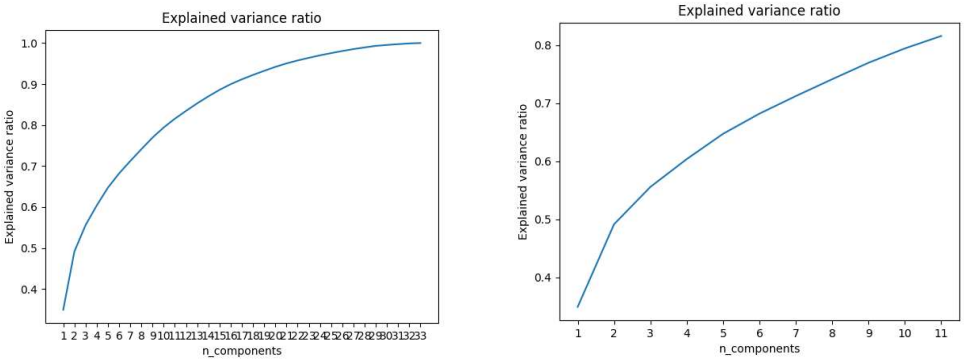


Figura 15 - Somatório das taxa de variância explicada ao longo dos componentes para os dados de acilcarnitinas com tratamento *MaxAbs*

Para a transformação de Máximo Absoluto, *MaxAbs*, consideraram-se apenas 11 componentes para explicar 81,5% dos dados.

Avaliação de resultados de PCA

Na Tabela 12 são apresentados os melhores resultados para cada um dos algoritmos, combinados com diferentes tratamentos de dados.

Tabela 12 - Melhores resultados obtidos para cada algoritmo, utilizando redução de dimensionalidade PCA, em combinação com os tratamentos de dados de acilcarnitinas

	Standard (14 componentes)	*MinMax (11 componentes)*	*Robust (5 componentes)*	*MaxAbs (11 componentes)*
Kmeans	Clusters: 6	Clusters: 4	**Clusters: 4**	Clusters: 4
	ET: 0.055s	ET: 0.046s	**ET: 0.039s**	ET: 0.046s
	SC: 0.224	SC: 0.275	**SC: 0.633**	SC: 0.274
	CH: 38.916	CH: 84.213	**CH: 198.241**	CH: 84.993
	DB: 1.466	DB: 1.693	**DB: 0.533**	DB: 1.711
Agglomerative	Clusters: 5	Clusters: 4	**Clusters: 8**	Clusters: 5
	ET: 0.001s	ET: 0.001s	**ET: 0.001s**	ET: 0.001s
	SC: 0.254	SC: 0.351	**SC: 0.448**	SC: 0.332
	CH: 40.186	CH: 74.971	**CH: 163.32**	CH: 67.175
	DB: 1.414	DB: 1.168	**DB: 0.696**	DB: 1.401
Birch	Clusters: 5	Clusters: 9	Clusters: 6	Clusters: 4
	ET: 0.011s	ET: 0.004s	ET: 0.008s	ET: 0.003s
	SC: 0.259	SC: 0.321	SC: 0.534	SC: 0.393
	CH: 40.525	CH: 41.587	CH: 174.585	CH: 64.797
	DB: 1.372	DB: 1.26	DB: 0.766	DB: 1.144
GMM	Clusters: 5	Clusters: 4	Clusters: 13	Clusters: 4
	ET: 0.013s	ET: 0.013s	ET: 0.029s	ET: 0.011s
	SC: 0.172	SC: 0.228	SC: 0.252	SC: 0.244
	CH: 31.643	CH: 72.716	CH: 145.447	CH: 74.721
	DB: 1.884	DB: 1.949	DB: 0.989	DB: 1.897
Spectral	Clusters: 4	Clusters: 10	Clusters: 4	Clusters: 4
	ET: 0.027s	ET: 0.035s	ET: 0.024s	ET: 0.026s
	SC: 0.1	SC: 0.107	SC: 0.188	SC: 0.15
	CH: 31.799	CH: 43.58	CH: 99.055	CH: 74.851
	DB: 2.112	DB: 1.681	DB: 1.3	DB: 1.723
Meanshift	Clusters: 16	Clusters: 9	**Clusters: 10**	Clusters: 9
	ET: 0.373s	ET: 0.425s	**ET: 0.341s**	ET: 0.344s
	SC: 0.248	SC: 0.333	**SC: 0.525**	SC: 0.335
	CH: 13.211	CH: 34.52	**CH: 125.763**	CH: 35.653
	DB: 0.805	DB: 0.793	**DB: 0.576**	DB: 0.936

Comparação de variâncias

De seguida, aplicar-se-á a técnica de VTFS a cada tratamento de dados. Para identificar o limite (*threshold)* apropriado para cada uma das transformações, é necessário avaliar a variância para cada uma das características, em cada uma das transformações, e só depois é possível definir o valor limite.

Na Tabela 13 serão demonstradas as variâncias para cada uma das características dos dados, em cada uma das transformações, tal como o limite identificado para ser utilizado na técnica de redução de dimensionalidade. Não foi considerada a uniformização (*Standard*) para esta técnica, dado que transforma a variância igual para todas as características.

Os limites definidos para esta técnica foram escolhidos com base na média e mediana das variâncias. As características que se mantiveram depois de aplicar o VTFS para a transformação de *MinMax* e a Escala Máximo Absoluto são as mesmas, logo, poderão ter resultados iguais nas experiências efetuadas. Na medida em que, a Escala Robusta procura manter os valores extremos através do uso da distância interquartil, conduz a resultados inesperados para as várias características, tendo em consideração as anomalias que possam existir.

Tabela 13 - Variâncias das características dos dados de acilcarnitinas

Características	Controlo	MinMax	Robust	MaxAbs
CARNITINE	84.270	**0.040**	0.528	**0.030**
C2	33.890	0.009	**1.256**	0.009
C3	0.403	0.023	0.783	0.023
C4	0.056	0.009	**3.090**	0.009
C5	0.004	0.023	0.537	0.023
C5DC	0.010	0.016	**1.338**	0.016
C6	0.000	0.016	**3.891**	0.016
C8	0.006	0.019	**2.426**	0.019
C10	0.014	0.012	**1.955**	0.012
C12	0.001	0.022	0.656	0.022
C14	0.001	0.019	0.836	0.019
C16	0.145	**0.030**	0.829	**0.030**
C18	0.038	**0.034**	0.846	**0.034**
C3DC	0.002	0.015	0.002	0.015
C40H	0.001	0.016	**52.705**	0.016
C4DC	0.174	**0.028**	**2.893**	**0.028**
C5:1	0.000	**0.028**	0.000	**0.028**
C50H	0.619	**0.139**	0.230	**0.139**
C6DC	0.002	0.013	**4.530**	0.013
C8:1	0.006	0.011	**1.387**	0.011
C10:1	0.011	0.021	**1.096**	0.021
C10:2	0.004	0.013	0.004	0.013
C12:1	0.931	0.005	**29.913**	0.005
C14:1	0.010	**0.027**	**1.371**	**0.027**
C14:10H	0.000	0.014	0.696	0.014
C14:2	0.008	0.024	**2.395**	0.024
C140H	0.000	**0.087**	0.347	**0.087**
C16:1	0.001	**0.038**	0.965	**0.038**
C160H	0.000	**0.077**	0.253	**0.077**
C18:1	0.256	0.023	**1.185**	0.023
C18:10H	0.000	**0.048**	0.356	**0.048**
C18:2	0.014	**0.032**	0.561	**0.032**
C180H	0.000	**0.091**	0.318	**0.091**
Threshold:	--------	0.025	1.000	0.250
Componentes:	--------	13	15	13

Avaliação dos resultados de Variance Threshold

Na Tabela 14 são apresentados os melhores resultados obtidos, por cada algoritmo, para cada tratamento de dado, utilizando redução de dimensionalidade VTFS.

