COIMBRA BUSINESS SCHOOL

iscac Politécnico de Coimbra

100 ANOS

Similaridade em Linhas Celulares nos Sistemas de Recomendação Farmacológicos para o Tratamento Oncológico

Bruno Alexandre Cordeiro Bento

ISCAC | 2023

COIMBRA BUSINESS SCHOOL

iscac Politécnico de Coimbra

100 ANOS

Bruno Alexandre Cordeiro Bento

Similaridade em Linhas Celulares nos Sistemas de Recomendação Farmacológicos para o Tratamento Oncológico

Coimbra, outubro de 2023

Bruno Alexandre Cordeiro Bento

Similaridade em Linhas Celulares nos Sistemas de Recomendação Farmacológicos para o Tratamento Oncológico

Trabalho de projeto submetido ao Instituto Superior de Contabilidade e Administração de Coimbra para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Análise de Dados e Sistemas de Apoio à Decisão, realizado sob a orientação do Professor Doutor Fernando Paulo Belfo e coorientação do Professor Doutor António Trigo.

Coimbra, outubro de 2023

TERMO DE RESPONSABILIDADE

Declaro ser o autor deste projeto, que constitui um trabalho original e inédito, que nunca foi submetido a outra Instituição de ensino superior para obtenção de um grau académico ou outra habilitação. Atesto ainda que todas as citações estão devidamente identificadas e que tenho consciência de que o plágio constitui uma grave falta de ética, que poderá resultar na anulação do presente projeto.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: presidencia@iscac.pt; www.iscac.pt

PENSAMENTO

"Só é útil o conhecimento que nos torna melhores."

Sócrates

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: presidencia@iscac.pt; www.iscac.pt

AGRADECIMENTOS

A toda a minha família e amigos, pelo apoio e paciência que tiveram comigo ao longo desta jornada, apoiando e celebrando cada passo do meu caminho, motivando e incentivando a minha caminhada para conseguir atingir todos os meus objetivos, pois sem eles nada disto tinha sido possível. Com todas as suas palavras de encorajamento, motivação e força, foram verdadeiramente o alicerce que impulsionou a manutenção do foco nos meus objetivos.

Ao meu orientador, Professor Doutor Fernando Paulo Belfo, e coorientador, Professor Doutor António Trigo, por toda a ajuda, disponibilidade, persistência e partilha de conhecimento em todas as fases deste estudo e por toda a confiança depositada no meu trabalho e reconhecimento de todo o meu esforço.

A todos os professores da parte letiva do Mestrado, por toda a partilha de conhecimento e inspiração que me deram para seguir em frente na obtenção dos meus objetivos.

Aos meus colegas de Mestrado, por toda a motivação e partilha de experiências ao longo de toda a parte letiva deste Mestrado.

À Coimbra Business School – ISCAC e a toda a sua estrutura, pondo à disposição dos seus discentes todos os meios necessários para a obtenção do conhecimento.

A cada uma destas pessoas, não posso deixar de expressar a minha mais profunda gratidão, por fazerem parte desta minha jornada, e por estarem sempre presentes, apoiando e celebrando cada passo deste meu percurso.

RESUMO

Nas últimas décadas a área da saúde tem-se focado na busca de respostas, cada vez mais personalizadas, para o tratamento das mais variadas patologias. Neste caminho encontra-se o doente oncológico, diferenciando-se dos demais pela complexidade da sua patologia. Neste sentido têm surgido novas disciplinas como: a Bioinformática, a Farmacogenómica, o *Machine Learning*, o Data Mining, a Genómica, entre outras. A descoberta do sequenciamento genético tem avanços muito significativos nestas áreas, permitindo cada vez mais praticar a chamada medicina de precisão e individualizada para cada doente. Ou seja, cada vez mais o doente é tratado de forma individualizada, com uma determinada patologia, e não um grupo de doentes com características distintas, que detêm a mesma patologia.

Será estudada a similaridade entre linhas celulares, tendo por base os Sistemas de Recomendação (RecSys), para o tratamento do doente oncológico. Na implementação deste projeto usar-se-á a metodologia Cross-Industry Standard Process for Data Mining (CRISP-DM), onde serão abordadas métricas de similaridade e algoritmos de *machine learning*, por forma a responder à identificação da similaridade entre linhas celulares. O *dataset* usado foi o do *Genomics of Drug Sensitivity in Cancer (GDSC1)*, tendo-se selecionado uma amostra de 20 linhas celulares (10 amostras referentes à patologia da mama e 10 amostras referentes a patologias da pele), com 49386 genes cada, dado os recursos de *hardware*. Para avaliar a similaridade da expressão génica entre estas linhas celulares, serão aplicadas métricas de similaridade, para avaliar 3 genes de uma amostra das 20 linhas celulares, e por outro lado os algoritmos de *machine learning* onde serão avaliados os 49386 genes de cada amostra das 20 linhas celulares. Assim as métricas de similaridade testadas foram as distâncias de Dice, Jaccard, Sorensen, Czekanowski, Minkowski, Pearson, Intersection, Manhattan, Tanimoto e Euclideana. Na parte dos algoritmos de *machine learning* foram testados*:* Rede Neural Artificial, *Logistic regression, Linear discriminant analysis, K-Nearest Neighbors, DecisionTreeClassifier, Gaussian NB e Support vector machine*. Como conclusão dos resultados obtidos, as distâncias de similaridade com melhores resultados foram Jaccard e Dice, uma vez que apresentaram os resultados mais consistentes para os dois genes selecionados sendo que num dos genes os resultados ainda foram mais consistentes, já os

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: presidencia@iscac.pt; www.iscac.pt

algoritmos que apresentaram uma melhor *accuracy* foram *Logistic Regression*, *Linear Discriminant Analysis* e *Gaussian NB.*

Palavras-chave: algoritmo de similaridade; distâncias de similaridade; linha celular; tratamento oncológico; sistema de recomendação; GDSC; DNA; microarray; machine learning.

ABSTRACT

In recent decades, the health sector has focused on finding increasingly personalised responses to the treatment of the most varied pathologies. On this path lies the cancer patient, distinguished from others by the complexity of their pathology. In this sense, new disciplines have emerged, such as Bioinformatics, Pharmacogenomics, Machine Learning, Data Mining, Genomics, among others. The discovery of genetic sequencing has led to very significant advances in these areas, making it increasingly possible to practise so-called precision medicine that is individualised for each patient. In other words, more and more patients are treated on an individual basis, with a specific pathology, rather than a group of patients with different characteristics who have the same pathology.

Similarity between cell lines will be studied, based on Recommendation Systems (RecSys) for the treatment of cancer patients. This project will be implemented using the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, which will use similarity metrics and machine learning algorithms to identify similarities between cell lines. The dataset used was from the Genomics of Drug Sensitivity in Cancer Project (GDSC1). A sample of 20 cell lines was selected (10 samples relating to breast pathology and 10 samples relating to skin pathologies), with 49386 genes each, given the limited hardware resources. To assess the similarity of gene expression between these cell lines, similarity metrics will be applied to assess 3 genes from a sample of the 20 cell lines, while machine learning algorithms will be used to assess the 49386 genes from each sample of the 20 cell lines. The similarity metrics tested were Dice, Jaccard, Sorensen, Czekanowski, Minkowski, Pearson, Intersection, Manhattan, Tanimoto and Euclidean distances. The following machine learning algorithms were tested: Artificial Neural Network, Logistic regression, Linear discriminant analysis, K-Nearest Neighbours, DecisionTreeClassifier, Gaussian NB and Support vector machine. As a conclusion of the results obtained, the similarity distances with the best results were Jaccard and Dice, since they showed the most consistent results for the two selected genes, with one of the genes still showing the best results were more consistent, while the algorithms with the best accuracy were Logistic Regression, Linear Discriminant Analysis and Gaussian NB.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: presidencia@iscac.pt; www.iscac.pt

Keywords: similarity algorithm; similarity distances; cell line; cancer treatment; recommendation system; GDSC; DNA; microarray; machine learning.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: presidencia@iscac.pt; www.iscac.pt

ÍNDICE GERAL

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: presidencia@iscac.pt; www.iscac.pt

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: presidencia@iscac.pt; www.iscac.pt

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: presidencia@iscac.pt; www.iscac.pt

ÍNDICE DE TABELAS

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: presidencia@iscac.pt; www.iscac.pt

ÍNDICE DE FIGURAS

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: presidencia@iscac.pt; www.iscac.pt

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: presidencia@iscac.pt; www.iscac.pt

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: presidencia@iscac.pt; www.iscac.pt

LISTA DE ABREVIATURAS, ACRÓNIMOS E SIGLAS

cDNA	Complementary DNA
CRISP-DM	Cross-Industry Standard Process for Data Mining
Cy	Cyanine dye
DNA	DeoxyriboNucleic Acid
EBI	European Bioinformatics Institute
EFO	Experimental Factor Ontology
GDSC	Genomics of Drug Sensitivity in Cancer
KNN	K-Nearest Neighbors
LDA	Linear Discriminant Analysis
MCDT	Meios Complementares de Diagnóstico e Terapêutico
MM	Mismatch
mRNA	Messenger RNA
MSE	Mean Square Error
ML	Machine learning
OMS	Organização Mundial da Saúde
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PET-CT	Positron Emission Tomography – Computed Tomography
PM	Perfect match
RecSys	Recommender Systems
RMA	Robust Multi-array Average
RNA	Ribonucleic Acid
RNA-Seq	RNA sequencing
SSIM	Structural Similarity Index Measure
SVM	Support Vector Machine

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: presidencia@iscac.pt; www.iscac.pt

1 INTRODUÇÃO

Ao longo dos tempos, a comunidade médica e científica tem-se focado na procura de novas abordagens clínicas cada vez mais personalizadas, precisas e sem efeitos adversos para o doente. Subsistindo um esforço em maximizar a compreensão da área genómica e potenciar as ferramentas criadas a partir da bioinformática. Nestas áreas, a capacidade de processamento de grandes quantidades de dados biológicos, através do *Big Data*, contou com uma evolução das tecnologias de sequenciamento, o crescimento da capacidade de processamento e a redução do seu custo (Bento et al., 2023). Por outro lado, a existência de maior facilidade na investigação dos mecanismos da genómica, proteómica, transcriptómica, levou ao surgimento de novos fármacos e terapias utilizadas nas áreas da saúde e da biomedicina (Subramanian et al., 2020).

A bioinformática tem um papel crucial nos avanços da ciência biológica e da biotecnologia. Esta disciplina consiste na abordagem de métodos computacionais no processamento de dados biológicos, focando-se principalmente naqueles derivados de técnicas de biologia molecular e bioquímica, como o sequenciamento de genomas, análise de estruturas de biomoléculas, compreensão de rotas metabólicas e regulação génica (Pevsner, 2015). Neste sentido destacam-se duas abordagens, por um lado a bioinformática estrutural tratando problemas biológicos e por outro os dados de prossecuções de nucleotídeos e aminoácidos (Pevsner, 2015). Na última década, a bioinformática tornou-se um ramo imprescindível na investigação científica moderna ("ciência de *Big Data*"), conseguindo-se extrair informação/conhecimento de grandes quantidades de dados através de supercomputadores com plataformas computacionais cada vez maiores (Pérez-Wohlfeil et al., 2018).

Para este paradigma muito têm contribuído várias disciplinas da ciência de dados, como é o caso do *Machine Learning* (ML) e do *Data Mining*, através da aplicação de algoritmos de ML à bioinformática abrangendo a genómica, a proteómica, os *microarrays*, a biologia de sistemas, e a mineração de texto (Libbrecht & Noble, 2015). Nesta área encontram-se estudos desenvolvidos para encontrar melhores tratamentos e novos fármacos

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

promissores, onde foram aplicadas algoritmos de ML com técnicas de predição (Mostavi et al., 2020). No entanto as adoções destas técnicas ainda encontram alguns desafios, pois está-se a tratar dados biológicos. Contrariamente a outras áreas aqui existe a dificuldade em interpretar os resultados que dai advêm, pois, os resultados têm de estar associados ao seu significado biológico. Outro problema a ter ainda em linha de conta é a alta dimensionalidade dos dados genómicos, dado o baixo número de amostras e alto número de *features* (Hambali et al., 2020).

Por outro lado, a farmacogenómica é uma das abordagens emergentes da medicina de precisão, adaptando e selecionando a dosagem de medicamento às características genéticas de cada paciente (Bento et al., 2023). O estudo da genética humana tem sido alimentado por tecnologias de sequenciamento de ponta que conduzem a uma compreensão mais precisa da relação entre variação genética e saúde humana. Nesta área novas abordagens serão necessárias investigar, incluindo o estudo do papel farmacogenómico da genética do sistema imunológico e de variantes genéticas raras anteriormente descuradas, muitas vezes relatados para responder por uma grande parte da variabilidade inter-individual no metabolismo de drogas (Cecchin & Stocco, 2020). Na atualidade existem vários projetos de investigação que mapeiam a resposta de linhas celulares relacionadas a uma determinada patologia, para uma vasta coleção de fármacos utilizados no seu tratamento. Nesse sentido são desenvolvidos diretrizes e projetos que visam aproximar estudos da prática clínica e explorar novas abordagens para a descoberta de informações sobre a interação entre genes e fármacos e os seus efeitos nos doentes. As instituições públicas e privadas também começam a explorar mais a farmacogenómica, uma vez que o progresso dessa investigação poderá não só trazer benefícios para os doentes como reduzir o custo dos tratamentos a longo prazo (Guo et al., 2019).

Uma das áreas da saúde que mais pode beneficiar desse tipo de investigação é a oncologia. As doenças oncológicas continuam a ter um grande peso no número total de mortes no mundo. Segundo a Organização Mundial da Saúde (OMS) o cancro continua a ser sem dúvida uma das principais causas de morte em todo o mundo, tendo contribuído em 2020 com quase 10 milhões de mortes (Gourd, 2020). A International Agency for Research on

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Cancer, da OMS, prevê que em 2070, os casos de cancro dupliquem face a 2020, ou seja este flagelo continuará a crescer nos próximos anos (Observador, 2022). Neste contexto será fulcral atuar de uma forma preventiva, de modo a garantir que os diagnósticos para a doença oncológica sejam realizados num estádio precoce, identificando o tamanho do tumor, a sua localização e envolvimento ganglionar, aumentando assim a taxa de sobrevida dos doentes. Como o doente deve estar no centro de qualquer sistema de saúde, as terapêuticas devem cada vez mais ser individualizadas, dirigidas e aceites numa decisão interdisciplinar, planeada e integrada.

Na área da saúde, e em particular na área oncológica, os Sistemas de Recomendação (RecSys) têm cada vez mais um papel preponderante no tratamento destas patologias, auxiliando na tomada de decisões clínicas, fornecendo informações relevantes aos profissionais de saúde e ajudando os pacientes a receberem cuidados de saúde cada vez mais personalizados. Estes sistemas podem ser aplicados em várias áreas da saúde, incluindo diagnóstico médico, tratamento, prevenção, monitorização de saúde assim como na gestão de dados clínicos. Através de técnicas de ML a área da saúde tem aumentado a sua precisão no tratamento de doentes com variadíssimas patologias (Ihnaini et al., 2021). Um dos temas fulcrais na abordagem dos RecSys, é a forma como será calculada a similaridade, ou seja, a similaridade desempenha um papel crucial nos RecSys, permitindo identificar itens ou utilizadores semelhantes e utilizar essa informação para fazer recomendações precisas e personalizadas. Ao encontrar a melhor métrica de similaridade, é possível aumentar a precisão nas recomendações, proporcionando maior fiabilidade, confiabilidade e relevância nas recomendações encontradas para os seus utilizadores.

Este projeto é composto por uma revisão da literatura, onde serão abordados temas como: genoma humano, imagem médica, RecSys, similaridade e imagens médicas, algoritmos de ML e *microarrays*. Seguindo-se a apresentação da metodologia usada. O entendimento dos dados começa com a apresentação do *dataset* "Projeto *Genomics of Drug Sensitivity in Cancer (GDSC1)*", de onde se selecionou aleatoriamente uma amostra de 20 linhas celulares com 49386 genes cada, das 970 linhas celulares cancerígenas, com 10 amostras

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

referentes à patologia da mama e 10 amostras referentes a patologias da pele (Cancer Genome Project et al., 2022). Para a análise desta amostra inicialmente obteve-se a expressão génica de cada amostra, elaborando-se seguidamente uma estatística descritiva para os dados. Testando-se ainda algumas técnicas de agrupamento como clusters, análise das componentes principais e diagrama de Venn. Na parte da modelação inicialmente foram encontrados os genes diferencialmente expressos através de um teste-t. Como se pretendeu encontrar uma métrica eficaz na avaliação da similaridade entre estas linhas celulares, foram seguidas duas abordagens, uma com as distâncias de Dice, Jaccard, Sorensen, Czekanowski, Minkowski, Pearson, Intersection, Manhattan, Tanimoto e Euclideana, para avaliar 3 genes de uma amostra das 20 linhas celulares de similaridade, e outra com os algoritmos de ML Rede Neural Artificial, *Logistic regression, Linear discriminant analysis (LDA), K-Nearest Neighbors (KNN), DecisionTreeClassifier, Gaussian NB e Support vector machine (SVM),* para avaliar os 49386 genes de cada amostra das 20 linhas celulares. Posteriormente seguiu-se a avaliação dos resultados apresentados, com as métricas mais relevantes para o presente estudo. Finalizando-se com a conclusão de todo este estudo. Em suma, este projeto tem como objetivo demonstrar a importância dos RecSys no tratamento do doente oncológico baseado na similaridade das suas linhas celulares. Neste sentido o foco está na compreensão e avaliação das métricas de similaridade ligadas aos RecSys.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

2 REVISÃO DE LITERATURA

A revisão da literatura tem como objetivo expor o estado da arte nas várias temáticas abordadas neste estudo. Este capítulo aborda temas desde o genoma humano, a imagem médica, os RecSys, a similaridade entre imagens médicas, os algoritmos de ML até aos *microarrays*.

2.1 Genoma humano

Nas últimas décadas o forte investimento na genómica e biologia molecular tem aumentado as fortes expectativas na medicina de precisão e no tratamento personalizado de patologias. Em 1953 a descoberta da estrutura química do DeoxyriboNucleic Acid (DNA), por James D. Watson, Francis Crick e Maurice Wilkins, abriu portas para o novo mundo na área da Biologia. Hoje existe a possibilidade de tratamento e prevenção de patologias, cujo organismo futuramente tenha tendência para desenvolver, através do sequenciamento genético (Albert Einstein, 2021). Esta descoberta revelou um passo importante para a descoberta dos segredos da vida humana (Ariel & Moraes, 2016). Penchaszadeh (2004) defendeu que "a descoberta do genoma humano abre uma série de possibilidades para melhorar a compreensão das interações ambiente-genes na causa de doenças humanas".

Por outro lado, Fogle (1990), efetuou uma diferenciação do conceito mendeliano de unidade hereditária do conceito molecular clássico. Assim segundo o Projeto Genoma Humano, o gene pode ser definido como uma unidade hereditária possuidora de estrutura, função e localização (Fogle, 2010). Neste sentido o genoma humano detém toda a informação genética que cada individuo carrega. Em suma a análise genética vai para além de toda a carga hereditária de cada individuo, incluindo também toda a propensão para determinadas doenças. Nesta linha Born & Oliveira (2001) evidenciou ainda que:

"[...] é o compêndio de toda a herança genética herdada de seus pais, pelo ser humano, no momento da conceção. É a herança genética que contém as instruções do que virá a acontecer ao longo da existência do ser humano: qual será sua estatura, a cor da sua pele,

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

seus cabelos, se será afligido por doenças fatais e por quais delas, dentre outras tantas informações".

Passando agora para o sequenciamento completo do genoma, sendo o processo de determinar a sequência completa de DNA do genoma de um organismo num determinado instante. Envolvendo o sequenciamento de todo o DNA cromossômico do organismo, bem como o DNA contido nas mitocôndrias e, o DNA do cloroplasto. Por outro lado, o sequenciamento do genoma completo tem sido amplamente utilizado pela comunidade científica, como uma ferramenta de pesquisa em estudos funcionais de associação, melhorando o conhecimento disponível para os investigadores na área da biologia, tendo sido introduzido na prática clínica (Ng et al., 2009; Rauch et al., 2006; D. G. Wang et al., 1998). No futuro numa medicina personalizada, os dados de sequenciamento do genoma completo, podem ser uma ferramenta orientadora para as abordagens de tratamento (Mooney, 2015). Ferramentas de sequenciamento de genes no nível de *Single-Necleotide Plymorphism*, também são usadas para identificar variantes funcionais de estudos de associação e melhorar o conhecimento disponível para pesquisadores interessados em biologia.

As patologias oncológicas, dado o elevado número de óbitos, têm tido prioridade na implementação da genómica ligada ao seu diagnóstico. Nesta área a medicina de precisão ganha cada vez mais adeptos, pois são várias as vantagens na sua aplicação. De acordo com o estudo *Global Cancer Statistics* (Hyuna et al., 2021), estima-se um aumento global de quase 50% de novos casos de cancro entre 2020 e 2046. Sendo este um preço a pagar pelo envelhecimento geral da população nos países desenvolvidos. O diagnóstico de um doente oncológico na medicina dita tradicional, é feito através de Meios Complementares de Diagnóstico Terapêutico (MCDT) laboratoriais, histológicos e biópsias, sendo posteriormente e segundo as *guidelines* desta área, encaminhados para cirurgia, tratamento quimioterapia e/ou radioterapia. Por outro lado, a medicina de precisão/personalizada detém uma abordagem mais emergente olhando ao mesmo tempo para o diagnóstico, o tratamento e o perfil genético do doente (sinais, sintomas, histórico pessoal e familiar, e exames complementares). Esta abordagem proporciona uma

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

capacidade de predição, maior *accuracy* no diagnóstico, antevendo o surgimento de patologias hereditárias e com predisposição genética. Nestes pontos testes genéticos podem facilitar o diagnóstico de patologias oncológicas no futuro salvando vidas aos doentes. Pois a deteção de um tumor antecipadamente pode ser um fator decisor de salvar uma vida ou não…

2.2 Imagem médica

A imagem médica é uma técnica fiável e fundamental na obtenção de um diagnóstico atempado, permitindo ao médico determinar o estádio em que uma doença se encontra (Jacob et al., 2000). Nos últimos anos têm aparecido um número crescente de inovações na área dos meios complementares de diagnóstico, por meio de imagens médicas. Consequentemente, a comunidade científica tem verificado avanços tecnológicos que permitem a produção de imagens de maior resolução, bem como métodos de análise de imagens médicas que permitem a extração de novas e melhores informações. Uma das principais áreas de pesquisa é a aplicação da inteligência artificial nas imagens médicas, utilizando a racionalidade no diagnóstico médico, sendo as imagens médicas o ponto de partida (Pereira, 2021). Cada vez mais o médico é auxiliado por algoritmos de inteligência artificial que permitem efetuar um diagnóstico médico com maior precisão, melhorando assim a relação entre a Imagiologia e a ciência de dados. De acordo com a OMS a doença oncológica é a segunda maior causa de mortalidade no mundo (Hyuna et al., 2021), sendo assim de extrema importância estreitar cada vez mais a ligação entre as tecnologias de imagem médica e a inteligência artificial (Kuhl, 2015). Estas novas abordagens irão reduzir o tempo de espera do doente, aumentar a rapidez de reposta do médico a casos urgentes, agilizar a interpretação e emissão de relatórios médicos em tempo real, aumentando assim o grau de confiança no diagnóstico médico. Estes serão os desafios que a imagiologia irá e já está a passar, para que definitivamente seja inserida como uma disciplina da medicina de precisão na avaliação multidisciplinar do doente.

Nas últimas décadas a aplicação de novas tecnologias em saúde, tem melhorado a taxa de mortalidade, designadamente na área oncológica (Silva, 2003). Métricas quantitativas

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

como custo-benefício, eficácia, graus de evidências, e acurácia têm sido ferramentas metodológicas na ajuda no diagnóstico e tratamento de determinadas patologias, facilitando assim a aplicação e o desenvolvimento de tecnologia em dispositivos médicos. Neste contexto pode assegurar-se que o uso racional de métodos tecnológicos inovadores poderão aumentar a qualidade de vida dos doentes (Muller, 2017; Silveira et al., 2017). Por outro lado, e segundo Kadir e Gleeson (2018) os algoritmos de ML poderiam também ser usados para reduzir o número de nódulos benignos que são desnecessariamente acompanhados, consumindo escusadamente vários recursos aos sistemas de saúde.

No caso do cancro, as imagens radiológicas servem para identificar e localizar o tumor e determinar as zonas pelas quais o mesmo se disseminou. O diagnóstico radiológico complementa a informação obtida através de biópsias e de procedimentos cirúrgicos invasivos. Pois nenhum estudo radiológico do cancro poderá ser por si só um diagnóstico definitivo (Fundação Portuguesa do Pulmão, 2020). Posteriormente a classificação por tipo e estadio será efetuada através de estudos de patologia molecular, só depois disto é que o médico oncologista estará munido de toda a informação para empregar a terapêutica mais indicada para o doente, em consonância com as *guidelines* terapêuticas mais atuais. Por outro lado, o cancro poderá ser de difícil diagnóstico na sua fase inicial, uma vez que pode desenvolver-se durante muito tempo sem a manifestação de qualquer sintoma. Assim o surgimento dos primeiros sintomas pode resultar do crescimento do tumor, de uma invasão loco-regional, duma metastização sistémica, duma comorbilidade do doente ou de sintomas paraneoplásicos. Em suma, a imagem médica é e será uma abordagem não invasiva, que possibilita coadjuvar o médico no rápido e eficaz diagnóstico com redução de custos nos sistemas de saúde.

2.3 Sistemas de recomendação (RecSys)

Num mundo exigente e cada vez mais informação os dados têm aumentado o seu grau de acuidade, desenvolvendo-se assim mecanismos que simplifiquem a filtragem da informação, consoante as preferências/necessidades dos seus utilizadores. Os RecSys podem ser apresentados como softwares que sugerem itens aos seus utilizadores, tendo

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

como referência o histórico das interações ou através de métricas de similaridade entre os itens a sugerir, ou ainda uma combinação entre ambos, tal como ilustrado na Figura infra (Azambuja et al., 2021).

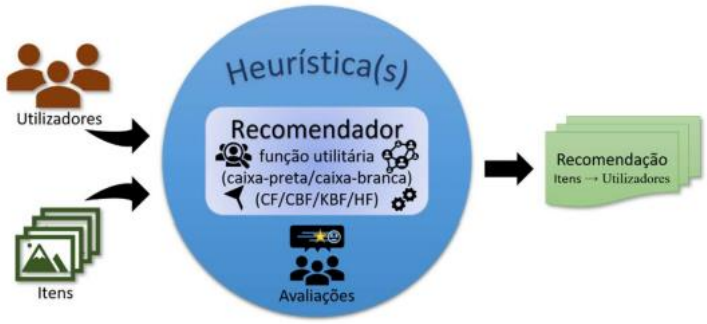


Figura 1 - Exemplo da estrutura básica em RecSys.

Fonte: Azambuja et al. (2021)

Estes sistemas estão muito centrados na web, designadamente no e-commerce, onde de acordo com as preferências de cada consumidor são "sugeridos" / recomendados determinados bens/serviços. Neste contexto as áreas de ML em cada vez mais aplicabilidade, tanto no sector público como no privado. Desde logo na educação, nas finanças, na saúde, no marketing, na indústria ou até mesmo nas áreas de entretenimento, contribuindo para o impulsionamento do conhecimento e para uma tomada de decisão cada vez mais informada, definindo estratégias que permitem aumentar o desempenho das entidades (Aggarwal, 2016).

RecSys podem ser apresentados como técnicas de aprendizagem de máquina que filtram uma grande quantidade de dados, tendo adjacente informações dos utilizadores e itens (Takahashi & Jr, 2015). A partir dessas técnicas são aconselhados determinados itens a cada utilizador. Ou seja, de uma forma mais simplista os sistemas de recomendação são técnicas que fornecem sugestões para que os seus utilizadores possam tomar melhores decisões (Gorakala & Usuelli, 2015), dependendo do contexto onde a recomendação seja aplicada. Por outro lado, RecSys são sistemas que fornecem sugestões personalizadas de acordo com os interesses particulares dos seus utilizadores. Um RecSys para fornecer a sugestão de um item terá que seguir os seguintes passos: desde recolher a informação do

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

utilizador, até à filtragem e processamento dos dados, consoante a abordagem seguida. As suas técnicas poder-se-ão basear em:

a) conteúdo onde é usada a correlação entre o perfil do utilizador e os itens recomendados;
b) filtragem colaborativa que utilizam a correlação entre perfis de utilizadores diferentes e entre itens da mesma classe. Por outro lado Weitzel et al. (2010) referiu ainda que são processos de filtragem de informação de várias fontes de dados.
c) híbridas que têm em consideração tanto a correlação entre perfis de utilizadores diferentes e correlação entre utilizadores e itens. A utilização de RecSys tem fornecido uma abordagem eficiente para reduzir o esforço do utilizador em encontrar informações do seu interesse.

Na saúde, os RecSys podem ter como objetivo auxiliar os profissionais na tomada de decisão e na realização de algumas tarefas. Estes sistemas são muitas vezes utilizados para a recomendação automática do diagnóstico médico. Por outro lado, em muitos estudos relacionados com RecSys a filtragem colaborativa é usada em combinação com outros métodos computacionais (Pinheiro, 2013). Na área oncológica, o estudo biológico do cancro é complexo, estando iminentemente relacionado com a genética de cada doente, o que leva a que determinada linha terapêutica seja mais eficaz num doente do que noutro.

Nesse sentido, o problema poderá ser abordado com um RecSys apoiado num algoritmo de ML, em que, para um novo doente (com uma determinada linha genética) seja proposta pelo sistema uma determinada linha terapêutica, com os fármacos mais apropriados.

2.4 Similidade e imagens médicas

Na área da saúde muitos autores referem a imagem médica como uma ferramenta de elevada relevância na área dos MCDT, para o diagnóstico médico, o que tem levado a muitos estudos e trabalhos científicos para o tratamento e identificação cada vez melhor da imagem médica. Aqui ter-se-á de referir que a área oncológica tem tido um papel preponderante no impulsionamento desta área de investigação. Ao longo dos tempos

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

muitos autores já se debruçaram sobre estas temáticas, como por exemplo Diniz et al. (2016) em "Análise Temporal de Lesões em Mamografias Utilizando Índices de Similaridade", Jbeli et al. (2018) em "*Detection and Characterization of Subsolid Juxta-pleural Lung Nodule from CT Images*", C. Vinod e D Menaka (2021) em "*Computer Aided Detection of Nodule from Computed Tomography Images of Lung*" ou Silveira et al. (2017) em "Diferença entre Tomografia Computadorizada, Ressonância Magnética e Positron Emission Tomography – Computed Tomography (PET-CT) na Identificação de Lesões Tumorais". Estes e outros estudos usam a imagem médica para detetar nódulos tumorais, tendo para isso que analisar as várias características das imagens médicas. Um dos passos nestes métodos será a comparação e deteção de semelhanças entre imagens. Ou seja, há que detetar a similaridade entre imagens para fazer agrupamentos, por exemplo para separar imagens com e sem patologia oncológica. Segundo Maheswari e Geetha (2019), as métricas de similaridade são uma boa opção para tratar grandes volumes de dados, sobretudo quando as matrizes de avaliação são dispersas.

Passando agora à temática da similaridade entre imagens médicas, vários foram os autores que já usaram os índices/coeficientes de similaridade para comparar imagens médicas. Diniz et al. (2016) no seu estudo sobre "Análise Temporal de Lesões em Mamografias Utilizando Índices de Similaridade" usou os seguintes índices de similaridade:

- Jaccard
- Anderberg
- Czekanowsky
- Kulcyznski
- Ochiai

Já Karakus & Avci (2020) no seu estudo "A new image steganography method with optimum pixel similarity for data hiding in medical images" a fim de avaliar o desempenho da sua proposta *Genetic Algorithm-Optimum Pixel Similarity* (GA-OPS) foram utilizadas as seguintes métricas de similaridade:

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

- Mean Square Error
- Root Mean Square Error
- Peak Signal to Noise Ratio
- Structural Similarity Index Measure
- Universal Quality Index

Já Martínez-Martínez et al. (2013), no seu estudo "Analysis of several biomechanical models for the simulation of lamb liver behaviour using similarity coefficients from medical image", utilizaram apenas os seguintes coeficientes de similaridade:

- Jaccard similarity coefficient
- Hausdorff coefficient

Estes e outros autores usaram variadíssimas métricas de similaridade em seus estudos. O presente trabalho tem como objetivo principal identificar os índices de similaridades mais populares e estudar o índice de similaridade mais apropriado, para ser usado na identificação de linhas celulares num RecSys farmacológico para o tratamento oncológico. A fase seguinte consistirá em testar estes índices de similaridade, com o intuito de estudar o melhor índice para o nosso RecSys.

2.4.1 Grau de semelhança entre objetos

As técnicas de agrupamento têm como finalidade agrupar elementos por conjuntos, de acordo com o seu grau de semelhança. Mais em concreto os índices de similaridade são muito utilizados na biologia para fazer comparações entre os vários grupos de espécies. Apesar de muito se falar nestes índices, dada a necessidade de quantificar as relações entre conjuntos, estes foram já criados no século passado (Meyer, 2002). Sendo que alguns têm a abordagem da presença ou ausência de determinadas características, enquanto outros dão relevo à abundância dessas mesmas características (J. O. B. Diniz, 2015).

Alguns autores defendem que as semelhanças entre objetos são representadas em modelos geométricos, em que as dissemelhanças são representadas pelas distâncias entre pontos

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

(Tversky, 1977). Os índices de semelhança devem respeitar as propriedades das seguintes métricas: simetria, desigualdade triangular, diferenciabilidade de não idênticos e a indiferenciabilidade de idênticos. Segundo Aldenderfer e Blashfield (1984) os índices de semelhança/dissemelhança podem ser apresentados nas quatro seguintes categorias:

- Coeficientes de correlação:

Os coeficientes de correlação são das medidas de semelhança mais usadas nas ciências sociais, e em particular o coeficiente de correlação de Pearson (Reis, 2001):

$$r_{ij} = \frac{\sum_{v=1}^{p} (X_{iv} - \bar{X}_i)(X_{jv} - \bar{X}_j)}{\sqrt{\sum_v (X_{iv} - \bar{X}_i)^2 \sum_v (X_{jv} - \bar{X}_j)^2}}$$

O valor do coeficiente varia entre -1 e +1, onde zero indica que não existe correlação entre os objetos

- Medidas de distância:

As medidas de distância ou dissemelhança podem ser podem ser representadas por várias medidas, onde se desta a distância Euclideana (Reis, 2001):

$$d_{ij} = \sqrt{\sum_{v=1}^{p} (X_{iv} - X_{jv})^2}$$

- Coeficientes de associação:

Os coeficientes de associação são muito usados para avaliar a semelhança entre dois objetos, sendo que a literatura apresenta variadíssimos exemplos propostos por vários autores. Estes variam entre 0 e 1, sendo que 0 significa ausência de semelhança e 1 perfeitamente semelhantes. Um exemplo é o coeficiente de Jaccard (Reis, 2001):

$$J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

- Medidas de semelhança probabilística:

Nas medidas de semelhança probabilística medem-se os ganhos probabilísticos da informação, a partir das variáveis iniciais, e agrupam-se os dois indivíduos que menos

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

ganho de informação provoquem (Reis, 2001). Estas diferem das medidas anteriores, pois as anteriores não calculam um valor para a semelhança entre os objetos.

Os índices de similaridade têm um grande grau de importância quando estamos perante um grande número de algoritmos de processamento de imagens e reconhecimento de padrões (Bento et al., 2023). Estas métricas de análise de imagens podem ser abordadas de dois pontos de vista, por um lado as variações de intensidade e por outro as distorções geométricas. Ou seja, quando se refere que os índices estão baseados na intensidade assume-se que as imagens estão a ser analisadas na mesma escala, sendo a sua similaridade determinada a partir da intensidade dos pixéis correspondentes, por outro lado quando estão baseados na geometria a similaridade pode ser determinado pelas transformações geométricas entre os pixéis correspondentes (Sampat et al., 2009).

2.4.2 Índices de similaridade baseados na intensidade

Os índices de similaridade baseados na intensidade são usados para comparar duas imagens, quantificando a "sobreposição espacial" onde são aplicadas operações booleanas nas intensidades dos pixéis correspondentes. Estes índices penalizam as imagens que divergem apenas por um pixel. Se por um lado esta característica é desejável, no caso da segmentação de imagens muito densas, com muitos pixéis, o mesmo não acontece quando estamos a analisar estruturas lineares com poucos pixéis (Sampat et al., 2009).

2.4.2.1 Coeficiente de Dice

O Coeficiente de Similaridade de Dice é também conhecido por Índice de Sorensen–Dice ou simplesmente Coeficiente de Dice, designa-se como uma função estatística usada para medir a similaridade entre dois conjuntos de dados. Muitas vezes ouve-se falar neste índice para validar algoritmos de Inteligência Artificial na segmentação de imagens, porem este é um conceito muito mais amplo que poderá ser aplicado a conjuntos de dados para uma diversidade de aplicações, compreendendo o processamento de linguagem natural (Yao et al., 2020). O coeficiente de Dice calculado pela fórmula:

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

$$SDC(X,Y) = \frac{2\,|X \cap Y|}{|X| + |Y|}$$

O índice pode variar de 0 (quando X e Y são separados) a 1 (quando X e Y são iguais) onde X *e* Y são dois conjuntos e |X| é o número de itens de X.

2.4.2.2 Coeficiente de Jaccard

O Coeficiente de Similaridade de Jaccard ou simplesmente Índice de Jaccard é uma função estatística usada na avaliação da sobreposição em dois conjuntos de dados. Este coeficiente é semelhante ao Coeficiente de Dice, normalmente com aplicações diferentes e distinta representação matemática. Este índice é usado para variadíssimas tarefas de Inteligência Artificial, bem como para a avaliação da deteção de objetos, onde a área detetada pode ser entendida como um conjunto de pixéis. Quando usada para deteção de objetos, em algoritmos de Inteligência Artificial, toma a designação de métrica de intersecção sobre união (Yao et al., 2020). O coeficiente de Jaccard calculado pela fórmula:

$$J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

O índice pode variar de 0 (quando X e Y são separados) a 1 (quando X e Y são iguais) onde X *e* Y são dois conjuntos.