Tabela 14 - Melhores resultados obtidos para cada algoritmo utilizando redução de dimensionalidade VTFS em combinação com os tratamentos de dados de acilcarnitinas

Algoritmos	*MinMax (13 componentes)*	*Robust (15 componentes)*	*MaxAbs (13 componentes)*
Kmeans	Clusters: 6	**Clusters: 4**	Clusters: 6
	ET: 0.055s	**ET: 0.049s**	ET: 0.050s
	SC: 0.202	**SC: 0.615**	SC: 0.205
	CH: 71.83	**CH: 151.957**	CH: 73.04
	DB: 1.662	**DB: 0.551**	DB: 1.645
Agglomerative	Clusters: 5	**Clusters: 4**	Clusters: 5
	ET: 0.002s	**ET: 0.002s**	ET: 0.002s
	SC: 0.203	**SC: 0.577**	SC: 0.242
	CH: 74.683	**CH: 137.036**	CH: 75.341
	DB: 1.566	**DB: 0.503**	DB: 1.429
Birch	Clusters: 4	**Clusters: 5**	Clusters: 4
	ET: 0.004s	**ET: 0.013s**	ET: 0.004s
	SC: 0.364	**SC: 0.58**	SC: 0.374
	CH: 69.937	**CH: 124.575**	CH: 73.507
	DB: 1.485	**DB: 0.536**	DB: 1.182
GMM	Clusters: 4	Clusters: 11	Clusters: 12
	ET: 0.015s	ET: 0.028s	ET: 0.038s
	SC: 0.162	SC: 0.249	SC: 0.108
	CH: 68.535	CH: 98.357	CH: 42.286
	DB: 2.735	DB: 1.082	DB: 1.918
Spectral	Clusters: 4	Clusters: 4	Clusters: 4
	ET: 0.031s	ET: 0.028s	ET: 0.026s
	SC: 0.199	SC: 0.221	SC: 0.2
	CH: 89.892	CH: 81.558	CH: 90.982
	DB: 1.548	DB: 1.522	DB: 1.536
Meanshift	Clusters: 4	Clusters: 18	Clusters: 4
	ET: 0.447s	ET: 0.304s	ET: 0.388s
	SC: 0.358	SC: 0.436	SC: 0.363
	CH: 67.159	CH: 70.098	CH: 68.4
	DB: 1.009	DB: 0.619	DB: 1.009

- **Discussão de resultados**

Independentemente do método de redução, a técnica de transformação de dados que melhor resultados apresentou, foi, sem dúvida, a escala Robusta. Esta transformação foi a melhor, dado ter sido utilizada distância interquartil, continuando assim a existir uma grande diferença de variâncias nas características dos dados, as quais se devem à existência de valores extremos (considerados as anomalias).

O melhor resultado obtido das avaliações do PCA, foi o algoritmo K-means, quando os dados se encontravam transformados através da escala robusta. Esta avaliação apresentou resultados de SC de 0.633, sendo dos valores mais altos obtidos em todas as experiências, demonstrando que os 4 *clusters* formados estão separados e distantes entre si. A avaliação de CH de 198 define que os *clusters* estão densos, bem definidos e existe separação dos restantes. A avaliação de Davies-Bouldin apresenta um valor baixo, de 0.533, o que permite concluir que existe baixa semelhança nos elementos dos grupos, em comparação com os elementos dos outros grupos. A segunda melhor avaliação nesta técnica de redução de dimensionalidade, foi alcançada com o algoritmo Agglomerative, também com a escala robusta, para a formação de 8 clusters e valores de SC de 0.448, de CH de 163 e de DB de 0.696. Ambos os resultados são extremamente promissores na formação de clusters para os dados de acilcarnitinas.

Os melhores resultados obtidos para a redução de dimensionalidade utilizando VTFS, foram obtidos com os algoritmos de K-means, de Agglomerative e de Birch, utilizando a escala robusta. A avaliação de Silhouette apresentou valores de 0.615, 0.577 e 0.58, e a de Calinski, respetivamente, de, 151, 137 e 124 e a de Davis-Bouldin, respetivamente, de 0.551, 0.503 e 0.536.

Não obstante, os resultados obtidos com a Variance Threshold serem muito bons e ter sido possível concluir a formação de *clusters* densos, distantes entre si e com bom nível de semelhança entre instâncias pertencentes ao mesmo grupo, os resultados de K-means, com escala robusta e utilização de PCA, foram os melhores resultados alcançados.

5.2.4.2 Aminoácidos

Seguidamente, começar-se-á com a técnica PCA de redução de dimensionalidade para os dados de aminoácidos, tal como foi apresentado anteriormente para os dados de acilcarnitinas.

Com o objetivo de identificar um número de componentes que permita explicar mais de 80% dos dados, de acordo com [45], serão apresentados vários gráficos com o somatório das Taxas de Variância Explicada de ordem crescente dos componentes, para cada uma das normalizações utilizadas nos dados de aminoácidos. Cada figura vai apresentar, no primeiro gráfico o somatório de todos os componentes, logo o valor total é 100%, e no segundo gráfico, apenas é apresentado o valor mínimo de componentes que permita alcançar um somatório de 80%.

- **PCA com Standard**

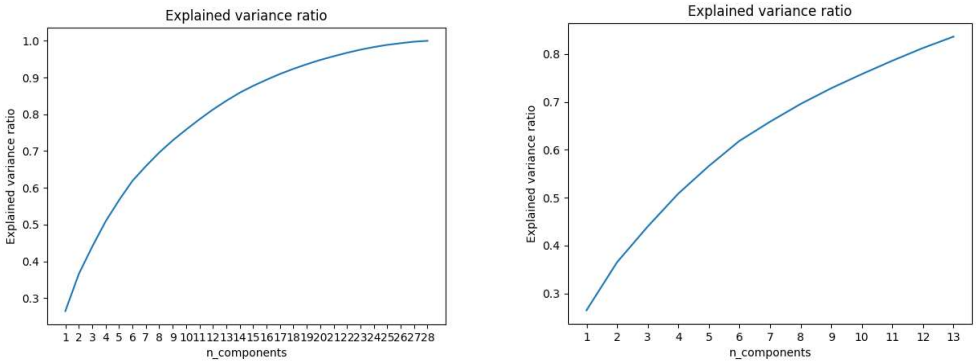


Figura 16 - Somatórios das taxas de variância explicada ao longo dos componentes para os dados de aminoácidos com tratamento *Standard*

Para a transformação de dados de Uniformização, *Standard*, 13 componentes foram satisfatórios para explicar 82% dos dados, reduzindo a dimensionalidade das 28 concentrações para 13 componentes.

- **PCA com Normalização (MinMax)**

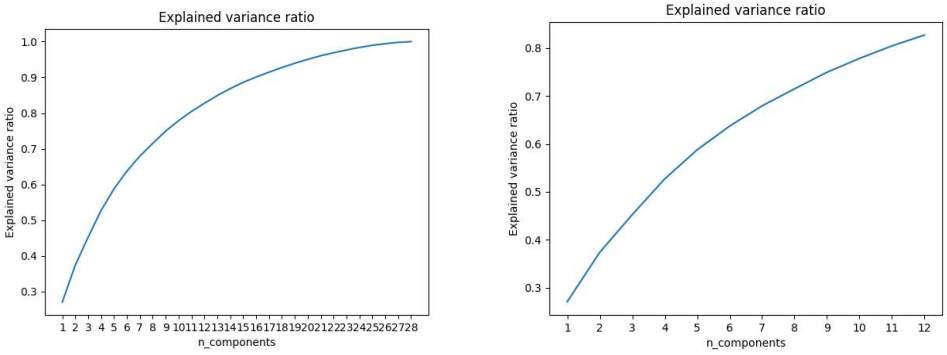


Figura 17 - Somatórios das taxas de variância explicada ao longo dos componentes para os dados de aminoácidos com tratamento *MinMax*

Para a transformação de dados de Normalização, *MinMax*, o número de componentes considerado foi de 12, permitindo explicar 82,7% dos dados.

- **PCA com Escala Robusta**

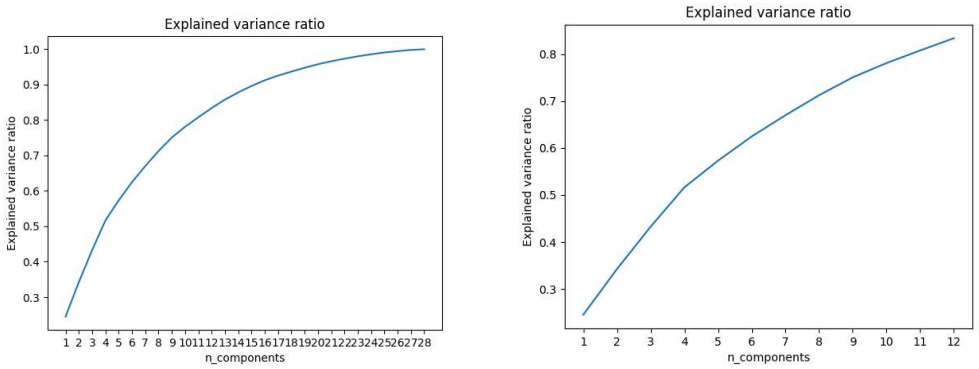


Figura 18 - Somatórios das taxas de variância explicada ao longo dos componentes para os dados de aminoácidos com tratamento *Robust*

Para a transformação de Escala Robusta formaram-se apenas 12 componentes para explicar 83,3% dos dados.

- **PCA com Máximo Absoluto**

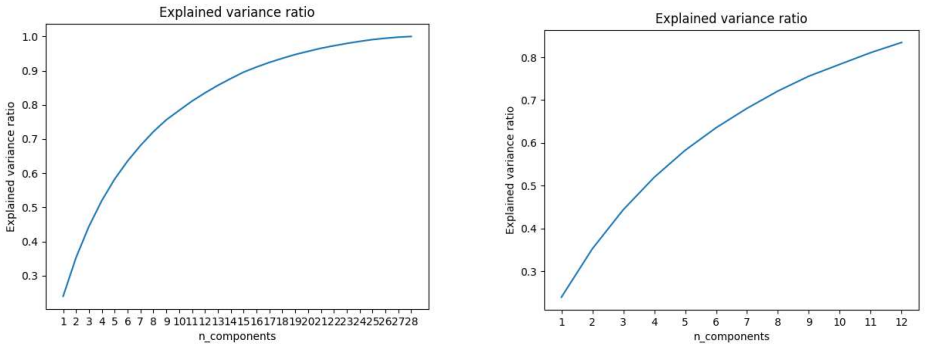


Figura 19 - Somatórios das taxas de variância explicada ao longo dos componentes para os dados de aminoácidos com tratamento MaxAbs

Para a transformação de Máximo Absoluto, *MaxAbs*, consideraram-se apenas 12 componentes para explicar 83,4% dos dados.

- **Avaliação de resultados de PCA**

Na Tabela 15 é possível verificar os melhores resultados obtidos para cada um dos algoritmos em estudo, combinados com diferente tratamento de dados de aminoácidos, utilizando a técnica de redução de dimensionalidade PCA.

Tabela 15 - Melhores resultados obtidos para cada algoritmo utilizando redução de dimensionalidade PCA em combinação com os tratamentos de dados de aminoácidos

Algoritmos	*Standard (13 componentes)*	*MinMax (12 componentes)*	*Robust (12 componentes)*	*MaxAbs (12 componentes)*
Kmeans	Clusters: 5	Clusters: 3	**Clusters: 6**	Clusters: 5
	ET: 0.038s	ET: 0.035s	**ET: 0.046s**	ET: 0.045s
	SC: 0.113	SC: 0.127	**SC: 0.125**	SC: 0.125
	CH: 28.389	CH: 41.584	**CH: 26.671**	CH: 28.32
	DB: 2.086	DB: 2.111	**DB: 1.868**	DB: 1.947
Agglomerative	Clusters: 13	Clusters: 7	Clusters: 10	Clusters: 8
	ET: 0.001s	ET: 0.001s	ET: 0.001s	ET: 0.001s
	SC: 0.094	SC: 0.122	SC: 0.111	SC: 0.119
	CH: 16.75	CH: 22.627	CH: 18.998	CH: 20.146
	DB: 1.804	DB: 1.941	DB: 1.812	DB: 1.93
Birch	Clusters: 5	Clusters: 8	Clusters: 7	**Clusters: 4**
	ET: 0.008s	ET: 0.003s	ET: 0.007s	**ET: 0.003s**
	SC: 0.1	SC: 0.109	SC: 0.116	**SC: 0.212**
	CH: 24.169	CH: 17.065	CH: 21.926	**CH: 15.174**
	DB: 2.032	DB: 1.64	DB: 1.83	**DB: 1.476**
GMM	Clusters: 4	Clusters: 11	Clusters: 4	Clusters: 18
	ET: 0.016s	ET: 0.017s	ET: 0.014s	ET: 0.023s
	SC: 0.089	SC: 0.09	SC: 0.09	SC: 0.096
	CH: 24.534	CH: 19.076	CH: 23.813	CH: 14.676
	DB: 2.493	DB: 1.936	DB: 2.436	DB: 1.708
Spectral	Clusters: 4	Clusters: 17	Clusters: 6	Clusters: 7
	ET: 0.024s	ET: 0.039s	ET: 0.024s	ET: 0.025s
	SC: 0.093	SC: 0.1	SC: 0.082	SC: 0.1
	CH: 28.979	CH: 15.938	CH: 22.116	CH: 22.328
	DB: 2.384	DB: 1.751	DB: 2.009	DB: 1.988
Meanshift	**Clusters: 5**	Clusters: 4	Clusters: 10	Clusters: 3
	ET: 0.367s	ET: 0.433s	ET: 0.406s	ET: 0.380s
	SC: 0.274	SC: 0.206	SC: 0.22	SC: 0.244
	CH: 9.502	CH: 9.075	CH: 7.896	CH: 5.81
	DB: 1.035	DB: 1.392	DB: 0.994	DB: 1.35

Comparação de variâncias

Seguidamente, aplicar-se-á a técnica de VTFS aos diferentes tratamentos de dados de aminoácidos, de acordo com o referido na experiência anterior.

Com o objetivo de definir um limite de variância de características, que permita reduzir a dimensionalidade, é necessário identificar um limite para cada característica. Na Tabela 16 avaliaram-se as variâncias de cada uma das características para cada uma das transformações.

Não se considerou a transformação *Standard*, pois transforma os desvios padrões num único valor.

Tabela 16 - Variâncias das características dos dados de aminoácidos

Características	Controlo	MinMax	Robust	MaxAbs
TAU	820.672	0.026	0.528	**0.023**
ASP	11.108	**0.041**	0.472	**0.041**
HYP	75.191	0.020	**1.914**	**0.020**
THR	1609.062	**0.042**	0.544	**0.025**
SER	651.542	**0.033**	0.599	0.020
ASN	304.475	0.019	0.639	0.014
GLU	2491.844	**0.039**	0.673	**0.035**
GLN	18252.070	0.013	**0.914**	0.008
AAA	17.432	0.023	**1.108**	**0.023**
PRO	8014.318	0.030	0.696	**0.030**
GLY	4721.158	**0.032**	**1.160**	0.020
ALA	10552.820	**0.040**	0.627	**0.026**
CIT	262.797	**0.040**	0.517	**0.039**
ABU	107.216	0.029	**0.707**	**0.025**
VAL	2456.374	0.026	0.674	0.017
CYS2	265.746	0.017	0.617	0.010
MET	66.478	0.029	**0.798**	0.019
CYSTA	1.692	**0.037**	**1.331**	**0.037**
ILE	519.621	0.028	0.677	**0.022**
LEU	1384.136	0.024	0.696	0.017
TYR	351.339	0.030	0.537	0.020
PHE	141.125	0.028	0.660	0.014
ORN	1461.986	**0.041**	0.657	**0.026**
LYS	1632.011	**0.035**	0.700	0.017
1MHIS	180.559	**0.032**	**0.800**	**0.032**
HIS	257.692	**0.038**	0.639	0.018
3MHIS	14.278	0.015	**1.496**	0.015
ARG	543.665	**0.032**	0.596	**0.032**
Threshold:	--------	0.03	0.7	0.02
Componentes:	--------	13	9	15

As características que foram conservadas para cada uma das transformações encontram-se realçadas. Os limites (*threshold*) definidos para esta técnica foram escolhidos com base na média e mediana das variâncias.