2.4.2.3 Índice de Simpson

O índice de Simpson é uma função estatística usada em ecologia para determinar a biodiversidade de espécies numa determinada região. Poder-se-á definir que a sua principal funcionalidade é resumir a representação dessa diversidade num único valor hábil de qualificar esta região como muito heterogênea ou uniforme. Segundo LYONS (2009).

Para uma amostra finita o índice de Simpson pode ser obtido ainda através da expressão infra:

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

$$D = \frac{\sum_{i=1}^{J} n_i\,(n_i - 1)}{N\,(N - 1)}$$

Os valores obtidos para o Índice de Simpson estão no intervalo entre 0 que representa diversidade infinita na amostra e 1 que representa que não há diversidade na amostra, onde ni é o número de indivíduos na espécie i, e N é o número total de indivíduos.

2.4.2.4 Mean Square Error

O Erro Quadrático Médio, conhecido em inglês como *Mean Squared Error* (MSE), é um método bastante simples, que tem a sua origem na probabilidade estatística. Segundo Wang e Bovik (2009) é uma métrica que serve para avaliar a qualidade de um sinal, entre imagens, comparando-os e atribuindo-lhes uma pontuação quantitativa que descreva a similaridade entre eles. O MSE calculado pela seguinte fórmula:

$$MSE(x, y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2$$

Para o MSE os valores de $x\ e\ y$ são imagens, N é o número de pixéis e onde $x_i\ e\ y_i$ são os valores dos i-ésimos pixéis em x e y.

2.4.2.5 Structural Similarity Index Meaure

O índice de similaridade estrutural, em inglês *Structural Similarity Index Measure* (SSIM), permite medir a semelhança entre duas imagens. Este índice surge em 2004 no artigo "Image Quality Assessment: From Error Visibility to Structural Similarity" (Wang et al., 2004). Quando se comparam duas imagens o erro quadrático médio (MSE) não será a melhor alternativa, apesar da simplicidade a sua implantação, este não é sinónimo de semelhança. O SSIM visa resolver esta deficiência considerando a extração da luminância, contraste e estrutura, tal como está ilustrado na Figura 2 - Diagrama do Sistema de Similaridade Estrutural. A maioria das técnicas de avaliação de qualidade de imagem baseia-se na quantificação do erro entre as imagens de referência e de amostra. Uma medida comum é quantificar a diferença no valor de cada pixel correspondente entre a amostra e a imagem de referência (por exemplo, usando o erro quadrático médio).

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

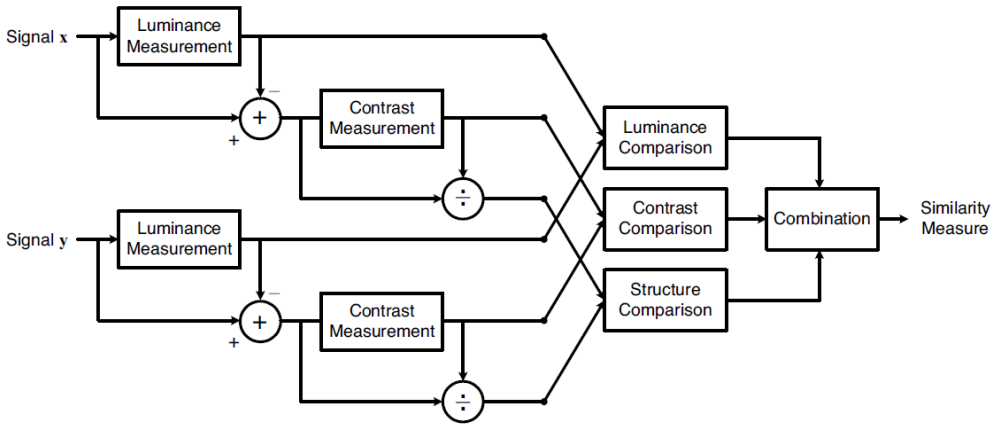


Figura 2 - Diagrama do Sistema de Similaridade Estrutural
Fonte: Wang et al. (2004)

Os sistemas de perceção visual humana são altamente capazes de reconhecer informações estruturais numa imagem, identificando assim a diferença entre as informações extraídas da imagem de referência e da imagem de amostra. Portanto, uma métrica que replica esse comportamento terá melhor desempenho em tarefas que envolvem a distinção entre amostras e imagens de referência. O SSIM calculado pela fórmula:

$$SSIM(X,Y) = \frac{(2\mu_X\mu_Y + C_1)\,(2\sigma_{XY} + C_2)}{(\mu_X^2 + \mu_Y^2 + C_1)\,(\sigma_X^2 + \sigma_Y^2 + C_2)}$$

O cálculo do SSIM varia num valor entre -1 e +1, sendo que um valor -1 indica que as duas imagens fornecidas são muito diferentes, e um valor de +1 indica que as duas imagens são semelhantes ou iguais. No entanto muitas vezes os valores são ajustados para [0, 1] sendo que os extremos têm o mesmo significado aos anteriormente explanados. Onde μ_X é a média de X, μ_Y é a média de Y, σ_X^2 é a variação de X, σ_y^2 é a variação de Y, σ_{XY} é a covariância de X e Y, e onde $C_1 = (K_1\,L)^2 e\;\; C_2 = (K_2\,L)^2$ *são* duas variáveis para estabilizar a divisão com denominador fraco.

2.4.2.6 Distâncias de similaridade

As medidas de distância desempenham um papel crucial em diversas áreas da matemática e da informática. Algumas dessas áreas incluem a Geometria, a Probabilidade e a

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Estatística, Teoria dos Grafos, Redes e Agrupamentos (*Clustering*), RecSys, Reconhecimento de Padrões, Visão Computacional, Computação Gráfica, Astronomia, Biologia Molecular, Física, entre outras (Deza & Deza, 2009).

Distância Dice	$D(P,Q) = 2\frac{\sum_{i=1}^{n}(P_i,Q_i)}{\sum_{i=1}^{n} P_i^2+\sum_{i=1}^{n} Q_i^2)}$
Distância Jaccard	$D(P,Q) = 2\frac{\sum_{i=1}^{n}(P_i,Q_i)}{\sum_{i=1}^{n} P_i^2+\sum_{i=1}^{n} Q_i^2-\sum_{i=1}^{n} P_iQ_i}$
Distância Sorensen	$D(P,Q) = \frac{2\lvert P\cap Q\rvert}{\lvert P\rvert+\lvert Q\rvert}$
Distância Czekanowski	$D(P,Q) = 2\frac{\sum_{i=1}^{n} min(P_i,Q_i)}{\sum_{i=1}^{n}(P_i+Q_i)}$
Distância Minkowski	$D(P,Q) = (\sum_{i=1}^{n} \lvert P_i - Q_i\rvert^p)^{1/p}$
Distância Pearson	$D(P,Q) = \frac{\sum_{i=1}^{n}(P_i-Q_i)^2}{Q_i}$
Distância Intersection	$D(P,Q) = \sum_{i=1}^{n} min(P_i,Q_i)$
Distância Manhattan	$D(P,Q) = \sum_{i=1}^{n} \lvert P_i - Q_i\rvert$
Distância Tanimoto	$D(P,Q) = 1 - \frac{\lvert P \cap Q\rvert}{\lvert P \cup Q\rvert}$
Distância Euclideana	$D(P,Q) = \sqrt{\sum_{i=1}^{n}(P_i - Q_i)^2}$

Figura 3 - Distâncias de similaridade.
Fonte: Adaptado de Bento et al. (2023)

Segundo Bento et al. (2023) a Figura 3 apresenta as distâncias de similaridade mais relevantes que serão usadas na parte prática deste estudo.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Apesar de ser facilmente usado o conceito de distância, nem sempre é interpretado da melhor forma os conceitos de similaridade associados. No entanto se estiverem a ser considerados n pontos, poder-se-á concluir que a sua similaridade será a distância entre eles.

2.5 Algoritmos de Machine Learning

Na atualidade os algoritmos de ML estão por todo o lado, a partir da inferência de dados descobrem-se padrões / conhecimento para qualquer questão diária. Segundo Fabiano (2017) a palavra "algoritmo" entrou no nosso quotidiano, fazendo parte e estando associado a várias áreas da nossa sociedade. Quando se está a trabalhar com algoritmos de ML, um dos principais fatores que influenciam a sua escolha é o tipo e a quantidade de dados que se está a usar.

É importante destacar que nas últimas décadas, a aplicação de novas tecnologias na área da saúde tem melhorado as taxas de mortalidade, especialmente no campo da oncologia (Silva, 2003). Métricas quantitativas, como custo-benefício, eficácia, graus de evidência e acurácia, são ferramentas metodológicas que auxiliam no diagnóstico e tratamento de várias doenças, facilitando assim a utilização e o desenvolvimento de dispositivos médicos inovadores. Nesse contexto, pode-se afirmar que o uso racional de métodos tecnológicos avançados pode aumentar a qualidade de vida dos pacientes (Muller, 2017; Silveira et al., 2017). Por outro lado, de acordo com Kadir & F. (2018), os algoritmos de ML também podem ser utilizados para reduzir o número de nódulos benignos que são desnecessariamente monitorados, evitando o consumo desnecessário de recursos nos sistemas de saúde.

2.5.1 Redes Neurais Artificiais

As redes neurais são algoritmos de ML que se inspiram no sistema nervoso central de animais, como o cérebro, para reconhecer padrões. Elas são representadas como sistemas de "neurônios interconectados, capazes de processar valores de entrada", simulando o comportamento das redes neurais biológicas (Shah, 2020). Segundo Freeman & García-

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Bermejo (1993), as redes neurais artificiais são um paradigma específico dentro da inteligência artificial, representando um avanço em relação aos conceitos tradicionais baseados em lógica computacional e heurística, permitindo a incorporação de inteligência em artefactos. No final da década de 1950, um grupo de cientistas de inteligência artificial conseguiu desenvolver artefactos inteligentes que podiam aprender de forma autônoma, sem a necessidade de informações prévias sobre todos os detalhes lógicos e heurísticos necessários para cumprir sua missão (Freedman & Freedman, 1996). Segundo Bigus (1996), muitas das funções básicas desempenhadas pelas redes neurais artificiais refletem as capacidades humanas. Uma Rede Neural Artificial é uma estrutura computacional composta por unidades de processamento, ou neurônios, que estão interconectados e organizados em camadas. Essa organização permite o processamento simultâneo de informações por vários neurônios. Cada neurônio possui conexões de entrada e saída, e cada conexão possui um peso associado.

2.5.2 Regressão Logística

Segundo Shipe et al. (2019) uma regressão logística poder-se-ia definir como um método estatístico amplamente utilizado para analisar e modelar uma variável dependente que possui apenas duas categorias possíveis. Por outro lado, a regressão linear é usada para modelar uma variável dependente contínua. Na análise multivariada, são estimados coeficientes para cada variável preditora incluída no modelo final e esses coeficientes são ajustados em relação às outras variáveis preditoras. Esses coeficientes dão informações sobre a contribuição de cada variável preditora na estimativa do risco do resultado em questão.

2.5.3 Análise Discriminante Linear

Por outro lado, e de acordo com Ali et al. (2019) o LDA é um método de redução de dimensionalidade que é aplicado na etapa inicial de um modelo preditivo usado para classificação de padrões. Sua principal função é encontrar vetores no espaço de características que permitam uma melhor separação entre as classes dos dados. A separabilidade das classes pode ser avaliada projetando os pontos de dados originais

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

nesses vetores. Portanto, se as classes estiverem sobrepostas para um determinado ponto de dados, o LDA procurará separá-las de maneira mais eficiente, aplicando algum tipo de transformação.

2.5.4 K-Nearest Neighbors

De acordo com Mailagaha Kumbure et al. (2020) o algoritmo KNN e suas disseminações são utilizados para realizar classificação. O objetivo desses algoritmos é identificar a classe à qual um novo objeto ou amostra não classificada pertence. Na classificação baseada em ML supervisionado, o algoritmo é treinado com dados previamente classificados para realizar a classificação. O classificador KNN aborda o problema de classificação medindo inicialmente a similaridade entre a nova amostra a ser classificada e as amostras de treino. Seguidamente, o KNN mais próximos da nova amostra e determina a associação da nova amostra à classe que possui o maior número de vizinhos/amostras próximos.

2.5.5 Decision Tree Classifier

Segundo Tangirala et al. (2020) uma arvore de decisão é um método simples que utiliza um fluxograma para atribuir rótulos de classe a uma variável de saída com base nos valores de uma ou mais variáveis de entrada. O processo de classificação começa no nó raiz da árvore de decisão e avança recursivamente até chegar a um nó folha, que contém os rótulos de classe. Em cada nó, uma condição de divisão é aplicada para decidir se o valor de entrada deve seguir para a subárvore esquerda ou direita, até chegar aos nós folha. A condição de divisão aplicada em cada nó deve resultar em subconjuntos homogêneos, ou seja, subconjuntos nos quais todos os registos possuem o mesmo rótulo de classe. No entanto, é difícil obter subconjuntos homogêneos puros com dados do mundo real, pois sempre haverá algum tipo de mistura. Portanto, ao construir a árvore de decisão, o objetivo em cada nó é selecionar as condições de divisão que melhor dividem o conjunto de dados em subconjuntos homogêneos. Para avaliar a qualidade de uma condição de divisão, é utilizado um critério chamado “impureza”, que é calculado

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

matematicamente para cada condição de divisão. A condição de divisão que resulta na menor impureza é escolhida como a melhor.

2.5.6 Gaussian NB

Os classificadores Naive Bayes de acordo com Rodrigues et al. (2020) são um tipo básico de classificação probabilística que se baseiam na aplicação do Teorema de Bayes. Esses classificadores não realizam um único cálculo, mas sim utilizam vários cálculos. Estes dividem o conjunto de dados em dois conjuntos: a matriz de recursos e a matriz de resposta. A matriz de recursos contém as linhas que possuem as características dependentes, enquanto a matriz de resposta contém os valores da variável de classe, que representa a saída para cada linha da matriz de recursos. Poder-se-á afirmar que o Teorema de Bayes é utilizado para calcular a probabilidade de uma ocorrência, dado que outra ocorrência tenha ocorrido anteriormente.

2.5.7 Support Vector Machine

O SVM e de acordo com Rodrigues et al. (2020) é uma fórmula utilizada para resolver problemas de classificação ou regressão, mas é principalmente empregada em problemas de classificação. O SVM é considerado um classificador binário, o que significa que assume que os dados fornecidos possuem dois valores-alvo possíveis. Nessa fórmula, cada item de dados é apresentado como um ponto em um espaço n-dimensional (sendo n o número de recursos), onde cada recurso representa uma coordenada específica. O objetivo é encontrar um hiperplano que melhor separe as duas classes, classificando assim os dados. Os vetores de suporte são os pontos de dados que estão mais próximos do hiperplano e eles definem a fronteira de separação entre as classes.

Existem conceitos importantes no SVM, como:

- Vetores de suporte: são os pontos de dados mais próximos do hiperplano. A linha de separação é definida com a ajuda desses pontos.
- Hiperplano: é um plano ou espaço que divide objetos com categorias diferentes.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

- Margem: é o espaço entre duas linhas que estão próximas aos pontos de dados de diferentes categorias. É a distância perpendicular dos vetores de suporte em relação à linha. Uma margem maior é considerada melhor, enquanto uma margem menor é considerada pior.

O objetivo principal do SVM é separar os conjuntos de dados em diferentes classes, procurando um hiperplano que minimize a margem. Isso é feito em duas etapas: o SVM gera hiperplanos iterativamente que separam as classes da melhor maneira possível e, em seguida, escolhe o hiperplano que separa as classes adequadamente.

2.6 Microarrays

De acordo com Miller e Tang (2009) um *microarray* de DNA ou chip de DNA é uma coleção de micro spots, geralmente preenchidos com DNA, que contém sondas para determinadas moléculas "alvo", onde são produzidos resultados quantitativos, como a expressão génica. A natureza da sonda, o suporte sólido usado e o método usado para direcionar a sonda são algumas das características desta técnica.

No início da década de 70, a técnica de estudar em simultâneo um elevado número de genes, através do uso de arranjos de ácidos nucleicos, ficou conhecida como *Dot-Blot* (Kafatos et al., 1979). No entanto só na década de 90, com a evolução de técnicas como a confeção de arranjos de alta densidade e a implementação da técnica de deteção de fluorescência, nas medidas de intensidade, é que começou a estruturar a técnica tal e qual a conhecemos hoje.

Os chips de DNA são compostos por coleções organizadas de segmentos de DNA que são distribuídos de forma ordenada em uma superfície sólida, lembrando a disposição de milhões de transístores num componente eletrónico em miniatura. Para que um arranjo de DNA seja possível, é necessário que cada componente da coleção possua um endereço exclusivo, ou seja, uma posição individual dentro do arranjo (Chaudhuri, 2005).

Segundo Bento et al. (2023) a tecnologia de *microarrays* consiste num arranjo de pontos em uma plataforma sólida, normalmente uma lâmina de vidro, com quantidades pequenas

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

de DNA imobilizado, podendo designar-se por sonda ou *probe*. Cada sonda é complementar a uma sequência de nucleotídeos específica, conhecida como alvo ou target, que representa um gene do genoma. A hibridização ocorre quando as sondas se ligam aos seus alvos por meio da complementaridade das cadeias de nucleotídeos.

A confeção dos arranjos pode ser feita de duas formas. Através de robôs que depositam na superfície da lâmina de vidro as amostras de DNA, também conhecido como "impressão do slide". Essas amostras são constituídas de oligonucleotídeos pré-sintetizados, *Complementary* DNA (cDNAs) produzidos em projetos de sequenciamento, ou ainda produtos de amplificação por Polymerase Chain Reaction (PCR). A outra forma utiliza processos especiais para realizar a síntese química de oligonucleotídeos diretamente sobre a superfície da lâmina de vidro.

A tecnologia de microarray é utilizada para fazer a hibridização com um pool de Messenger RNA (mRNAs) extraídos de amostras biológicas, que foram previamente marcados com fluoróforos. A maioria dos protocolos laboratoriais utiliza o processo de transcrição reversa para converter os mRNAs nos correspondentes cDNAs durante o processo de marcação. Após a hibridização, cada lâmina é lavada para remover os "alvos" excedentes e exposta à ação de raios laser que estimulam os fluoróforos incorporados aos "alvos", fazendo com que emitam luz (fluorescência). Quanto maior for a expressão de um determinado gene, maior será a quantidade de "alvos", e consequentemente maior será a intensidade da fluorescência do complexo alvo-sonda após a hibridização (N. A. M. Silva et al., 2013).

Assim, a tecnologia de *microarrays* fornece uma medida indireta do nível de expressão génica, mediante a quantificação da abundância dos Ribonucleic Acid (RNA) transcritos. Segundo a FGED Society - History (2023) onde se pode consultar um pouco da história desta tecnologia, organizações como a *Microarray Gene Expression Data Society* e o *European Bioinformatics Institute* (EBI) estabeleceram guias de orientação para auxiliar os pesquisadores a planear e implementar suas experiências com *microarrays*, por forma a padronizar esta tecnologia. O guia *Minimum Information About* a *Microarray Experiment* contém diversas recomendações e padrões para coleta e análise de dados

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

provenientes de experiências com *microarrays*, para que estes possam ser corretamente interpretados e reproduzidos (Brazma et al., 2001).

Além disso, iniciativas como o *Gene Expression Omnibus*, do *National Center for Biotechnology Information*, e o *ArrayExpress*, mantido pelo EBI, foram criadas para compartilhar dados brutos obtidos em experiências com *microarrays*, uma vez que essas informações não podem ser incluídas nas publicações. Dessa forma, os resultados obtidos em diferentes experiências podem ser comparados e utilizados como base para o planeamento de novas pesquisas sobre o mesmo tema.

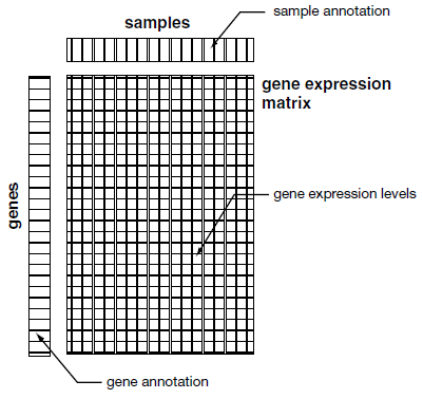


Figura 4 - Explicação de um microarray.

Fonte: Brazma et al. (2001)

De uma forma geral, uma coleção de dados de expressão génica é representada como uma tabela com linhas representando os genes e colunas representando as amostras. Cada posição na tabela descreve a medição de um gene específico em uma amostra particular, formando assim uma matriz de expressão génica. Para descrever completamente uma experiência de microarray, é necessário fornecer informações sobre os genes medidos, as amostras recolhidas e a matriz de expressão génica (Brazma et al., 2001).

Embora seja ideal medir a expressão génica em unidades naturais e ter uma estimativa de erro ou confiabilidade, existem vários desafios experimentais que tornam a medição direta da expressão génica difícil. Os dados brutos de experiências de microarray são imagens que precisam ser analisadas para identificar e quantificar cada recurso (ponto)

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

na imagem. Por outro lado, é ainda importante observar que uma sequência de DNA pode ser encontrada várias vezes em um microarray, e várias sequências distintas de DNA podem ser mapeadas para o mesmo gene. Para interpretar esses dados, é importante ter informações sobre as condições experimentais em que as amostras foram obtidas e sobre as anotações dos genes medidos. Idealmente, os dados de expressão génica serão medidos em unidades naturais, como cópias de mRNA por célula, e acompanhados de uma estimativa de erro ou indicador de confiabilidade (Brazma et al., 2001).

No entanto, a medição direta da expressão génica pode ser desafiadora e os dados brutos obtidos em experiências de microarray são imagens que precisam ser processadas para identificar e quantificar cada recurso (ponto) na imagem. Além disso, uma mesma sequência de DNA pode estar presente em várias regiões do microarray, e diferentes sequências podem ser mapeadas para o mesmo gene, o que pode complicar a análise dos dados de expressão génica. Ou seja, a expressão génica deverá ser observada como um todo e não em partes, gene a gene.

2.6.1 Plataformas e fabricantes de microarray

Nesta secção serão abordadas algumas tecnologias de microarray, nomeadamente a Affymetrix GeneChip, Illumina BeadChip, Agilent DNA Microarray, Roche NimbleGen e Thermo Fisher Scientific.

Affymetrix GeneChip é uma tecnologia de microarray que permite a análise simultânea de um milhão de genes em uma única amostra. A sua principal característica diferenciadora é que na sua construção ser dirigida por síntese fotoquímica. Essa tecnologia usa pequenos fragmentos de DNA conhecidos como sondas para detetar a presença e a quantidade de RNA ou DNA em uma amostra biológica. Ao longo dos anos esta plataforma potenciou várias descobertas, devido ao seu sistema ser robusto e confiável (Dalma-Weiszhausz et al., 2006).

Illumina BeadChip é uma tecnologia de microarray muito popular e fácil de usar, baseando-se na conversão de bissulfito de sódio do DNA, de genotipagem de base única de locais CpG direcionais onde usa sondas num microarray. Estas plataformas são de uso

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

facilitado, eficientes tanto em tempo como em custo-benefício, sendo ainda uma boa opção para as medições de metilação do DNA de outras plataformas (Pidsley et al., 2016).

Agilent DNA Microarray é uma tecnologia que utiliza sondas de DNA imobilizadas numa matriz para detetar a presença e a quantidade de RNA ou DNA na amostra. Os microarrays Agilent permitem a análise de milhares de sondas em uma única amostra sendo ainda indicados para estudos de expressão génica, genotipagem e análise de variações genéticas (Miller & Tang, 2009).

Roche NimbleGen é uma tecnologia que permite uma interrogação precisa, sensível e especifica da expressão génica para qualquer genoma anotado e sequenciado, fornecendo ainda *microarrays* de alta densidade sondas de oligo-longo e capacidade de design flexível para uma análise avançada da expressão génica. Cada tipo de microarray tem seu próprio conjunto de arquivos usados para análise de dados (SelectScience, 2023).

Thermo Fisher Scientific é uma tecnologia que oferece vários tipos de *microarrays* para análise genómica, incluindo *microarrays* de expressão génica, genotipagem, análise de variantes e sequenciamento de RNA. Cada tipo de microarray tem seu próprio conjunto de arquivos usados para análise de dados (Thermo Fisher Scientific, 2023).

Essas são apenas algumas de muitas técnicas de *microarrays* utilizadas em genómica, proteómica e biologia molecular. Cada plataforma tem seus próprios pontos fortes e limitações, e a escolha da plataforma adequada depende do objetivo da pesquisa e das características da amostra em questão. Aqui importa ainda clarificar, que em 2016 a Affymetrix foi adquirida pela Thermo Fisher Scientific, sendo desde então os microarrays fabricados pela Affymetrix comercializados pela marca Thermo Fisher Scientific. No entanto, estas duas tipologias de microarray continuam a ter diferenciações em alguns aspetos técnicos, como o número e a qualidade das sondas de DNA, a plataforma de deteção, a metodologia de preparação da amostra, entre outros pontos.

Neste ponto, importa ainda salientar que a análise de um microarray a partir dos seus dados brutos, será uma abordagem bastante interessante, principalmente quando não foi fornecida a estatística deste conjunto de dados. No mercado existem várias opções de

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

plataformas dedicadas a esta análise, no então cada fornecedor tem um formato padrão para os seus arquivos de dados. Na tabela infra, estão compilados os vários tipos de dados disponibilizados pelos principais fabricantes de *microarrays*.

Tabela 1 - Tipos de dados disponibilizados pelos principais fabricantes de microarrays.

Nome	Extensão	Fabricante
Arquivo de design de microarray	CDF (.cdf)	Affymetrix (www.affymetrix.com)
Arquivo de imagem de microarray	CEL (.cel)	
Arquivo de design de microarray	Ilmn (.ilmn)	Illumina (www.illumina.com)
Arquivo de resultados de análise de expressão génica	Texto (.txt)	
Arquivo de imagem de microarray	IDAT (.idat)	
Arquivo de design de microarray	Agilp (.agilp)	Agilent (www.agilent.com.br)
Arquivo de imagem de microarray	Tiff (.tiff)	
Arquivo de design de microarray	Ndf (.ndf)	Roche NimbleGen (www.roche.com)
Arquivo de imagem de microarray	Gpr (.gpr)	
Arquivo de resultados de ChIP-chip	Pair (.pair)	
Arquivo de resultados de sequenciamento de DNA	Fasta (.fasta)	
Arquivo de design de microarray	Gal (.gal)	Thermo Fisher Scientific (www.thermofisher.com)
Arquivo de imagem de microarray	CEL (.cel)	
Arquivo de resultados de análise de expressão génica	Texto (.txt)	

Fontes: Sites dos vários fornecedores

Os ficheiros de design e implementação de microarray de DNA são um arquivo eletrónico que contém informações detalhadas sobre a construção do microarray, como a sequência das sondas de DNA, anotação a nível de sonda (*probe-level annotation*), a disposição dos spots no chip, as anotações dos genes e outras informações relacionadas. Por outro lado, os ficheiros de imagem de microarray são arquivos digitais que contêm a imagem do chip de microarray após a realização do procedimento. Estes arquivos são gerados por scanners de microarray que capturam as imagens dos spots do chip de microarray que foram hibridizados com as sondas de DNA fluorescentes.

2.6.2 Microarrays na Oncologia

A ocorrência da patologia oncológica deve-se ao acumular de alterações genéticas ao longo da vida do ser humano. Ao longo dos anos e segundo vários estudos efetuados a patologia oncológica é geneticamente heterogénea, ou seja, para doentes com

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

diagnósticos semelhantes poder-se-á obter diversas respostas em relação à mesma terapia. A identificação de biomarcadores específicos, como alterações genéticas, será fulcral para a medicina oncológica de precisão, permitindo estratégias eficazes na seleção de pacientes para a terapia a ser aplicada (Cokelaer et al., 2018).

É de salientar ainda que o cancro é uma doença que ocorre devido à acumulação de mudanças genéticas e epigenéticas em células relacionadas a esta patologia, como oncogenes e genes supressores de tumores, bem como genes que controlam o ciclo celular, a morte celular programada, a adesão celular, o reparo do DNA e a formação de novos vasos sanguíneos. Nesta área da medicina, a análise dos perfis de expressão génica pode fornecer informações sobre as mudanças na expressão génica, entre a comparação das células tumorais e o perfil de expressão das células saudáveis, que estão associadas à disfunção celular do tumor e às vias regulatórias afetadas (Silva, 2003).

A tecnologia de microarray tem sido amplamente usada nos últimos anos, para estudar a classificação e a progressão das neoplasias, bem como a resistência, recorrência local, metástases, sensibilidade à quimioterapia e prognóstico pós-operatório (Cassali et al., 2007). Os microarrays também podem ajudar a entender os mecanismos de resistência aos medicamentos anticancerígenos e prever a sensibilidade aos medicamentos e efeitos colaterais inesperados. A técnica de microarray é útil para comparar amostras com doença oncológica versus amostras saudáveis (sem doença oncológica), a fim de gerar perfis de expressão génica e discriminar entre diferentes tipos de células ou condições biológicas. Os genes podem ser agrupados em classes existentes ou descobertas de novas classes usando métodos supervisionados ou não-supervisionados, prevendo novos biomarcadores através de técnicas de ML (Zhang et al., 2022).

"*Esquema de hibridização e análise dos dados. A tecnologia de microarray é baseada na hibridização entre alvos livres marcados derivados de amostras biológicas e um array de muitas sondas de DNA imobilizadas em uma matriz sólida. Os alvos são produzidos a partir de RNA de amostras biológicas por transcrição reversa, marcados com fluoróforos específicos (Cyanine dye3 (Cy3) e Cyanine dye5 (Cy5)) e co-hibridizados com as sondas de DNA por 16 horas. Os sinais resultantes da hibridização são detetados por laser, para*

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

a aquisição da imagem. Posteriormente, os sinais são quantificados, integrados e normalizados com softwares específicos e refletem o perfil de expressão para cada amostra biológica" (Colombo & Rahal, 2009).

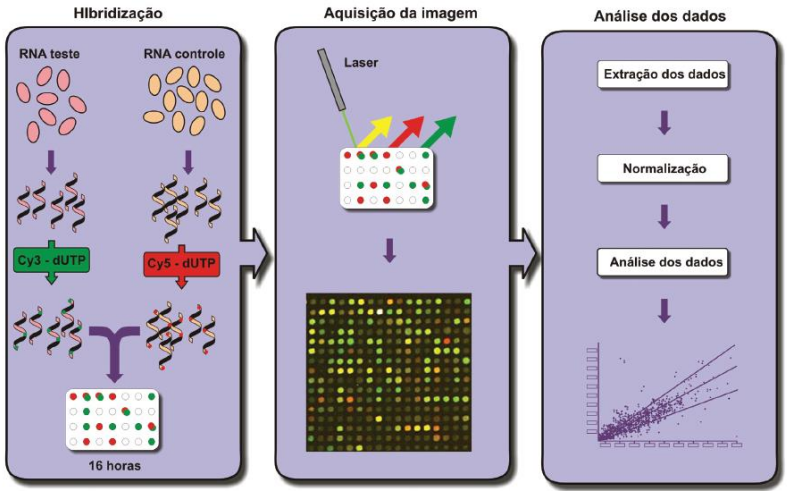


Figura 5 - Esquema de hibridização e análise dos dados para a tecnologia de microarray.

Fonte: Colombo & Rahal (2009)

2.6.3 Análise estatística de dados de microarray

Segundo Rosa et al. (2007) a análise estatística de microarray envolve pelo menos três passos: a obtenção dos dados, a normalização dos dados e a análise estatística dos dados, sendo que esta última envolve desde testes de significância, analises discriminante ou de agrupamento. A análise das imagens das laminas é um processo de obtenção de dados a partir das sondas de genes, onde é usado o sistema de hibridização competitiva. Os valores de intensidade fluorescente são extraídos de cada pixel da imagem e combinados em uma medida resumo para obter a expressão relativa de cada gene nas duas amostras hibridizadas em cada lâmina. Existem vários procedimentos e programas computacionais disponíveis para a leitura destas imagens para os diferentes tipos de lâminas no mercado. Seguidamente, para cada lamina, os pixels de *foreground* correspondentes a cada spot são combinados de forma resumida, como por exemplo média, mediana ou intensidade total, sendo muitas vezes são ajustados para valores de background. Esse mesmo procedimento é realizado tanto para as intensidades relativas ao Cy3 quanto ao Cy5, de modo que para

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

cada ponto há duas medidas de intensidade, das quais se obtém uma medida da expressão relativa de cada gene nas duas amostras hibridizadas em cada lâmina.

A Figura 6 apresenta um esquema de *microarrays* retirado da obra "*Platelet Genomics and Proteomics*" (García et al., 2007), onde estão apresentados dois principais tipos de *microarrays* de DNA.

No Ponto A - Microarray de oligonucleotídeos de DNA complementar impresso (cDNA) *"Neste exemplo, 60 pontos de dados mostram a eficiência de hibridização para 60 diferentes cDNAs impressos ou oligonucleotídeos representando 60 genes. Um sinal verde representa uma expressão génica comparativamente alta na amostra 1, e um sinal vermelho indica expressão génica alta na amostra 2, um sinal amarelo indica expressão génica equivalente em ambas as amostras, sendo a cinza a falta de expressão génica em ambas as amostras."*

No Ponto B - Microarray de oligonucleótidos sintéticos "*Neste exemplo do tipo de microarray da Affymetrix, cada gene no microarray é representado por uma linha de (22-25 bases) oligonucleotídeos, cada um com seu próprio controle de incompatibilidade na linha abaixo. Os dados para quatro genes diferentes são mostrados, com amarelo indicando hibridização detetável e cinza indicando sem hibridização. As análises computacionais determinam se os genes individuais estão "presentes" na amostra de mRNA (por exemplo, o gene no topo do microarray), "equívoco" ou "ausente" (por exemplo, o gene na parte inferior do microarray)."*

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

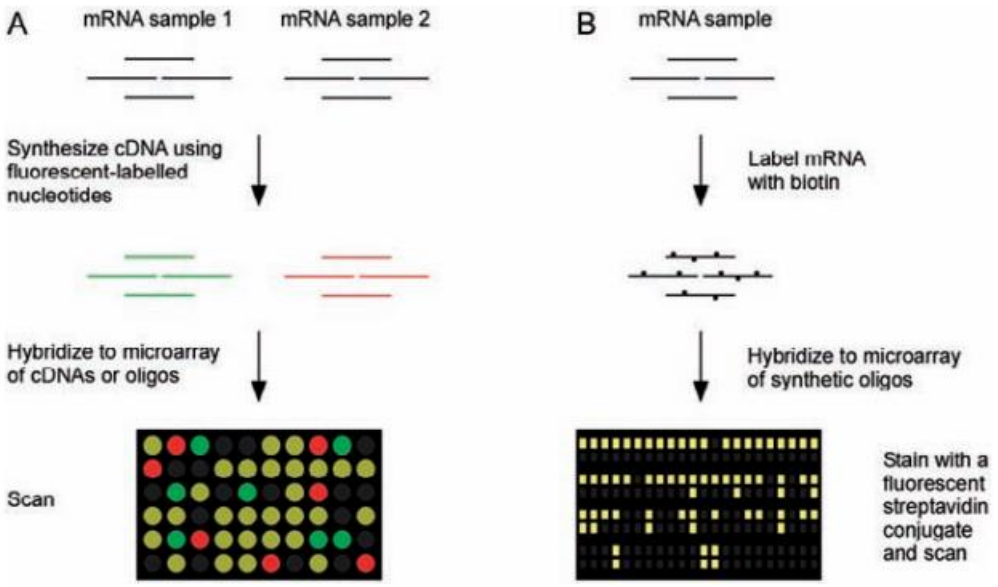


Figura 6 - Esquema de microarrays.

Fonte: García et al. (2007)

De acordo com Rosa et al. (2007), num Affymetrix *GeneChip* cada gene seria representado por um grupo de sondas de cadeias curtas de oligonucleotídeos (25 bases), geralmente com cerca de 16 a 20 pares de sondas por gene. Cada grupo inclui uma sonda perfeitamente compatível com o gene (chamada de "perfect match" (PM)) e uma sonda com uma mudança de base na 13ª posição (chamada de "mismatch" (MM)). Cada oligonucleotídeo no chip é quase idêntico, diferindo apenas por uma incompatibilidade central de uma única base, o que permite a determinação do grau de ligação não específica. Os *GeneChips* da Affymetrix são capazes de medir a expressão absoluta dos genes em células ou tecidos, sendo que os valores de intensidade observados para cada gene são combinados em uma única medida resumo, geralmente utilizando a média das diferenças entre PM e MM para cada gene, dada pela seguinte formula:

$$AvDiff_g = \frac{1}{K}\sum_{i=1}^{K}(PM_{gi} - MM_{gi})$$

Onde: $AvDiff_g$ *é a medida de expressão génica, relativa ao gene* g, *e* PM_{gi} *e* MM_{gi} são as intensidades PM e MM relativas ao j.ésimo par de sondas (i =1,2,…,K) do gene g.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

A medida resumo AvDiff foi proposta pela Affymetrix para resumir as intensidades observadas para cada gene, no entanto outras metodologias alternativas foram surgindo como o MAS5.0 indicado pela Affymetrix, o MBEI (*Multiplicative Model-Based Expression Index*) proposto por (C. Li & Wong, 2001) e o *Robust Multi-array Average (*RMA) proposto por (Irizarry, 2003).

2.6.4 Affymetrix Human Genome U219 Array

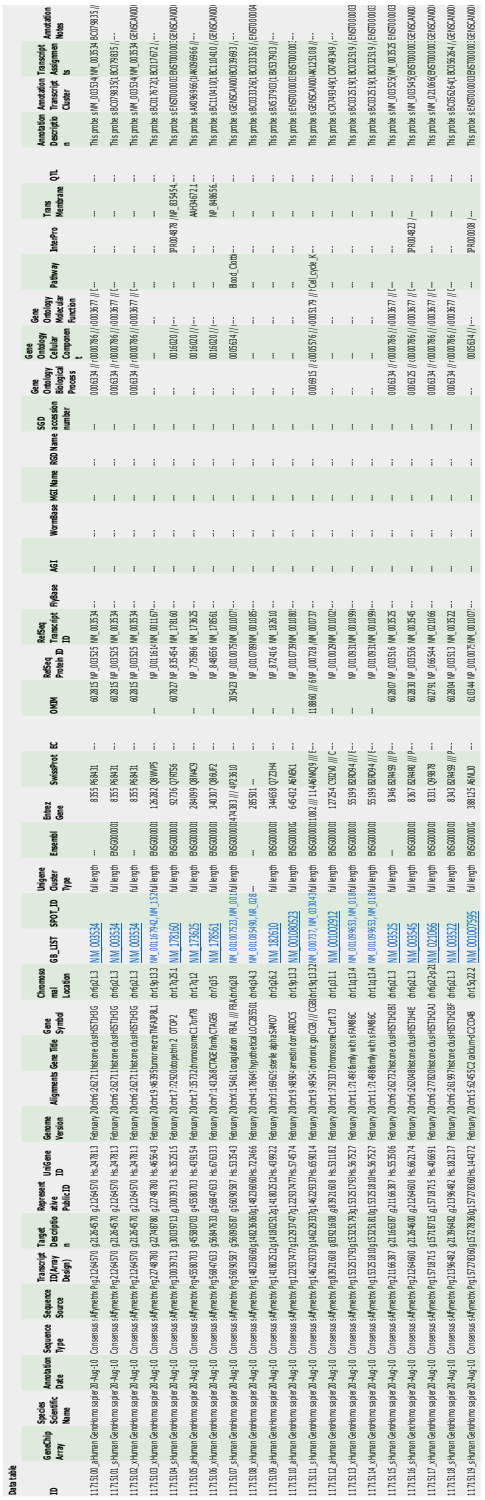
Segundo a Thermo Fisher Scientific, os microarrays U219 de Genoma Humano são constituídos por mais de 530 000 sondas sendo cobertas por mais de 36 000 transcrições e variantes, onde mais de 20 000 genes são mapeados através do *Unigene* ou da anotação RefSeq (PrimeView Human Genome U219 Array Plate). Ainda segundo o seu fornecedor este tipo de *array* apresenta ainda as seguintes vantagens:

- Permitir uma maior produtividade e eficiência através do processamento paralelo;
- Processar 16, 24 ou 96 amostras em uma única placa de matriz (*array*);
- Oferecer uma excelente precisão e reprodução da expressão génica
- Ser capaz de medições múltiplas e independentes por transcrição para maior confiança nos seus resultados;
- Apresenta ainda 11 sondas por conjunto para sequências bem anotadas ou 9 sondas por conjunto para o restante.

Seguidamente apresentam-se o tipo de dados que se conseguem ler deste tipo de *array*, como iremos usar os dados do Projeto GDSC1

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Tabela 2 - [HG-U219] Affymetrix Human Genome U219 Array.



Fonte: Biotechnology Information (2011)

3 METODOLOGIA

No desenvolvimento do presente estudo irá ser usada a metodologia *Cross-Industry Standard Process for Data Mining (*CRISP-DM), por forma a facilitar a compreensão, implementação e desenvolvimento do projeto. Segundo Chapman (2000) a metodologia CRISP-DM poderá ser traçada através de processos hierárquicos, consistindo em conjuntos de tarefas descritas em quatro níveis de abstração (da generalidade à especificidade).

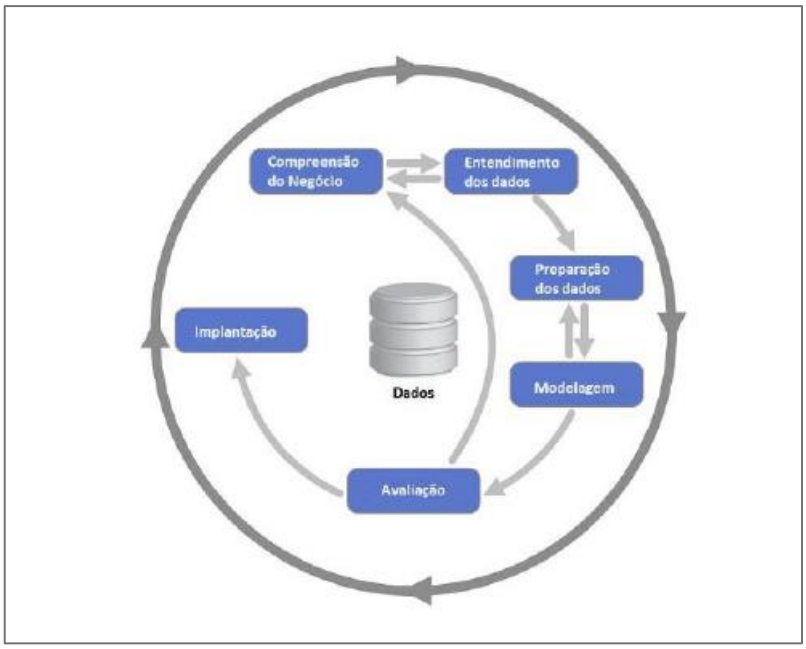


Figura 7 - Fases do CRISP-DM.

Fonte: Chapman et al. (2000)

Assim, será possível compreender que estamos perante um processo contínuo onde as fases são realimentadas de forma contínua, consoante se descobrem novas informações ou melhorias do processo.

Na literatura esta metodologia é amplamente utilizada no desenvolvimento de projetos de ciência de dados, pois fornece uma estrutura organizada e sistemática para guiar as etapas do processo, desde o entendimento do negócio até à implementação das soluções (Ramos et al., 2020).

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

A metodologia CRISP-DM é composta por seis fases principais, conforme explanadas infra:

1. **Compreensão do Negócio**

Nesta fase inicial, o objetivo é obter uma compreensão abrangente dos objetivos e requisitos do negócio que se está a abordar, sendo identificadas as metas e critérios de sucesso, bem como a análise do contexto e das restrições envolvidas. Esta fase é crucial para a compreensão das fases seguintes (Chapman et al., 2000). Esta foi uma das fases mais importantes para a compreensão dos dados genómicos a estudar. Assim, na revisão da literatura houve um esforço alargado para o entendimento de todas as temáticas relacionadas com este estudo, como o genoma humano, a imagem médica, os RecSys, a similaridade de imagens médicas, os algoritmos de ML e os *microarrays*.

2. **Entendimento dos Dados**

Nesta fase, o foco está na exploração dos dados disponíveis. Isso inclui a identificação das fontes de dados relevantes, a obtenção dos conjuntos de dados necessários e a realização de uma análise exploratória dos dados, descrição e formulação de hipóteses, explorando a qualidade e potencial para resolver o problema em questão (Chapman et al., 2000). Nesta fase será elaborado um estudo alargado do *dataset* "Projeto *GDSC1*", de onde se selecionará aleatoriamente uma amostra de 20 linhas celulares (10 amostras referentes à patologia da mama e 10 amostras referentes a patologias da pele), com 49386 genes cada, do universo das 970 linhas celulares cancerígenas do *dataset*. Para o estudo destas linhas celulares, inicialmente obter-se-á a expressão génica de cada amostra, elaborando-se seguidamente uma análise descritiva dos seus dados. Serão ainda testadas algumas técnicas de agrupamento como clusters, análise das componentes principais e diagrama de Venn. Para além desta abordagem, existirá ainda a necessidade de um entendimento mais profundo sobre os dados genómicos e o próprio fornecedor do tipo de *microarray* a estudar.

3. **Preparação dos Dados**

Esta fase consiste em atividades ligadas ao tratamento dos dados, ou seja, uma vez coletados os dados é necessário prepará-los para a etapa seguinte. Aqui são realizadas

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

tarefas como a limpeza dos dados, tratamento de valores ausentes, remoção de *outliers*, seleção de variáveis relevantes e transformação dos dados, se necessário. O objetivo será obter um conjunto de dados pronto para aplicar vários modelos (Chapman et al., 2000). Nesta fase e tendo como objetivo o estudo de similaridade será aplicada a abordagem mais apropriada ao presente estudo, tendo presente a capacidade do *hardware*. Tendo-se já selecionado as 20 linhas celulares, será extraída a expressão génica de cada amostra (nos seus ficheiros CEL), para posteriormente elaborar um *dataset* com toda esta informação para os 49386 genes. No entanto como a fase da modelagem terá duas abordagens, será necessário criar dois *datasets*, um com toda a informação das 20 linhas celulares para os 49386 genes, que será usado para os algoritmos de ML, e outro com os 3 genes selecionados para as 20 linhas celulares, que será usado para as distâncias de similaridade.

4. Modelação

Nesta fase, os modelos de mineração de dados são construídos e avaliados. Diferentes técnicas e algoritmos podem ser aplicados para explorar os dados e desenvolver modelos preditivos ou descritivos, dependendo dos objetivos do projeto. Os modelos são ajustados e refinados com base na avaliação dos resultados obtidos (Chapman et al., 2000). Nesta fase poderá existir a necessidade de retornar à fase três (Preparação dos Dados) do modelo selecionado. Aqui serão adotadas duas abordagens, por um lado o estudo das distâncias de similaridade de Dice, Jaccard, Sorensen, Czekanowski, Minkowski, Pearson, Intersection, Manhattan, Tanimoto e Euclideana, e por outro o estudo dos algoritmos de ML de Rede Neural Artificial, *Logistic regression, LDA, KNN, DecisionTreeClassifier, Gaussian NB e SVM*, para avaliar as melhores métricas para o presente estudo

5. Avaliação

Após a criação dos modelos, é necessário avaliar sua qualidade e eficácia. Isso envolve a aplicação de critérios de avaliação pré-definidos para avaliar o desempenho dos modelos em relação aos objetivos do projeto em questão definidos na primeira fase. Sendo também importante verificar se os modelos atendem aos requisitos do projeto e se são úteis para a tomada de decisões (Chapman et al., 2000). Se a resposta for positiva poder-se-á seguir

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

para a fase seguinte. A fase de avaliação servirá para avaliar os resultados obtidos e escolher a melhor abordagem. Podendo, no entanto, existir alguma dificuldade na obtenção/eleição da melhor abordagem ao presente problema.

6. Implementação

Nesta última fase, os modelos são implementados em ambiente de produção, integrados aos sistemas existentes e disponibilizados para uso prático. Também é importante desenvolver um plano de acompanhamento e monitorização contínuo, por forma a garantir que os modelos estão operando corretamente e que vão sendo atualizados consoante a sua necessidade (Chapman et al., 2000). Nesta fase poderá ainda ser importante a elaboração de relatórios / *dashboards* para melhorar o entendimento e compreensão dos dados em questão.

É importante destacar que o CRISP-DM é uma metodologia iterativa, ou seja, as etapas poderão ser revistas e repetidas ao longo do processo, conforme seja necessário. Por outro lado, a colaboração entre diferentes partes interessadas, como especialistas que entendam o negócio e cientistas de dados, é fulcral para o sucesso da metodologia CRISP-DM. Neste estudo não existira a oportunidade de implementação dos resultados obtidos, no entanto esta fase poderá ser aplicada em estudos futuros.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

4 ENTENDIMENTO DOS DADOS

O presente estudo, através de dados biológicos humanos, visa confirmar que a similaridade entre linhas celulares implica semelhanças na escolha da linha terapêutica a adotar, ou seja, linhas celulares cancerígenas similares partilham de tratamento similar. Para tal, e existindo doença oncológica, o doente será sujeito a uma biópsia onde será extraída uma amostra da linha celular, sendo posteriormente analisada microscopicamente através da tecnologia *microarray* DNA (Miller & Tang, 2009), sendo extraída a expressão genómica do doente. Sendo este o cerne do presente estudo, e depois de obtida a imagem da linha celular num formato *microarray*, pretende-se que a mesma seja comparada com o conjunto de dados do Projeto GDSC1, onde existem cerca de 1 000 imagens de linhas celulares (Cancer Genome Project et al., 2022). Esta análise de similaridade entre a imagem com as demais, poderia criar um ranking com as 10 imagens de linhas celulares mais similares relativamente a esta nova. O presente estudo pretende encontrar a similaridade entre uma nova linha celular em comparação com as restantes do conjunto de dados através da sua expressão génica. Num próximo estudo, tendo por base as conclusões do presente estudo, e com a base nos dados das linhas celulares testadas com as substâncias farmacológicas, poderá ser criado um algoritmo para identificar as linhas celulares mais similares relativamente à eficácia dos tratamentos adotados.

O presente estudo utiliza o *dataset* do Projeto GDSC1 onde foi testada a sensibilidade de 970 linhas celulares cancerígenas com 403 compostos na sua fase 1 (GDSC1) (Cancer Genome Project et al., 2022). No entanto, o projeto supracitado tem ainda outro *dataset* com 969 linhas celulares cancerígenas adicionais com 297 compostos na fase 2 (GDSC2) (Cancer Genome Project et al., 2022). Sendo que o GDSC1 (fase 1) atualiza versões anteriores com dados adicionais de triagem de medicamentos do Wellcome Sanger Institute e do Massachusetts General Hospital, e o GDSC2 (fase 2) é o novo da última triagem no Wellcome Sanger Institute usando procedimentos experimentais aprimorados. Os dados foram concebidos por criação de perfil de transcrição por *array* usando a plataforma Affymetrix. O tipo de estudo usado foi a criação de perfil de transcrição por

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

array, utilizando o sistema Experimental Factor Ontology (EFO). A EFO é um sistema estruturado que organiza conceitos e suas relações em um domínio específico, padronizando e descrevendo de maneira consistente os fatores experimentais, facilitando a compreensão, replicação e integração de dados entre estudos (EMBL-EBI, 2023). O organismo estudado foi o Homo sapiens (ser humano). Este conjunto de dados é útil para identificar marcadores genómicos que possam prever a resposta das linhagens celulares a diferentes drogas, o que pode levar ao desenvolvimento de terapias personalizadas para pacientes com cancro. A sensibilidade de cada linha de células cancerígenas em relação aos fármacos está representada como um valor de IC50 (a concentração na qual uma linha de células exibiu uma inibição absoluta no crescimento de 50%; IC50 mais baixo implica maior sensibilidade). GDSC também quantificou a expressão génica de nível basal de muitas das linhagens de células cancerígenas usando a tecnologia microarray (Y. Li et al., 2021). Seguidamente apresentam-se os dados que serão usados no âmbito deste projeto.

4.1 Ficheiros associados ao projeto

Um ficheiro CEL é um arquivo de dados criado por um software de análise de imagens de microarray da Affymetrix DNA. Neste ficheiro estão descritos milhares de pontos de dados, extraídos das sondas de um GeneChipe da Affymetrix, onde são armazenados os cálculos de intensidade, desvio padrão da intensidade, o número de pixels usados para calcular o valor da intensidade. Segue-se infra o detalhe do tipo de dados presentes nestes ficheiros por secção. Existem assim dois ficheiros para o entendimento destes dados, por um lado o ficheiro "E-MTAB-3610.idf.txt" contendo o comentário (ArrayExpress Accession) de todos os dados referentes a este projeto de investigação. Por outro lado, o ficheiro "E-MTAB-3610.sdrf.txt" é o ficheiro que contem todas anotações de cada ficheiro CEL do Projeto GDSC1 (Cancer Genome Project et al., 2022).

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Comment[ArrayExpressAccession]	E-MTAB-3610
MAGE-TAB Version	1.1
Investigation Title	Transcriptional Profiling of 1,000 human cancer cell lines
Comment[Submitted Name]	Transcriptional Profiling of 1,000 human cancer cell lines
Experiment Description	Basal expression profiles of 1,000 human cancer cell lines in the Genomics of Drug Sensitivity in Cancer (GDSC) panel [upcoming version], profiled using a diverse collection of 265 compounds. We have carried out an extensive computational exploration of the data to determine (1) to what extent does the mutational landscape of cancer cell lines recapitulate that seen in primary tumours, (2) what effect the status of these genomic features have on the variation in drug response; (3) whether genomic alterations acting in concert explain more of the variation in drug response; and (4) what is the predictive ability of these individual data-omics and at what extent this is improved when they are combined. [See publication]
Experimental Factor Name	cell_line
Experimental Factor Type	cell_line
Experimental Factor Term Source REF	
Experimental Factor Term Accession Number	
Person Last Name	Iorio
Person First Name	Francesco
Person Mid Initials	
Person Email	iorio@ebi.ac.uk
Person Phone	
Person Fax	
Person Address	
Person Affiliation	European Molecular Biology Laboratory - European Bioinformatics Institute
Person Roles	submitter
Date of Experiment	2014-03-19
Public Release Date	2015-07-01
PubMed ID	
Publication DOI	
Publication Author List	
Publication Title	A landscape of pharmacogenomic interactions in cancer
Publication Status	in preparation
Publication Status Term Source REF	EFO
Publication Status Term Accession Number	EFO_0001795
Protocol Name	P-MTAB-44937 P-MTAB-44938 P-MTAB-44939 P-MTAB-44940 P-MTAB-44941
Protocol Type	nucleic acid labeling protocol nucleic acid hybridization to array protocol array scanning and feature extraction protocol nucleic acid extraction protocol growth protocol
Protocol Term Source REF	EFO EFO EFO EFO EFO
Protocol Term Accession Number	EFO_0003813 EFO_0003815 EFO_0003814 EFO_0002944 EFO_0003789
Protocol Description	3'-IVT Express (Affymetrix) Affymetrix Affymetrix RNeasy Mini kit (Qiagen) followed by DNase digestion in solution, Qiagen RNeasy cleanup and elution cell lines cultured in RPMI or DMEM/F12 medium and collected as flash-frozen pellets when growing in log phase
Protocol Hardware	Gene Titan Gene Titan
Protocol Software	
Term Source Name	ArrayExpress EFO
Term Source File	http://www.ebi.ac.uk/arrayexpress/ http://www.ebi.ac.uk/efo/
Term Source Version	
Comment[AEExperimentType]	transcription profiling by array
SDRF File	E-MTAB-3610.sdrf.txt

Figura 8 - Comment[ArrayExpressAccession] - BioStudies

Fonte: European Bioinformatics Institute (2023)

Source Name	Characteristics[organism]	Characteristics[cell line]	Material Type	Protocol REF	Protocol REF	Extract Name	Protocol REF	Labeled Extract Name	Label	Protocol REF	Assay Name	Technology Type	Array Design REF	Term Source REF	Protocol REF	Array Data File	Comment [ArrayExpress FTP file]	Factor Value[cell_line]
J132_EPA01P10_UACC-812_breast_910910	Homo sapiens	UACC-812	cell	P-MTAB-44941	P-MTAB-44940	J132_EPA01P10_UACC-812_breast_910910	P-MTAB-44937	J132_EPA01P10_UACC-812_breast_910910:Biotin labeled streptavidin	Biotin labeled streptavidin	P-MTAB-44938	5500994173212120213068_A01	array assay	A-GEOD-13667	ArrayExpress	P-MTAB-44939	5500994173212120213068_A01.cel	ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/MTAB/E-MTAB-3610/E-MTAB-3610.raw.1.zip	UACC-812
J132_EPA01P2_201T_Lung:NSCLC_1287381	Homo sapiens	201T	cell	P-MTAB-44941	P-MTAB-44940	J132_EPA01P2_201T_Lung:NSCLC_1287381	P-MTAB-44937	J132_EPA01P2_201T_Lung:NSCLC_1287381:Biotin labeled streptavidin	Biotin labeled streptavidin	P-MTAB-44938	5500994158987071513209_A01	array assay	A-GEOD-13667	ArrayExpress	P-MTAB-44939	5500994158987071513209_A01.cel	ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/MTAB/E-MTAB-3610/E-MTAB-3610.raw.1.zip	201T
J132_EPA01P3_EVSA-T_Breast_906862	Homo sapiens	EVSA-T	cell	P-MTAB-44941	P-MTAB-44940	J132_EPA01P3_EVSA-T_Breast_906862	P-MTAB-44937	J132_EPA01P3_EVSA-T_Breast_906862:Biotin labeled streptavidin	Biotin labeled streptavidin	P-MTAB-44938	5500994172383112813929_A01	array assay	A-GEOD-13667	ArrayExpress	P-MTAB-44939	5500994172383112813929_A01.cel	ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/MTAB/E-MTAB-3610/E-MTAB-3610.raw.1.zip	EVSA-T
J132_EPA01P4_KYSE-520_Esophagus_753575	Homo sapiens	KYSE-520	cell	P-MTAB-44941	P-MTAB-44940	J132_EPA01P4_KYSE-520_Esophagus_753575	P-MTAB-44937	J132_EPA01P4_KYSE-520_Esophagus_753575:Biotin labeled streptavidin	Biotin labeled streptavidin	P-MTAB-44938	5500994158987071513202_A01	array assay	A-GEOD-13667	ArrayExpress	P-MTAB-44939	5500994158987071513202_A01.cel	ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/MTAB/E-MTAB-3610/E-MTAB-3610.raw.1.zip	KYSE-520
J132_EPA01P5_MS751_Cervix_1240179	Homo sapiens	MS751	cell	P-MTAB-44941	P-MTAB-44940	J132_EPA01P5_MS751_Cervix_1240179	P-MTAB-44937	J132_EPA01P5_MS751_Cervix_1240179:Biotin labeled streptavidin	Biotin labeled streptavidin	P-MTAB-44938	5500994158987071513207_A01	array assay	A-GEOD-13667	ArrayExpress	P-MTAB-44939	5500994158987071513207_A01.cel	ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/MTAB/E-MTAB-3610/E-MTAB-3610.raw.1.zip	MS751
J132_EPA01P6_MEL-JUSO_Skin_908125	Homo sapiens	MEL-JUSO	cell	P-MTAB-44941	P-MTAB-44940	J132_EPA01P6_MEL-JUSO_Skin_908125	P-MTAB-44937	J132_EPA01P6_MEL-JUSO_Skin_908125:Biotin labeled streptavidin	Biotin labeled streptavidin	P-MTAB-44938	5500994172383112813928_A01	array assay	A-GEOD-13667	ArrayExpress	P-MTAB-44939	5500994172383112813928_A01.cel	ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/MTAB/E-MTAB-3610/E-MTAB-3610.raw.1.zip	MEL-JUSO
J132_EPA01P8_NCI-H1869_Lung_1240183	Homo sapiens	NCI-H1869	cell	P-MTAB-44941	P-MTAB-44940	J132_EPA01P8_NCI-H1869_Lung_1240183	P-MTAB-44937	J132_EPA01P8_NCI-H1869_Lung_1240183:Biotin labeled streptavidin	Biotin labeled streptavidin	P-MTAB-44938	5500994172383112813930_A01	array assay	A-GEOD-13667	ArrayExpress	P-MTAB-44939	5500994172383112813930_A01.cel	ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/MTAB/E-MTAB-3610/E-MTAB-3610.raw.1.zip	NCI-H1869
J132_EPA01P9_CMK_haematopoietic_and_lymphoid_tissue_910566	Homo sapiens	CMK	cell	P-MTAB-44941	P-MTAB-44940	J132_EPA01P9_CMK_haematopoietic_and_lymphoid_tissue_910566	P-MTAB-44937	J132_EPA01P9_CMK_haematopoietic_and_lymphoid_tissue_910566:Biotin labeled streptavidin	Biotin labeled streptavidin	P-MTAB-44938	5500994172948120113978_A01	array assay	A-GEOD-13667	ArrayExpress	P-MTAB-44939	5500994172948120113978_A01.cel	ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/MTAB/E-MTAB-3610/E-MTAB-3610.raw.1.zip	CMK
J132_EPA02P10_SNU-61_large_intestine_1660035	Homo sapiens	SNU-61	cell	P-MTAB-44941	P-MTAB-44940	J132_EPA02P10_SNU-61_large_intestine_1660035	P-MTAB-44937	J132_EPA02P10_SNU-61_large_intestine_1660035:Biotin labeled streptavidin	Biotin labeled streptavidin	P-MTAB-44938	5500994173212120213068_A02	array assay	A-GEOD-13667	ArrayExpress	P-MTAB-44939	5500994173212120213068_A02.cel		

Figura 9 - Characteristics[cell line]

Fonte: European Bioinformatics Institute (2023)

Não fazendo parte do Projeto GDSC1, mas sendo um ficheiro muito importante noutros projetos científicos, fica ainda aqui a nota da existência do ficheiro ADF, onde se pode encontrar todos os genes com as devidas posições na sonda da Affimetrix Human Genome U219 Array.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

```
Array Design Name       [HG-U219] Affymetrix Human Genome U219 Array
Provider        " Affymetrix Inc.(geo@ncbi.nlm.nih.gov, support@affymetrix.com)"
Comment[ArrayExpressAccession]  A-GEOD-13667
Comment[SecondaryAccession]     GPL13667
Comment[Description]    "Array Manufacturer: Affymetrix, Distribution: commercial, Technology: in situ oligonucleotide<br>Reporter Database Entry [genbank] = GenBank accession numbers
LINK_PRE:http://www.ncbi.nlm.nih.gov/sites/entrez?db=nucleotide&cmd=search&term= DELIMIT:,"
Comment[SubmittedName]  [HG-U219] Affymetrix Human Genome U219 Array
Comment[Organism]       Homo sapiens
Comment[ArrayExpressReleaseDate]        2011-05-31
Printing Protocol       see manufacturer's website
Surface Type
Substrate Type
Term Source Name        genbank
Comment[AdditionalFile:txt]     A-GEOD-13667_comments.txt

[main]
Reporter Name   Reporter Database Entry [genbank]
11715100_at     NM_003534
11715101_s_at   NM_003534
11715102_x_at   NM_003534
11715103_x_at   NM_001167942;NM_152362
11715104_s_at   NM_178160
11715105_at     NM_173625
11715106_x_at   NM_178561
11715107_s_at   NM_001007523;NM_001007524;NM_012151
11715108_x_at   NM_001085490;NR_028342
11715109_at     NM_182610
11715110_at     NM_001080523
11715111_s_at   NM_000737;NM_033043;NM_033142;NM_033183;NM_033377;NM_033378
11715112_at     NM_001002912
11715113_x_at   NM_001099653;NM_018172;NM_152563
11715114_x_at   NM_001099653;NM_018172;NM_152563
11715115_s_at   NM_003525
11715116_s_at   NM_003545
11715117_x_at   NM_021066
11715118_s_at   NM_003522
11715119_s_at   NM_001007595
11715120_s_at   NM_003523
11715121_s_at   NM_003529
11715122_at     NM_004283
```

Figura 10 - Array Design Name.

Fonte: European Bioinformatics Institute (2023)

O ficheiro CEL engloba todos os cálculos da intensidade nos valores do pixel. Segundo a Affymetrix Developer Network - Affymetrix CEL Data File Format este ficheiro contém "...um valor de intensidade, desvio padrão da intensidade, o número de pixels usados para calcular o valor de intensidade, um sinalizador para indicar um *outlier* conforme calculado pelo algoritmo e um sinalizador definido pelo utilizador indicando que o recurso deve ser excluído de análises futuras." Assim o ficheiro armazena os detalhes acima declarados para cada característica na matriz de testes.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Por outro lado, e para uma melhor compreensão o ficheiro CEL está dividido nas seguintes secções header, intensidade, masks e outliers.

Secção "Header" com os campos apresentados na Figura 11.

MARCAÇÃO	Descrição
Cols	O número de colunas na matriz (de células).
Linhas	O número de linhas na matriz (de células).
TotalX	O mesmo que Col.
TotalY	O mesmo que Linhas.
Offset X	Não usado, sempre 0.
OffsetY	Não usado, sempre 0.
GridCornerUL	Coordenadas XY do canto superior esquerdo da grade em coordenadas de pixel.
GridCornerUR	Coordenadas XY do canto superior direito da grade em coordenadas de pixel.
GridCornerLR	Coordenadas XY do canto inferior direito da grade em coordenadas de pixel.
GridCornerLL	Coordenadas XY do canto inferior esquerdo da grade em coordenadas de pixel.
Axis-invertX	Não usado, sempre 0.
AxisInvertY	Não usado, sempre 0.
trocaXY	Não usado, sempre 0.
DatHeader	O cabeçalho do arquivo DAT.
Algoritmo	O nome do algoritmo usado para criar o arquivo CEL.
Parâmetros do Algoritmo	Os parâmetros usados pelo algoritmo. O formato é pares TAG:VALUE separados por ponto e vírgula ou pares TAG=VALUE separados por espaços.

Figura 11 - Secção "HEADER" contém diversas informações de cabeçalho.

Fonte: Affymetrix (2009)

Secção "Intensidade" na Figura 12.

MARCAÇÃO	Descrição
NúmeroCélulas	O número total de células na matriz (Rows*Cols)
CellHeader	O cabeçalho para o restante dos dados nesta seção. O cabeçalho é sempre definido como: " XY MEAN STDV NPIXELS"
N / D	As linhas restantes nesta seção contêm a intensidade, o valor do desvio padrão e o número de pixels usados para calcular o valor da intensidade para cada célula na matriz. A ordem é definida pelo cabeçalho.

Figura 12 - Secção "INTENSIDADE" contém informações de intensidade.

Fonte: Affymetrix (2009)

Secção "Masks" na Figura 13.

MARCAÇÃO	Descrição
NúmeroCélulas	O número de células mascaradas.
CellHeader	O cabeçalho para o restante dos dados nesta seção. O cabeçalho é sempre definido como: "X Y".
N / D	As linhas restantes nesta seção contêm as coordenadas XY dessas células mascaradas pelo usuário.

Figura 13 - Secção "MASKS" especifica quais células foram mascaradas pelo utilizador.

Fonte: Affymetrix (2009)

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Secção "Outliers" na Figura 14.

MARCAÇÃO	Descrição
NúmeroCélulas	O número de células atípicas.
CellHeader	O cabeçalho para o restante dos dados nesta seção. O cabeçalho é sempre definido como: "X Y".
N / D	As linhas restantes nesta seção contêm as coordenadas XY dessas células chamadas de outliers pelo software.

Figura 14 - Secção "OUTLIERS" especifica as células que foram identificadas como outliers pelo software.

Fonte: Affymetrix (2009)

Assim para a análise dos dados serão usados ficheiros CEL individualmente, fornecidos por este banco de dados genómicos, onde serão obtidas informações sobre a expressão génica de uma amostra específica, com os níveis de expressão de um conjunto de genes específicos. Estas informações poderão ser usadas para identificar padrões de expressão génica que possam estar relacionados à sensibilidade ou resistência a determinadas drogas numa determinada amostra de célula cancerígena. Estes poderão ser usados para apresentar terapias personalizadas a doentes com patologia oncológica. Segue infra um exemplo da informação que consta no interior de um ficheiro CEL.

```
[CEL]
Version=3

[HEADER]
Cols=744
Rows=744
TotalX=744
TotalY=744
OffsetX=0
OffsetY=0
GridCornerUL=0 0
GridCornerUR=743 0
GridCornerLR=743 743
GridCornerLL=0 743
Axis-invertX=0
AxisInvertY=0
swapXY=0
DatHeader= □ □ HG-U219.1sq □ □ □ □ □ □ □ □ □ □
Algorithm=Percentile
AlgorithmParameters=Percentile:75;CellMargin:2;OutlierHigh:1.5
00000;OutlierLow:1.004000;NumPixelsToUse:0;ImageCalibration:FA
LSE;FeatureExtraction:TRUE;IgnoreShiftRowOutliers:FALSE;FixedC
ellSize:TRUE;UseSubgrids:FALSE;RandomizePixels:FALSE;ErrorBasi
s:StdvMean;PercentileSpread:15.000000;StdMult:1.000000;ExtendP
oolWidth:2;ExtendPoolHeight:2;OutlierRatioLowPercentile:55;Out
lierRatioHighPercentile:75;HalfCellRowsDivisor:5;HalfCellRowsR
emainder:4;CellIntensityCalculationType:Percentile;HighCutoff:
3500;LowCutoff:4096;PairCutoff:2.500000;featureRows:109;featur
eColumns:127;featureWidth:8.000000;featureHeight:8.000000;Full
FeatureWidth:8;FullFeatureHeight:8

[INTENSITY]
NumberCells=553536
CellHeader=X    Y    MEAN  STDV  NPIXELS
  0    0  251.5 80.3   36
  1    0  5576.8     1024.6      36
  2    0  711.5 167.6  36
  3    0  5725.3     969.0  36
  4    0  197.0 37.6   36
  5    0  183.5 22.1   36
  6    0  161.3 32.1   36
  7    0  6646.3     1220.2      36
  8    0  187.5 32.6   36
  9    0  189.5 17.4   36
 10    0  692.3 107.8  36
 11    0  195.0 15.0   36
 12    0  108.8 8.6    36
```

Figura 15 - Exemplo de Ficheiro CEL

Fonte: Cancer Genome Project et al. (2022)

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Para analisar cada ficheiro CEL, será necessário utilizar um software de análise de microarray, que no caso deste trabalho será o software RStudio Desktop, instalado na máquina, usando várias bibliotecas como por exemplo a Affy da Biocondutor para pré-processar estes dados. Seguidamente após o processamento dos dados poder-se-á realizar várias análises como por exemplo análises de expressão génica. A Biocondutor é um projeto que visa apoiar e disseminar a produção de software de código aberto, oferecendo uma análise rigorosa e replicável de dados biológicos.

Neste presente *Dataset* cada ficheiro CEL representa uma amostra, ou seja, ir-se-ão fazer análises para cada amostra ou comparação entre elas. Por outro lado, será ainda importante compreender o protocolo usado para cada uma destas amostras. Assim por exemplo para a amostra do ficheiro "55009941732121202130 68_A01.cel" o protocolo usado foi "J132_EPA01P10_UACC-812_breast_910910: Biotin", ou seja, este nome refere-se ao nome do arquivo de imagem gerado pelo scanner do microarray.

Assim as informações sobre esta amostra, a plataforma e a data do microarray é:

- "**J132**" - identificador do scanner usado neste microarray.
- "**EPA01P10**" - nome da plataforma de microarray usada para produzir o microarray.
- "**UACC-812_breast_910910**" - informações sobre a amostra, incluindo o nome da célula (UACC-812) e o tecido de origem (Breast). O "910910" poder-se-á referir a uma data, como "10-09-1991".
- "**Biotin**" - tipo de corante usado no microarray para marcar as sondas.

Neste ponto serão analisadas 20 amostras, 10 amostras referentes à patologia da Mama (Breast) e 10 amostras referentes a patologias da Pele (Skin). Estas imagens foram extraídas aleatoriamente do *Dataset* Projeto *GDSC1* conforme a Figura 16.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

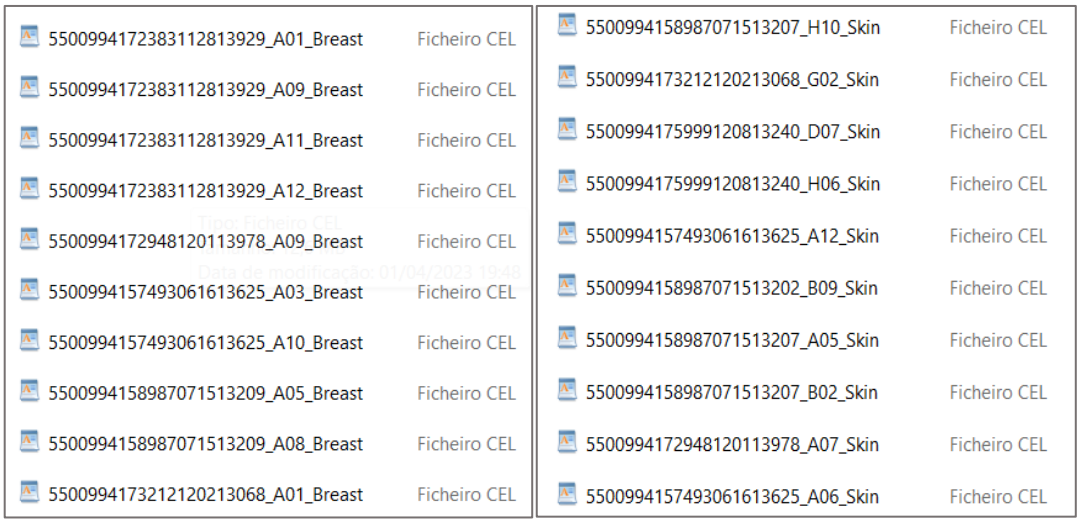


Figura 16 - Ficheiros CEL estudados (20 amostras)

Fonte: Cancer Genome Project et al. (2022)

Seguidamente os ficheiros CEL foram carregados no software RStudio Desktop através da biblioteca Affy. Na Figura 17 podemos visualizar através da AffyBatch que estamos perante matrizes de 744 x 744, microarrays da Affimetrix U219, com 49386 genes, e onde para simplificar, a anotação destes ficheiros de intensidade pode ser obtida através da biblioteca "hug219". Salienta-se ainda que todo este código se encontra no apêndice 1 deste estudo.

```
> sampleNames(arrays)
 [1] "A01.5500_Breast.cel" "A01_Breast.cel"      "A03_Breast.cel"      "A05_Breast.cel"
 [5] "A05_Skin.cel"        "A06_Skin.cel"        "A07_Skin.cel"        "A08_Breast.cel"
 [9] "A09.3978_Breast.cel" "A09_Breast.cel"      "A10_Breast.cel"      "A11_Breast.cel"
[13] "A12_Breast.cel"      "A12_Skin.cel"        "B02_Skin.cel"        "B09_Skin.cel"
[17] "D07_Skin.cel"        "G02_Skin.cel"        "H06_Skin.cel"        "H10_Skin.cel"
```

```
> arrays
AffyBatch object
size of arrays=744x744 features (23 kb)
cdf=HG-U219 (49386 affyids)
number of samples=20
number of genes=49386
annotation=hgu219
notes=
```

Figura 17 - Características das 20 amostras

Na Figura 18 pode-se observar um *Boxplot* com a distribuição das intensidades das 20 amostras.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

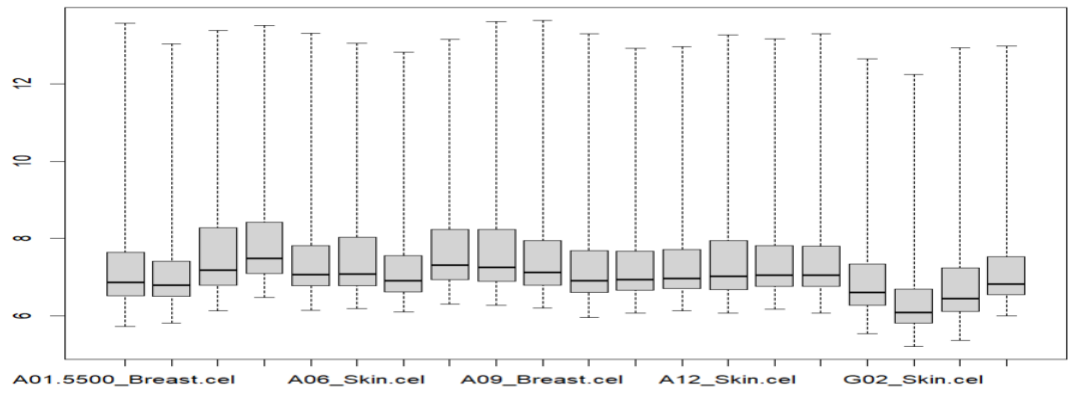


Figura 18 - Boxplot com a distribuição das intensidades das 20 amostras

Fonte: Bento et al. (2023)

Na Figura 18 podem-se verificar pequenas variações na distribuição das intensidades das 20 amostras.