Avaliação dos resultados de Variance Threshold

Na Tabela 17 é possível verificar os melhores resultados dos algoritmos, utilizando a redução de dimensionalidade VTFS, para cada uma das transformações aplicadas.

Tabela 17 - Melhores resultados obtidos para cada algoritmo utilizando redução de dimensionalidade VTFS em combinação com os tratamentos de dados

	MinMax (13 componentes)	*Robust (9 componentes)*	*MaxAbs (15 componentes)*
Kmeans	Clusters: 8	**Clusters: 5**	Clusters: 5
	ET: 0.055s	**ET: 0.046s**	ET: 0.044s
	SC: 0.117	**SC: 0.181**	SC: 0.125
	CH: 24.15	**CH: 31.977**	CH: 25.124
	DB: 1.848	**DB: 1.565**	DB: 1.975
Agglomerative	Clusters: 4	Clusters: 5	Clusters: 5
	ET: 0.002s	ET: 0.002s	ET: 0.002s
	SC: 0.122	SC: 0.121	SC: 0.111
	CH: 28.621	CH: 24.25	CH: 22.785
	DB: 2.174	DB: 1.729	DB: 2.184
Birch	Clusters: 5	Clusters: 11	Clusters: 5
	ET: 0.004s	ET: 0.009s	ET: 0.004s
	SC: 0.126	SC: 0.123	SC: 0.146
	CH: 22.195	CH: 21.957	CH: 12.463
	DB: 1.694	DB: 1.604	DB: 1.705
GMM	Clusters: 16	Clusters: 3	Clusters: 6
	ET: 0.020s	ET: 0.014	ET: 0.017s
	SC: 0.098	SC: 0.199	SC: 0.091
	CH: 16.07	CH: 24.746	CH: 20.805
	DB: 1.787	DB: 2.301	DB: 2.12
Spectral	Clusters: 5	**Clusters: 5**	Clusters: 9
	ET: 0.026s	**ET: 0.025s**	ET: 0.033s
	SC: 0.109	**SC: 0.147**	SC: 0.085
	CH: 26.535	**CH: 28.155**	CH: 17.389
	DB: 2.112	**DB: 1.656**	DB: 1.962
Meanshift	Clusters: 3	**Clusters: 11**	Clusters: 3
	ET: 0.487s	**ET: 0.383s**	ET: 0.414s
	SC: 0.229	**SC: 0.243**	SC: 0.28
	CH: 5.257	**CH: 12.137**	CH: 7.399
	DB: 1.253	**DB: 0.897**	DB: 1.354

- **Discussão de resultados**

Relativamente aos dados de perfis metabólicos de aminoácidos, os resultados não foram tão elevados como os de acilcarnitinas, tanto para as experiências de PCA ou de VTFS, uma vez que os valores mais altos de SC tiveram perto de alcançar 0.3, sendo um valor bom, mas baixo em

comparação com os resultados obtidos nas acilcarnitinas. Os valores de DB também não foram o esperado, poucas foram as avaliações que tiveram valores abaixo de 1.

Para ambas as experiências de redução de dimensionalidade, o melhor algoritmo foi o Meanshift. Na experiência de PCA, as melhores avaliações ocorreram com o tratamento de dados *Standard*, com o algoritmo Meanshift, e de escala Máximo Absoluto, com o algoritmo Birch. Para o tratamento *Standard*, a melhor avaliação foi com a formação de 5 grupos, teve valores de SC de 0.274, sendo melhor que o algoritmo de Birch com a formação de 4 grupos com valor de 0.212. Os valores de CH foram de 9 e 15 respetivamente, sendo muito baixos, e os de DB 1.035 e 1.447. Os melhores resultados para a técnica de PCA tiveram valores abaixo do esperado para SC e CH e altos para DB.

Para a técnica de VTFS, a experiência que apresentou melhores resultados para este conjunto de dados foi a de tratamento com escala robusta e com a utilização do algoritmo MeanShift, obtendo-se com SC de 0.243, CH de 12.137 e DB de 0.897. Ambas as situações apresentaram resultados similares e, por essa razão, vão ser utilizados na seguinte experiência de otimização de parâmetros.

5.2.5 Otimização de parâmetros dos algoritmos

Os *Hyperparameters* em ML são os parâmetros que são explicitamente definidos pelo utilizador para controlar o processo de aprendizagem. Estes parâmetros são usados para melhorar a aprendizagem dos modelos e os seus respetivos valores são definidos antes de começar o processo. Dependendo do algoritmo de aprendizagem que se esteja a usar, os seus *hyperparameters* podem variar.

De um modo geral, definir *hyperparameters* para os algoritmos de aprendizagem é complicado, dado não existir uma verdade absoluta para validação dos resultados obtidos. Por essa razão, a comparação dos resultados obtidos vai ser idêntica às práticas utilizadas anteriormente para validar os grupos formados, uma vez que o sucesso dos métodos de agrupar os dados depende sobretudo da escolha correta destes parâmetros.

Nas experiências anteriores, o parâmetro que permite definir o número de grupos a se formar foi, sempre que possível, utilizado. No entanto, nesta subsecção, todos os parâmetros restantes possíveis dos algoritmos de *Clustering*, dos melhores resultados das experiências passadas, serão otimizados com o objetivo de encontrar a melhor combinação de valores para alcançar os melhores resultados possíveis. O número de *clusters,* para cada uma das melhores avaliações anteriormente identificadas, será mantido durante a otimização dos restantes parâmetros.

A otimização de *hyperparameters* foi efetuada através de métodos que testavam as melhores combinações possíveis, apesar do alto nível de computação, uma vez que os treinos efetuados previamente são relativamente rápidos.

O método desenvolvido para otimizar os algoritmos, é um processo de pesquisa exaustiva em subconjuntos especificados manualmente para os vários parâmetros do algoritmo pretendido.

Este método desenvolvido verifica todas as combinações do conjunto de dados apresentados para cada um dos algoritmos na Tabela 18 e avalia os grupos formados consoante as métricas de avaliação anteriormente apresentadas. Os algoritmos apresentados na tabela correspondem aos algoritmos que apresentaram melhores resultados para as acilcarnitinas, K-means, e para os aminoácidos, Meanshift e Birch.

Tabela 18 - Hyperparameters existentes para cada um dos algoritmos em estudo

Algoritmos	**Hyperparameters**
Kmeans	Random_state: range (1,200) Init: [“k-means++”; “random”], padrão: “k-means++” n_init: [10,20,50,100,200,250], padrão: 10 max_iter: [100;200;300;350;400], padrão: 300 algorithm: [“lloyd”, “elkan”], padrão: “lloyd”
Meanshift	Quantile = [0.1,0.2,0.3,0.4,0.5], padrão: 0.3 max_iter: [100,200,300,350,400,450], padrão: 300 bin_seeding: [True, False], padrão: False
Birch	Branching_factor: [20,25,35,50,75,80,100,200], padrão: 50 Threshold: [0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9], padrão: 0.5

O conjunto de dados para cada parâmetro do algoritmo, foi definido, de acordo com o respetivo valor padrão que foi utilizado nas experiências anteriores.

Para o algoritmo de *K-means*, os parâmetros permitem definir o tipo de algoritmo a ser utilizado, o número de tentativas efetuadas com diferentes centros de grupos para serem testados, o número máximo de iterações efetuadas para cada tentativa e o número da semente (`random_state`).

Tanto o algoritmo *MeanShift* como o *Birch,* não permitem definir o valor da *seed*. Para o algoritmo *MeanShift* define-se o conjunto de valores para o `quantile` que permite calcular a `bandwidth` do método, o número máximo de iterações até a operação de agrupar terminar e `bin_seeding` que está relacionado com o processo de inicialização de *seeds*. O algoritmo Birch apenas permite definir os parâmetros de `branching_factor`, relacionados com o número máximo de novas amostras que podem ser adicionadas aos *clusters*, até serem criados *subclusters*, e `threshold` que permite definir o raio dos *subclusters* para adicionar novas amostras, caso este valor seja demasiado baixo o algoritmo forma mais grupos.

5.2.5.1 Comparação de algoritmos

De seguida, será apresentada a comparação dos resultados obtidos nas experiências passadas, para os melhores resultados para os dados de acilcarnitinas, algoritmo K-means com escala Robusta e redução PCA e VTFS sem otimização e com o resultado depois de efetuado as otimizações aos algoritmos.