4.2 Relatório de qualidade dos dados

Segundo Kauffmann et al (2009) a análise de microarrays, tem como uma das suas principais preocupações a garantia da qualidade dos dados. Assim, para ajudar neste processo foi usado o pacote Bioconductor "arrayQualityMetrics", que apresenta um relatório completo com gráficos e diagnósticos para os dados destes *microarrays* do presente estudo. Este pacote utiliza métricas de qualidade para avaliar a reprodutibilidade, detetar matrizes atípicas e calcular medidas de relação sinal-ruído. Embora o diagnóstico de qualidade seja sempre dependente do contexto, o pacote Bioconductor "arrayQualityMetrics" oferece instrumentos interessantes, objetivos, automatizados e abrangentes para auxiliar a tomada de decisões nesta área da biologia.

Seguidamente, na Figura 19 apresenta-se o relatório de qualidade através do pacote Bioconductor "arrayQualityMetrics" para os nossos dados, ou seja, para as intensidades dos vinte ficheiros CEL.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

array	sampleNames	*1	*2	*3	*4
1	A01.5500_Breast.cel				
2	A01_Breast.cel				x
3	A03_Breast.cel				
4	A05_Breast.cel	x			
5	A05_Skin.cel				
6	A06_Skin.cel				
7	A07_Skin.cel				x
8	A08_Breast.cel			x	
9	A09.3978_Breast.cel	x			
10	A09_Breast.cel				
11	A10_Breast.cel				
12	A11_Breast.cel				
13	A12_Breast.cel				
14	A12_Skin.cel				
15	B02_Skin.cel				
16	B09_Skin.cel				
17	D07_Skin.cel				
18	G02_Skin.cel	x	x		x
19	H06_Skin.cel				
20	H10_Skin.cel				

Figura 19 - Métricas de qualidade das 20 amostras (output)

As colunas numeradas de *1 a *4 indicam os diferentes métodos para a deteção de valores atípicos:

*1. Deteção de valores atípicos por distâncias entre arrays

*2. Deteção de valores atípicos por *Boxplots*

*3. Deteção de valores atípicos por RLE (*Relative Log Expression*)

*4. Deteção de valores atípicos por NUSE (*Normalized unscaled standard errors*)

Alguns dos critérios de deteção destes valores atípicos são explicados nas próximas secções. As matrizes que foram chamadas de *outliers* por pelo menos um critério são marcadas pela seleção na linha de seleção nesta tabela e são indicadas por linhas ou pontos destacados em alguns dos gráficos infra.

4.2.1 Comparação entre arrays

A Figura 20 expõe um mapa de calor de cores com as distâncias entre as matrizes. A escala de cores é escolhida para cobrir o intervalo de distâncias encontradas no conjunto

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

de dados. Segundo o pacote Bioconductor "arrayQualityMetrics" padrões neste gráfico poderá indicar o agrupamento das matrizes devido a fatores biológicos ou experimentais não intencionais pretendidos. A distância entre duas matrizes A e B é calculada como a diferença absoluta média (L1-distância) entre os dados das matrizes (usando os dados de todas as sondas sem filtragem). Na fórmula, DAB = média | M Ai - M Bi |, onde M Ai é o valor da i-ésima sonda na a-ésima matriz. A deteção de *outliers* foi realizada procurando matrizes para as quais a soma das distâncias para todas as outras matrizes, SA = ΣB DAB era excecionalmente grande. Três matrizes foram detetadas, sendo marcadas no gráfico por um asterisco (*).

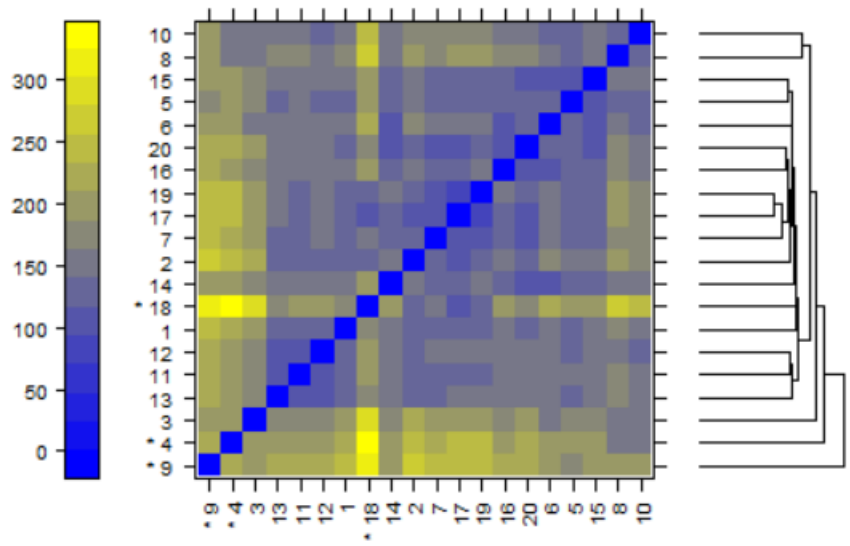


Figura 20 - Mapa de calor de distâncias entre matrizes

Destas três matrizes destacam-se desde já a número 9 e 4, por serem *outliers* do grupo mama, o que também se irá confirmar mais a frente com a análise da expressão génica.

4.2.2 Distribuições de intensidade da matriz

A Figura 21 *Boxplots* representa o resumo das distribuições de intensidade do sinal das matrizes. Cada caixa corresponde a uma matriz. Normalmente, espera-se que as caixas tenham posições e larguras semelhantes. Se a distribuição de uma matriz é muito diferente das outras, isso pode indicar um problema experimental. Segundo o pacote Bioconductor

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

arrayQualityMetrics deteção de *outliers* foi realizada calculando-se a estatística K de Kolmogorov-Smirnov entre a distribuição de cada matriz e a distribuição dos dados agrupados**.**

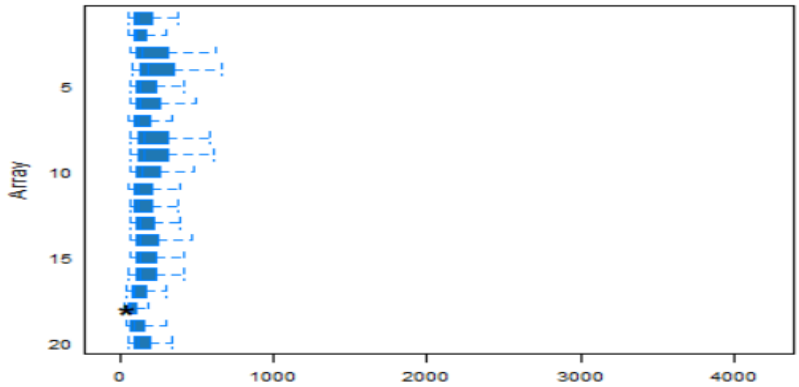


Figura 21 - Boxplots com as distribuições de intensidade do sinal das matrizes

Da análise da Figura 21 verifica-se que a distribuição das intensidades destas 20 amostras são bastante semelhantes. Ou seja, ter-se-á que optar por outra análise que não a análise de intensidades, para conseguir aferir a semelhança entre as amostras nestes dois grupos.

4.2.3 Dependência média da variância - Desvio padrão versus classificação da média

A Figura 22 exibe um gráfico de densidade do desvio padrão das intensidades entre as matrizes no eixo y versus a classificação de sua média no eixo x. De acordo com o pacote Bioconductor "arrayQualityMetrics" os pontos vermelhos, conectados por linhas, mostram a mediana do desvio padrão. Após a normalização e transformação para uma escala logarítmica, normalmente espera-se que a linha vermelha seja aproximadamente horizontal, ou seja, não mostre nenhuma tendência substancial. Em alguns casos, uma ligeira saliência cónica à direita do eixo x pode ser observada como sintoma de uma saturação das intensidades.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

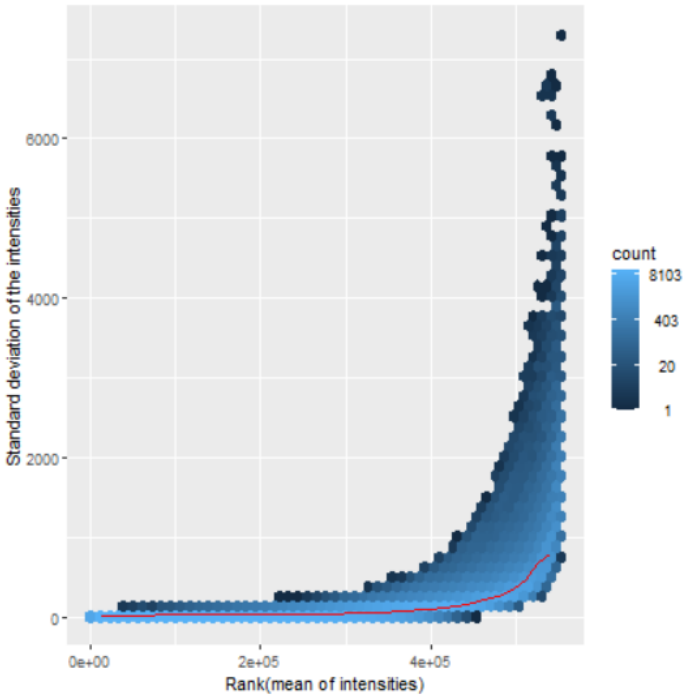


Figura 22 - Densidade do desvio padrão das intensidades

A ligeira saturação de intensidades, observada na Figura 22, é mais um sinal de que esta poderá não será a melhor opção para o nosso problema. Sendo ainda importante identificar e corrigir esta saturação de intensidades.

4.3 Análise de expressão génica (RNA-Seq)

Na análise da expressão génica foi usada a expressão da Figura 23 - Código para a análise da expressão génica, onde se pode visualizar que as vinte amostras foram classificadas com “B” Mama e “S” Pele.

```
dados <- matrix(as.integer(arraysRMAtable), ncol=20)
condition <- factor(c("B","B","B","B","S","S","S","B","B","B","B","B","B","S","S","S","S","S","S","S"))

dds <- DESeqDataSetFromMatrix(countData = dados, DataFrame(condition), design = ~ 1)
dds <- DESeq(dds)
res <- results(dds)
res

res_df <- as.data.frame(res)
```

Figura 23 - Código para a análise da expressão génica

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Os resultados apresentados são do pacote DESeq2 para análise de expressão génica diferencial. A Figura 24 apresenta a tabela com os resultados da análise de RNA sequencing (RNA-Seq). Cada linha representa um gene, e cada coluna fornece informações sobre a expressão do gene. A primeira linha mostra informações sobre o modelo ajustado pelo pacote, a segunda linha mostra o resultado do Teste de Wald para testar a hipótese nula de que não há diferença significativa na expressão entre grupos de tratamento e controle.

```
log2 fold change (MLE): Intercept
Wald test p-value: Intercept
DataFrame with 49386 rows and 6 columns
       baseMean log2FoldChange     lfcSE      stat      pvalue        padj
      <numeric>      <numeric> <numeric> <numeric>   <numeric>   <numeric>
1          3.65        1.86790  0.178197  10.48222 1.04269e-25 1.97987e-25
2          3.75        1.90689  0.176274  10.81774 2.83684e-27 5.54919e-27
3          3.50        1.80735  0.181808   9.94103 2.75952e-23 4.99548e-23
4          3.55        1.82782  0.179806  10.16553 2.82590e-24 5.23501e-24
5          2.95        1.56071  0.195650   7.97709 1.49826e-15 1.79520e-15
...         ...            ...       ...       ...         ...         ...
49382      5.15        2.36457  0.152374  15.51817 2.61395e-54 7.60666e-54
49383      6.10        2.60881  0.140920  18.51274 1.63000e-76 6.41735e-76
49384      2.10        1.07039  0.228536   4.68367 2.81780e-06 2.82966e-06
49385      2.55        1.35050  0.209268   6.45343 1.09347e-10 1.15076e-10
49386      2.85        1.51096  0.198829   7.59930 2.97733e-14 3.37244e-14
```

Figura 24 -Rresultados do Pacote DESeq2

A terceira linha apresenta os resultados da análise para o DataFrame com 49386 linhas e 6 colunas.

Análise das colunas são:

- **baseMean**: a média das contagens para todas as amostras;
- **log2FoldChange**: a mudança média na expressão génica (em escala log2) entre os vários grupos, patologia da mama e da pele;
- **lfcSE**: o erro padrão da mudança na expressão génica (em escala log2) estimada;
- **stat**: o valor estatístico do teste de Wald para a hipótese nula de que a mudança na expressão génica é zero;
- **pvalue**: o valor p do teste de Wald;

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

- **padj**: o valor p ajustado para múltiplas comparações usando um procedimento de correção de FDR (*false discovery rate*).

Na Figura 25 os valores de log2FoldChange representam a direção e a magnitude da mudança na expressão génica entre os dois grupos. Um valor positivo indica que o gene está aumentando na expressão do grupo, enquanto um valor negativo indica que o gene está diminuindo na expressão do grupo. Os valores de padj são frequentemente usados para filtrar genes diferencialmente expressos, com um valor de corte típico de 0,05. Ao analisar esses dados, é importante considerar o valor de log2FoldChange, o valor p e o valor padj para cada gene individualmente, bem como o conjunto de genes em geral. Além disso, é necessário levar em conta a biologia do sistema estudado e as limitações técnicas do RNA-Seq para interpretar corretamente esses dados e tirar conclusões sobre a regulação genética.

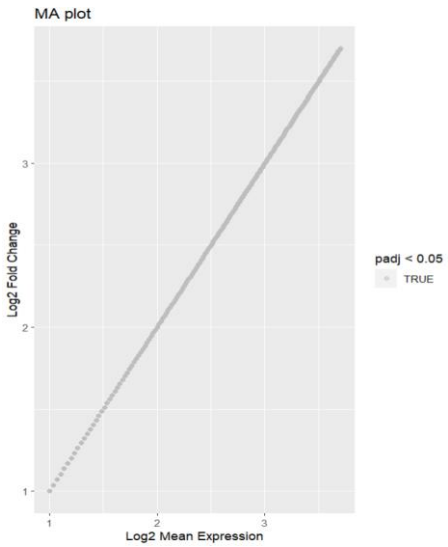


Figura 25 - MA plot na análise da expressão génica

Na Figura 25 pode-se aferir que todos esses genes têm uma expressão significativamente aumentada entre os dois grupos (Mama e Pele), sendo que essa diferença é estatisticamente significativa.

Uma maneira comum de visualizar os resultados da análise de expressão génica diferencial é por meio de um gráfico de *Volcano plot*. Na Figura 26, os valores de p são

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

apresentados em escala logarítmica no eixo y, e as diferenças na expressão génica (log2FoldChange) são apresentados em escala logarítmica no eixo x.

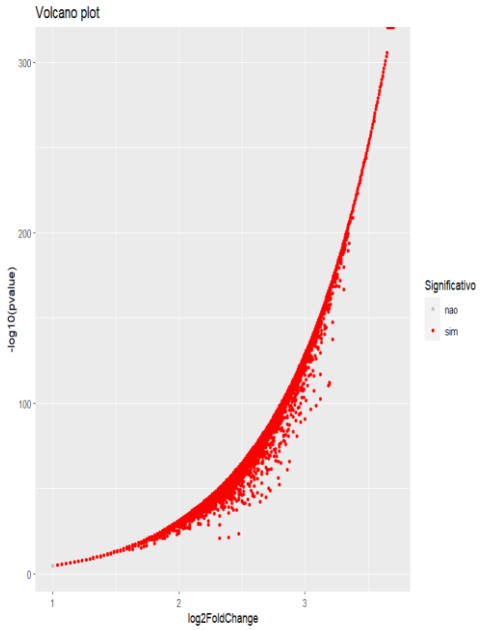


Figura 26 - Volcano plot na análise da expressão génica

A análise da expressão génica, através do pacote DESeq2, da Figura 26 apresenta também diferenças significativas entre os vários genes nos dois grupos (Mama e Pele). Ao longo deste projeto irão ser propostas outras abordagens fora deste pacote DESeq2 para testar e encontrar os genes diferencialmente expressos nestes dois grupos de maior relevo.

4.4 Diagrama de Venn

O Diagrama de Venn é frequentemente usado pela matemática, estatística, lógica, entre outras áreas, para ajudar a visualizar a relação entre os conjuntos de dados. Esta representação gráfica permite comparar conjuntos de elementos, mostrando as relações de inclusão e exclusão entre eles. Assim os círculos representam os conjuntos e as áreas onde os círculos se sobrepõem mostram os elementos que pertencem a mais de um conjunto. Na análise de microarrays, o Diagrama de Venn, serve para comparar diferentes conjuntos de genes e identificar aqueles que são exclusivos para cada conjunto. Ou seja, serve para comparar os conjuntos de genes que são diferencialmente expressos nos

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

diferentes tipos de cancro, e identificar os genes que são específicos para cada tipo de cancro, bem como aqueles que são comuns a mais de um tipo de cancro. Seguidamente o Na Figura 27 o Diagrama de Venn apresenta relativamente aos primeiros 10 000 genes dos microarrays, das quatro amostras apresentadas, o número de genes expressos nos dois tipos de cancro, cancro da mama e cancro da pele, e identificando o número de genes que são específicos para cada tipo de cancro, bem como aqueles que são comuns a mais do que um tipo de cancro. Na secção da modelação ir-se-á realizar um teste t, para estas e outras amostras, onde se poderá concluir quais os genes diferencialmente expressos.

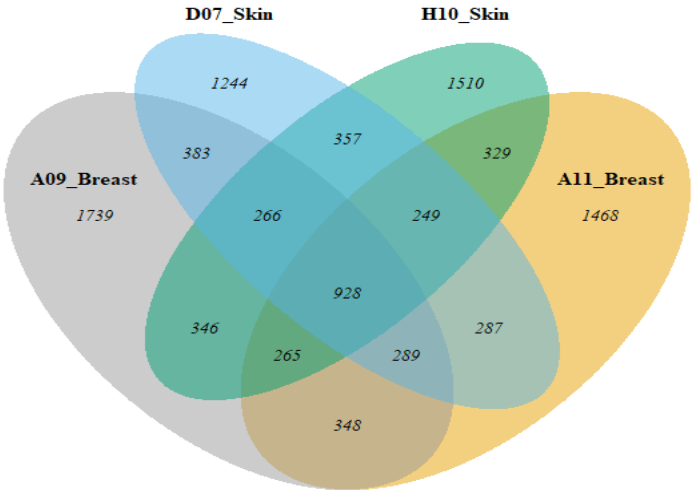


Figura 27 - Diagrama de Venn (para os 10000 primeiros genes do microarray)

Poder-se-á concluir que nestas quatro amostras a maioria dos genes não são comuns, conforme Diagrama Venn da Figura 27. Ou seja, pretendeu-se ilustrar quantos genes estão presentes em cada amostra e quantos são exclusivos de cada uma destas quatro amostras. Com esta análise e na observação de que a maioria dos genes não é comum entre as quatro amostras, conclui-se que há uma diversidade genética ou de expressão gênica considerável entre estas amostras. Podendo indicar diferenças biológicas significativas em cada amostra.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

4.5 Cluster

Seguidamente e para verificar a similaridade das vinte amostras, ir-se-á usar uma técnica de clusterização aplicada aos dados obtidos da expressão génica, por forma a verificar os agrupamentos de forma visual. Segundo Kassambara (2017) que sugere o uso da função "fviz_cluster" do pacote "factoextra", esta função utiliza os dados originais e os clusters encontrados para mostrar os resultados num gráfico usando a técnica de componentes principais, explicada mais detalhadamente no ponto 2.1.1.7. Assim, fica nítida a proximidade entre as várias amostras nos respetivos clusters.

A visualização dos resultados do cluster é uma abordagem útil para avaliar o número de clusters e comparar diferentes análises de cluster. Para conseguir isso, um gráfico de dispersão pode ser criado onde cada ponto de dados é colorido de acordo com seu cluster atribuído. No entanto, ao lidar com dados contendo mais de duas variáveis, torna-se um desafio decidir quais variáveis usar para um gráfico de dispersão x y. Uma solução para este problema é aplicar um algoritmo de redução de dimensionalidade como a Análise de Componentes Principais (PCA) para reduzir o número de dimensões para dois. Esse algoritmo opera em todas as variáveis e produz duas novas variáveis que podem ser usadas para criar o gráfico de dispersão. Em suma, para visualizar um conjunto de dados multidimensional com atribuições de cluster, podemos executar o PCA e plotar os pontos de dados de acordo com as duas primeiras coordenadas dos componentes principais. A função fviz_cluster(), pode ser usada para criar um gráfico de dispersão de clusters k-means. Essa função usa os resultados k-means e os dados originais como argumentos e, no gráfico resultante, as observações são representadas por pontos, usando componentes principais se o número de variáveis for maior que dois. Além disso, elipses de concentração podem ser desenhadas em torno de cada cluster para fornecer uma melhor compreensão dos dados. Assim foram criados dois cenários para dois e quatro clusters, Figura 28 e 29, para analisar os dois grupos de dados.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

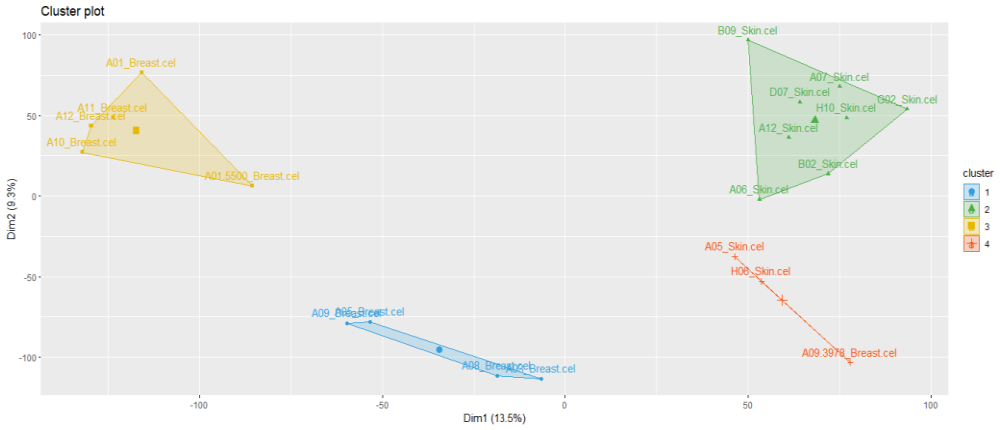


Figura 28 - Análise dados para 4 Clusters

Fonte: Bento et al. (2023)

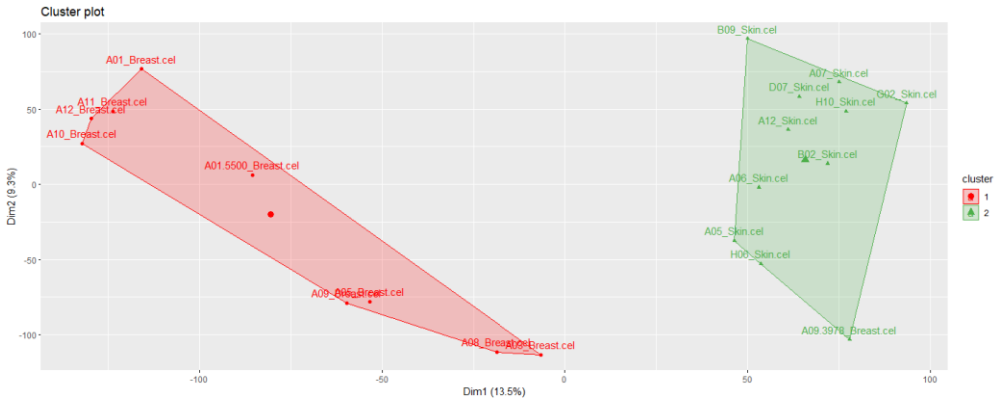


Figura 29 - Análise dados para 2 Clusters

Na Figura 28, dos 4 clusters conclui-se que o 4º cluster ter uma amostra da Mama misturada com as outras amostras da Pele, o mesmo é analisado na Figura 29 no 2º cluster.

4.6 Análise de Componentes Principais

Principal Component Analysis é uma técnica de análise multivariada utilizada para explorar e identificar padrões em conjuntos de dados com muitas variáveis. A análise PCA busca encontrar uma nova base de dados que represente a maior parte da variação dos dados originais, transformando as variáveis originais em um conjunto menor de variáveis não correlacionadas chamadas de componentes principais. Esses componentes são ordenados em termos de importância, sendo que o primeiro componente principal é

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

aquele que explica a maior parte da variação nos dados. Os componentes seguintes são escolhidos de forma que sejam ortogonais (não correlacionados) ao componente anterior e expliquem a maior parte da variação remanescente.

A análise PCA é amplamente utilizada em diversas áreas, incluindo estatística, ciência de dados, análise financeira e bioinformática, entre outras. Ela é útil para reduzir a dimensionalidade dos dados, identificar padrões e tendências, explorar relações entre variáveis e melhorar a visualização de dados complexos. A Figura 30 apresenta a análise da expressão génica para os dois grupos criados (Mama e Pele).

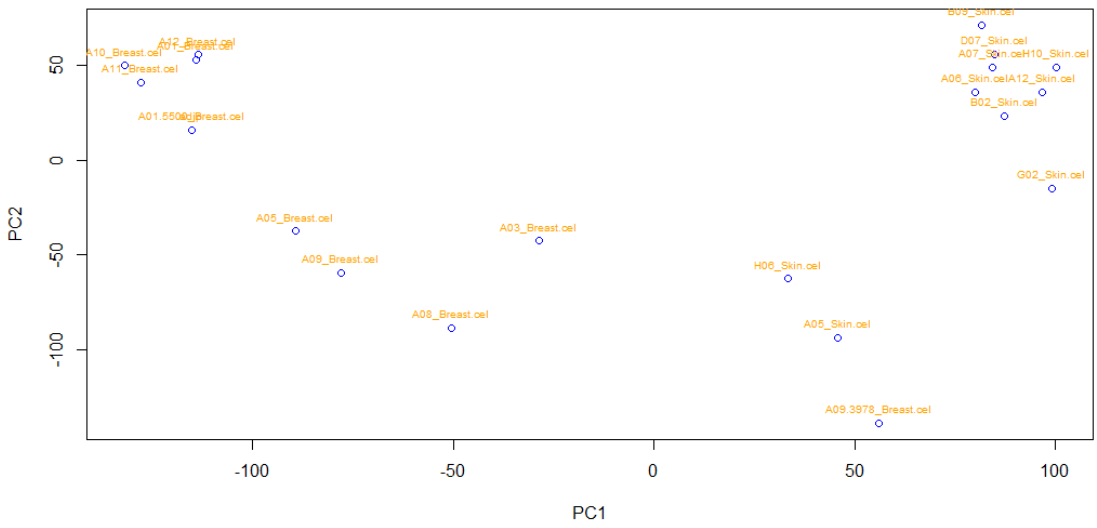


Figura 30 - Análise PCA

Aqui conclui-se, pela análise PCA, na Figura 35 que valores PC2 superior a 0 encontramos dois grupos de Mama e Pele com características muito semelhantes. Por outro lado, ao analisar-se PC1 menor que 0 temos amostras da Mama, sendo que algumas estão dispersas, ou com PC1 maior que 0 temos as amostras da Pele, com algumas mais dispersas, mas aqui temos também uma amostra da Mama. Em suma, estes resultados estão em consonância com os resultados da análise de clusters.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

4.7 Correlação de Pearson

A correlação é uma medida estatística usada para medir o grau de associação entre duas variáveis emparelháveis, ou seja, obtidas sobre a mesma unidade.

Segundo Afonso & Nunes (2019) o coeficiente de Pearson mede o grau de associação linear numa amostra bivariada quantitativa. A Figura 31 infra apresenta o mapa de calor dos dados da expressão génica, com a representação das vinte amostras.

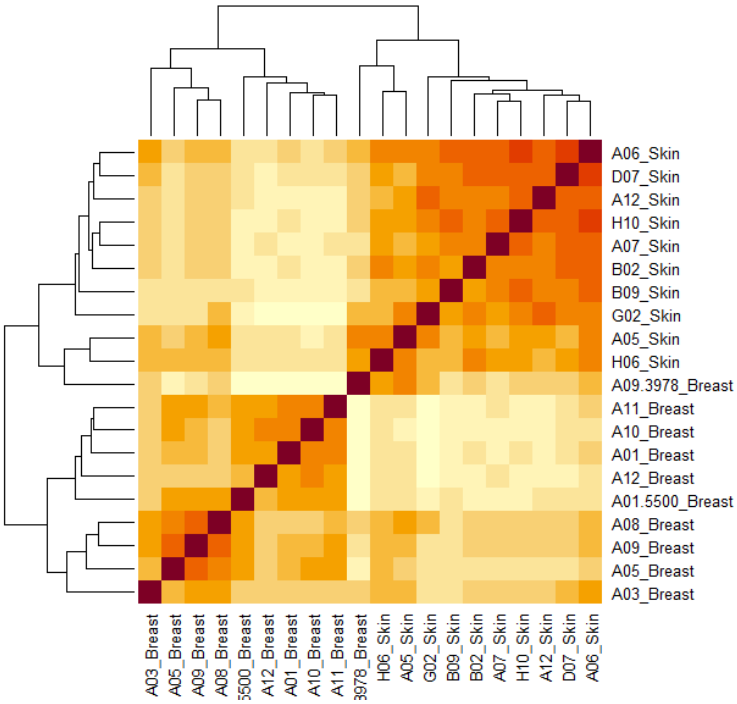


Figura 31 - Mapa de calor da corelação de Pearson

O mapa de calor apresentado na Figura 36 teve como pressupostos o facto de que uma maior correlação entre amostras está associada a cores mais escuras e de que uma menor correlação entre amostras está associada a cores mais claras. Observando a figura 31, constata-se que entre as amostras do mesmo grupo (mama ou pele, em inglês, “breast” ou “skin”) existe tipicamente maior correlação.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

5 PREPARAÇÃO DOS DADOS

Inicialmente, foi adotada uma abordagem para analisar e investigar a totalidade das amostras das 970 linhas celulares. No entanto, devido a limitações de *hardware*, a execução completa dessa abordagem abrangente não pôde ser realizada. Diante dessa limitação, teve-se de adotar outra alternativa: concentrar a análise na extração e avaliação dos dados contidos no arquivo "E-MTAB-3610.sdrf" *Dataset,* do Projeto *GDSC1*. Adotando uma abordagem que permita dividir o processo de análise em etapas, ou seja, implementar uma estratégia prática para lidar com grandes volumes de dados ou, por outro lado, quando há limitações tecnológicas. Desta forma, tentou-se otimizar o uso de recursos, explorando os dados de maneira mais focalizada e gradualmente aumentar a complexidade da análise à medida que soluções tecnológicas ou recursos adicionais se tornem disponíveis no futuro. Portanto, o enfoque inicial na análise de apenas 20 arquivos representa uma estratégia sensata para avançar na investigação, mesmo diante das limitações de processamento que impossibilitaram a análise completa da totalidade das amostras das 970 linhas celulares. Assim, foram extraídos 20 ficheiros CEL (10 referentes a amostras de mama e 10 referentes a amostras de Pele) da base de dados para estudar as suas intensidades e as imagens. Inicialmente as imagens apresentam semelhanças visuais consideráveis e poucas diferenças aparentes. Dada a necessidade de explorar um nível mais profundo da informação contida nas amostras, que pode não ser percetível a olho nu, mas que por outro lado poderá ter relevância significativa em termos de características genéticas e biológicas, optou-se por extrair a expressão génica das imagens. A extração da expressão génica das amostras permitirá a obtenção de dados quantitativos sobre a atividade dos genes em cada amostra. Permitindo visualizar padrões de expressão que não seriam facilmente identificáveis através de uma análise puramente visual. O próximo passo para analisar a similaridade entre as amostras, através do cálculo e análise das distâncias, foi usar o software RStudio Desktop comumente utilizado em estudos de análise de dados biológicos. O uso do software RStudio Desktop permite a aplicação de métodos estatísticos e algoritmos específicos para calcular as distâncias entre as expressões génicas das 20 amostras. Esta análise de similaridade pode revelar

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

agrupamentos naturais, padrões de expressão compartilhados ou ainda diferenças entre as amostras. Neste contexto, e segundo Gysi et al., (2018), adotou-se a biblioteca "affy" para calcular as intensidades de expressão génica. Esta biblioteca "affy" dispõe de vários métodos para o cálculo dos valores de expressão génica a partir dos dados de intensidade obtidos dos microarrays. Um desses métodos é o método RMA, que será explorado neste contexto. A escolha deste método, de pré-processamento e cálculo de expressão génica, deveu-se às particularidades e características dos dados e aos objetivos de análise do presente estudo. Por outro lado, através da função "rma()" disponível na biblioteca "affy", é possível realizar uma série de etapas de pré-processamento e cálculo para obter os valores de expressão génica com base no método RMA. Sendo que estas etapas incluem:

- Correção de Fundo: A correção de fundo visa ajustar as intensidades de sinal para remover ou atenuar o ruído de fundo presente nos *microarrays*. Isto é fundamental para garantir que as intensidades de sinal estejam mais próximas das verdadeiras intensidades de expressão génica.
- Normalização: A normalização é um passo importante para tornar os dados comparáveis entre os diferentes *microarrays*. Ajustando as intensidades de sinal para minimizar as variações técnicas, permitindo a comparação de expressão entre as amostras.
- Cálculo de Valores de Expressão: O cálculo dos valores de expressão propriamente dito é realizado com base no método RMA. Esse método utiliza informações de várias sondas que correspondem a um mesmo gene para obter uma estimativa mais precisa da expressão génica. As intensidades de sinal das sondas são resumidas para produzir um valor representativo da expressão do gene.

A abordagem RMA é valorizada pela sua robustez contra variações técnicas e sua capacidade de lidar com dados de *microarry* de alta dimensão. Ao resumir as intensidades de várias sondas em um único valor de expressão para cada gene, RMA contribui para reduzir o efeito do ruído e aumentar a confiabilidade nas estimativas de expressão génica. Em suma, o uso da função "rma()" da biblioteca "affy", no calculo dos valores de expressão génica com base no método RMA é uma etapa essencial na análise de dados

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

de *microarry*, visando obter informações precisas e confiáveis sobre os níveis de expressão dos genes nas 20 amostras estudadas.

Seguidamente, e para testar os algoritmos de ML, onde será usado o software Python, teve-se que inicialmente carregar o arquivo "RMAvalues.xlsx", que contém os valores de expressão génica calculados com base no método RMA. O passo seguinte passa por converter os dados em um DataFrame através da biblioteca "Pandas". Por forma a reorganizar os dados de acordo com as necessidades dos algoritmos de ML a implementar, e usando a biblioteca "NumPy" e a função np.transpose() aplicada ao DataFrame, esta operação irá transformar as linhas do DataFrame em colunas e vice-versa. Assim com o DataFrame transposto, ir-se-á criar uma coluna com a classificação de cada amostra. Ou seja, as amostras "Mama" devem ser rotuladas com 0 e as amostras "Pele" com 1. Esta coluna de classificação é crucial para o treino e avaliação dos algoritmos de ML. Destaca-se ainda que a etapa de pré-processamento e preparação dos dados é essencial para o sucesso da análise dos algoritmos de ML. O passo de transformação do DataFrame usando a função np.transpose() e a criação da coluna de classificação são etapas cruciais para garantir que os dados estejam no formato correto para que os algoritmos de ML possam aprender com eficácia os padrões presentes nos dados em estudo.

Tabela 3 - DataFrame para algoritmos de ML

	0	1	2	3	4	5	6	7	8	9	...	49377	49378	49379	49380	49381	49382	49383	49384	49385	PATOLOGIA
Breast	3.506779	3.837062	3.034804	3.320748	3.207577	3.406380	4.459605	3.899979	3.106215	2.982536	...	2.939977	3.330262	2.980037	8.291359	6.427310	7.112765	2.737219	2.918268	3.184849	0
Breast.1	3.606053	3.614073	3.263720	3.542708	3.283971	3.508140	4.415589	3.636535	3.593522	3.804533	...	3.325428	3.077175	3.113228	9.006563	6.744123	7.805018	3.013973	3.231146	3.140215	0
Breast.2	3.399144	3.809148	3.085871	4.240574	3.063888	2.932110	3.584010	3.569080	2.865101	3.193683	...	3.008207	2.938454	3.142097	7.675911	4.932720	6.045096	2.810140	2.924056	2.920566	0
Breast.3	6.868023	6.746078	6.328479	4.918580	3.118956	3.091272	4.885773	4.149956	3.054706	3.264133	...	2.988093	3.020701	2.807498	7.780543	4.961774	5.953088	2.906912	3.066872	3.055157	0
Skin	3.687865	3.699593	3.239385	4.630733	3.174554	3.180487	4.402737	3.815917	3.108357	3.169952	...	3.180734	3.156036	2.962118	7.635859	5.273305	6.013062	2.771389	3.046477	3.002211	1
Skin.1	3.625251	3.442291	3.122627	3.932745	3.384378	2.947452	3.981439	3.724892	2.790038	3.186700	...	2.896712	3.221304	3.108655	7.893739	5.079450	6.138925	2.909423	2.894133	3.133448	1
Skin.2	3.211393	3.557406	3.059871	4.309129	3.055184	3.551850	4.599679	4.304261	5.425494	3.295593	...	2.801132	3.212712	3.150708	7.724544	4.935027	5.584038	2.961398	3.128022	3.093553	1

A última transformação do conjunto de dados "RMAvalues.xlsx" constituiu em dividi-lo em duas partes, uma correspondente a 80% dos dados para treino e a outra a 20% dos dados para teste dos algoritmos de ML.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Em suma, os dados foram preparados para realizar os vários testes de similaridade com a finalidade de verificar qual seria a melhor opção para comparar as expressões génicas destas duas patologias e aferir a que grupo pertencem.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

6 MODELAÇÃO

6.1 Identificação de genes diferencialmente expressos

Em testes estatísticos paramétricos, uma das suas condicionantes, é que a variável em estudo seja quantitativa e tenha uma distribuição normal, ou pelo menos aproximadamente do normal. Para aferir esta condição existem vários procedimentos, como por exemplo o teste *Kolmogorov-Smirnov*, utilizado para amostras de maior dimensão, e o teste de *Shapiro-Wilk*, utilizado para amostras de pequena dimensão.

No presente estudo e acordo com Raul Laureano (2020), sendo a amostra em análise superior a 30 genes (n > 30), pela aplicação do Teorema do Limite Central, poder-se-á aferir que a distribuição da média amostral é aproximadamente normal. Assim, irá ser executado um teste t usando a função mt.teststat() implementada pela biblioteca multtest da Biocondutor do R. Esta função executará um teste t para cada conjunto de sondas. Posteriormente salvar-se-á o resultado desta função na variável stats. Essa variável será um vetor com os resultados das estatísticas do teste t para cada conjunto de sondas.

A Tabela 3 mostra na coluna "adjp" os resultados do Teste t para cada gene nas 20 amostras.

Tabela 4 - Teste t para cada gene nas 20 amostras

	A01.55	A01_Br	A03_Br	A05_Br	A05_Sk	A06_Sk	A07_Sk	A08_Br	A09.39	A09_Br	A10_Br	A11_Br	A12_Br	A12_Sk	B02_Sk	B09_Sk	D07_Sk	G02_Sk	H06_Sk	H10_Sk	adjp
11715100_at	3.5067	3.6060	3.3991	6.8680	3.6878	3.6252	3.2113	4.1301	3.4776	5.8372	5.7414	6.8003	5.4232	3.3765	3.6494	3.7427	3.7508	3.7298	3.6226	3.6546	0.047652
11715101_s_at	3.8370	3.6140	3.8091	6.7460	3.6995	3.4422	3.5574	4.0416	3.3907	6.5239	6.0282	6.9303	5.1076	3.6381	3.4844	3.3344	3.6055	3.4537	3.4472	3.7645	0.019870
11715102_x_at	3.0348	3.2637	3.0858	6.3284	3.2393	3.1226	3.0598	3.5893	3.2484	5.9044	5.8864	6.6449	4.8327	3.1343	3.2286	3.1223	2.9493	3.3594	3.1194	3.2124	0.031589
11715103_x_at	3.3207	3.5427	4.2405	4.9185	4.6307	3.9327	4.3091	3.9984	5.2538	4.2384	3.6119	3.7032	3.8499	4.6159	4.0122	3.8659	3.6704	3.5179	4.3814	4.3338	0.910321
11715104_s_at	3.2075	3.2839	3.0638	3.1189	3.1745	3.3843	3.0551	3.1743	3.1709	3.6241	3.2865	3.3706	3.2282	3.1806	3.1980	2.9569	3.1205	3.0413	3.2041	3.3251	0.424227
11715105_at	3.4063	3.5081	2.9321	3.0912	3.1804	2.9474	3.5518	3.1630	3.4067	3.3095	3.1485	3.6564	3.2567	3.2741	3.3902	3.1302	3.1443	3.5976	3.3341	3.0994	0.916101
11715106_x_at	4.4596	4.4155	3.5840	4.8857	4.4027	3.9814	4.5996	4.3099	4.3847	4.2920	4.3789	4.3943	4.3395	4.2837	4.2242	4.2644	4.1713	4.9542	4.2768	4.4190	0.966921
11715107_s_at	3.8999	3.6365	3.5690	4.1499	3.8159	3.7248	4.3042	3.4635	3.3879	3.4355	3.3613	3.6251	3.8507	3.2942	3.7622	4.0106	4.0054	3.8485	3.5361	3.5916	0.477905
11715108_x_at	3.1062	3.5935	2.8651	3.0547	3.1083	2.7900	5.4254	3.0230	3.2118	3.1021	2.9612	3.2242	3.1471	3.0268	3.2137	3.1944	3.0648	3.1562	2.8714	3.0791	0.742984
11715109_at	2.9825	3.8045	3.1936	3.2641	3.1699	3.1866	3.2955	3.1445	3.2575	3.3164	3.3254	3.5373	3.2792	3.3261	3.1949	3.3994	3.2603	3.2134	3.3936	3.3191	0.830455
11715110_at	2.9316	2.9220	3.0097	3.1230	2.7517	2.9171	3.1623	3.0264	2.7882	3.0249	3.2745	3.1211	3.0513	3.1140	3.1369	3.1449	3.4411	3.1411	2.8676	3.2804	0.636454
11715111_s_at	4.0049	4.2003	3.7279	4.4965	4.0856	3.8073	3.8810	4.8031	4.5359	4.3655	4.2350	3.9498	3.8546	4.0019	4.0306	4.1022	3.9103	4.2364	4.2016	3.7342	0.258066
11715112_at	3.0196	3.0362	2.9719	2.7961	3.0151	3.0727	3.1700	2.8394	3.0209	3.0543	3.1604	3.1517	9.2106	3.1447	2.8540	3.0091	2.9850	2.9754	2.8808	3.0502	0.599938
11715113_x_at	5.6088	5.6377	5.7385	4.4177	5.6227	5.2718	6.7068	4.9595	4.3497	5.1112	6.5821	5.2646	5.5667	6.1547	6.5034	6.0906	5.5165	5.2443	5.0097	6.5639	0.233390
11715114_x_at	5.0106	5.3358	5.5175	4.2486	5.4550	4.7932	6.5567	5.1498	4.3331	5.1725	6.3709	5.2217	5.6268	5.8608	6.2339	5.8023	5.5488	5.1838	4.8002	6.1104	0.332967
11715115_s_at	2.8678	3.1944	2.8600	2.9146	2.7386	2.8467	2.7418	2.7729	2.8306	3.0676	3.1480	3.1609	2.8799	2.6901	2.7552	2.7577	2.9971	2.9598	2.8837	2.8570	0.095540

O teste t compara os dois grupos, Mama representado por 1 e Pele representado por 0, representadas pelo vetor "groups" e a matriz "dados1". O vetor "*groups*" indica a qual

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

grupo cada observação pertence, onde 1 representa o primeiro grupo e 0 representa o segundo grupo. A função "mt.teststat" calcula a estatística t para cada gene (coluna da matriz) usando os dois grupos. A seguir, o valor de p bruto é calculado com base na distribuição t usando a função "pnorm" (que calcula a probabilidade acumulada da distribuição normal) e o valor absoluto da estatística t. Seguidamente, a função "p.adjust" é usada para ajustar os valores de p brutos para corrigir o problema de múltiplas comparações (usando o método Benjamini-Hochberg). O resultado é uma matriz "arraysRMAstats" que vai conter as colunas da matriz "dados1" concomitantemente com os valores ajustados de p para cada teste t realizado. Em suma, todo o código realiza um teste t das duas amostras independentes para cada coluna da matriz "dados1", comparando as amostras nos grupos indicados pelo vetor "*groups*" e corrigindo os valores de p para múltiplas comparações. Assim, para os valores na coluna "adjp" ajustados, com valor de p abaixo de 0,05, existe evidência estatística suficiente para rejeitar a hipótese nula, de que não há diferença real entre os dois grupos. Ou seja, existe diferença estatística significante destes genes entre os dois grupos. Nesta situação encontram-se 4050 genes, conforme se pode constatar no gráfico da Figura 32, ou seja, poderemos estar a falar em genes relevantes ou biomarcadores nestas duas patologias. Este gráfico apresenta no eixo dos y os valores de p-value e o eixo dos x a quantidade de genes.

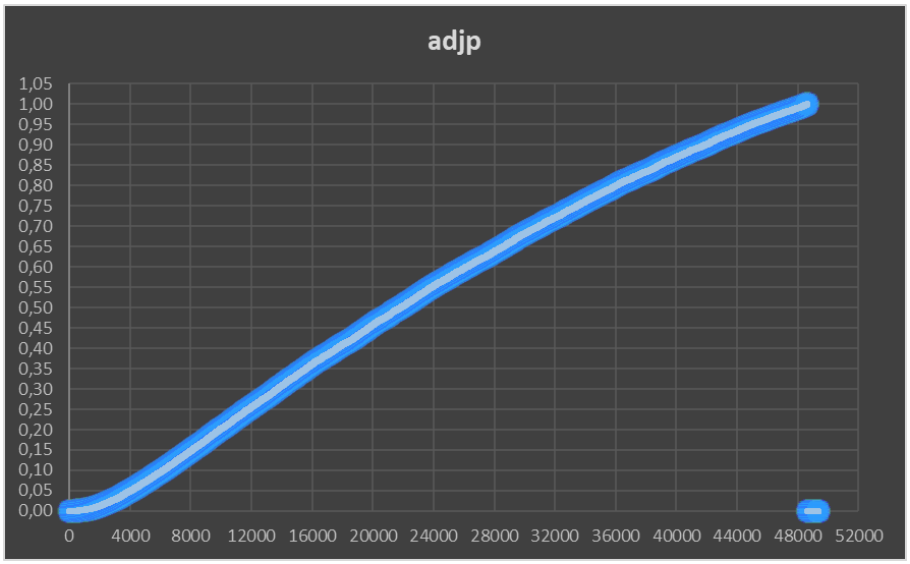


Figura 32 - Resultados Teste t

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

6.2 Índices de similaridade

A análise da expressão génica tem sido uma área de grande interesse na pesquisa biomédica, pois estes dados fornecem informações valiosas sobre a regulação genética em diferentes condições e doenças. No entanto, devido à complexidade e heterogeneidade desses dados, tem sido cada vez mais necessário o desenvolvimento de técnicas estatísticas adequadas para esta complexa análise. A quantificação de expressão génica em larga escala permite monitorizar em simultâneo milhares de genes em diferentes amostras. Essas amostras podem ser de origens variadas, incluindo diferentes tecidos, tipos de células ou doenças, além de envolver diferentes condições experimentais. A análise estatística de dados de expressão génica busca extrair informações relevantes desses perfis e identificar padrões ocultos. Essa análise pode ser realizada de forma supervisionada, incorporando informações biológicas prévias sobre genes e amostras, ou de forma não supervisionada, onde as informações são ignoradas e os padrões são identificados sem prévias informações. Assim e numa primeira abordagem para as vinte amostras foram calculados nove índices de similaridade para a sua expressão génica. Para o efeito usou-se o software *Python* dado as limitações do RStudio Desktop no processamento desta grande quantidade de dados (49386 genes x 20 amostras).

Tabela 5 - Cálculos de nove índices de similaridade para a expressão génica (49386 genes de cada amostra)

	Jaccard	Dice	MSE	RMSE	Cosine	Czekanowski	Pearson	Euclidiana	Simpson
Breast	[1.0]	[1.0]	[0.0]	[0.0]	[6.2379]	1.0000	[(1.0, 0.0)]	[0.0]	[0.0]
Breast.1	[0.00070356218948553337]	[0.0014061350754986373]	[1.1178486029422068]	[1.0572835962702753]	[6.0726]	0.9321	[(0.8935765176193363, 0.0)]	[234.9597223034276]	[0.0]
Breast.2	[0.00091836293703271154]	[0.0018350406407530145]	[1.4461583423773203]	[1.2025632384109037]	[6.1227]	0.9314	[(0.8689128191025959, 0.0)]	[267.2451606608553]	[0.0]
Breast.3	[0.00077555285069534047]	[0.0015499036691816721]	[1.0961467663135764]	[1.0469702795751064]	[6.1401]	0.9378	[(0.899506108378282, 0.0)]	[232.66779794626132]	[0.0]
Skin	[0.00075550751839147412]	[0.0015090109511080453]	[1.6207911145672418]	[1.2731029473562778]	[6.0817]	0.9293	[(0.8499959005672604, 0.0)]	[282.92117273901187]	[0.0]
Skin.1	[0.00089475329603397911]	[0.0017879068565150897]	[1.6008960239526049]	[1.265265199060104]	[6.0969]	0.9280	[(0.8537726363266447, 0.0)]	[281.17939298412915]	[0.0]
Skin.2	[0.00063781025685375771]	[0.0012748074284541341]	[1.7504884831772862]	[1.323060271936727]	[6.0571]	0.9237	[(0.8361912854342219, 0.0)]	[294.02316954654011]	[0.0]
Breast.4	[0.00094822477237217831]	[0.0018946529878462316]	[1.2044564087701937]	[1.097477293054482]	[6.1342]	0.9369	[(0.8896388826970532, 0.0)]	[243.89195190396254]	[0.0]
Breast.5	[0.0010025657057846963]	[0.002003123148995746]	[2.011612398591679]	[1.4183132230194002]	[6.057]	0.9186	[(0.8168176815639594, 0.0)]	[315.19119581111277]	[0.0]
Breast.6	[0.00076635796472594611]	[0.0015315422198733783]	[1.181302221319744]	[1.0868772797881756]	[6.1275]	0.9368	[(0.8906129952080812, 0.0)]	[241.53631507932067]	[0.0]
Breast.7	[0.00066990091949303631]	[0.0013389049053588588]	[1.0787113195708349]	[1.0386102828158572]	[6.1324]	0.9378	[(0.9001840197182269, 0.0)]	[230.80995911859014]	[0.0]
Breast.8	[0.0006816190074328931]	[0.0013623094388582552]	[1.1065177812668119]	[1.0519114892740795]	[6.1148]	0.9359	[(0.896801204452452, 0.0)]	[233.7658810554756]	[0.0]
Breast.9	[0.00057292342283910581]	[0.0011451907390801742]	[1.2564010241586831]	[1.1208929583857161]	[6.1128]	0.9334	[(0.8834490514599957, 0.0)]	[249.09560610155432]	[0.0]
Skin.3	[0.00067982432476179171]	[0.0013587249552483446]	[1.7174240185967058]	[1.3105052531740213]	[6.0779]	0.9262	[(0.841916994563399, 0.0)]	[291.2330726109535]	[0.0]
Skin.4	[0.00073422231819899591]	[0.0014673672626047927]	[1.7758506826106217]	[1.3326104767000078]	[6.0708]	0.9246	[(0.8366017194733748, 0.0)]	[296.1455078359423]	[0.0]
Skin.5	[0.00088568219130735341]	[0.0017697969050137054]	[1.6729152481670606]	[1.2934122498905987]	[6.0689]	0.9264	[(0.8446368027869078, 0.0)]	[287.43450114065723]	[0.0]
Skin.6	[0.00056217431728253591]	[0.0011237169097784983]	[1.7066896183358065]	[1.3064033138107873]	[6.0666]	0.9247	[(0.8420672460595664, 0.0)]	[290.32150022196447]	[0.0]
Skin.7	[0.0003252173535979881]	[0.00065022324331353771]	[1.8358917696008212]	[1.3549508365991814]	[6.0377]	0.9215	[(0.827100207752336, 0.0)]	[301.11019732567365]	[0.0]
Skin.8	[0.00063776199587076131]	[0.0012747110294911957]	[1.6231049523720635]	[1.2740113627327128]	[6.083]	0.9248	[(0.8503597832215084, 0.0)]	[283.12304953473273]	[0.0]
Skin.9	[0.00062670859131036131]	[0.0012526321472922631]	[1.7669140891637096]	[1.3292532073174432]	[6.0698]	0.9250	[(0.8368910389312554, 0.0)]	[295.3994231670722]	[0.0]

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Seguidamente ir-se-á calcular algumas distâncias de similaridade por forma a encontrar a métrica mais apropriada ao presente estudo. Assumindo os resultados do teste t realizado na secção anterior, e dado que estamos a observar a expressão génica de 20 amostras distribuídas por dois grupos (patologia da mama e pele), ir-se-á começar por analisar uma amostra de 2 genes (identificados pelos seus ID - 11715918_s_at, 11716887_a_at) escolhendo aleatoriamente com um valor de $p \approx 0$, sendo contrastado com 1 gene (identificados pelo seu ID - 11724186_a_at) com um valor $p \approx 0.7209$. Ou seja, irá ser calculada a distância entre as várias amostras para estes genes escolhidos.

De acordo com Bass et al. (2013) e Torrente (2021) os índices de similaridade mais usados na expressão génica são: Jaccard, Dice, Simpson, Cosine e Pearson. No entanto, e para alargar o espetro da análise, foram ainda escolhidos outros índices de similaridade para validar o desempenho dos anteriormente mencionados.

6.2.1 Gene Sonda ID - 11715918_s_at

Seguem-se os *Heatmaps* das Figuras 33 até à 37, com as distâncias de similaridade Dice, Jaccard, Sorensen, Czekanowski, Minkowski, Pearson, Intersection, Manhattan, Tanimoto e Euclideana, para a análise do gene identificado com a Sonda ID "11715918_s_at" nas 20 amostras (10 patologia da mama e 10 patologia da pele), estando os cálculos nas matrizes de similaridade no apêndice 3.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Distância de similaridade de Dice

Distância de similaridade de Jaccard

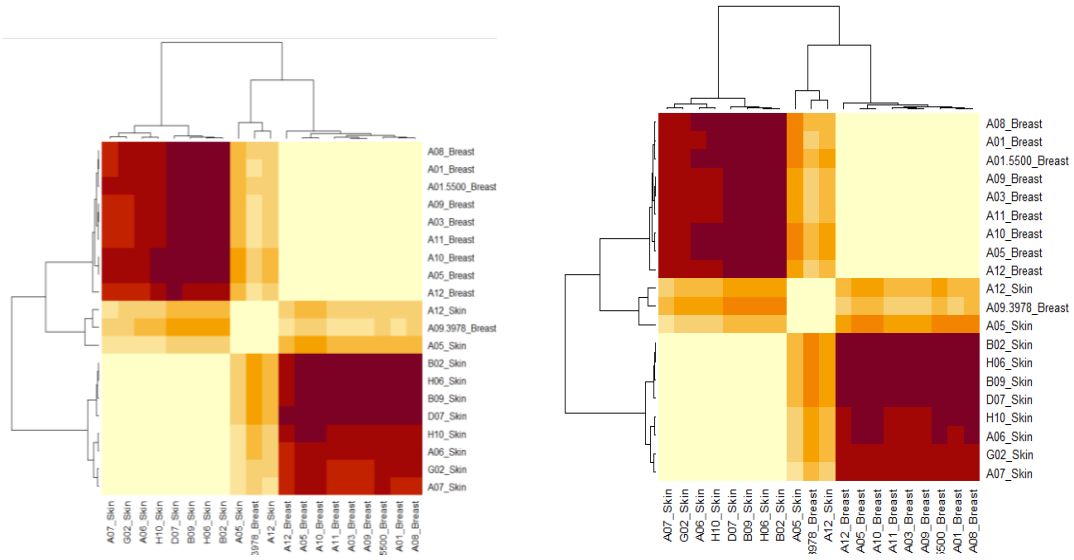


Figura 33 - a) Distância similaridade Dice_ID:11715918_s_at, b) Distância similaridade Jaccard_ID:11715918_s_at

Fonte: Bento et al. (2023)

Distância de similaridade de Sorensen

Distância de similaridade de Czekanowski

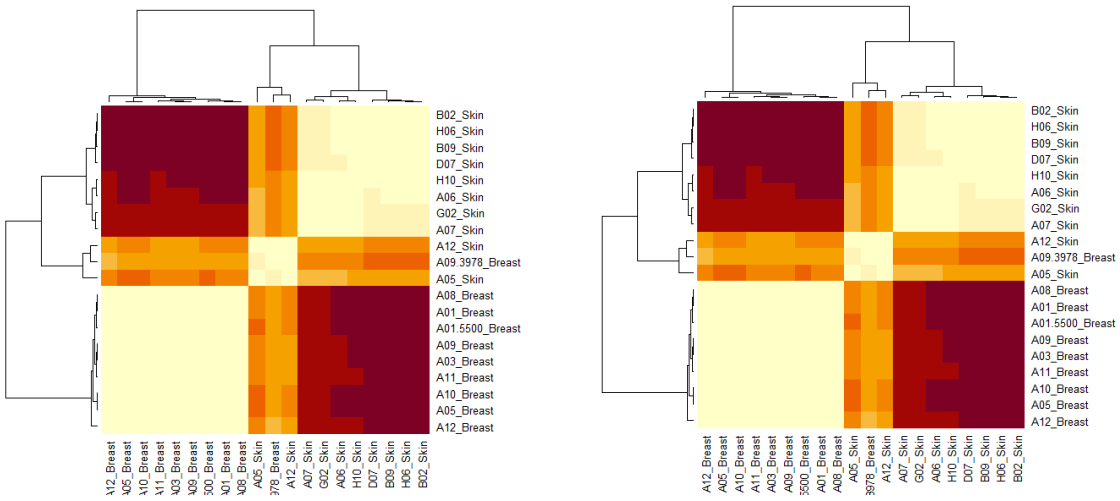


Figura 34 - a) Distância similaridade Sorensen_ID:11715918_s_at, b) Distância similaridade Czekanowski_ID:11715918_s_at

Fonte: Bento et al. (2023)

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Distância de similaridade de Minkowski **Distância de similaridade de Pearson**

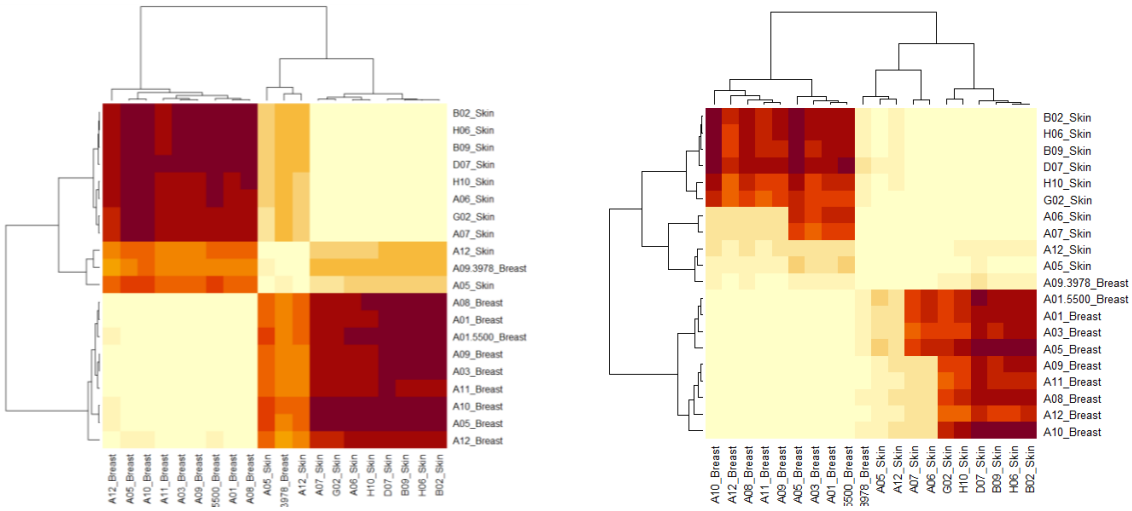


Figura 35 - a) Distância similaridade Minkowski_ID:11715918_s_at, b) Distância similaridade Pearson_ID:11715918_s_at

Fonte: Bento et al. (2023)

Distância de similaridade de Intersection **Distância de similaridade de Manhattan**

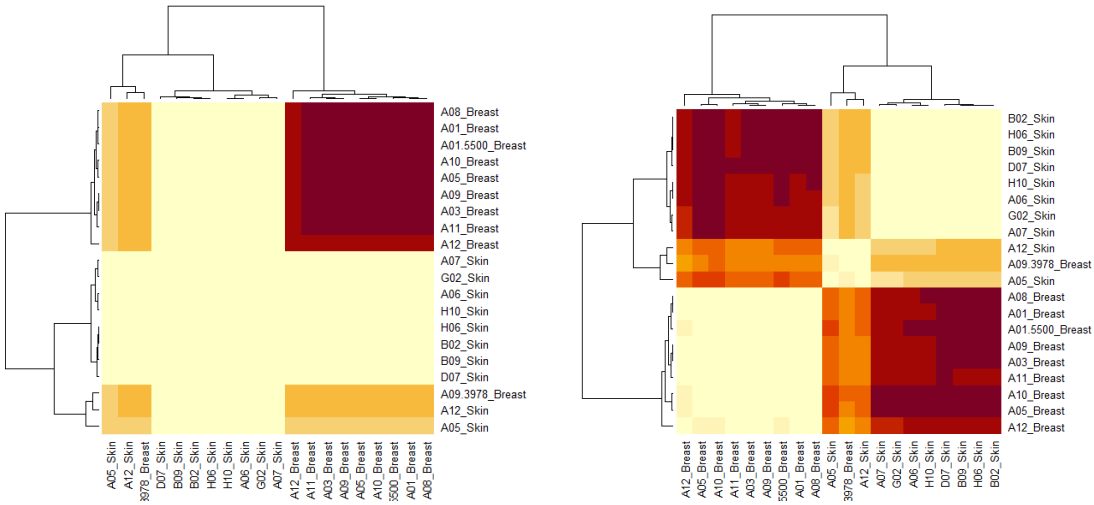


Figura 36 - a) Distância similaridade Intersection_ID:11715918_s_at, b) Distância similaridade Manhattan_ID:11715918_s_at

Fonte: Bento et al. (2023)

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Distância de similaridade de Tanimoto **Distância de similaridade de Euclideana**

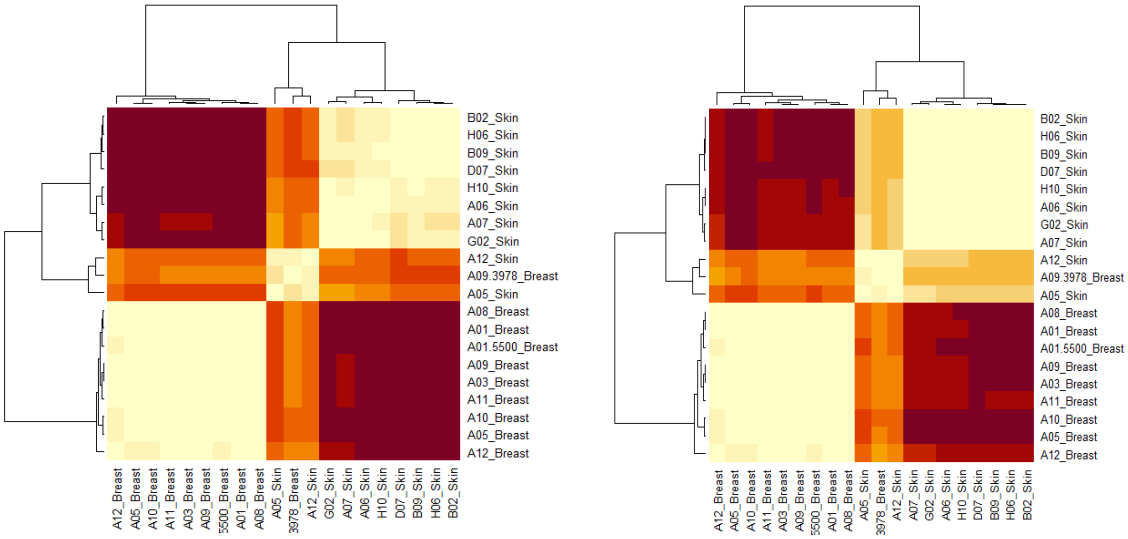


Figura 37 - a) Distância similaridade Tanimoto_ID:11715918_s_at, b) Distância similaridade Euclideana_ID:11715918_s_at

Fonte: Bento et al. (2023)

6.2.2 Gene Sonda ID - 11716887_a_at

Seguem-se os *Heatmaps* das Figuras 38 até à 42, com as distâncias de similaridade Dice, Jaccard, Sorensen, Czekanowski, Minkowski, Pearson, Intersection, Manhattan, Tanimoto e Euclideana, para a análise do gene identificado com a Sonda ID "11716887_a_at" nas 20 amostras (10 patologia da mama e 10 patologia da pele), estando os cálculos nas matrizes de similaridade no apêndice 3.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Distância de similaridade de Dice **Distância de similaridade de Jaccard**

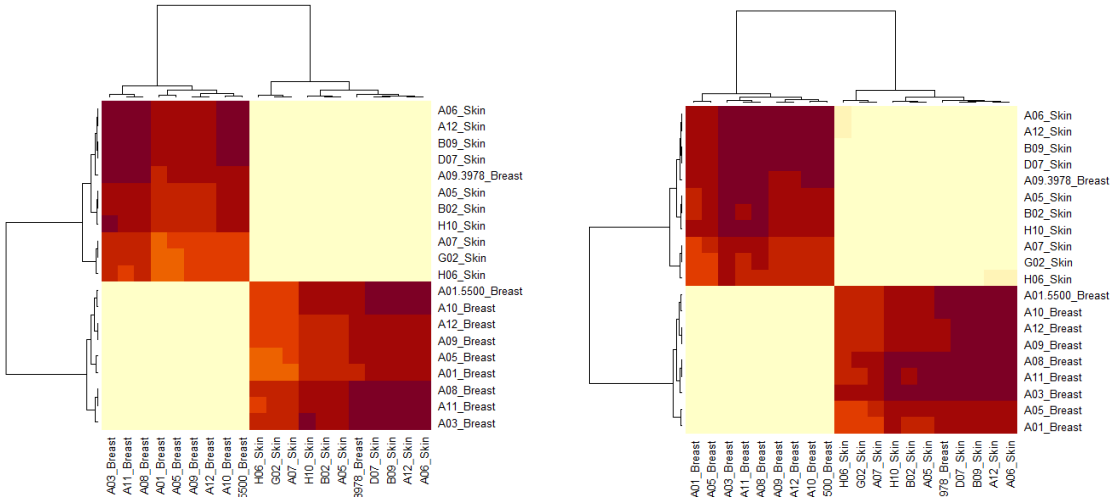


Figura 38 - Distância similaridade Dice_ID: 11716887_a_at, b) Distância similaridade Jaccard_ID: 11716887_a_at

Fonte: Bento et al. (2023)

Distância de similaridade de Sorensen **Distância de similaridade de Czekanowski**

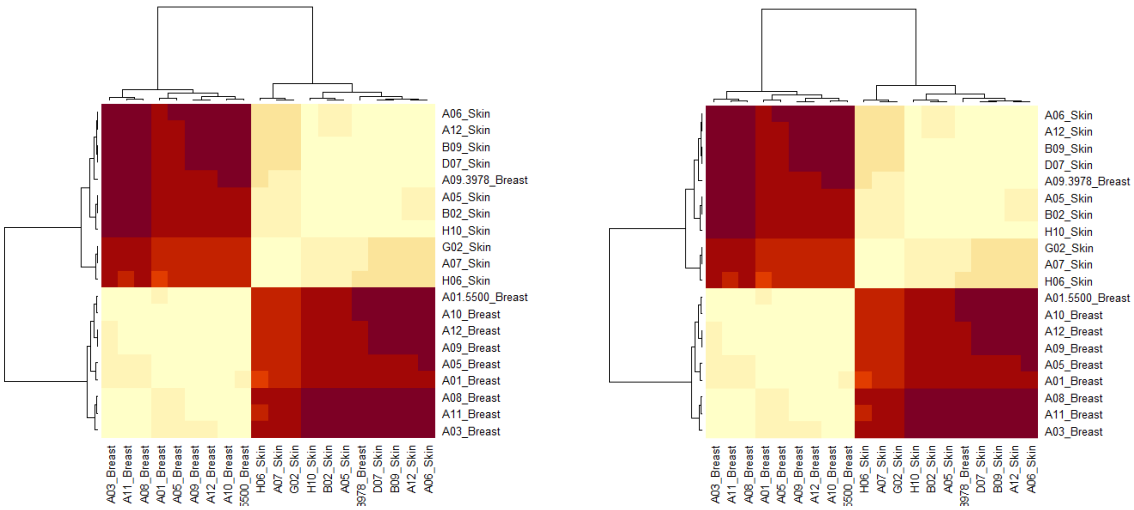


Figura 39 - a) Distância similaridade Sorensen_ID: 11716887_a_at, b) Distância similaridade Czekanowski_ID: 11716887_a_at

Fonte: Bento et al. (2023)

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Distância de similaridade de Minkowski **Distância de similaridade de Pearson**

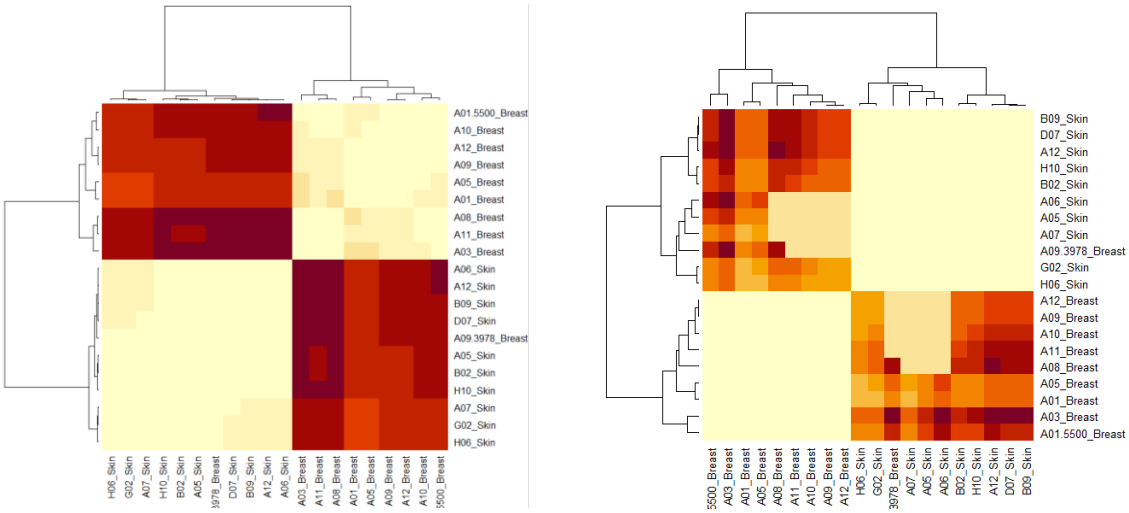


Figura 40 - a) Distância similaridade Minkowski_ID: 11716887_a_at, b) Distância similaridade Pearson_ID: 11716887_a_at

Fonte: Bento et al. (2023)

Distância de similaridade de Intersection **Distância de similaridade de Manhattan**

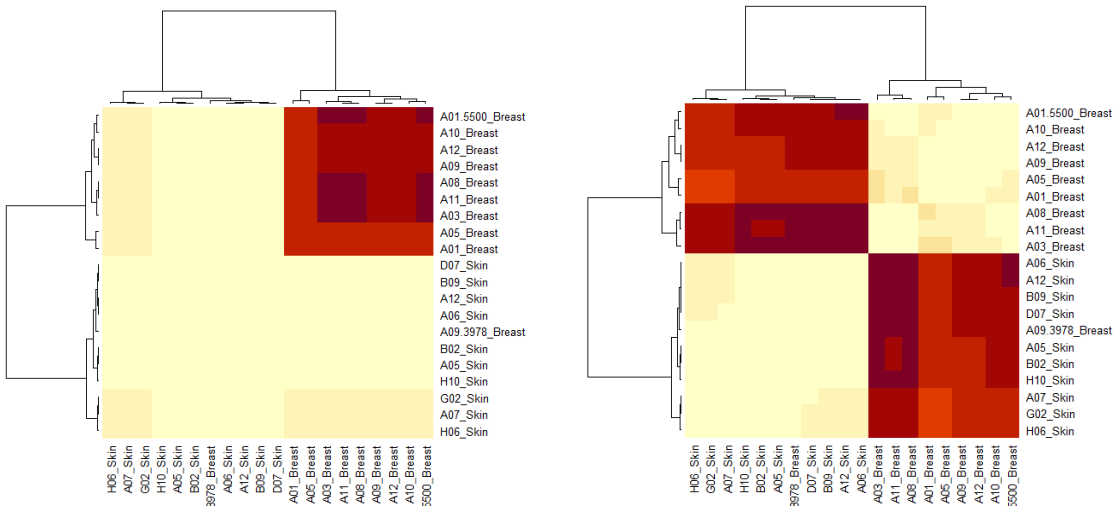


Figura 41 - a) Distância similaridade Intersection_ID: 11716887_a_at, b) Distância similaridade Manhattan_ID: 11716887_a_at

Fonte: Bento et al. (2023)

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Distância de similaridade de Tanimoto **Distância de similaridade de Euclideana**

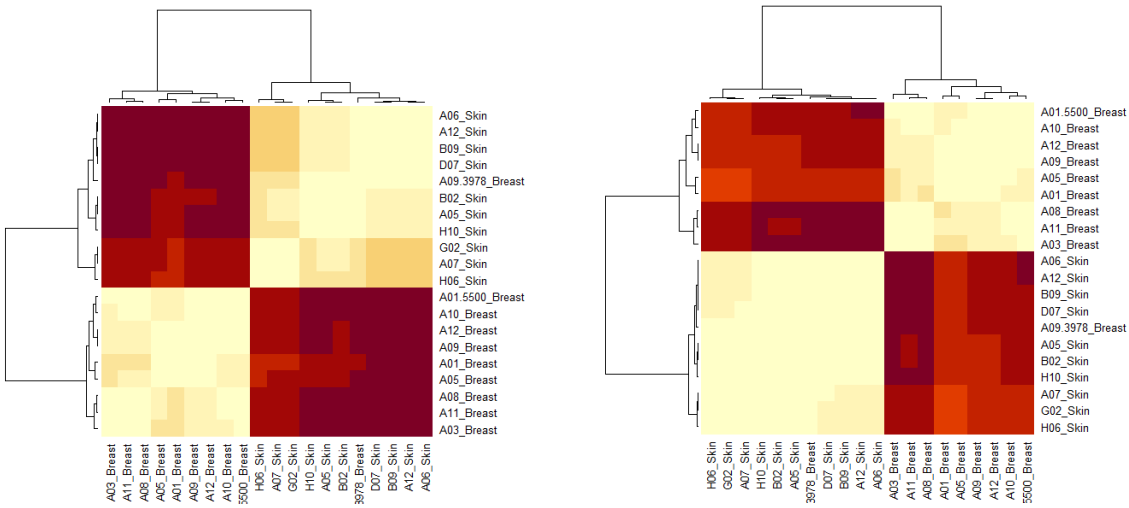


Figura 42 - a) Distância similaridade Tanimoto_ID: 11716887_a_at, b) Distância similaridade Euclideana_ID: 11716887_a_at