Acilcarnitinas

Na Tabela 19 são apresentados os resultados das melhores combinações de parâmetros dos algoritmos, para os dados de acilcarnitinas:

Tabela 19 - Otimização dos algoritmos para os dados de acilcarnitinas

Algoritmo	Resultado sem otimização	Resultados com otimização	Melhores parâmetros identificados
K-means com PCA e Robust Scale	Cluster: 4 ET: 0.039 s SC: 0.633 CH: 198.241 DB: 0.533	**Cluster: 4** **ET: 0.017 s** **SC: 0.670** **CH: 192.352** **DB: 0.380**	Seed = 44 Init = "k-means++" n_init = 10 max_iter = 100 algorithm = "lloyd"
K-means com VTFS e Robust Scale	Cluster: 4 ET: 0.049s SC: 0.615 CH: 151.957 DB:0.551	Cluster:4 ET: 0.018 s SC: 0.629 CH: 147.590 DB: 0.431	Seed = 35 init = k-means++ n_init = 10 max_iter = 100 algorithm = "elkan"

- **Discussão de resultados**

O melhor resultado identificado para agrupar os dados de acilcarnitinas, foi o de redução de dimensionalidade de PCA e Escala Robusta com o algoritmo K-means. A melhor combinação foi alcançada com inicialização (*init*) de "k-means++" e para a semente (*seed*) 44. O resultado desta experiência repetiu-se várias vezes com diferentes combinações, pois quanto maior o número

de iterações(*max_iter*), mais facilmente o algoritmo acaba por descobrir o melhor conjunto de centros dos grupos.

Nas experiências efetuadas anteriormente, foram utilizados valores padrões, que eram relativamente baixos, impedindo assim o algoritmo de alcançar os melhores resultados. Os valores obtidos depois da otimização de parâmetros foram superiores comparados com a experiência inicial, com um aumento nos valores de SC para 0.670, um valor de CH muito perto do anterior com 192.352 e uma diminuição de DB para 0.380. Considerou-se uma das melhores experiências efetuadas para dados de acilcarnitinas, com bons resultados e com bom prognóstico de dispersão entre os grupos formados e grupos compostos por pontos semelhantes.

Para a combinação de VTFS com Robust Scale e K-means, o melhor resultado foi obtido para os parâmetros de *init* com "k-means++", *n_init* de 10, com um máximo de 100 iterações e foi apenas alcançado na *seed* 35. O valor de SC sofreu um pequeno aumento em comparação com a primeira experiência, situando-se em 0.629. Porém, o valor de CH desceu ligeiramente como na experiência de PCA. Apesar da descida de CH, considerou-se ser este o melhor resultado para a redução utilizando VTFS, uma vez que, a descida do valor de DB para 0.431, foi uma melhoria significativa para a formação dos grupos.

Conclui-se que o parâmetro de máximo de iterações (*max_iter*) a 100 foi o suficiente uma vez que o valor da *seed* permitiu aleatoriamente selecionar os melhores centros dos grupos, conduzindo às melhores avaliações em ambos os casos. Não obstante, ambos os resultados serem extraordinários e os grupos gerados com boas características, o melhor dos dois foi alcançado através do uso do algoritmo de *K-means,* com redução de dimensionalidade utilizando PCA para 5 componentes com os dados tratados com escala Robusta.

Aminoácidos

Na Tabela 20 são apresentadas as comparações dos melhores algoritmos, Birch e Meanshift, para os dados de aminoácidos identificados na experiência passada com e sem as suas otimizações:

Tabela 20 - Otimização dos algoritmos para os dados de aminoácidos

Algoritmo	Resultado sem otimização	Resultados com otimização	Melhores hyperparameter
Meanshift com PCA e Standard	Cluster: 5 ET: 0.367s SC: 0.274 CH: 9.502 DB: 1.035	Cluster: 3 ET: 0.317 s SC: 0.314 CH:4.699 DB: 0.850	Bandwidth = 5.935 Quantile = 0.4 Bin_seeding = False Max_iter = 100
Birch com PCA e MaxAbs	Cluster: 4 ET: 0.003s SC: 0.212 CH: 15.174 DB: 1.476	Cluster: 4 ET: 0.003 s SC: 0.227 CH: 11.006 DB: 1.239	Branching_factor = 20 Threshold = 0.7
Meanshift com Variance Threshold e Robust Scale	Cluster: 11 ET: 0.383s SC: 0.243 CH: 12.137 DB: 0.897	**Cluster: 5** **ET: 0.059s** **SC: 0.391** **CH: 15.270** **DB: 0.743**	Bandwidth = 3.8495 Quantile = 0.5 Bin_seeding = True Max_iter = 100

A otimização do algoritmo Birch permitiu melhorar a primeira avaliação dos dados de aminoácidos, com tratamento de escala de Máximo Absoluto, embora, 0.227 de SC ser baixo, tal como o de CH de 11. O valor de DB também foi alto com 1.239, o que significa que os grupos formados apresentam alguma semelhança e não se encontram tão dispersos e separados entre si, como era de esperar.

A otimização do algoritmo MeanShift, para ambas as técnicas de redução de dimensionalidade, produziu bons resultados, principalmente com VTFS e escala robusta, sendo a melhor experiência efetuada para os dados de aminoácidos. Os valores para o SC foram de 0.391, não sendo este o melhor de todas as experiências, mas o melhor para este conjunto de dados; CH de 15 foi melhor que os antigos valores obtidos; e um valor baixo de DB de 0.743.

Em geral, para os algoritmos de MeanShift e Birch, a otimização dos *hyperparameters* teve resultados positivos em todas as experiências.

5.2.6 Análise dos resultados dos algoritmos otimizados

A visualização de dados de alta dimensionalidade é um dos problemas mais recorrentes em diferentes setores. De forma a resolver este problema, recorrer-se-á à utilização da técnica de redução de dimensionalidade t-SNE, referenciada anteriormente, para reduzir os diversos componentes que ainda existem nos dados tratados pelas técnicas de PCA e VTFS para os dados de perfis metabólicos.

A técnica t-SNE permite definir o número de componentes para o qual se pretende reduzir os dados e a perplexidade. A perplexidade consiste no número alvo de vizinhos esperados para os pontos centrais durante o processo de redução de dimensionalidade. Diferentes perplexidades permitem diferentes reduções de dimensionalidade, logo diferentes visualizações.

O processo de reduzir a alta dimensionalidade para apenas dois componentes, através do uso de diferentes perplexidades para visualizar os dados, pode induzir o utilizador a diferentes conclusões.

Para os dados de acilcarnitinas, os dados foram reduzidos para 5 componentes, através da utilização da técnica PCA e normalização Robusta e para os dados de aminoácidos para apenas 13 características através da utilização da técnica VTFS e normalização Robusta. Estes dados necessitam de ser reduzidos para apenas 2 componentes, permitindo assim a sua visualização num plano 2D.

Acilcarnitinas

Na Tabela 21 são apresentados o número dos *clusters* e a quantidade de instâncias que pertencem a cada grupo. Na Figura 20 é apresentada a dispersão dos dados depois de terem sido reduzidos a um plano de 2 componentes com uma perplexidade de 200.