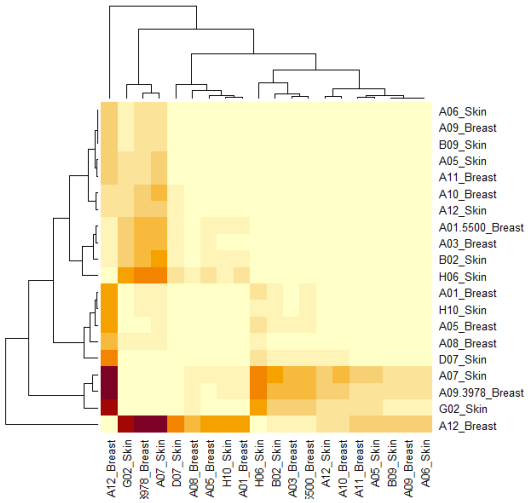
Fonte: Bento et al. (2023)

6.2.3 Gene Sonda com ID - 11724186_a_at

Seguem-se os Heatmaps das das Figuras 43 até à 47, com as distâncias de similaridade Dice, Jaccard, Sorensen, Czekanowski, Minkowski, Pearson, Intersection, Manhattan, Tanimoto e Euclideana, para a análise do gene identificado com a Sonda ID "11724186_a_at" nas 20 amostras (10 patologia da mama e 10 patologia da pele), estando os cálculos nas matrizes de similaridade no apêndice 3.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Distância de similaridade de Dice



Distância de similaridade de Jaccard

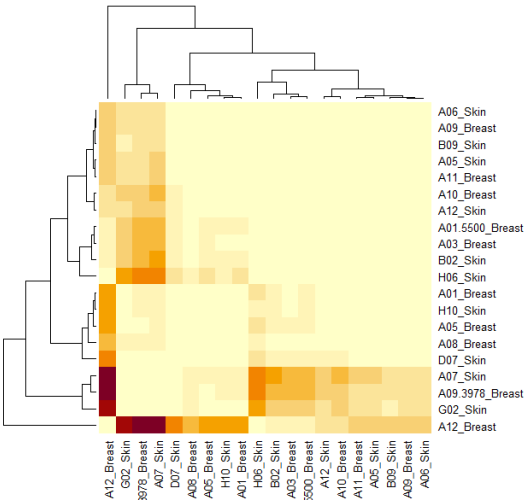
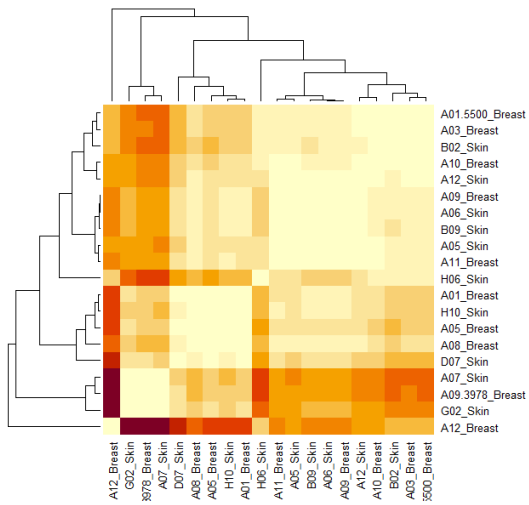


Figura 43 - a) Distância similaridade Dice_ID: 11724186_a_at, b) Distância similaridade Jaccard_ID: 11724186_a_at

Fonte: Bento et al. (2023)

Distância de similaridade de Sorensen



Distância de similaridade de Czekanowski

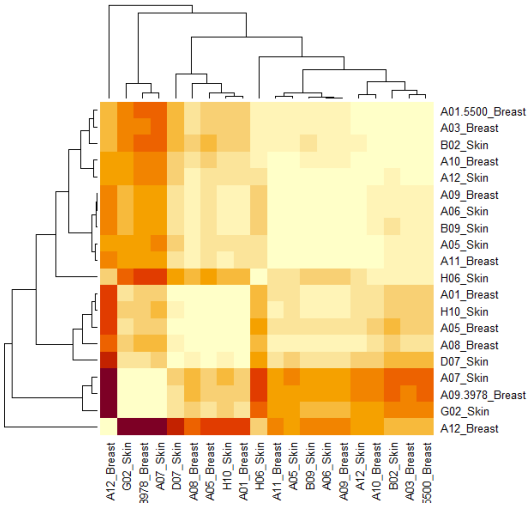


Figura 44 - a) Distância similaridade Sorensen_ID: 11724186_a_at, b) Distância similaridade Czekanowski_ID: 11724186_a_at

Fonte: Bento et al. (2023)

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Distância de similaridade de Minkowski **Distância de similaridade de Pearson**

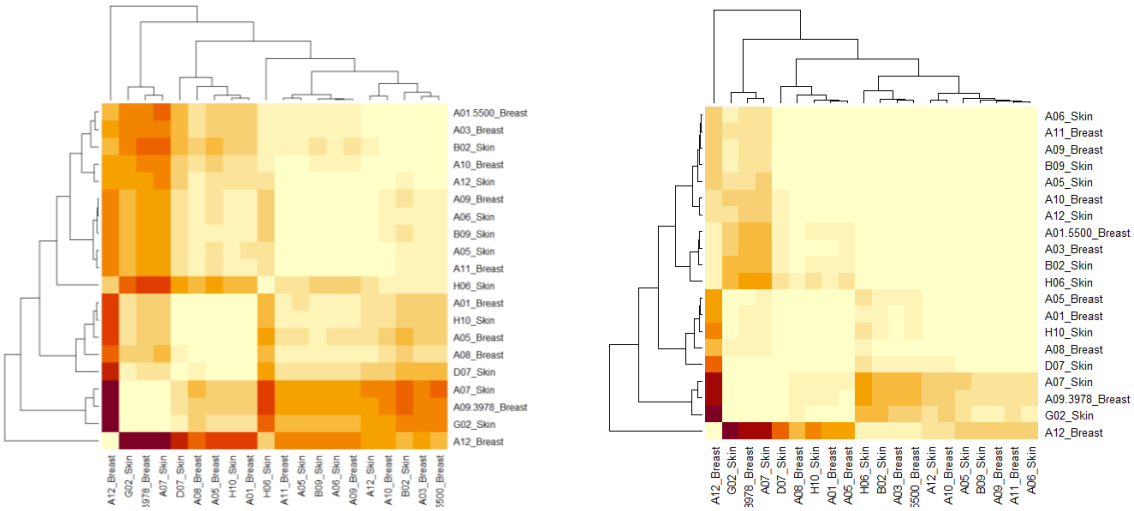


Figura 45 - a) Distância similaridade Minkowski_ID: 11724186_a_at, b) Distância similaridade Pearson_ID: 11724186_a_at

Fonte: Bento et al. (2023)

Distância de similaridade de Intersection **Distância de similaridade de Manhattan**

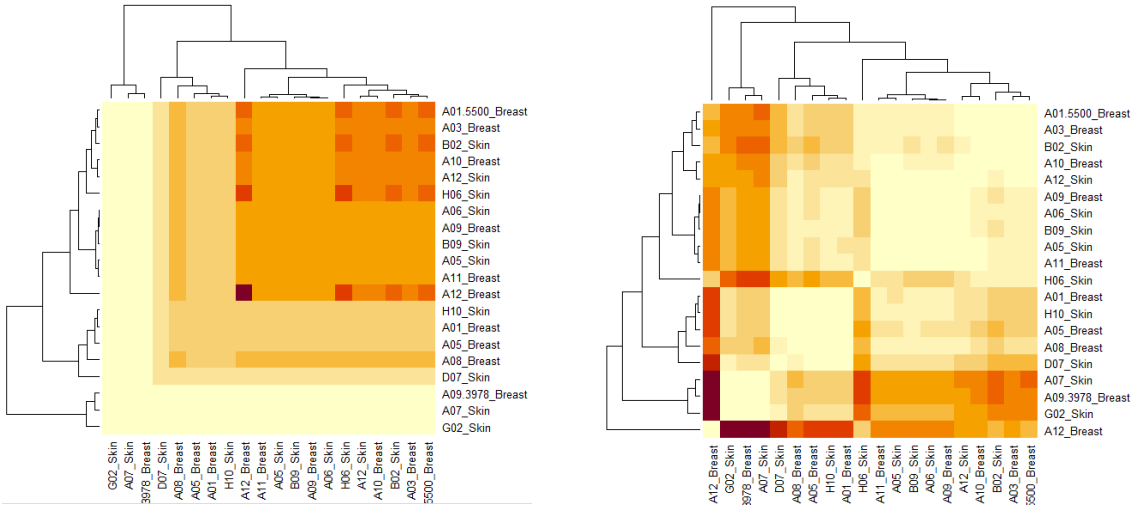


Figura 46 - a) Distância similaridade Intersection_ID: 11724186_a_at, b) Distância similaridade Manhattan_ID: 11724186_a_at

Fonte: Bento et al. (2023)

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Distância de similaridade de Tanimoto **Distância de similaridade de Euclideana**

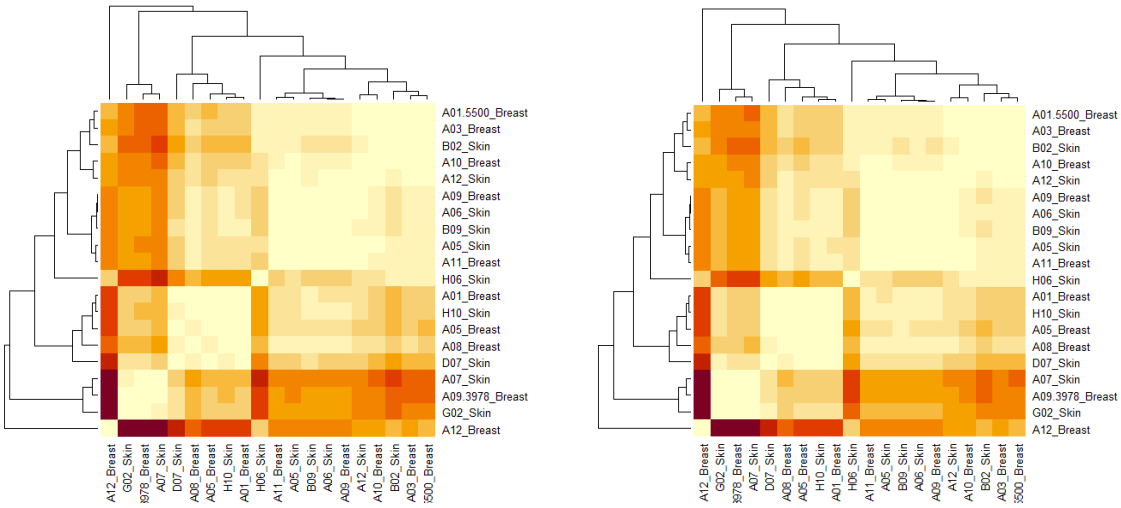


Figura 47 - a) Distância similaridade Tanimoto_ID: 11724186_a_at, b) Distância similaridade Euclideana_ID: 11724186_a_at

Fonte: Bento et al. (2023)

Em suma, um valor 0 numa distância de similaridade génica, pode indicar que duas amostras ou grupos de amostras são idênticos, ou seja, não existe grandes diferenças genéticas entre elas. Por outro lado, o valor 1 numa distância de similaridade génica pode indicar que duas amostras ou grupos de amostras são diferentes, ou seja, existe diferenças genéticas significativas entre as mesmas.

6.3 Algoritmos de machine learning

Nesta secção ir-se-ão aplicar alguns algoritmos de ML para diferenciar os dois grupos de patologia da mama e da pele. Ou seja, o objetivo destes algoritmos será, quando for apresentada uma nova amostra, o algoritmo identificar a que grupo é que esta nova amostra pertencerá.

6.3.1 Rede Neural Artificial

Na contrução de Rede Neural Artificial, foram efetuados vários teste para criar a que melhor se adequa aos dados. Assim foi criada uma rede neural com três camadas e seis neurónios, na primeira e segunda camada e um neuronio na última camada, ou seja na sua

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

saída. Por outro lado, a escolha da função de ativação recaiu sobre a sigmoid, sendo a mais adequada para saídas binárias, dado que se pretende classificar cada amostra de linhas celulares nos dois grupos (patologia da mama ou pele). Optou-se ainda pelo otimizador 'adam', que se ajusta automaticamente à taxa de aprendizagem durante o treino da rede neural. Além disso e dado que as classes do dataset estudado estarem bem balanceadas, a métrica de precisão escohida foi a accuracy, conforme se pode observar no treino da rede neural, na Figura 48. Salienta-se ainda que todo este código se encontra no apêndice 1 deste estudo.

```
model = Sequential()
model.add(Dense(units=6, kernel_initializer='uniform', activation='relu', input_dim=X_train.shape[1]))
model.add(Dense(units=6, kernel_initializer='uniform', activation='relu'))
model.add(Dense(units=1, kernel_initializer='uniform', activation='sigmoid'))

# compila a rede neural
model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])

# treina a rede neural
history=model.fit(X_train, y_train, batch_size=64, epochs=70, verbose=1)

# avalia a precisão da rede neural no conjunto de teste
loss, accuracy = model.evaluate(X_test, y_test, verbose=0)
print('Test accuracy:', accuracy)
```

Figura 48 - Rede Neural Artificial (algoritmo)

Neste ponto conclui-se que este algoritmo apresenta um valor baixo para a *accuracy* de 0.75.

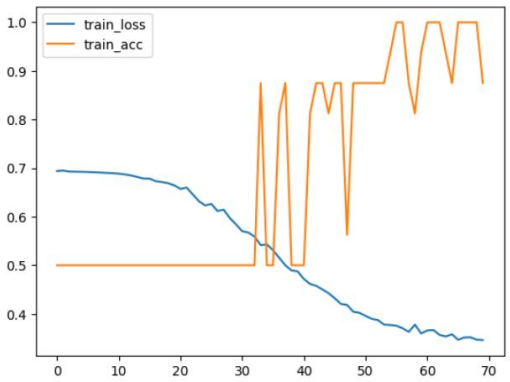


Figura 49 - Evolução da perda precisão e da accuracy

O gráfico da Figura 49 que mostra a evolução da perda (train_loss) e da precisão *accuracy* (train_acc) durante o treino da rede neural. Assim tem-se:

- O Eixo X representa o número de épocas. Uma época é uma passagem completa pelos dados de treino durante o processo de treino da rede neural.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

- No Eixo Y esta linha que representa 'train_loss' mostra a perda de treino (loss) ao longo das épocas. A perda é uma medida da diferença entre as saídas reais e as saídas previstas pela rede neural. Idealmente, essa linha deve diminuir à medida que as épocas passam, indicando que o algoritmo está se ajustando melhor aos dados de treino.

Por outro lado, a linha que representa "train_acc" mostra a precisão de treino "accuracy" ao longo das épocas. A precisão é a proporção de predições corretas em relação ao total de predições feitas pela rede neural. Idealmente, essa linha deve aumentar à medida que as épocas passam, indicando que o algoritmo está melhorando na classificação correta dos dados de treino. Em suma, o gráfico da Figura 69 permite observar como a perda e a precisão da rede neural evoluíram durante o seu treino. Dado que a perda diminuiu e a precisão aumentou ao longo das épocas, isto indica que o treino progrediu de uma maneira positiva.

6.3.2 Algoritmos Logistic regression, Linear discriminant analysis K-Nearest Neighbors, DecisionTreeClassifier, Gaussian NB e Support vector machine

Nesta secção serão testados os algoritmos *Logistic regression, LDA, KNN, DecisionTreeClassifier, Gaussian NB e SVM*, para encontrar o que melhor responderá ao problema apresentado neste estudo.

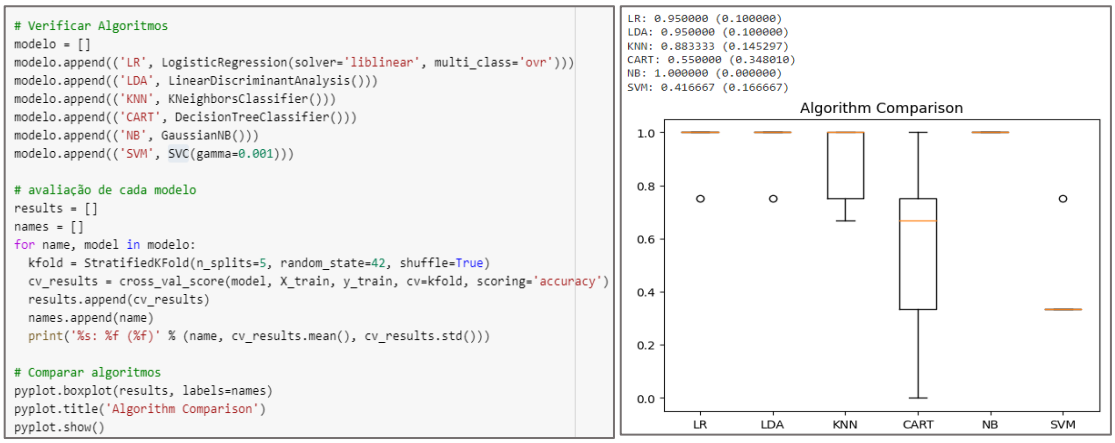
A escolha da métrica de avaliação é crucial para avaliar algoritmos de ML, dependendo do tipo de problema a estudar. Seguem-se algumas notas sobre estas métricas:

- *Accuracy* - mede a proporção de predições corretas em relação ao total;
- *Precision* - avalia a precisão das predições positivas, sendo a proporção de verdadeiros positivos em relação à soma de verdadeiros positivos e falsos positivos;
- *Recall* - mede a capacidade do modelo em capturar todos os casos positivos, sendo a proporção de verdadeiros positivos em relação à soma de verdadeiros positivos e falsos negativos;

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

- F1-Score - é a média harmônica entre *precision* e *recall*, proporcionando um equilíbrio entre estas duas métricas.

Assim, dado o *dataset* estudado e o problema em estudo, optou-se a pela *accuracy* como a métrica de avaliação para estes seis algoritmos de ML. A *accuracy* é uma das métricas mais utilizadas na avaliação de algoritmos de ML, no entanto tem algumas limitações, como por exemplo: não lidar bem com conjuntos de dados não balanceados, não diferenciar tipos de erros em classes distintas, perda de detalhes sobre diferentes tipos de erros, não considerar as probabilidades associadas às previsões, ou poder não indicar o desempenho absoluto em diferentes contextos.



```
# Verificar Algoritmos
modelo = []
modelo.append(('LR', LogisticRegression(solver='liblinear', multi_class='ovr')))
modelo.append(('LDA', LinearDiscriminantAnalysis()))
modelo.append(('KNN', KNeighborsClassifier()))
modelo.append(('CART', DecisionTreeClassifier()))
modelo.append(('NB', GaussianNB()))
modelo.append(('SVM', SVC(gamma=0.001)))

# avaliação de cada modelo
results = []
names = []
for name, model in modelo:
  kfold = StratifiedKFold(n_splits=5, random_state=42, shuffle=True)
  cv_results = cross_val_score(model, X_train, y_train, cv=kfold, scoring='accuracy')
  results.append(cv_results)
  names.append(name)
  print('%s: %f (%f)' % (name, cv_results.mean(), cv_results.std()))

# Comparar algoritmos
pyplot.boxplot(results, labels=names)
pyplot.title('Algorithm Comparison')
pyplot.show()
```

```
LR: 0.950000 (0.100000)
LDA: 0.950000 (0.100000)
KNN: 0.883333 (0.145297)
CART: 0.550000 (0.348810)
NB: 1.000000 (0.000000)
SVM: 0.416667 (0.166667)
```


Figura 50 - Avaliação da accuracy nos algoritmos de ML

Todo este código encontra-se no apêndice 1 deste estudo. O Boxplot da Figura 50 representa os valores de *accuracy* para cada um dos algoritmos estudados, sendo que alguns apresentam bons resultados para esta métrica.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

7 AVALIAÇÃO

A avaliação servirá para concluir quais serão os índices de similaridade mais capazes de responder ao nosso caso, ou seja, a comparação de um gene nos dois grupos das vinte amostras, assim como quais os algoritmos de ML que apresentaram melhor performance.

Relativamente às distâncias de similaridade estudadas, e de acordo com a literatura Bass et al. (2013) e Torrente (2021), as que melhor responderam foram: Jaccard e o Dice, uma vez que apresentaram os resultados mais consistentes 2 genes (identificados pelos seus genes com as sondas ID - 11715918_s_at e 11716887_a_at). Tendo-se efetivamente constatado que este gene com a sonda ID - 11716887_a_at identifica melhor estas amostras de Mama e Pele, embora exista uma amostra de Mama (A09 3978_Breast) no meio do grupo das de Pele (Bento et al., 2023). De acordo com Bento et al. (2023) a explicação para este acontecimento em termos de expressão do gene em questão (sonda ID - 11716887_a_at) poderá estar ligada a algumas razões plausíveis, como por exemplo:

- O gene em questão poderá estar em processos biológicos comuns, nestas duas patologias, ou seja, embora as patologias sejam diferentes, pode haver mecanismos moleculares sobrepostos que resultam em padrões de expressão semelhantes para esse gene específico.
- Se houver uma sobreposição significativa de fatores como por exemplo: fatores não relacionados à patologia, como idade, sexo, características genéticas ou ambientais, entre as amostras, isso pode levar a uma maior similaridade na expressão desse gene, independentemente da patologia em questão.
- Uma outra justificação, poderá ser a de que os doentes metastizados poderão eventualmente ter o mesmo gene em duas patologias distintas em consequência dessa mesma metastização.
- Um outro fator poderá estar relacionado com a expressão génica ser altamente variável, mesmo em indivíduos saudáveis.
- Por fim, esta ocorrência poderá não ter nenhuma razão genética e estar apenas relacionado com um ruído nos dados de expressão génica. Se o ruído for

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

particularmente pronunciado para este gene em específico, poderá haver uma sobreposição aparente entre as amostras nos dois grupos.

Em qualquer dos casos, será importante realizar uma análise cuidadosa e considerar outras informações clínicas, para obter uma compreensão abrangente e contextualizada da expressão génica, por forma a obter a sua relação com as patologias em estudo.

Por outro lado, os resultados obtidos para o gene da sonda ID: 11724186_a_at foram os esperados, dado se tratar de um gene com um valor p $\approx$ 0.7209 em que indica uma evidência estatística suficiente para aceitar a hipótese nula, de que não há diferença real entre os dois grupos, ou seja, não se conseguindo distinguir os dois grupos, Mama e Pele, na análise deste gene nas 20 linhas celulares estudadas.

Relativamente aos algoritmos de ML que avaliaram toda a expressão génica, para as vinte amostras, os que apresentaram melhores resultados para a acurácia foram: *Logistic Regression, LDA* e *Gaussian NB*. No entanto estes algoritmos necessitavam ainda de testar outras abordagens, como por exemplo abordar e testar outras métricas de avaliação e avaliar se os mesmo não estarão em *overfitting*. Podendo-se para o efeito ajustar a complexidade destes modelos, aplicar técnicas de regularização, aumentar a quantidade de dados de treino ou até mesmo explorar diferentes arquiteturas dos modelos.

Ao longo deste estudo, houve um esforço em explicar quais serão as abordagens possíveis, em estudos que utilizem dados da expressão génica, testando a presença de um gene nos dois grupos de amostras de DNA. Os indicadores mencionados anteriormente poderão fornecer uma base sólida em estudos futuros de similaridade entre linhas celulares. No entanto, é fundamental reconhecer que antes de serem adotados ou considerados como critérios confiáveis em estudos genéticos, estes indicadores necessitam ainda de ser testados e estudados. A validação rigorosa dos indicadores de similaridade entre patologias é crucial para garantir a precisão e a robustez das conclusões obtidas a partir de estudos genéticos. Isso envolve a realização de análises em diferentes conjuntos de dados, a comparação dos resultados com outras abordagens de avaliação de similaridade e a investigação minuciosa dos seus potenciais suas limitações.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

8 CONCLUSÃO

O presente trabalho teve como propósito demonstrar a importância dos RecSys no tratamento do doente oncológico, baseado na similaridade entre linhas celulares. Como contributo científico foi elaborado o artigo "*Comparative Analysis of Cell Line Similarity Algorithms in Oncology Treatment*", que será apresentado na conferencia da *HCist - International Conference on Health and Social Care Information Systems and Technologies*, em novembro de 2023. Este artigo tem por base o trabalho desenvolvido ao longo deste projeto, visando demonstrar a importância da similaridade entre linhas celulares, nos sistemas de recomendação, para o tratamento do doente oncológico.

Ao longo deste projeto foram encontradas várias dificuldades, desde logo na compreensão de alguns conceitos desta área até à compreensão dos dados contidos nos ficheiros CEL. Por outro lado, adotar a melhor abordagem para avaliar a similaridade entre linhas celulares. Quais os índices a usar e qual o software que melhor responderia ao problema em estudo. De acordo com a literatura consultada, muitos autores estudaram linhas celulares através da avaliação da sua expressão génica. Assim, empregou-se esta abordagem, usando a técnica do RNA-Seq, para interpretar corretamente esses dados e tirar as ilações sobre a regulação genética. Os genes estatisticamente relevantes foram encontrados através de um teste-t, já que o pacote DESeq2 não apresentou consistentemente essa informação. Apresentando-se várias abordagens para avaliar a similaridade entre as 20 amostras de linhas celulares, patologia Mama e patologia Pele selecionadas aleatoriamente, com o intuito de selecionar as melhores abordagens que permitissem distinguir as amostras pertencentes a estes dois grupos. Por um lado, as distâncias de similaridade de Jaccard e Dice apresentaram os melhores resultados, uma vez que apresentaram os resultados mais consistentes para os dois genes selecionados, sendo que num dos genes os resultados ainda foram mais consistentes. Por outro lado, os algoritmos de ML *Logistic Regression, LDA* e *Gaussian NB* apresentaram os melhores valores de acurácia, tendo sido avaliada toda a expressão génica para as 20 amostras.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Em abordagens futuras, poder-se-á ter uma abordagem mais exaustiva, tratando por exemplo todo o *dataset*, através de técnicas de *Big Data*, com a finalidade de alargar o número de grupos e avaliar o comportamento dos resultados obtidos neste trabalho. Uma destas técnicas de *Big Data* que poderá ser adotada será o processamento distribuído por várias máquinas, tendo em atenção que os dados de cada patologia deverão ficar todos na mesma máquina para facilitar a avaliação final.

Tendo-se apresentado a metodologia que poderá ser usada pela comunidade médica, ou em estudos similares, assim como as técnicas possíveis para abordar esta problemática. No entanto, para a validação de um método suficientemente robusto poder-se-á ter a necessidade de envolver a colaboração entre pesquisadores de diferentes disciplinas, como genética, biologia molecular, estatística e medicina clínica. Garantindo assim que todas as nuances e complexidades destas patologias, e dos dados genéticos, seriam devidamente consideradas.

Contudo a grande mais-valia para toda esta indústria, é ter-se apresentado a possibilidade de encontrar os genes mais relevantes através de um teste-t, aplicando duas abordagens distintas com índices de similaridade e algoritmos de ML. Podendo uma ser usada como validação ou complemento da outra, para criar um algoritmo com toda esta informação. Por outro lado, embora estas métricas ofereçam um ponto de partida promissor, será essencial submetê-las a testes rigorosos e avaliações detalhadas para confirmar a sua utilidade e confiabilidade em estudos genéticos futuros. Assim como estudar outras abordagens, como por exemplo os genes estatisticamente mais relevantes, a extração da sua expressão génica através de outros métodos e métricas que esta indústria detenha como referência. E ainda, fazer um estudo mais aprofundado sobre outras métricas de avaliação dos algoritmos de ML e avaliar a existência de *overfitting* nos mesmos. Somente por meio desse processo de validação minucioso, poder-se-á obter insights sólidos, robustos e confiáveis sobre a similaridade entre diferentes patologias e sua base genética subjacente.

Em trabalhos futuros outras questões poderão ainda ser levantadas, como por exemplo qual o grau de confiança nestes índices de similaridade por parte da classe médica? E qual

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

o papel destes algoritmos de ML. Poderão ajudar a tratar de uma forma mais rápida, eficaz e eficiente esta montanha de dados? Estas abordagens serão uma mais-valia eficaz para auxiliar a classe médica a tomar decisões nas terapêuticas e tratamentos a eleger para os seus doentes? E por outro lado, qual a opinião do doente sobre estas temáticas? No entanto, é possível discutir outras questões, como, por exemplo, se os médicos estarão dispostos a compartilhar suas práticas clínicas com máquinas. Poderá ainda haver preocupação de que as máquinas possam substituir os seres humanos, como já ocorreu em outros setores de atividade. Essas e outras questões poderão ser exploradas e abordadas em trabalhos futuros.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

REFERÊNCIAS

Affymetrix. (2009). *Affymetrix Developer Network*. https://www.affymetrix.com/support/developer/powertools/changelog/gcos-agcc/cel.html#V3

Afonso, A., & Nunes, C. (2019). *Versão revista e aumentada PROBABILIDADES E ESTATÍSTICA Aplicações e Soluções em SPSS*.

Aggarwal, C. C. (2016). *Recommender Systems*. https://doi.org/10.1007/978-3-319-29659-3

Albert Einstein, H. I. (2021). *Você sabe o que é sequenciamento genético? Aprenda, agora! - Vida Saudável o blog do Einstein*. https://vidasaudavel.einstein.br/sequenciamento-genetico/

Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster Analysis. Sage University Paper Series On Quantitative Applications in the Social Sciences 07-044*. Sage Publications.

Ali, L., Zhu, C., Zhang, Z., & Liu, Y. (2019). Automated Detection of Parkinson's Disease Based on Multiple Types of Sustained Phonations Using Linear Discriminant Analysis and Genetically Optimized Neural Network. *IEEE Journal of Translational Engineering in Health and Medicine*, *7*. https://doi.org/10.1109/JTEHM.2019.2940900

Ariel, S., & Moraes, J. (2016). *Juruá Editora - Acesso às Informações Genéticas do Trabalhador - Discriminação Genética e o Livre Consentimento Esclarecido* (pp. 37–42). https://www.jurua.com.br/shop_item.asp?id=24863

Azambuja, R. X. de, Morais, A. J., & Filipe, V. (2021). Teoria e Prática em Sistemas de Recomendação. *Revista de Ciências Da Computação*. https://grouplens.org.

Bass, J. I. F., Diallo, A., Nelson, J., Soto, J. M., Myers, C. L., & Walhout, A. J. M. (2013). Using networks to measure similarity between genes: Association index selection.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

In *Nature Methods* (Vol. 10, Issue 12, pp. 1169–1176). https://doi.org/10.1038/nmeth.2728

Bento, B., Paulo Belfo, F., & Trigo, A. (2023, November 8). Comparative analysis of cell line similarity algorithms in oncology treatment. *HCist 2023*.

Bigus, J. P. (1996). *Data Mining With Neural Networks - Solving Business Problems - From Application Development To Decision support-McGraw-Hill*. https://pt.scribd.com/document/468727267/Joseph-P-Bigus-Data-mining-with-neural-networks-solving-business-problems-from-application-development-to-decision-support-McGraw-Hill-1996-pdf#

Biotechnology Information, N. C. (2011). *Gene Expression Omnibus*. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL13667

Born, S., & Oliveira, D. E. (2001). *MANIPULAÇÃO GENÉTICA E DIGNIDADE HUMANA: DA BIOÉTICA AO DIREITO* (p. 72).

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C. P., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., … Vingron, M. (2001). *Minimum information about a microarray experiment (MIAME)-toward standards for microarray data*. http://www.ncbi.nlm.nih.gov/Taxon-

C, V., & D, M. (2021). Computer Aided Detection of Nodule from Computed Tomography Images of Lung. *International Research Journal on Advanced Science Hub*, *3*(Special Issue ICARD 3S), 96–100. https://doi.org/10.47392/irjash.2021.073

Cancer Genome Project, Wellcome Sanger Institute (UK), Center for Molecular Therapeutics, M. G., & Hospital Cancer Center (USA). (2022). *Genomics of Drug Sensitivity in Cancer*. https://www.cancerrxgene.org/

Cassali, G. D., Bertagnolli, A. C., Silva, A. E., & Ferreira, E. (2007). *Microarrays em Câncer de Mama*. www.sboc.org.br/

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Cecchin, E., & Stocco, G. (2020). Pharmacogenomics and Personalized Medicine. *Genes*, *11*(6), 1–5. https://doi.org/10.3390/GENES11060679

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., & Shearer, C. (2000). *CRISP-DM 1.0 Step-by-step data mining guide*. DaimlerChrysler.

Chaudhuri, J. (2005). Genes arrayed out for you: the amazing world of microarrays. *Medical Science Monitor : International Medical Journal of Experimental and Clinical Research*.

Cokelaer, T., Chen, E., Iorio, F., Menden, M. P., Lightfoot, H., Saez-Rodriguez, J., & Garnett, M. J. (2018). GDSCTools for mining pharmacogenomic interactions in cancer. *Bioinformatics*, *34*(7), 1226–1228. https://doi.org/10.1093/bioinformatics/btx744

Colombo, J., & Rahal, P. (2009). *A Tecnologia de Microarray no Estudo do câncer de Cabeça e Pescoço*. *1*, 64–72. http://www.ufrgs.br/seerbio/ojshttp://www.ufrgs.br/seerbio/ojs/index.php/rbb/article/view/1268

Dalma-Weiszhausz, D. D., Warrington, J., Tanimoto, E. Y., & Miyada, C. G. (2006). The affymetrix GeneChip platform: an overview. *Methods in Enzymology*, *410*, 3–28. https://doi.org/10.1016/S0076-6879(06)10001-4

Deza, M. M., & Deza, E. (2009). *Encyclopedia of Distances Third Edition*.

Diniz, J. O. B. (2015). *Análise Temporal de Lesão em Mamografias Digitalizadas Usando Índices de Similaridade*.

Diniz, P. H. B., Diniz, J. O. B., Silva, A. C., Paiva, A. C., & Gattas, M. (2016). *ANÁLISE TEMPORAL DE LESÕES EM MAMOGRAFIAS UTILIZANDO ÍNDICES DE SIMILARIDADE*. www.jhi-sbis.saude.ws/ojs-jhi/index.php/jhi-sbis

do Pulmão, F. P. (2020). *ONDR 2020*.

EMBL-EBI. (2023). *Experimental Factor Ontology*. https://www.ebi.ac.uk/efo/

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

European Bioinformatics Institute. (2023a). *BioStudies – one package for all the data supporting a study*. https://www.ebi.ac.uk/biostudies/files/E-MTAB-3610/E-MTAB-3610.idf.txt

European Bioinformatics Institute. (2023b). *BioStudies – one package for all the data supporting a study*. https://ftp.ebi.ac.uk/biostudies/nfs/E-MTAB-/610/E-MTAB-3610/Files/E-MTAB-3610.sdrf.txt

European Bioinformatics Institute. (2023c). *BioStudies – one package for all the data supporting a study*. https://ftp.ebi.ac.uk/biostudies/nfs/A-GEOD-/667/A-GEOD-13667/Files/A-GEOD-13667.adf.txt

Fabiano, W. (2017). *O Algoritmo Mestre Como a busca pelo algoritmo de machine learning*.

FGED Society - History. (2023). https://www.fged.org/history#h.a887fllea086

Fogle, T. (1990). *Are Genes Units of Inheritance?* (pp. 349–371).

Fogle, T. (2010). The Dissolution of Protein Coding Genes in Molecular Biology. *The Concept of the Gene in Development and Evolution*, 3–25. https://doi.org/10.1017/CBO9780511527296.003

Freedman, D. H., & Freedman, D. H. (1996). *Los hacedores de cerebros : cómo los científicos están perfeccionando las computadoras, creando un rival del cerebro humano*. 272.

Freeman, J. A., & García-Bermejo, R. (1993). Redes neuronales : algoritmos, aplicaciones y técnicas de programación /. In *Redes neuronales : algoritmos, aplicaciones y técnicas de programación*. Addison-Wesley Iberoamericana,. http://fama.us.es/record

García, Á., Senis, Y., Tomlinson, M. G., & Watson, S. P. (2007). Platelet Genomics and Proteomics. *Platelets, Second Edition*, 99–116. https://doi.org/10.1016/B978-012369367-9/50767-9

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Gorakala, S. K., & Usuelli, M. (2015). *Building a Recommendation System with R - Google Livros*. https://books.google.pt/books?hl=pt-PT&lr=&id=V9VOCwAAQBAJ&oi=fnd&pg=PP1&dq=Building+a+recommendation+system+with+R&ots=xYJI1mYeEr&sig=qeXDl7-ZAG6aMXzTBFF5ahMDo5E&redir_esc=y#v=onepage&q=Building%20a%20recommendation%20system%20with%20R&f=false

Gourd, E. (2020). Lung cancer control in the UK hit badly by COVID-19 pandemic. *The Lancet. Oncology*, *21*(12), 1559. https://doi.org/10.1016/S1470-2045(20)30691-4

Guo, C., Xie, X., Li, J., Huang, L., Chen, S., Li, X., Yi, X., Wu, Q., Yang, G., Zhou, H., Liu, J. P., & Chen, X. (2019). Pharmacogenomics guidelines: Current status and future development. *Clinical and Experimental Pharmacology and Physiology*, *46*(8), 689–693. https://doi.org/10.1111/1440-1681.13097

Gysi, D. M., Voigt, A., Fragoso, T. de M., Almaas, E., & Nowick, K. (2018). wTO: An R package for computing weighted topological overlap and a consensus network with integrated visualization tool. *BMC Bioinformatics*, *19*(1). https://doi.org/10.1186/s12859-018-2351-7

Hambali, M. A., Oladele, T. O., & Adewole, K. S. (2020). Microarray cancer feature selection: Review, challenges and research directions. *International Journal of Cognitive Computing in Engineering*, *1*, 78–97. https://doi.org/10.1016/J.IJCCE.2020.11.001

Hyuna, S., Jacques, F., & siegel Rebecca L. (2021). *Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries*. https://doi.org/10.3322/caac.21660

Ihnaini, B., Khan, M. A., Abbas Khan, T., Abbas, S., Sh Daoud, M., Ahmad, M., & Adnan Khan, M. (2021). *A Smart Healthcare Recommendation System for Multidisciplinary Diabetes Patients with Data Fusion Based on Deep Ensemble Learning*. https://doi.org/10.1155/2021/4243700

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Irizarry, R. A. (2003). *Measures of Gene Expression for Affymetrix High Density Oligonucleotide Arrays*.

Jacob, B., L., K. H., Yongmin, K., Van, M. R. L., & C., H. S. (2000). *Handbook of Medical Imaging: Display and PACS - Google Livros*. https://books.google.pt/books?hl=pt-PT&lr=&id=YKVULpCZ_iEC&oi=fnd&pg=PR11&dq=Handbook+of+Medical+Imaging.+Bellingham,+Washington+Spie+Press&ots=aJQP2KFJaW&sig=Jys7MOPmIxxHbuHDKdbS9NNb9zU&redir_esc=y#v=onepage&q=Handbook%20of%20Medical%20Imaging.%20Bellingham%2C%20Washington%20Spie%20Press&f=false

Jbeli, N., Mastouri, R., Neji, H., Hantous-Zannad, S., & Khlifa, N. (2018). Detection and Characterization of Subsolid Juxta-pleural Lung Nodule from CT Images. *2018 5th International Conference on Control, Decision and Information Technologies, CoDIT 2018*, 1117–1121. https://doi.org/10.1109/CoDIT.2018.8394965

Kadir, T., & F., G. (2018). *Lung cancer prediction using machine learning and advanced imaging techniques*. https://doi.org/10.21037/tlcr.2018.05.15

Kafatos, F. C., Jones, C. W., & Efstratiadis, A. (1979). Determination of nucleic acid sequence homologies and relative concentrations by a dot hybridization procedure. *Nucleic Acids Research*, *7*(6), 1541–1552. https://doi.org/10.1093/NAR/7.6.1541

Karakus, S., & Avci, E. (2020). A new image steganography method with optimum pixel similarity for data hiding in medical images. *Medical Hypotheses*, *139*. https://doi.org/10.1016/J.MEHY.2020.109691

Kassambara, A. (2017). *Multivariate Analysis I Practical Guide To Cluster Analysis in R Unsupervised Machine Learning*. http://www.sthda.com

Kauffmann, A., Gentleman, R., & Huber, W. (2009). arrayQualityMetrics—a bioconductor package for quality assessment of microarray data. *Bioinformatics*, *25*(3), 415. https://doi.org/10.1093/BIOINFORMATICS/BTN647

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Kuhl, C. K. (2015). *The Changing World of Breast Cancer A Radiologist's Perspective*. 615–628. https://doi.org/10.1097/RLI.0000000000000166

Li, C., & Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Sciences*, *98*(1), 31–36. https://doi.org/10.1073/PNAS.98.1.31

Li, Y., Umbach, D. M., Krahn, J. M., Shats, I., Li, X., & Li, L. (2021). Predicting tumor response to drugs based on gene-expression biomarkers of sensitivity learned from cancer cell lines. *BMC Genomics*, *22*(1). https://doi.org/10.1186/s12864-021-07581-7

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning in genetics and genomics. *Nature Reviews. Genetics*, *16*(6), 321. https://doi.org/10.1038/NRG3920

LYONS, D. J., DUNWORTH, P. M., TILBURY, & D. W. (2009). *Simpsons Diversity Index*. http://www.countrysideinfo.co.uk/simpsons.htm

Maheswari, M., Geetha, S., & Selva kumar, S. (2019). Adaptable and proficient Hellinger Coefficient Based Collaborative Filtering for recommendation system. *Cluster Computing*, *22*, 12325–12338. https://doi.org/10.1007/s10586-017-1616-7

Mailagaha Kumbure, M., Luukka, P., & Collan, M. (2020). A new fuzzy k-nearest neighbor classifier based on the Bonferroni mean. *Pattern Recognition Letters*, *140*, 172–178. https://doi.org/10.1016/J.PATREC.2020.10.005

Martínez-Martínez, F., Lago, M. A., Rupérez, M. J., & Monserrat, C. (2013). Analysis of several biomechanical models for the simulation of lamb liver behaviour using similarity coefficients from medical image. *Computer Methods in Biomechanics and Biomedical Engineering*, *16*(7), 747–757. https://doi.org/10.1080/10255842.2011.637492

Meyer, A. da S. (2002). *Comparação de coeficientes de similaridade usados em análises de agrupamento com dados de marcadores moleculares dominantes.* https://doi.org/10.11606/D.11.2002.tde-24072002-165250

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Miller, M. B., & Tang, Y. W. (2009). Basic concepts of microarrays and potential applications in clinical microbiology. *Clinical Microbiology Reviews*, *22*(4), 611–633. https://doi.org/10.1128/CMR.00019-09/ASSET/D3C1254D-77C0-41A1-8C5F-BF70D6D592A2/ASSETS/GRAPHIC/ZCM0040922940008.JPEG

Mooney, S. D. (2015). Progress towards the integration of pharmacogenomics in practice. *Hum Genet*, *134*, 459–465. https://doi.org/10.1007/s00439-014-1484-7

Mostavi, M., Chiu, Y.-C., Huang, Y., & Chen, Y. (2020). Convolutional neural network models for cancer type prediction based on gene expression. *BMC Medical Genomics*. https://doi.org/10.1186/s12920-020-0677-2

Muller, C. B. (2017). *Potencial Preditivo em Câncer de Pulmão de Não-Pequenas Células. Tese (Doutorado em Ciências Biológicas). Departamento de Bioquímica Professor Tuiskon Dick. Instituto de Ciências Básicas da Saúde. Universidade Federal do Rio Grande do Sul, Porto Alegre*.

Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., Huff, C. D., Shannon, P. T., Jabs, E. W., Nickerson, D. A., Shendure, J., & Bamshad, M. J. (2009). Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics*. https://doi.org/10.1038/ng.499

Observador. (2022, February 14). *Cancro: o que esperar do futuro*. https://observador.pt/especiais/cancro-o-que-esperar-do-futuro/

Pereira, S. P. (2021). *Análise de imagens médicas com recurso a metodologias de deep learning*. http://repositorio.ulusiada.pt

Pérez-Wohlfeil, E., Torreno, O., Bellis, L. J., Fernandes, P. L., Leskosek, B., & Trelles, O. (2018). Training bioinformaticians in High Performance Computing. *Heliyon*, *4*(12). https://doi.org/10.1016/J.HELIYON.2018.E01057

Pevsner, J. (2015). *Bioinformatics and functional genomics*.

Pidsley, R., Zotenko, E., Peters, T. J., Lawrence, M. G., Risbridger, G. P., Molloy, P., Van Djik, S., Muhlhausler, B., Stirzaker, C., & Clark, S. J. (2016). Critical

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biology*, *17*(1). https://doi.org/10.1186/s13059-016-1066-1

Pinheiro, S. M. C. (2013). *Recomendação de Cuidados de Saúde*.

PrimeView Human Genome U219 Array Plate. (2023). https://www.thermofisher.com/order/catalog/product/901604

Ramos, J. L. C., Rodrigues, R. L., Silva, J. C. S., & Oliveira, P. L. S. de. (2020). *CRISP-EDM: uma proposta de adaptação do Modelo CRISP-DM para mineração de dados educacionais*. 1092–1101. https://doi.org/10.5753/cbie.sbie.2020.1092

Rauch, A., Hoyer, J., Guth, S., Zweier, C., Kraus, C., Becker, C., Zenker, M., Hüffmeier, U., Thiel, C., Rüschendorf, F., Nürnberg, P., Reis, A., & Trautmann, U. (2006). Diagnostic Yield of Various Genetic Approaches in Patients With Unexplained Developmental Delay or Mental Retardation. *American Journal of Medical Genetics Part A*, *140*, 2063–2074. https://doi.org/10.1002/ajmg.a.31416

Raul M. S. Laureano. (2020). *Testes de Hipóteses e Regressão - O Meu Manual de Consulta Rápida*. Edições Sílabo, Lda.

Reis, Elizabeth. (2001). *Estatística multivariada aplicada*.

Rodrigues, I., Mirza, I., Parayil, A., & Shetty, T. (2020). *Use of Linear Discriminant Analysis (LDA), K Nearest Neighbours (KNN), Decision Tree (CART), Random Forest (RF), Gaussian Naive Bayes (NB), Support Vector Machines (SVM) to Predict Admission for Post Graduation Courses*. https://ssrn.com/abstract=3683065

Rosa, G. J. de M., Bernardes Da Rocha, L., & Furlan, L. R. (2007). *Estudos de expressão gênica utilizando-se microarrays: delineamento, análise, e aplicações na pesquisa zootécnica*. 185–209. www.sbz.org.br

Sampat, M. P., Wang, Z., Gupta, S., Bovik, A. C., & Markey, M. K. (2009). Complex wavelet structural similarity: A new image similarity index. *IEEE Transactions on Image Processing*, *18*(11), 2385–2401. https://doi.org/10.1109/TIP.2009.2025923

SelectScience. (2023). *NimbleGen Gene Expression Microarray da Roche Applied Science - um membro do Grupo Roche | SelecioneCiência*. https://www.selectscience.net/products/nimblegen-gene-expression-microarray/?prodID=85094

Shah, H. (2020). *A Full Overview of Artificial Neural Networks (ANN)*. https://learn.g2.com/artificial-neural-network

Shipe, M. E., Deppen, S. A., Farjah, F., & Grogan, E. L. (2019). Developing prediction models for clinical use using logistic regression: An overview. In *Journal of Thoracic Disease* (Vol. 11, pp. S574–S584). AME Publishing Company. https://doi.org/10.21037/jtd.2019.01.25

Silva, L. K. (2003). Avaliação tecnológica e análise custo-efetividade em saúde: a incorporação de tecnologias e a produção de diretrizes clínicas para o SUS. *Ciência & Saúde Coletiva*, *8*(2), 501–520. https://doi.org/10.1590/S1413-81232003000200014

Silva, N. A. M., Poeta, P. A. C. Q. D., & Igrejas, G. P. P. (2013). *Bioinformática Aplicada Ao Estudo Da Resistência Aos Antibióticos V : Transcriptoma-Microarrays De DNA: Vol. V*.

Silveira, H. C., Defaveri, Q., & Cabanellos, V. (2017). *DIFERENÇA ENTRE TOMOGRAFIA COMPUTADORIZADA, RESSONÂNCIA MAGNÉTICA E PET-CT NA IDENTIFICAÇÃO DE LESÕES TUMORAIS*. http://ojs.fsg.br/index.php/pesquisaextensao

Subramanian, I. Verma, S Kumar, S Jere, & A e Anamika K. (2020). Multi-omics Data Integration, Interpretation, and Its Application. *Bioinformatics and Biology Insights*, 1–24. https://doi.org/10.1177/1177932219899051

Takahashi, M. M., & Jr, R. H. (2015). *Estudo comparativo de Algoritmos de Recomendação. Universidade de São Paulo.*

Tangirala, S. (2020). Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm*. *IJACSA) International Journal of Advanced Computer Science and Applications*, *11*(2). www.ijacsa.thesai.org

Thermo Fisher Scientific. (2023). *Ensaios de RNA QuantiGene para criação de perfis de expressão gênica*. https://www.thermofisher.com/pt/en/home/life-science/gene-expression-analysis-genotyping/quantigene-rna-assays.html?gclid=Cj0KCQjwi46iBhDyARIsAE3nVrapkGNWsVbHgU75KdxaP0ijKecVONgtOeh74dD5WGuzXr-wB2U-9z4aAsrPEALw_wcB&ef_id=Cj0KCQjwi46iBhDyARIsAE3nVrapkGNWsVbHgU75KdxaP0ijKecVONgtOeh74dD5WGuzXr-wB2U-9z4aAsrPEALw_wcB:G:s&s_kwcid=AL!3652!3!572685280554!p!!g!!thermo%20gene%20expression%20assay!18557644741!81491820470&cid=bid_pca_iqg_r01_co_cp1359_pjt0000_bid00000_0se_gaw_nt_pur_con

Torrente, A. (2021). Band-based similarity indices for gene expression classification and clustering. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-00678-9

Tversky, A. (1977). *Features of Similarity* (Vol. 84).

Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M. S., Shen, N., … Lander, E. S. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Undefined*, *280*(5366), 1077–1082. https://doi.org/10.1126/SCIENCE.280.5366.1077

Wang, Z., & Bovik, A. C. (2009). Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures. *IEEE Signal Processing Magazine*. https://doi.org/10.1109/MSP.2008.930649

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE TRANSACTIONS ON IMAGE PROCESSING*, *13*(4). https://doi.org/10.1109/TIP.2003.819861

Weitzel, L., Palazzo, J., De Oliveira, M., Zanon Boito, F., Dias, H., Santos, P. Dos, Campos Nobre, J., Adolfo, J., Lutz, F., Santos Dos Anjos, J. C., Yamashita, M. C., Muccillo Sklar, M., Astiazara, M. V., & Guimarães Moraes, T. (2010). *Proposta de métricas de avaliação da qualidade da informação médica para Sistemas de Recomendação baseados no perfil do usuário*.

Yao, A. D., Cheng, D. L., Pan, I., & Kitamura, F. (2020). Deep learning in neuroradiology: A systematic review of current algorithms and approaches for the new wave of imaging technology. *Radiology: Artificial Intelligence*, *2*(2). https://doi.org/10.1148/RYAI.2020190026/ASSET/IMAGES/LARGE/RYAI.2020190026.FIG3.JPEG

Zhang, X., Zhang, H., Fan, C., Hildesjö, C., Shen, B., & Sun, X. F. (2022). Loss of CHGA Protein as a Potential Biomarker for Colon Cancer Diagnosis: A Study on Biomarker Discovery by Machine Learning and Confirmation by Immunohistochemistry in Colorectal Cancer Tissue Microarrays. *Cancers*, *14*(11). https://doi.org/10.3390/CANCERS14112664/S1

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

APÊNDICES

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

APÊNDICE 1. CÓDIGO SOFTWARE RStudio Desktop

```
getwd()
setwd("C:/Users/fhole/Documents/TDNA/PAT")
library(affy)
library(arrayQualityMetrics)
library(DESeq2)
library(multtest)
library(ggplot2)
#Quality_Data
myData <- ReadAffy()
arrayQualityMetrics(myData, outdir="quality_assesment", force = T)
browseURL(file.path("quality_assesment", "index.html"))
Dataset <- ReadAffy()
Index <- c(1,2,3,100,1000,2000)
Dataset1 <- pm(Dataset)[Index,]
dim(exprs(Dataset))
#image
ids <- geneNames(Dataset)
ids[1:1000]
arrays=Dataset
arraysRMA=rma(arrays)
arraysRMA
arraysRMAtable=exprs(arraysRMA)
arraysRMAtable
dim(arraysRMAtable)
head(arraysRMAtable)
write.exprs(arraysRMA, "RMAvalues.txt")
dados <- matrix(as.integer(arraysRMAtable), ncol=20)
condition <- factor(c("B","B","B","B","S","S","S","B","B","B","B","B","B","S","S","S","S","S","S","S"))
dds <- DESeqDataSetFromMatrix(countData = dados, DataFrame(condition), design = ~ 1)
dds <- DESeq(dds)
res <- results(dds)
res
res_df <- as.data.frame(res)
# use 'fortify' function on the data frame
ggplot(fortify(res_df), aes(x=log2FoldChange, y=-log10(pvalue))) +
geom_point() +
ggtitle("Volcano plot")
ggplot(res_df, aes(x = log2(baseMean), y = log2FoldChange)) +
geom_point(aes(color = padj < 0.05), alpha = 0.5) +
scale_color_manual(values = c("gray", "red")) +
xlab("Log2 Mean Expression") +
ylab("Log2 Fold Change") +
ggtitle("MA plot")
```


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

```
#install.packages("pheatmap")
library(pheatmap)
mat <- assay(rlog(dds))
mat <- mat[rownames(res_df), ]
pheatmap(mat, show_rownames = FALSE, cluster_rows = FALSE, cluster_cols = FALSE,
annotation_col = res_df, annotation_colors = list("group" = c("gray", "red")))
# Gráfico de dispersão
ggplot(res, aes(x = baseMean, y = log2FoldChange)) +
geom_point(aes(color = padj < 0.05), alpha = 0.5) +
scale_color_manual(values = c("grey", "red")) +
labs(title = "Gráfico de Dispersão", x = "Média de Expressão", y = "log2 Fold Change")
library(pheatmap)
pheatmap(res_df, scale="none", cluster_cols=FALSE)
res1 <- results(dds, alpha = 0.05, lfcThreshold = 0.5)
res1
library(ggplot2)
resultados <- as.data.frame(res1)
p_corte <- 0.05
resultados$significativo <- ifelse(resultados$padj < p_corte & abs(resultados$log2FoldChange) > 1, "sim", "nao")
# Crie o gráfico
ggplot(resultados, aes(x = log2FoldChange, y = -log10(pvalue), color = significativo)) +
geom_point(size = 1) +
scale_color_manual(values = c("grey", "red")) +
labs(x = "log2FoldChange", y = "-log10(pvalue)", color = "Significativo") +
ggtitle("Volcano plot")
theme_classic()
groups <- c(1,1,1,1,0,0,0,1,1,1,1,1,1,0,0,0,0,0,0,0)
dados1 <- matrix(arraysRMAtable, ncol=20)
tmp_arraysRMA0.05=arraysRMAtable[,1:20]
stats=mt.teststat(tmp_arraysRMA0.05, groups, test="t")
rawp=2*(1-pnorm(abs(stats)))
adjp=p.adjust(rawp, method="BH")
arraysRMAstats=cbind(arraysRMAtable, adjp)
write.table(arraysRMAstats, "stats.txt", col.names=TRUE, row.names=TRUE, sep="\t")
df <- as.data.frame(arraysRMAstats[,1:21])
df1=df['adjp']
df1
library(pheatmap)
pheatmap(df1, scale="none", cluster_cols=FALSE)
#ANÁLISE PCA
# Transpor os dados
dados_transp <- t(dados1)
# Normalizar os dados
```


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

```
dados_norm <- scale(dados_transp)
# Realizar a análise de componentes principais
pca <- prcomp(dados_transp)
#Gráfico
plot(pca$x[,1], pca$x[,2], col = “blue”, xlab = “PC1”, ylab = “PC2”)
text(pca$x[,1], pca$x[,2], labels = rownames(dados_norm), pos=3, cex = 0.6, col = “orange”)
# clusterização
library(oligo)
clusters <- kmeans(arraysRMAtable, centers = 4)
clusters
plot(dados, col = clusters$cluster)
PCAprobes<-prcomp(exprs(arraysRMA))
PCAprobes
library(ggplot2)
library(factoextra)
arraysRMAtransp <- t(arraysRMAtable)
set.seed(123)
km.res <- kmeans(arraysRMAtransp, 2, nstart=25)
fviz_cluster(km.res, data=arraysRMAtransp,
palette = c(“red”, “#4DAF4A”, “#E7B800”, “#FC4E07”))
# Criar um objeto de dados com o resultado da clusterização
cluster_table <- data.frame(gene = names(km.res$cluster), cluster = km.res$cluster)
print(cluster_table)
# Salvar as expressões gênicas dos clusters em arquivos de texto separados
write.table(cluster_table, file=“cluster_table.txt”, quote=F, sep='\t')
getwd()
setwd(“C:/Users/fhole/Documents/TDNA/PAT”)
library(affy)
library(DESeq2)
library(philentropy)
library(tibble)
library(vegan)
install.packages(“vegan”)
##Gene CA 15.3 e CA 27-29
Dataset <- ReadAffy()
Index <- c(1,2,3,100,1000,2000)
Dataset1 <- pm(Dataset)[Index,]
dim(exprs(Dataset))
#image
ids <- geneNames(Dataset)
ids[1:1000]
arrays=Dataset
arraysRMA=rma(arrays)
arraysRMA
arraysRMAtable=exprs(arraysRMA)
arraysRMAtable
A <- arraysRMAtable[819:819, 1:20]
```


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

```
dados <- tibble(A)
dados
dist_matrix <- distance(x=dados, method = “dice”)
colnames(dist_matrix) <- c(“A01.5500_Breast”,”A01_Breast”,”A03_Breast”,”A05_Breast”,”A05_Skin”,”A06_Skin”,”A07_Skin”,”A08_Breast”,”A09.3978_Breast”,”A09_Breast”,“A10_Breast”,”A11_Breast”,”A12_Breast”,”A12_Skin”,”B02_Skin”,”B09_Skin”,”D07_Skin”,”G02_Skin”,”H06_Skin”,”H10_Skin”)
rownames(dist_matrix) <- c(“A01.5500_Breast”,”A01_Breast”,”A03_Breast”,”A05_Breast”,”A05_Skin”,”A06_Skin”,”A07_Skin”,”A08_Breast”,”A09.3978_Breast”,”A09_Breast”,“A10_Breast”,”A11_Breast”,”A12_Breast”,”A12_Skin”,”B02_Skin”,”B09_Skin”,”D07_Skin”,”G02_Skin”,”H06_Skin”,”H10_Skin”)

dist_matrix
write.table(dist_matrix, file=“Dice.txt”, quote=F, sep='\t')
heatmap(dist_matrix, scale = “none”)
#########################################################
dist_matrix1 <- distance(x=dados, method = “jaccard”)
colnames(dist_matrix1) <- c(“A01.5500_Breast”,”A01_Breast”,”A03_Breast”,”A05_Breast”,”A05_Skin”,”A06_Skin”,”A07_Skin”,”A08_Breast”,”A09.3978_Breast”,”A09_Breast”,“A10_Breast”,”A11_Breast”,”A12_Breast”,”A12_Skin”,”B02_Skin”,”B09_Skin”,”D07_Skin”,”G02_Skin”,”H06_Skin”,”H10_Skin”)
rownames(dist_matrix1) <- c(“A01.5500_Breast”,”A01_Breast”,”A03_Breast”,”A05_Breast”,”A05_Skin”,”A06_Skin”,”A07_Skin”,”A08_Breast”,”A09.3978_Breast”,”A09_Breast”,“A10_Breast”,”A11_Breast”,”A12_Breast”,”A12_Skin”,”B02_Skin”,”B09_Skin”,”D07_Skin”,”G02_Skin”,”H06_Skin”,”H10_Skin”)
dist_matrix1
write.table(dist_matrix1, file=“Jaccard.txt”, quote=F, sep='\t')
heatmap(dist_matrix1, scale = “none”)
#########################################################
dist_matrix2 <- distance(x=dados, method = “sorensen”)
colnames(dist_matrix2) <- c(“A01.5500_Breast”,”A01_Breast”,”A03_Breast”,”A05_Breast”,”A05_Skin”,”A06_Skin”,”A07_Skin”,”A08_Breast”,”A09.3978_Breast”,”A09_Breast”,“A10_Breast”,”A11_Breast”,”A12_Breast”,”A12_Skin”,”B02_Skin”,”B09_Skin”,”D07_Skin”,”G02_Skin”,”H06_Skin”,”H10_Skin”)
rownames(dist_matrix2) <- c(“A01.5500_Breast”,”A01_Breast”,”A03_Breast”,”A05_Breast”,”A05_Skin”,”A06_Skin”,”A07_Skin”,”A08_Breast”,”A09.3978_Breast”,”A09_Breast”,“A10_Breast”,”A11_Breast”,”A12_Breast”,”A12_Skin”,”B02_Skin”,”B09_Skin”,”D07_Skin”,”G02_Skin”,”H06_Skin”,”H10_Skin”)
dist_matrix2
write.table(dist_matrix2, file=“Sorensen819.txt”, quote=F, sep='\t')
heatmap(dist_matrix2, scale = “none”)
#########################################################
```


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

```
dist_matrix3 <- distance(x=dados, method = “czekanowski”)
colnames(dist_matrix3) <- c(“A01.5500_Breast”,”A01_Breast”,”A03_Breast”,”A05_Breast”,”A05_Skin”,”A06_Skin”,”A07_Skin”,”A08_Breast”,”A09.3978_Breast”,”A09_Breast”,“A10_Breast”,”A11_Breast”,”A12_Breast”,”A12_Skin”,”B02_Skin”,”B09_Skin”,”D07_Skin”,”G02_Skin”,”H06_Skin”,”H10_Skin”)
rownames(dist_matrix3) <- c(“A01.5500_Breast”,”A01_Breast”,”A03_Breast”,”A05_Breast”,”A05_Skin”,”A06_Skin”,”A07_Skin”,”A08_Breast”,”A09.3978_Breast”,”A09_Breast”,“A10_Breast”,”A11_Breast”,”A12_Breast”,”A12_Skin”,”B02_Skin”,”B09_Skin”,”D07_Skin”,”G02_Skin”,”H06_Skin”,”H10_Skin”)
dist_matrix3
write.table(dist_matrix3, file=“czekanowski819.txt”, quote=F, sep='\t')
heatmap(dist_matrix3, scale = “none”)
#########################################################
dist_matrix4 <- distance(x=dados, method = “minkowski”, p = 2 )
colnames(dist_matrix4) <- c(“A01.5500_Breast”,”A01_Breast”,”A03_Breast”,”A05_Breast”,”A05_Skin”,”A06_Skin”,”A07_Skin”,”A08_Breast”,”A09.3978_Breast”,”A09_Breast”,“A10_Breast”,”A11_Breast”,”A12_Breast”,”A12_Skin”,”B02_Skin”,”B09_Skin”,”D07_Skin”,”G02_Skin”,”H06_Skin”,”H10_Skin”)
rownames(dist_matrix4) <- c(“A01.5500_Breast”,”A01_Breast”,”A03_Breast”,”A05_Breast”,”A05_Skin”,”A06_Skin”,”A07_Skin”,”A08_Breast”,”A09.3978_Breast”,”A09_Breast”,“A10_Breast”,”A11_Breast”,”A12_Breast”,”A12_Skin”,”B02_Skin”,”B09_Skin”,”D07_Skin”,”G02_Skin”,”H06_Skin”,”H10_Skin”)
dist_matrix4
write.table(dist_matrix4, file=“minkowski819.txt”, quote=F, sep='\t')
heatmap(dist_matrix4, scale = “none”)
#########################################################
dist_matrix5 <- distance(x=dados, method = “pearson”)
colnames(dist_matrix5) <- c(“A01.5500_Breast”,”A01_Breast”,”A03_Breast”,”A05_Breast”,”A05_Skin”,”A06_Skin”,”A07_Skin”,”A08_Breast”,”A09.3978_Breast”,”A09_Breast”,“A10_Breast”,”A11_Breast”,”A12_Breast”,”A12_Skin”,”B02_Skin”,”B09_Skin”,”D07_Skin”,”G02_Skin”,”H06_Skin”,”H10_Skin”)
rownames(dist_matrix5) <- c(“A01.5500_Breast”,”A01_Breast”,”A03_Breast”,”A05_Breast”,”A05_Skin”,”A06_Skin”,”A07_Skin”,”A08_Breast”,”A09.3978_Breast”,”A09_Breast”,“A10_Breast”,”A11_Breast”,”A12_Breast”,”A12_Skin”,”B02_Skin”,”B09_Skin”,”D07_Skin”,”G02_Skin”,”H06_Skin”,”H10_Skin”)
dist_matrix5
write.table(dist_matrix5, file=“pearson819.txt”, quote=F, sep='\t')
heatmap(dist_matrix5, scale = “none”)
#########################################################
dist_matrix6 <- distance(x=dados, method = “intersection”)
colnames(dist_matrix6) <- c(“A01.5500_Breast”,”A01_Breast”,”A03_Breast”,”A05_Breast”,”A05_Skin”,”A
```


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

```
06_Skin",”A07_Skin",”A08_Breast",”A09.3978_Breast",”A09_Breast",
“A10_Breast",”A11_Breast",”A12_Breast",”A12_Skin",”B02_Skin",”B09_Skin",
”D07_Skin",”G02_Skin",”H06_Skin",”H10_Skin")
181.  rownames(dist_matrix6) <-
c(“A01.5500_Breast",”A01_Breast",”A03_Breast",”A05_Breast",”A05_Skin",”A
06_Skin",”A07_Skin",”A08_Breast",”A09.3978_Breast",”A09_Breast",
“A10_Breast",”A11_Breast",”A12_Breast",”A12_Skin",”B02_Skin",”B09_Skin",
”D07_Skin",”G02_Skin",”H06_Skin",”H10_Skin")
182.  dist_matrix6
183.  write.table(dist_matrix6, file=“intersection819.txt", quote=F, sep='\t')
184.  heatmap(dist_matrix6, scale = “none")
185.  ########################################################
186.  dist_matrix7 <- distance(x=dados, method = “manhattan")
187.  colnames(dist_matrix7) <-
c(“A01.5500_Breast",”A01_Breast",”A03_Breast",”A05_Breast",”A05_Skin",”A
06_Skin",”A07_Skin",”A08_Breast",”A09.3978_Breast",”A09_Breast",
“A10_Breast",”A11_Breast",”A12_Breast",”A12_Skin",”B02_Skin",”B09_Skin",
”D07_Skin",”G02_Skin",”H06_Skin",”H10_Skin")
188.  rownames(dist_matrix7) <-
c(“A01.5500_Breast",”A01_Breast",”A03_Breast",”A05_Breast",”A05_Skin",”A
06_Skin",”A07_Skin",”A08_Breast",”A09.3978_Breast",”A09_Breast",
189.  “A10_Breast",”A11_Breast",”A12_Breast",”A12_Skin",”B02_Skin",”B09_Skin",
”D07_Skin",”G02_Skin",”H06_Skin",”H10_Skin")
190.  dist_matrix7
191.  write.table(dist_matrix7, file=“manhattan819.txt", quote=F, sep='\t')
192.  heatmap(dist_matrix7, scale = “none")
193.  ########################################################
194.  dist_matrix8 <- distance(x=dados, method = “tanimoto")
195.  colnames(dist_matrix8) <-
c(“A01.5500_Breast",”A01_Breast",”A03_Breast",”A05_Breast",”A05_Skin",”A
06_Skin",”A07_Skin",”A08_Breast",”A09.3978_Breast",”A09_Breast",
“A10_Breast",”A11_Breast",”A12_Breast",”A12_Skin",”B02_Skin",”B09_Skin",
”D07_Skin",”G02_Skin",”H06_Skin",”H10_Skin")
196.  rownames(dist_matrix8) <-
c(“A01.5500_Breast",”A01_Breast",”A03_Breast",”A05_Breast",”A05_Skin",”A
06_Skin",”A07_Skin",”A08_Breast",”A09.3978_Breast",”A09_Breast",
“A10_Breast",”A11_Breast",”A12_Breast",”A12_Skin",”B02_Skin",”B09_Skin",
”D07_Skin",”G02_Skin",”H06_Skin",”H10_Skin")
197.  dist_matrix8
198.  write.table(dist_matrix8, file=“tanimoto819.txt", quote=F, sep='\t')
199.  heatmap(dist_matrix8, scale = “none")
200.  ########################################################
201.  dist_matrix9 <- distance(x=dados, method = “sorensen")
202.  colnames(dist_matrix9) <-
c(“A01.5500_Breast",”A01_Breast",”A03_Breast",”A05_Breast",”A05_Skin",”A
06_Skin",”A07_Skin",”A08_Breast",”A09.3978_Breast",”A09_Breast",
203.  “A10_Breast",”A11_Breast",”A12_Breast",”A12_Skin",”B02_Skin",”B09_Skin",
”D07_Skin",”G02_Skin",”H06_Skin",”H10_Skin")
```


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

```
rownames(dist_matrix9) <- c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast","A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin","D07_Skin","G02_Skin","H06_Skin","H10_Skin")
dist_matrix9
write.table(dist_matrix9, file="sorensen819.txt", quote=F, sep='\t')
heatmap(dist_matrix9, scale = "none")
#########################################################
dist_matrix10 <- distance(x=dados, method = "euclidean")
colnames(dist_matrix10) <- c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast","A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin","D07_Skin","G02_Skin","H06_Skin","H10_Skin")
rownames(dist_matrix10) <- c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast","A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin","D07_Skin","G02_Skin","H06_Skin","H10_Skin")
dist_matrix10
write.table(dist_matrix10, file="euclidean819.txt", quote=F, sep='\t')
heatmap(dist_matrix10, scale = "none")

#############################################################
B <- arraysRMAtable[1788:1788, 1:20]
dados1 <- tibble(B)
dist_matrix11 <- distance(x=dados1, method = "dice")
colnames(dist_matrix11) <- c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast","A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin","D07_Skin","G02_Skin","H06_Skin","H10_Skin")
rownames(dist_matrix11) <- c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast","A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin","D07_Skin","G02_Skin","H06_Skin","H10_Skin")
dist_matrix11
write.table(dist_matrix11, file="Dice1788.txt", quote=F, sep='\t')
heatmap(dist_matrix11, scale = "none")
#########################################################
dist_matrix12 <- distance(x=dados1, method = "jaccard")
colnames(dist_matrix12) <- c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast","A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin","D07_Skin","G02_Skin","H06_Skin","H10_Skin")
```


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

```
rownames(dist_matrix12) <- c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast","A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin","D07_Skin","G02_Skin","H06_Skin","H10_Skin")
dist_matrix12
write.table(dist_matrix12, file="Jaccard1788.txt", quote=F, sep='\t')
heatmap(dist_matrix12, scale = "none")
#########################################################
dist_matrix13 <- distance(x=dados1, method = "sorensen")
colnames(dist_matrix13) <- c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast","A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin","D07_Skin","G02_Skin","H06_Skin","H10_Skin")
rownames(dist_matrix13) <- c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast","A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin","D07_Skin","G02_Skin","H06_Skin","H10_Skin")
dist_matrix13
write.table(dist_matrix13, file="Sorensen1788.txt", quote=F, sep='\t')
heatmap(dist_matrix13, scale = "none")
#########################################################
dist_matrix14 <- distance(x=dados1, method = "czekanowski")
colnames(dist_matrix14) <- c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast","A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin","D07_Skin","G02_Skin","H06_Skin","H10_Skin")
rownames(dist_matrix14) <- c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast","A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin","D07_Skin","G02_Skin","H06_Skin","H10_Skin")
dist_matrix14
write.table(dist_matrix14, file="czekanowski1788.txt", quote=F, sep='\t')
heatmap(dist_matrix14, scale = "none")
#########################################################
dist_matrix15 <- distance(x=dados1, method = "minkowski", p = 2)
colnames(dist_matrix15) <- c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast","A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin","D07_Skin","G02_Skin","H06_Skin","H10_Skin")
rownames(dist_matrix15) <- c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast",
```


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

```
"A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin",
"D07_Skin","G02_Skin","H06_Skin","H10_Skin")
dist_matrix15
write.table(dist_matrix15, file="cosine1788.txt", quote=F, sep='\t')
heatmap(dist_matrix15, scale = "none")
#########################################################
dist_matrix16 <- distance(x=dados1, method = "pearson")
colnames(dist_matrix16) <-
c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A
06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast",
"A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin",
"D07_Skin","G02_Skin","H06_Skin","H10_Skin")
rownames(dist_matrix16) <-
c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A
06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast",
"A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin",
"D07_Skin","G02_Skin","H06_Skin","H10_Skin")
dist_matrix16
write.table(dist_matrix16, file="pearson1788.txt", quote=F, sep='\t')
heatmap(dist_matrix16, scale = "none")
#########################################################
dist_matrix17 <- distance(x=dados1, method = "intersection")
colnames(dist_matrix17) <-
c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A
06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast",
"A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin",
"D07_Skin","G02_Skin","H06_Skin","H10_Skin")
rownames(dist_matrix17) <-
c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A
06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast",
"A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin",
"D07_Skin","G02_Skin","H06_Skin","H10_Skin")
dist_matrix17
write.table(dist_matrix17, file="intersection1788.txt", quote=F, sep='\t')
heatmap(dist_matrix17, scale = "none")
#########################################################
dist_matrix18 <- distance(x=dados1, method = "manhattan")
colnames(dist_matrix18) <-
c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A
06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast",
"A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin",
"D07_Skin","G02_Skin","H06_Skin","H10_Skin")
rownames(dist_matrix18) <-
c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A
06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast",
"A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin",
"D07_Skin","G02_Skin","H06_Skin","H10_Skin")
dist_matrix18
```


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

```
write.table(dist_matrix18, file="manhattan1788.txt", quote=F, sep='\t')
heatmap(dist_matrix18, scale = "none")
########################################################
dist_matrix19 <- distance(x=dados1, method = "tanimoto")
colnames(dist_matrix19) <- c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast", "A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin", "D07_Skin","G02_Skin","H06_Skin","H10_Skin")
rownames(dist_matrix19) <- c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast", "A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin", "D07_Skin","G02_Skin","H06_Skin","H10_Skin")
dist_matrix19
write.table(dist_matrix19, file="tanimoto1788.txt", quote=F, sep='\t')
heatmap(dist_matrix19, scale = "none")
########################################################
dist_matrix20 <- distance(x=dados1, method = "sorensen")
colnames(dist_matrix20) <- c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast", "A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin", "D07_Skin","G02_Skin","H06_Skin","H10_Skin")
rownames(dist_matrix20) <- c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast", "A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin", "D07_Skin","G02_Skin","H06_Skin","H10_Skin")
dist_matrix20
write.table(dist_matrix20, file="sorensen1788.txt", quote=F, sep='\t')
heatmap(dist_matrix20, scale = "none")
########################################################
dist_matrix21 <- distance(x=dados1, method = "euclidean")
colnames(dist_matrix21) <- c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast", "A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin", "D07_Skin","G02_Skin","H06_Skin","H10_Skin")
rownames(dist_matrix21) <- c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast", "A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin", "D07_Skin","G02_Skin","H06_Skin","H10_Skin")
dist_matrix21
write.table(dist_matrix21, file="euclidean1788.txt", quote=F, sep='\t')
heatmap(dist_matrix21, scale = "none")
########################################################
```