Tabela 21 - Número de instâncias pertencentes a cada grupo formado para os dados de acilcarnitinas

# Cluster	Número Instâncias
0	229
1	2
2	1
3	30

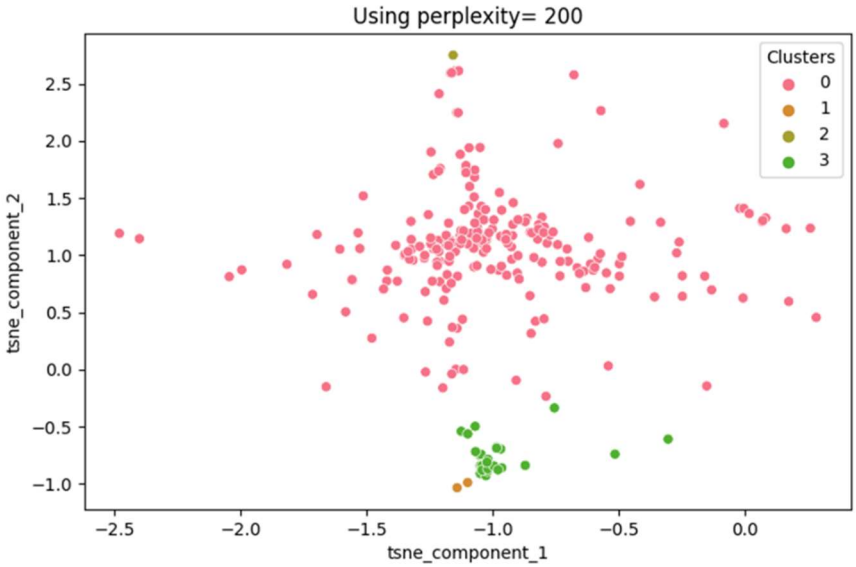


Figura 20 - Gráfico dos dados de acilcarnitinas

Primeiramente, reduzir dados de alta dimensionalidade para apenas 2 componentes pode induzir em erro nas conclusões. Contudo, com base na Figura 20, é possível verificar que os grupos se encontram densos e com instâncias semelhantes.

De igual modo, é possível verificar que, foram formados grupos somente com uma anomalia ou duas, estando isso relacionado com o tratamento de dados efetuado anteriormente. Os grupos foram formados após os dados serem tratados com a escala Robusta, preservando os pontos extremos que se apresentam como anomalias dos dados. Os dados anómalos foram mantidos no conjunto de dados, levando o algoritmo a formar grupos que apresentem as mesmas características, formando grupos com poucas instâncias. Com base na figura, é possível considerar a possibilidade da existência de mais grupos, mas uma vez que se iam encontrar demasiado próximos entre si, podem não ter sido bem avaliados durante as experiências acabando por não ter sido considerados.

Aminoácidos

Na Tabela 22 são apresentados o número dos *clusters* e a quantidade de instâncias que pertencem a cada grupo formado para os dados de aminoácidos. Na Figura 21 é apresentada a melhor dispersão dos dados depois de terem sido reduzidos a um plano de apenas dois componentes com uma perplexidade de 60. Considerou-se ser 60 a melhor perplexidade para esta técnica, pois permite a visualização dos dados e verificar as múltiplas anomalias existentes.

Tabela 22 - Número de instâncias pertencentes a cada grupo formado para os dados de aminoácidos

# Cluster	Número Instâncias
0	205
1	5
2	1
3	2
4	2

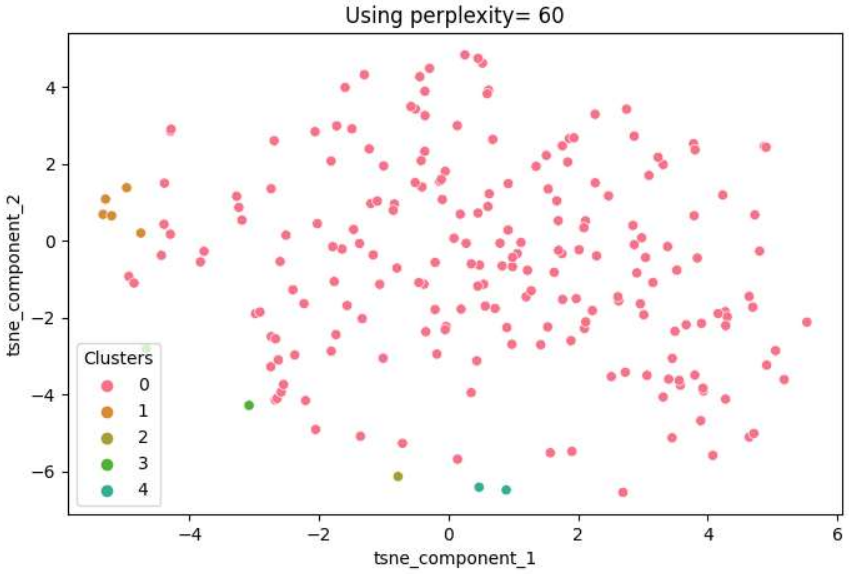


Figura 21 - Gráfico dos dados de aminoácidos

De todos os gráficos formados pela técnica de t-SNE, o com perplexidade 60, apresentado na Figura 21, mostrou ser o único caso onde foi possível verificar que os grupos formados eram referentes a anomalias nos dados. Assim como, nos dados de acilcarnitinas, a melhor avaliação resultou da combinação do algoritmo com o tratamento de dados de escala Robusta, visto que foram priorizadas as anomalias com valores extremos, contribuindo assim para a existência de tantos grupos com poucos elementos.

Na medida em que, os dados de aminoácidos apresentavam 13 componentes depois da redução de VTFS, a redução para 2 componentes não permitiu uma boa visualização dos dados. Apesar da visualização gerada não ter apresentado grupos densos e compactos, é possível verificar que os elementos dos mesmos grupos se mantiveram perto e separados das restantes instâncias.

6 Conclusão

Nesta secção, considerações gerais referentes à dissertação são apresentadas, tal como sugestões para as próximas etapas do trabalho no futuro, com o intuito de melhorar o trabalho desenvolvido.

6.1 Considerações Gerais

O objetivo principal proposto nesta dissertação consistia em analisar o conjunto de dados de perfis metabólicos de acilcarnitinas e aminoácidos, agrupar as instâncias semelhantes em clusters e detetar anomalias através de técnicas e algoritmos de aprendizagem não-supervisionada.

Os dados foram compilados e tratados para serem consumidos pelos algoritmos de ML, extraindo-se qualquer dado incompleto, dados duplicados ou dados que não iam de acordo com as regras apropriadas. Depois de efetuado o tratamento de dados, foi necessário processá-los através de técnicas de transformação que permitissem normalizar, de modo a não influenciar a aprendizagem dos algoritmos.

Começou-se por analisar as diferenças entre os dados com e sem tratamento de dados durante o processo de aprendizagem para cada um dos algoritmos não supervisionados. De seguida, uma vez que os dados apresentavam uma elevada dimensionalidade (demasiadas características/indicadores para cada instância) foi necessário analisar qual das técnicas de redução de dimensionalidade, PCA e VTFS, era melhor para cada um dos dados de perfis metabólicos. Contudo, depois de efetuadas as respetivas avaliações e de identificar o melhor conjunto de resultados, estes modelos foram otimizados para poderem apresentar resultados mais precisos e formar melhor grupos.

As análises efetuadas permitiram gerar grupos para cada experiência. Os resultados das experiências foram avaliados em função de validações internas dos clusters e entre clusters, de

modo a validar se os grupos teriam sido devidamente criados, se se encontravam dispersos e se os dados dos grupos eram semelhantes com os dados de outros grupos.

A melhor combinação de técnicas para os dados de acilcarnitinas ocorreu com a otimização do algoritmo K-means, utilizando a técnica de redução de dimensionalidade de PCA para formar cinco componentes principais, através de uma transformação Robusta dos dados. Para os dados de aminoácidos, a melhor combinação de técnicas foi alcançada utilizando uma otimização dos parâmetros do algoritmo MeanShift com a técnica de redução de VTFS, para uma escala Robusta.

Os resultados finais obtidos foram satisfatórios para ambos os dados de perfis metabólicos, com valores de SC e CH altos e DB baixos, tal como pretendido. Dos grupos formados pelos algoritmos, foi possível identificar dados anómalos que formaram pequenos grupos, alcançando-se assim um dos principais objetivos desta dissertação.

6.2 Limitações identificadas

Durante o desenvolvimento deste trabalho, as principais limitações identificadas prendem-se com a:

- Necessidade de mais uma iteração, de forma a permitir o feedback por parte dos profissionais de saúde para conseguir validar os grupos formados, assim como as anomalias identificadas;
- Falta de informação relativamente às várias etapas de tratamento efetuado a cada um dos pacientes.

6.3 Trabalho para o futuro

De forma a completar o trabalho desenvolvido até ao momento, poderão:

- Realizar mais experiências com novos dados de pacientes, de modo a identificar e validar os grupos a que são associados;
- Recolher mais dados sobre diferentes momentos dos pacientes com CRC e tratamentos efetuados, de forma a desenvolver modelos capazes de sugerir o tratamento correto para um novo paciente que faça parte do mesmo grupo de pacientes;
- Desenvolver um sistema de decisão, com base na arquitetura proposta, onde seja possível visualizar os diferentes grupos criados e combinar o trabalho desenvolvido com trabalhos passados efetuados.

Referências

[1] R. Bhardwaj, A. R. Nambiar, and D. Dutta, "A Study of Machine Learning in Healthcare," *Proceedings - International Computer Software and Applications Conference*, vol. 2, pp. 236–241, Sep. 2017, doi: 10.1109/COMPSAC.2017.164.

[2] H. Hart, "The healthcare industry focuses on new growth drivers and leadership requirements," Jul. 10, 2017. https://silo.tips/download/the-healthcare-industry-focuses-on-new-growth-drivers-and-leadership-requirement (accessed Jan. 18, 2023).

[3] "Machine Learning: What it is and why it matters | SAS." https://www.sas.com/en_us/insights/analytics/machine-learning.html (accessed Jan. 18, 2023).

[4] K. Shailaja, B. Seetharamulu, and M. A. Jabbar, "Machine Learning in Healthcare: A Review," *Proceedings of the 2nd International Conference on Electronics, Communication and Aerospace Technology, ICECA 2018*, pp. 910–914, Sep. 2018, doi: 10.1109/ICECA.2018.8474918.

[5] N. Hotz, "What is CRISP DM? - Data Science Process Alliance," Jan. 08, 2023. https://www.datascience-pm.com/crisp-dm-2/ (accessed Jan. 16, 2023).

[6] "O cancro colorretal ou cancro do cólon: sintomas, causas e tratamento | Médis." https://www.medis.pt/mais-medis/cancro/conheca-melhor-o-cancro-colorretal-ou-cancro-do-colon-guia-de-saude (accessed Jan. 16, 2023).

[7] "O que é o Cancro Colo-Rectal? - Europacolon Portugal - Apoio ao Doente com Cancro Digestivo." http://www.europacolon.pt/pagina/323-adoenca (accessed Jan. 16, 2023).

[8] C. Indiveri *et al.*, "The mitochondrial carnitine/acylcarnitine carrier: Function, structure and physiopathology," *Mol Aspects Med*, vol. 32, no. 4–6, pp. 223–233, Aug. 2011, doi: 10.1016/J.MAM.2011.10.008.

[9] M. Dambrova *et al.*, "Acylcarnitines: Nomenclature, Biomarkers, Therapeutic Potential, Drug Targets, and Clinical Trials," *Pharmacol Rev*, vol. 74, no. 3, pp. 506–551, Jul. 2022, doi: 10.1124/PHARMREV.121.000408.

[10] L. Vettore, R. L. Westbrook, and D. A. Tennant, "New aspects of amino acid metabolism in cancer," *British Journal of Cancer 2019 122:2*, vol. 122, no. 2, pp. 150–156, Dec. 2019, doi: 10.1038/s41416-019-0620-5.

[11] N. Okamoto, Y. Miyagi, and A. Chiba, "(4) (PDF) Diagnostic modeling with differences in plasma amino acid profiles between non-cachectic colorectal/breast cancer patients and healthy individuals," Nov. 2008.

https://www.researchgate.net/publication/253715455_Diagnostic_modeling_with_differences_in_plasma_amino_acid_profiles_between_non-cachectic_colorectalbreast_cancer_patients_and_healthy_individuals (accessed Feb. 19, 2023).

[12] K. Wakefield, "A guide to the types of machine learning algorithms | SAS UK." https://www.sas.com/en_gb/insights/articles/analytics/machine-learning-algorithms.html (accessed Jan. 19, 2023).

[13] I. Ibrahim and A. Abdulazeez, "The role of machine learning algorithms for diagnosing diseases," *jastt.org*, vol. 02, no. 01, pp. 10–19, 2021, doi: 10.38094/jastt20179.

[14] "Machine Learning basics – Brian Omondi Asimba," Apr. 18, 2019. https://brianasimba.github.io/MachineLearningblog//Introduction-post/ (accessed Jan. 19, 2023).

[15] "ML | Types of Learning - Part 2 - GeeksforGeeks," Jan. 11, 2023. https://www.geeksforgeeks.org/ml-types-learning-part-2/ (accessed Jan. 19, 2023).

[16] M. Fatima and M. Pasha, "Survey of Machine Learning Algorithms for Disease Diagnostic," *Journal of Intelligent Learning Systems and Applications*, vol. 09, no. 01, pp. 1–16, 2017, doi: 10.4236/JILSA.2017.91001.

[17] "What Is Deep Learning? | How It Works, Techniques & Applications - MATLAB & Simulink." https://www.mathworks.com/discovery/deep-learning.html (accessed Jan. 26, 2023).

[18] "What is Deep Learning? | IBM." https://www.ibm.com/topics/deep-learning (accessed Jan. 26, 2023).

[19] "What is Unsupervised Learning? | IBM." https://www.ibm.com/topics/unsupervised-learning (accessed Jan. 31, 2023).

[20] "Fuzzy C-Means Clustering - MATLAB & Simulink." https://www.mathworks.com/help/fuzzy/fuzzy-c-means-clustering.html (accessed Jan. 31, 2023).

[21] "Unsupervised Machine Learning: Algorithms, Types with Example." https://www.guru99.com/unsupervised-machine-learning.html (accessed Jan. 31, 2023).

[22] C. H. Chen, "A hybrid intelligent model of analyzing clinical breast cancer data using clustering techniques with feature selection," *Appl Soft Comput*, vol. 20, pp. 4–14, Jul. 2014, doi: 10.1016/J.ASOC.2013.10.024.

[23] K. Polat, “Classification of Parkinson’s disease using feature weighting method on the basis of fuzzy C-means clustering,” *http://dx.doi.org/10.1080/00207721.2011.581395*, vol. 43, no. 4, pp. 597–609, Apr. 2011, doi: 10.1080/00207721.2011.581395.

[24] M. Nilashi, O. Ibrahim, and A. Ahani, “Accuracy Improvement for Predicting Parkinson’s Disease Progression,” *Scientific Reports 2016 6:1*, vol. 6, no. 1, pp. 1–18, Sep. 2016, doi: 10.1038/srep34181.

[25] Y. Wu, H. Duan, and S. Du, “Multiple fuzzy c-means clustering algorithm in medical diagnosis,” *Technol Health Care*, vol. 23 Suppl 2, pp. S519–S527, Jun. 2015, doi: 10.3233/THC-150989.

[26] L. Trevithick, J. Painter, and P. Keown, “Mental health clustering and diagnosis in psychiatric in-patients,” *BJPsych Bull*, vol. 39, no. 3, pp. 119–123, Jun. 2015, doi: 10.1192/PB.BP.114.047043.

[27] N. Yilmaz, O. Inan, and M. S. Uzer, “A new data preparation method based on clustering algorithms for diagnosis systems of heart and diabetes diseases,” *J Med Syst*, vol. 38, no. 5, May 2014, doi: 10.1007/S10916-014-0048-7.

[28] J. B. Nikas and W. C. Low, “Application of clustering analyses to the diagnosis of Huntington disease in mice and other diseases with well-defined group boundaries,” *Comput Methods Programs Biomed*, vol. 104, no. 3, Dec. 2011, doi: 10.1016/J.CMPB.2011.03.004.

[29] H. Alashwal, M. El Halaby, J. J. Crouse, A. Abdalla, and A. A. Moustafa, “The application of unsupervised clustering methods to Alzheimer’s disease,” *Front Comput Neurosci*, vol. 13, p. 31, May 2019, doi: 10.3389/FNCOM.2019.00031/BIBTEX.

[30] V. Prabhu, “TechDay - Top 5 Machine Learning Libraries Today.” https://techdayhq.com/community/articles/top-5-machine-learning-libraries-today (accessed Dec. 11, 2022).

[31] M. N. Gevorkyan, A. V Demidova, T. S. Demidova, and A. A. Sobolev, “Computer Science and Computer Engineering Review and comparative analysis of machine learning libraries for machine learning”, doi: 10.22363/2658-4670-2019-27-4-305-315.

[32] M. N. Gevorkyan, A. V Demidova, T. S. Demidova, and A. A. Sobolev, “Computer Science and Computer Engineering Review and comparative analysis of machine learning libraries for machine learning”, doi: 10.22363/2658-4670-2019-27-4-305-315.

[33] P. Bellieveau, A. Griffin, and S. Somermeyer, *The PDMA ToolBook 1 for New Product Development - Google Livros*. Accessed: Jan. 06, 2023. [Online]. Available: https://books.google.pt/books?hl=pt-PT&lr=&id=kqX5EvT2U8AC&oi=fnd&pg=PA5&dq=Fuzzy+Front+End:+Effective+Methods,+Tools,+and+Techniques&ots=8Lpl18xUkg&sig=pY1UpWpy7quozYYZIHgG1weSj2E&r

edir_esc=y#v=onepage&q=Fuzzy%20Front%20End%3A%20Effective%20Methods%2C%20Tools%2C%20and%20Techniques&f=false

[34] "CHPORTO - Centro Hospitalar Universitário do Porto." https://www.chporto.pt/v0B0A/apresentacao (accessed Jan. 06, 2023).

[35] V. A. Zeithmal, "Zeithaml 1988 | PDF | Perception | Brand," 1988. https://pt.scribd.com/document/445042618/zeithaml1988# (accessed Jan. 13, 2023).

[36] Redator Rock Content, "Proposta de valor: o que é e como criar a proposta perfeita." https://rockcontent.com/br/blog/proposta-de-valor/ (accessed Jan. 13, 2023).

[37] P. Koen *et al.*, "Providing Clarity and A Common Language to the 'Fuzzy Front End,'" *http://dx.doi.org/10.1080/08956308.2001.11671418*, vol. 44, no. 2, pp. 46–55, 2016, doi: 10.1080/08956308.2001.11671418.

[38] A. Dalpati and A. A. Chouksey Dalpati, "AN APPLICATION OF QUALITY FUNCTION DEPLOYMENT: A CASE OF GYMNASIUM," 2018, doi: 10.26488/IEJ.10.8.58.

[39] Gomes. Pedro, "Pré-Processamento de Dados | Conheça as Técnicas e as Etapas!," Dec. 13, 2019. https://www.datageeks.com.br/pre-processamento-de-dados/ (accessed Jun. 01, 2023).

[40] M. Rajaratne, "Data Pre Processing Techniques You Should Know | by Maneesha Rajaratne | Towards Data Science," Dec. 02, 2018. https://towardsdatascience.com/data-pre-processing-techniques-you-should-know-8954662716d6 (accessed Jun. 01, 2023).

[41] E. Ryzhkov, "5 Stages of Data Preprocessing for K-means clustering | by Evgeniy Ryzhkov | Medium," Jun. 23, 2020. https://medium.com/@evgen.ryzhkov/5-stages-of-data-preprocessing-for-k-means-clustering-b755426f9932 (accessed Jun. 01, 2023).

[42] Y. Singh, "Robust Scaling: Why and How to Use It to Handle Outliers | Proclus Academy," Mar. 22, 2022. https://proclusacademy.com/blog/robust-scaler-outliers/ (accessed Jun. 15, 2023).

[43] S. Karanam, "Curse of Dimensionality — A 'Curse' to Machine Learning | by Shashmi Karanam | Towards Data Science," Aug. 11, 2021. https://towardsdatascience.com/curse-of-dimensionality-a-curse-to-machine-learning-c122ee33bfeb (accessed May 30, 2023).

[44] J. G. Dy and C. E. Brodley, "Feature Selection for Unsupervised Learning," *Journal of Machine Learning Research*, vol. 5, pp. 845–889, 2004.

[45] I. Lindgren, "Dealing with Highly Dimensional Data using Principal Component Analysis (PCA)," Apr. 24, 2020. https://towardsdatascience.com/dealing-with-highly-

dimensional-data-using-principal-component-analysis-pca-fea1ca817fe6 (accessed Jun. 28, 2023).

[46] C. Wijaya, “5 Feature Selection Method from Scikit-Learn you should know | by Cornellius Yudha Wijaya | Towards Data Science,” Mar. 08, 2021. https://towardsdatascience.com/5-feature-selection-method-from-scikit-learn-you-should-know-ed4d116e4172 (accessed Jun. 05, 2023).

[47] Bex T., “How to Use Variance Thresholding For Robust Feature Selection | by Bex T. | Towards Data Science,” Apr. 10, 2021. https://towardsdatascience.com/how-to-use-variance-thresholding-for-robust-feature-selection-a4503f2b5c3f (accessed Jun. 05, 2023).

[48] S. Loukas, “PCA clearly explained —When, Why, How to use it and feature importance: A guide in Python | by Serafeim Loukas, PhD | Towards Data Science,” May 30, 2020. https://towardsdatascience.com/pca-clearly-explained-how-when-why-to-use-it-and-feature-importance-a-guide-in-python-7c274582c37e (accessed May 30, 2023).

[49] “ML | T-distributed Stochastic Neighbor Embedding (t-SNE) Algorithm - GeeksforGeeks.” https://www.geeksforgeeks.org/ml-t-distributed-stochastic-neighbor-embedding-t-sne-algorithm/ (accessed May 30, 2023).

[50] K. J. Wong, “7 Evaluation Metrics for Clustering Algorithms | by Kay Jan Wong | Towards Data Science,” Dec. 09, 2022. https://towardsdatascience.com/7-evaluation-metrics-for-clustering-algorithms-bdc537ff54d2 (accessed May 30, 2023).

[51] D. Dey, “Dunn index and DB index - Cluster Validity indices | Set 1 - GeeksforGeeks,” Feb. 19, 2022. https://www.geeksforgeeks.org/dunn-index-and-db-index-cluster-validity-indices-set-1/ (accessed May 30, 2023).

[52] T. Dorfer, “How to Evaluate Clustering Performance without Ground Truth Labels | by Thomas A Dorfer | Towards Data Science,” Dec. 01, 2022. https://towardsdatascience.com/how-to-evaluate-clustering-performance-without-ground-truth-labels-9c9792ec1c54 (accessed May 30, 2023).

[53] S. Mehta, “A tutorial on various clustering evaluation metrics,” *A tutorial on various clustering evaluation metrics*, Mar. 23, 2022. Accessed: May 30, 2023. [Online]. Available: https://analyticsindiamag.com/a-tutorial-on-various-clustering-evaluation-metrics/

[54] jaintarun, “Clustering Performance Evaluation in Scikit Learn - GeeksforGeeks,” Nov. 07, 2022. https://www.geeksforgeeks.org/clustering-performance-evaluation-in-scikit-learn/ (accessed May 30, 2023).

[55] LEDU, “Understanding K-means Clustering in Machine Learning,” Sep. 12, 2018. https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1 (accessed May 20, 2023).

[56] H. Köhn and L. J. Hubert, “Hierarchical Cluster Analysis,” in *Wiley StatsRef: Statistics Reference Online*, Wiley, 2015, pp. 1–13. doi: 10.1002/9781118445112.stat02449.pub2.

[57] alokesh985, “ML | BIRCH Clustering.” https://www.geeksforgeeks.org/ml-birch-clustering/ (accessed May 20, 2023).

[58] C. Maklin, “https://towardsdatascience.com/machine-learning-birch-clustering-algorithm-clearly-explained-fb9838cbeed9 ,” Jul. 01, 2019. https://towardsdatascience.com/machine-learning-birch-clustering-algorithm-clearly-explained-fb9838cbeed9 (accessed May 20, 2023).

[59] A. Singh, “Build Better and Accurate Clusters with Gaussian Mixture Models,” Oct. 31, 2019. https://www.analyticsvidhya.com/blog/2019/10/gaussian-mixture-models-clustering/ (accessed May 20, 2023).

[60] N. Doshi, “Spectral clustering,” Feb. 04, 2019. https://towardsdatascience.com/spectral-clustering-82d3cff3d3b7 (accessed May 20, 2023).

[61] Yufeng, “Understanding Mean Shift Clustering and Implementation with Python,” Feb. 22, 2022. https://towardsdatascience.com/understanding-mean-shift-clustering-and-implementation-with-python-6d5809a2ac40 (accessed May 20, 2023).