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

```
C <- arraysRMAtable[9087:9087, 1:20]
dados2 <- tibble(C)
dist_matrix22 <- distance(x=dados2, method = “dice”)
colnames(dist_matrix22) <- c(“A01.5500_Breast”,”A01_Breast”,”A03_Breast”,”A05_Breast”,”A05_Skin”,”A06_Skin”,”A07_Skin”,”A08_Breast”,”A09.3978_Breast”,”A09_Breast”, “A10_Breast”,”A11_Breast”,”A12_Breast”,”A12_Skin”,”B02_Skin”,”B09_Skin”, ”D07_Skin”,”G02_Skin”,”H06_Skin”,”H10_Skin”)
rownames(dist_matrix22) <- c(“A01.5500_Breast”,”A01_Breast”,”A03_Breast”,”A05_Breast”,”A05_Skin”,”A06_Skin”,”A07_Skin”,”A08_Breast”,”A09.3978_Breast”,”A09_Breast”, “A10_Breast”,”A11_Breast”,”A12_Breast”,”A12_Skin”,”B02_Skin”,”B09_Skin”, ”D07_Skin”,”G02_Skin”,”H06_Skin”,”H10_Skin”)
dist_matrix22
write.table(dist_matrix11, file=“Dice9087.txt”, quote=F, sep='\t')
heatmap(dist_matrix22, scale = “none”)
#########################################################
dist_matrix23 <- distance(x=dados2, method = “jaccard”)
colnames(dist_matrix23) <- c(“A01.5500_Breast”,”A01_Breast”,”A03_Breast”,”A05_Breast”,”A05_Skin”,”A06_Skin”,”A07_Skin”,”A08_Breast”,”A09.3978_Breast”,”A09_Breast”, “A10_Breast”,”A11_Breast”,”A12_Breast”,”A12_Skin”,”B02_Skin”,”B09_Skin”, ”D07_Skin”,”G02_Skin”,”H06_Skin”,”H10_Skin”)
rownames(dist_matrix23) <- c(“A01.5500_Breast”,”A01_Breast”,”A03_Breast”,”A05_Breast”,”A05_Skin”,”A06_Skin”,”A07_Skin”,”A08_Breast”,”A09.3978_Breast”,”A09_Breast”, “A10_Breast”,”A11_Breast”,”A12_Breast”,”A12_Skin”,”B02_Skin”,”B09_Skin”, ”D07_Skin”,”G02_Skin”,”H06_Skin”,”H10_Skin”)
dist_matrix23
write.table(dist_matrix23, file=“Jaccard9087.txt”, quote=F, sep='\t')
heatmap(dist_matrix23, scale = “none”)
#########################################################
dist_matrix24 <- distance(x=dados2, method = “sorensen”)
colnames(dist_matrix24) <- c(“A01.5500_Breast”,”A01_Breast”,”A03_Breast”,”A05_Breast”,”A05_Skin”,”A06_Skin”,”A07_Skin”,”A08_Breast”,”A09.3978_Breast”,”A09_Breast”, “A10_Breast”,”A11_Breast”,”A12_Breast”,”A12_Skin”,”B02_Skin”,”B09_Skin”, ”D07_Skin”,”G02_Skin”,”H06_Skin”,”H10_Skin”)
rownames(dist_matrix24) <- c(“A01.5500_Breast”,”A01_Breast”,”A03_Breast”,”A05_Breast”,”A05_Skin”,”A06_Skin”,”A07_Skin”,”A08_Breast”,”A09.3978_Breast”,”A09_Breast”, “A10_Breast”,”A11_Breast”,”A12_Breast”,”A12_Skin”,”B02_Skin”,”B09_Skin”, ”D07_Skin”,”G02_Skin”,”H06_Skin”,”H10_Skin”)
dist_matrix24
write.table(dist_matrix24, file=“Sorensen9087.txt”, quote=F, sep='\t')
heatmap(dist_matrix24, scale = “none”)
#########################################################
dist_matrix25 <- distance(x=dados2, method = “czekanowski”)
```


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

```
colnames(dist_matrix25) <- c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast", "A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin", "D07_Skin","G02_Skin","H06_Skin","H10_Skin")
rownames(dist_matrix25) <- c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast", "A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin", "D07_Skin","G02_Skin","H06_Skin","H10_Skin")
dist_matrix25
write.table(dist_matrix25, file="czekanowski9087.txt", quote=F, sep='\t')
heatmap(dist_matrix25, scale = "none")
#########################################################
dist_matrix26 <- distance(x=dados2, method = "minkowski", p = 2)
colnames(dist_matrix26) <- c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast",
"A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin", "D07_Skin","G02_Skin","H06_Skin","H10_Skin")
rownames(dist_matrix26) <- c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast", "A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin", "D07_Skin","G02_Skin","H06_Skin","H10_Skin")
dist_matrix26
write.table(dist_matrix26, file="cosine9087.txt", quote=F, sep='\t')
heatmap(dist_matrix26, scale = "none")
#########################################################
dist_matrix27 <- distance(x=dados2, method = "pearson")
colnames(dist_matrix27) <- c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast", "A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin", "D07_Skin","G02_Skin","H06_Skin","H10_Skin")
rownames(dist_matrix27) <- c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast", "A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin", "D07_Skin","G02_Skin","H06_Skin","H10_Skin")
dist_matrix27
write.table(dist_matrix27, file="pearson9087.txt", quote=F, sep='\t')
heatmap(dist_matrix27, scale = "none")
#########################################################
dist_matrix28 <- distance(x=dados2, method = "intersection")
colnames(dist_matrix28) <- c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast",
```


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

```
"A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin",
"D07_Skin","G02_Skin","H06_Skin","H10_Skin")
348. rownames(dist_matrix28) <-
c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A
06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast",
"A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin",
"D07_Skin","G02_Skin","H06_Skin","H10_Skin")
349. dist_matrix28
350. write.table(dist_matrix28, file="intersection9087.txt", quote=F, sep='\t')
351. heatmap(dist_matrix28, scale = "none")
352. ########################################################
353. dist_matrix29 <- distance(x=dados2, method = "manhattan")
354. colnames(dist_matrix29) <-
c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A
06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast",
"A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin",
"D07_Skin","G02_Skin","H06_Skin","H10_Skin")
355. rownames(dist_matrix29) <-
c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A
06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast",
"A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin",
"D07_Skin","G02_Skin","H06_Skin","H10_Skin")
356. dist_matrix29
357. write.table(dist_matrix29, file="manhattan9087.txt", quote=F, sep='\t')
358. heatmap(dist_matrix29, scale = "none")
359. ########################################################
360. dist_matrix30 <- distance(x=dados2, method = "tanimoto")
361. colnames(dist_matrix30) <-
c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A
06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast",
"A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin",
"D07_Skin","G02_Skin","H06_Skin","H10_Skin")
362. rownames(dist_matrix30) <-
c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A
06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast",
"A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin",
"D07_Skin","G02_Skin","H06_Skin","H10_Skin")
363. dist_matrix30
364. write.table(dist_matrix30, file="tanimoto9087.txt", quote=F, sep='\t')
365. heatmap(dist_matrix30, scale = "none")
366. ########################################################
367. dist_matrix31 <- distance(x=dados2, method = "sorensen")
368. colnames(dist_matrix31) <-
c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A
06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast",
"A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin",
"D07_Skin","G02_Skin","H06_Skin","H10_Skin")
```


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

```
rownames(dist_matrix31) <- c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast",
"A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin",
"D07_Skin","G02_Skin","H06_Skin","H10_Skin")
dist_matrix31
write.table(dist_matrix31, file="sorensen9087.txt", quote=F, sep='\t')
heatmap(dist_matrix31, scale = "none")
#########################################################
dist_matrix32 <- distance(x=dados2, method = "euclidean")
colnames(dist_matrix32) <- c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast",
"A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin",
"D07_Skin","G02_Skin","H06_Skin","H10_Skin")
rownames(dist_matrix32) <- c("A01.5500_Breast","A01_Breast","A03_Breast","A05_Breast","A05_Skin","A06_Skin","A07_Skin","A08_Breast","A09.3978_Breast","A09_Breast",
"A10_Breast","A11_Breast","A12_Breast","A12_Skin","B02_Skin","B09_Skin",
"D07_Skin","G02_Skin","H06_Skin","H10_Skin")
dist_matrix32
write.table(dist_matrix32, file="euclidean9087.txt", quote=F, sep='\t')
heatmap(dist_matrix32, scale = "none")
#####################################################################
matriz <- matrix(c(12.23467, 12.10041, 11.9459, 12.49984, 6.001979, 3.679164, 3.898591, 12.14353, 7.120905, 11.95416, 12.55628, 11.81903, 11.43515, 6.592423, 3.372306, 3.421478, 3.321774, 3.825623, 3.378157, 3.637505),
nrow = 20, ncol = 20)
matriz
distancia <- dist(matriz)
similaridade <- as.matrix(1/(1+distancia))
distancia
x <- diversity(matriz, index = "simpson")
x
distancia <- diversity(teste, index = "simpson")
similaridade <- as.matrix(1/(1+x))
similaridade
distancia
teste
binario <- apply(matriz, 1, function(x) ifelse(x > 0, 1, 0))
n_linhas <- nrow(binario)
similaridade_simpson <- matrix(0, n_linhas, n_linhas)
for(i in 1:(n_linhas-1)) {
for(j in (i+1):n_linhas) {
div_i <- simpson_diversity(binario[i,])
div_j <- simpson_diversity(binario[j,])
div_comum <- simpson_diversity(binario[i,] & binario[j,])
similaridade_simpson[i,j] <- div_comum / min(div_i, div_j)
```


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

```
similaridade_simpson[j,i] <- similaridade_simpson[i,j]
}
}
similaridade_simpson
#####################################################################
similaridade_simpson <- function(x, y){
n <- length(intersect(x, y))
return(n / min(length(x), length(y)))
}
dados <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
matriz_distancia <- matrix(0, nrow = length(dados), ncol = length(dados))
for(i in 1:length(dados)){
for(j in 1:length(dados)){
sim <- similaridade_simpson(dados[i], dados[j])
matriz_distancia[i,j] <- 1 - sim
}
}
method <- “chebyshev”
matriz_distancia_chebyshev <- dist(x = matriz_distancia, method = method)
matriz_distancia_chebyshev
```


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

APÊNDICE 2. CÓDIGO SOFTWARE Python

```
[ ] from google.colab import drive
    drive.mount('/content/gdrive')
    path ='/content/gdrive/My Drive/ExpressãoGenica'
```

```
Drive already mounted at /content/gdrive; to attempt to forcibly remount, call drive.mount("/content/gdrive", force_remount=True).
```

```
[ ] !pip install pandas numpy tensorflow keras openpyxl
```

```
import pandas as pd

DF = pd.read_excel('/content/gdrive/My Drive/ExpressãoGenica/RMAvalues.xlsx')
```

```
[ ] import numpy as np
    import pandas as pd
    DF1 = np.transpose(DF)
```

```
[ ] #Breast=0 e Skin=1
    novos_valores = [0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1]
    #DF1 = DF1.drop('nova_coluna', axis=1)
    DF1['PATOLOGIA'] = novos_valores
```

```
[ ] DF1
```

```
[ ] import pandas as pd
    from sklearn.model_selection import train_test_split

    # separar a coluna de classes
    y = DF1['PATOLOGIA']

    # usar as colunas restantes como dados de entrada
    X = DF1[0:49385]
    X.columns = X.columns.astype(str)
    # dividir os dados em conjunto de treinamento e teste
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```
[ ] from sklearn.preprocessing import MinMaxScaler

    scaler = MinMaxScaler()
    X = scaler.fit_transform(X)
```

```
from keras.models import Sequential
from keras.layers import Dense, Dropout, Flatten, Conv1D, MaxPooling1D

model = Sequential()
model.add(Conv1D(filters=64, kernel_size=3, activation='relu', input_shape=(X.shape[1], 1)))
model.add(MaxPooling1D(pool_size=2))
model.add(Flatten())
model.add(Dense(128, activation='relu'))
model.add(Dropout(0.35))
model.add(Dense(len(y), activation='softmax'))
model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])
```


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

```
[ ] from keras.utils import to_categorical
    from matplotlib import pyplot as plt

    y_encoded = to_categorical(range(len(y)))
    model.fit(X.reshape(X.shape[0], X.shape[1], 1), y_encoded, epochs=15)

    plt.plot(history.history['loss'], label='train_loss')
    plt.plot(history.history['accuracy'], label='train_acc')
    plt.legend()
    plt.show()
```

```
[ ] from keras.utils import to_categorical
    from matplotlib import pyplot as plt

    y_encoded = to_categorical(range(len(y)))

    # Definir a métrica de acurácia no método compile()
    model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])

    # Treinar o modelo e obter o objeto History
    history = model.fit(X.reshape(X.shape[0], X.shape[1], 1), y_encoded, epochs=15)

    # Plotar as curvas de perda e acurácia
    plt.plot(history.history['loss'], label='train_loss')
    plt.plot(history.history['accuracy'], label='train_acc')
    plt.legend()
    plt.show()
```

```
Epoch 1/15
1/1 [==============================] - 6s 6s/step - loss: 0.5962 - accuracy: 0.7500
Epoch 2/15
1/1 [==============================] - 5s 5s/step - loss: 3.8145 - accuracy: 0.4000
Epoch 3/15
1/1 [==============================] - 4s 4s/step - loss: 1.4965 - accuracy: 0.6500
Epoch 4/15
1/1 [==============================] - 4s 4s/step - loss: 1.0393 - accuracy: 0.7500
Epoch 5/15
1/1 [==============================] - 6s 6s/step - loss: 1.8005 - accuracy: 0.6500
```


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

```
[ ] y_pred = model.predict(X.reshape(X.shape[0], X.shape[1], 1))
```

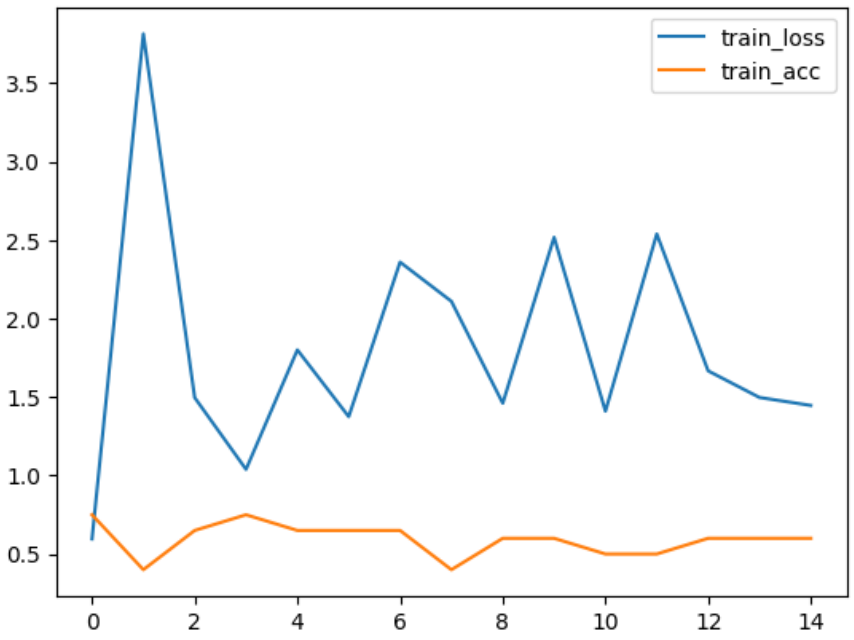
```
1/1 [==============================] - 1s 1s/step
```

```
yp_train = model.predict(X_train)
#yp_train = np.argmax(yp_train, axis = 1)

yp_test = model.predict(X_test)
#yp_test = np.argmax(yp_test, axis = 1)
```

```
1/1 [==============================] - 1s 652ms/step
1/1 [==============================] - 0s 263ms/step
```

```
from sklearn.metrics import accuracy_score

# Prever as saídas do modelo para o conjunto de teste
y_pred = model.predict(X_test)

# Transformar as saídas previstas em rótulos de classe
y_pred_classes = np.argmax(y_pred, axis=1)

# Calcular a acurácia média em todas as saídas do modelo
acc = accuracy_score(y_test, y_pred_classes)
acc
```

```
1/1 [==============================] - 0s 206ms/step
0.0
```

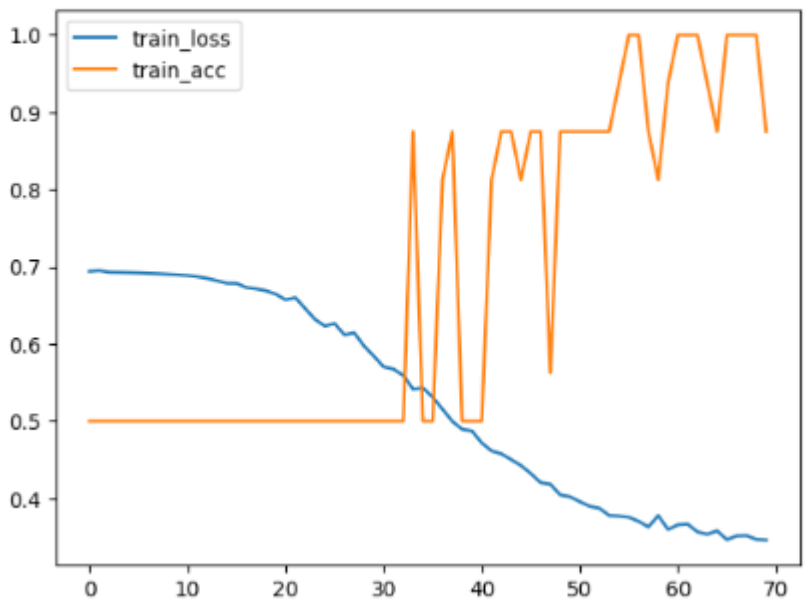


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

```python
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from keras.models import Sequential
from keras.layers import Dense

model = Sequential()
model.add(Dense(units=6, kernel_initializer='uniform', activation='relu', input_dim=X_train.shape[1]))
model.add(Dense(units=6, kernel_initializer='uniform', activation='relu'))
model.add(Dense(units=1, kernel_initializer='uniform', activation='sigmoid'))

# compila a rede neural
model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])

# treina a rede neural
history=model.fit(X_train, y_train, batch_size=64, epochs=70, verbose=1)

# avalia a precisão da rede neural no conjunto de teste
loss, accuracy = model.evaluate(X_test, y_test, verbose=0)
print('Test accuracy:', accuracy)
```

```
Epoch 1/70
1/1 [==============================] - 1s 1s/step - loss: 0.6940 - accuracy: 0.5000
Epoch 2/70
1/1 [==============================] - 0s 33ms/step - loss: 0.6949 - accuracy: 0.5000
Epoch 3/70
```

```python
from matplotlib import pyplot as plt
plt.plot(history.history['loss'], label='train_loss')
plt.plot(history.history['accuracy'], label='train_acc')
plt.legend()
plt.show()
```



Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

```python
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score

# Obter as previsões do modelo
y_pred = model.predict(X_test)

# Calcular as métricas
accuracy = accuracy_score(y_test, yp_test)
precision = precision_score(y_test, yp_test)
recall = recall_score(y_test, yp_test)
f1 = f1_score(y_test, yp_test)

print("Accuracy:", accuracy)
print("Precision:", precision)
print("Recall:", recall)
print("F1-score:", f1)
```

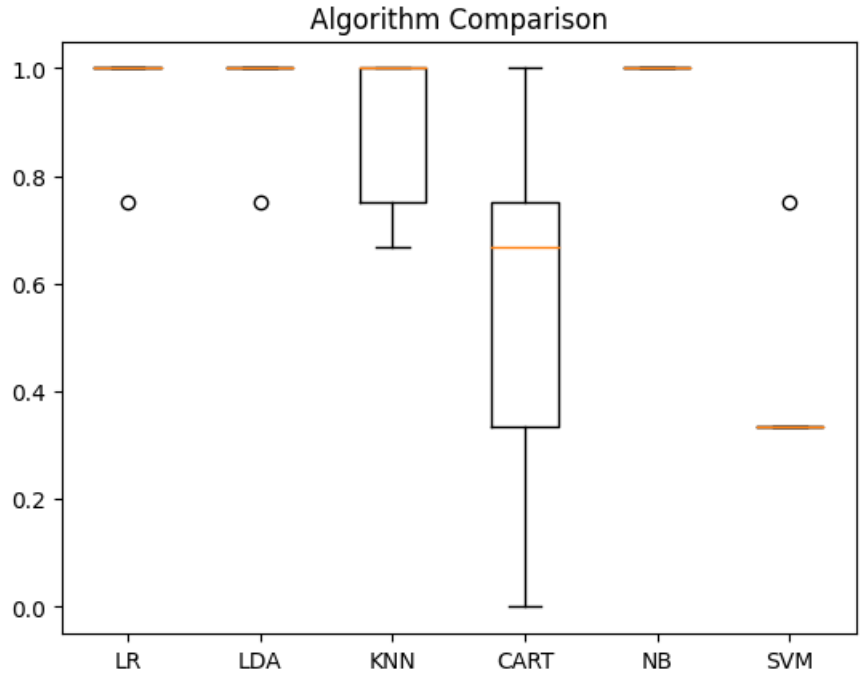
```
1/1 [==============================] - 0s 23ms/step
Accuracy: 0.5
Precision: 0.0
Recall: 0.0
F1-score: 0.0
```


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

```python
# comparar algoritmos - Machine Learning
from pandas import read_csv
from matplotlib import pyplot
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import StratifiedKFold
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC

# Verificar Algoritmos
modelo = []
modelo.append(('LR', LogisticRegression(solver='liblinear', multi_class='ovr')))
modelo.append(('LDA', LinearDiscriminantAnalysis()))
modelo.append(('KNN', KNeighborsClassifier()))
modelo.append(('CART', DecisionTreeClassifier()))
modelo.append(('NB', GaussianNB()))
modelo.append(('SVM', SVC(gamma=0.001)))

# avaliação de cada modelo
results = []
names = []
for name, model in modelo:
  kfold = StratifiedKFold(n_splits=5, random_state=42, shuffle=True)
  cv_results = cross_val_score(model, X_train, y_train, cv=kfold, scoring='accuracy')
  results.append(cv_results)
  names.append(name)
  print('%s: %f (%f)' % (name, cv_results.mean(), cv_results.std()))

# Comparar algoritmos
pyplot.boxplot(results, labels=names)
pyplot.title('Algorithm Comparison')
pyplot.show()
```


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

```
LR: 0.950000 (0.100000)
LDA: 0.950000 (0.100000)
KNN: 0.883333 (0.145297)
CART: 0.550000 (0.348010)
NB: 1.000000 (0.000000)
SVM: 0.416667 (0.166667)
```




```
model = LinearDiscriminantAnalysis()
model.fit(X_train, y_train)
predictions = model.predict(X_test)
```


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

```
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()
X = scaler.fit_transform(X)


from keras.models import Sequential
from keras.layers import Dense, Dropout, Flatten, Conv1D, MaxPooling1D

model = Sequential()
model.add(Conv1D(filters=64, kernel_size=3, activation='relu', input_shape=(X.shape[1], 1)))
model.add(MaxPooling1D(pool_size=2))
model.add(Flatten())
model.add(Dense(128, activation='relu'))
model.add(Dropout(0.35))
model.add(Dense(len(y), activation='softmax'))
model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])


from keras.utils import to_categorical
from matplotlib import pyplot as plt

y_encoded = to_categorical(range(len(y)))
model.fit(X.reshape(X.shape[0], X.shape[1], 1), y_encoded, epochs=15)

plt.plot(history.history['loss'], label='train_loss')
plt.plot(history.history['accuracy'], label='train_acc')
plt.legend()
plt.show()


from keras.utils import to_categorical
from matplotlib import pyplot as plt

y_encoded = to_categorical(range(len(y)))

# Definir a métrica de acurácia no método compile()
model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])

# Treinar o modelo e obter o objeto History
history = model.fit(X.reshape(X.shape[0], X.shape[1], 1), y_encoded, epochs=15)

# Plotar as curvas de perda e acurácia
plt.plot(history.history['loss'], label='train_loss')
plt.plot(history.history['accuracy'], label='train_acc')
plt.legend()
plt.show()
```


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

```
Epoch 1/15
1/1 [==============================] - 6s 6s/step - loss: 0.5962 - accuracy: 0.7500
Epoch 2/15
1/1 [==============================] - 5s 5s/step - loss: 3.8145 - accuracy: 0.4000
Epoch 3/15
1/1 [==============================] - 4s 4s/step - loss: 1.4965 - accuracy: 0.6500
Epoch 4/15
1/1 [==============================] - 4s 4s/step - loss: 1.0393 - accuracy: 0.7500
Epoch 5/15
1/1 [==============================] - 6s 6s/step - loss: 1.8005 - accuracy: 0.6500
Epoch 6/15
1/1 [==============================] - 4s 4s/step - loss: 1.3753 - accuracy: 0.6500
Epoch 7/15
1/1 [==============================] - 4s 4s/step - loss: 2.3599 - accuracy: 0.6500
Epoch 8/15
1/1 [==============================] - 6s 6s/step - loss: 2.1100 - accuracy: 0.4000
Epoch 9/15
1/1 [==============================] - 4s 4s/step - loss: 1.4606 - accuracy: 0.6000
Epoch 10/15
1/1 [==============================] - 4s 4s/step - loss: 2.5181 - accuracy: 0.6000
Epoch 11/15
```

```
y_pred = model.predict(X.reshape(X.shape[0], X.shape[1], 1))
```

```
1/1 [==============================] - 1s 1s/step
```

```
yp_train = model.predict(X_train)
#yp_train = np.argmax(yp_train, axis = 1)

yp_test = model.predict(X_test)
#yp_test = np.argmax(yp_test, axis = 1)
```

```
1/1 [==============================] - 1s 652ms/step
1/1 [==============================] - 0s 263ms/step
```

```
from sklearn.metrics import accuracy_score

# Prever as saídas do modelo para o conjunto de teste
y_pred = model.predict(X_test)

# Transformar as saídas previstas em rótulos de classe
y_pred_classes = np.argmax(y_pred, axis=1)

# Calcular a acurácia média em todas as saídas do modelo
acc = accuracy_score(y_test, y_pred_classes)
acc
```

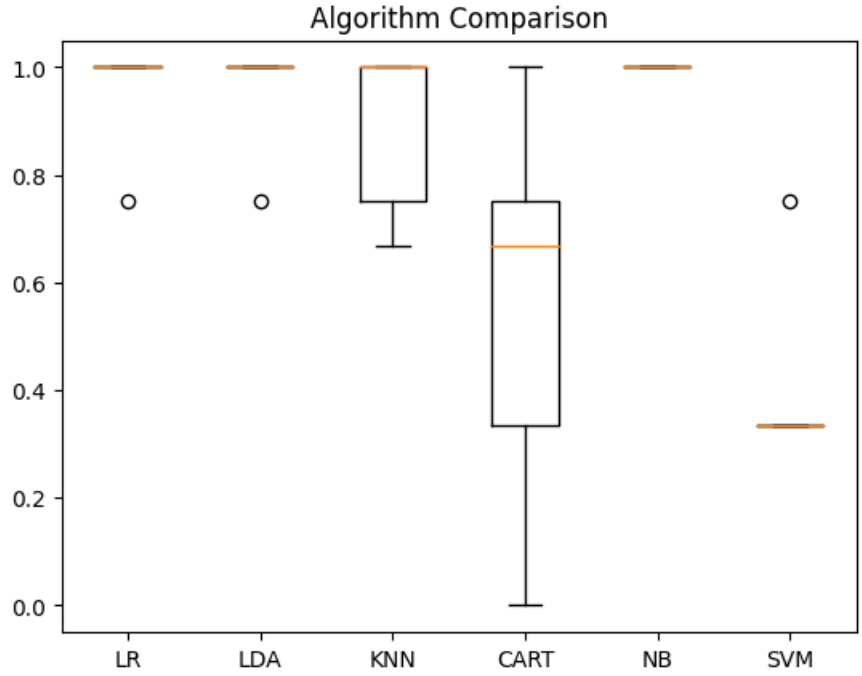
```
1/1 [==============================] - 0s 206ms/step
0.0
```

```
# comparar algoritmos - Machine Learning
from pandas import read_csv
from matplotlib import pyplot
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import StratifiedKFold
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC

# Verificar Algoritmos
modelo = []
modelo.append(('LR', LogisticRegression(solver='liblinear', multi_class='ovr')))
modelo.append(('LDA', LinearDiscriminantAnalysis()))
modelo.append(('KNN', KNeighborsClassifier()))
modelo.append(('CART', DecisionTreeClassifier()))
modelo.append(('NB', GaussianNB()))
modelo.append(('SVM', SVC(gamma='0.001')))

# avaliação de cada modelo
results = []
names = []
for name, model in modelo:
  kfold = StratifiedKFold(n_splits=5, random_state=42, shuffle=True)
  cv_results = cross_val_score(model, X_train, y_train, cv=kfold, scoring='accuracy')
  results.append(cv_results)
  names.append(name)
  print('%s: %f (%f)' % (name, cv_results.mean(), cv_results.std()))
```


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

```
[ ]  # Comparar algoritmos
     pyplot.boxplot(results, labels=names)
     pyplot.title('Algorithm Comparison')
     pyplot.show()
```

```
LR: 0.950000 (0.100000)
LDA: 0.950000 (0.100000)
KNN: 0.883333 (0.145297)
CART: 0.550000 (0.348010)
NB: 1.000000 (0.000000)
SVM: 0.416667 (0.166667)
```




```
model = LinearDiscriminantAnalysis()
model.fit(X_train, y_train)
predictions = model.predict(X_test)
```


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

APÊNDICE 3. MATRIZES DE DISTÂNCIAS DE SIMILARIDADE

Gene Sonda ID - 11715918_s_at

Matriz de distâncias de similaridade de Dice

	A01.5500_	A01_Breas	A03_Breas	A05_Breas	A05_Skin	A06_Skin	A07_Skin	A08_Breas	A09.3978_	A09_Breas	A10_Breas	A11_Breas	A12_Breas	A12_Skin	B02_Skin	B09_Skin	D07_Skin	G02_Skin	H06_Skin	H10_Skin
A01.5500_	0	6.0879572	0.0002851	0.0002298	0.2091769	0.4484447	0.4214435	2.7952149	0.1304956	0.0002689	0.0003365	0.0005969	0.0022793	0.1648222	0.4876548	0.4812602	0.4942698	0.4303249	0.4868918	0.4536698
A01_Breas	6.0879572	0	8.2567715	0.0005271	0.2038483	0.4433549	0.4162248	6.3286800	0.1257840	7.3924251	0.0006834	0.0002767	0.0015966	0.1597741	0.4827849	0.4763519	0.4894405	0.4251463	0.4820172	0.4486071
A03_Breas	0.0002851	8.2567715	0	0.0010264	0.1976754	0.4373948	0.4101185	0.0001346	0.1203679	2.3892961	0.0012403	5.6994636	0.0009539	0.1539483	0.4770765	0.4705995	0.4837788	0.4190856	0.4763035	0.4426780
A05_Breas	0.0002298	0.0005271	0.0010264	0	0.2195986	0.4582593	0.4315171	0.0004180	0.1398041	0.0009953	1.0149103	0.0015661	0.0039495	0.1747446	0.4970327	0.4907139	0.5035675	0.4403172	0.4962788	0.4634302
A05_Skin	0.2091769	0.2038483	0.1976754	0.2195986	0	0.1088673	0.0863725	0.2055635	0.0144353	0.1980065	0.2217988	0.1925757	0.1769887	0.0043860	0.1459015	0.1395128	0.1526521	0.0934979	0.1451321	0.1135055
A06_Skin	0.4484447	0.4433549	0.4373948	0.4582593	0.1088673	0	0.0016755	0.4449987	0.1843854	0.4377163	0.4603085	0.4324171	0.4168797	0.1489057	0.0037802	0.0026305	0.0051984	0.0007614	0.0036317	6.4836474
A07_Skin	0.4214435	0.4162248	0.4101185	0.4315171	0.0863725	0.0016755	0	0.4179098	0.1575464	0.4104478	0.4336221	0.4050228	0.3891411	0.1237102	0.0104238	0.0084605	0.0126830	0.0001784	0.0101782	0.0023976
A08_Breas	2.7952149	6.3286800	0.0001346	0.0004180	0.2055635	0.4449987	0.4179098	0	0.1272970	0.0001235	0.0005583	0.0003667	0.0018036	0.1613971	0.4843582	0.4779376	0.4910008	0.4268186	0.4835920	0.4502423
A09.3978_	0.1304956	0.1257840	0.1203679	0.1398041	0.0144353	0.1843854	0.1575464	0.1272970	0	0.1206572	0.1417846	0.1159284	0.1025662	0.0029659	0.2263539	0.2192745	0.2337714	0.1661836	0.2255044	0.1897766
A09_Breas	0.0002689	7.3924251	2.3892961	0.0009953	0.1980065	0.4377163	0.4104478	0.0001235	0.1206572	0	0.0012062	6.4613405	0.0009843	0.1542601	0.4773846	0.4709099	0.4840844	0.4194124	0.4766119	0.4429978
A10_Breas	0.0003365	0.0006834	0.0012403	1.0149103	0.2217988	0.4603085	0.4336221	0.0005583	0.1417846	0.0012062	0	0.0018279	0.0043579	0.1768474	0.4989888	0.4926860	0.5055065	0.4424047	0.4982368	0.4654678
A11_Breas	0.0005969	0.0002767	5.6994636	0.0015661	0.1925757	0.4324171	0.4050228	0.0003667	0.1159284	6.4613405	0.0018279	0	0.0005449	0.1491538	0.4723041	0.4657910	0.4790447	0.4140265	0.4715267	0.4377254
A12_Breas	0.0022793	0.0015966	0.0009539	0.0039495	0.1769887	0.4168797	0.3891411	0.0018036	0.1025662	0.0009843	0.0043579	0.0005449	0	0.1346092	0.4573777	0.4507565	0.4642335	0.3982507	0.4565873	0.4222620
A12_Skin	0.1648222	0.1597741	0.1539483	0.1747446	0.0043860	0.1489057	0.1237102	0.1613971	0.0029659	0.1542601	0.1768474	0.1491538	0.1346092	0	0.1891060	0.1822642	0.1962986	0.1317691	0.1882838	0.1540185
B02_Skin	0.4876548	0.4827849	0.4770765	0.4970327	0.1459015	0.0037802	0.0104238	0.4843582	0.2263539	0.4773846	0.4989888	0.4723041	0.4573777	0.1891060	0	0.0001047	0.0001139	0.0079013	1.5024638	0.0028585
B09_Skin	0.4812602	0.4763519	0.4705995	0.4907139	0.1395128	0.0026305	0.0084605	0.4779376	0.2192745	0.4709099	0.4926860	0.4657910	0.4507565	0.1822642	0.0001047	0	0.0004371	0.0062004	8.1179275	0.0018713
D07_Skin	0.4942698	0.4894405	0.4837788	0.5035675	0.1526521	0.0051984	0.0126830	0.4910008	0.2337714	0.4840844	0.5055065	0.4790447	0.4642335	0.1962986	0.0001139	0.0004371	0	0.0098896	0.0001416	0.0041081
G02_Skin	0.4303249	0.4251463	0.4190856	0.4403172	0.0934979	0.0007614	0.0001784	0.4268186	0.1661836	0.4194124	0.4424047	0.4140265	0.3982507	0.1317691	0.0079013	0.0062004	0.0098896	0	0.0076869	0.0012699
H06_Skin	0.4868918	0.4820172	0.4763035	0.4962788	0.1451321	0.0036317	0.0101782	0.4835920	0.2255044	0.4766119	0.4982368	0.4715267	0.4565873	0.1882838	1.5024638	8.1179275	0.0001416	0.0076869	0	0.0027293
H10_Skin	0.4536698	0.4486071	0.4426780	0.4634302	0.1135055	6.4836474	0.0023976	0.4502423	0.1897766	0.4429978	0.4654678	0.4377254	0.4222620	0.1540185	0.0028585	0.0018713	0.0041081	0.0012699	0.0027293	0

Matriz de distâncias de similaridade de Jaccard

	A01.5500_	A01_Breas	A03_Breas	A05_Breas	A05_Skin	A06_Skin	A07_Skin	A08_Breas	A09.3978_	A09_Breas	A10_Breas	A11_Breas	A12_Breas	A12_Skin	B02_Skin	B09_Skin	D07_Skin	G02_Skin	H06_Skin	H10_Skin
A01.5500_	0	0.0001217	0.0005702	0.0004595	0.3459823	0.6192086	0.5929796	5.5902736	0.2308644	0.0005377	0.0006728	0.0011932	0.0045483	0.2829998	0.6556021	0.6497983	0.6615536	0.6017163	0.6549122	0.6241717
A01_Breas	0.0001217	0	0.0001651	0.0010537	0.3386611	0.6143394	0.5877948	1.2657279	0.2234602	0.0001478	0.0013659	0.0005532	0.0031882	0.2755262	0.6511867	0.6453094	0.6572139	0.5966353	0.6504880	0.6193634
A03_Breas	0.0005702	0.0001651	0	0.0020507	0.3300985	0.6085938	0.5816795	0.0002691	0.2148721	4.7785912	0.0024776	0.0001139	0.0019060	0.2668201	0.6459740	0.6400104	0.6520902	0.5906418	0.6452650	0.6136893
A05_Breas	0.0004595	0.0010537	0.0020507	0	0.3601162	0.6285018	0.6028808	0.0008356	0.2453125	0.0019887	2.0298000	0.0031274	0.0078681	0.2975023	0.6640238	0.6583609	0.6698302	0.6114170	0.6633507	0.6333478
A05_Skin	0.3459823	0.3386611	0.3300985	0.3601162	0	0.1963577	0.1590108	0.3410248	0.0284599	0.3305600	0.3630693	0.3229577	0.3007484	0.0087338	0.2546494	0.2448639	0.2648711	0.1710071	0.2534766	0.2038706
A06_Skin	0.6192086	0.6143394	0.6085938	0.6285018	0.1963577	0	0.0033455	0.6159157	0.3113605	0.6089050	0.6304264	0.6037587	0.5884475	0.2592131	0.0075321	0.0052472	0.0103430	0.0015216	0.0072372	0.0001296
A07_Skin	0.5929796	0.5877948	0.5816795	0.6028808	0.1590108	0.0033455	0	0.5894731	0.2722075	0.5820106	0.6049323	0.5765356	0.5602615	0.2201818	0.0206325	0.0167792	0.0250484	0.0003568	0.0201513	0.0047838
A08_Breas	5.5902736	1.2657279	0.0002691	0.0008356	0.3410248	0.6159157	0.5894731	0	0.2258447	0.0002469	0.0011160	0.0007331	0.0036007	0.2779362	0.6526164	0.6467629	0.6586191	0.5982801	0.6519205	0.6209201
A09.3978_	0.2308644	0.2234602	0.2148721	0.2453125	0.0284599	0.3113605	0.2722075	0.2258447	0	0.2153330	0.2483562	0.2077704	0.1860500	0.0059143	0.3691494	0.3596803	0.3789541	0.2850042	0.3680189	0.3190121
A09_Breas	0.0005377	0.0001478	4.7785912	0.0019887	0.3305600	0.6089050	0.5820106	0.0002469	0.2153330	0	0.0024095	0.0001292	0.0019667	0.2672883	0.6462563	0.6402974	0.6523677	0.5909663	0.6455479	0.6139965
A10_Breas	0.0006728	0.0013659	0.0024776	2.0298000	0.3630693	0.6304264	0.6049323	0.0011160	0.2483562	0.0024095	0	0.0036491	0.0086781	0.3005444	0.6657672	0.6601335	0.6715434	0.6134266	0.6650975	0.6352481
A11_Breas	0.0011932	0.0005532	0.0001139	0.0031274	0.3229577	0.6037587	0.5765356	0.0007331	0.2077704	0.0001292	0.0036491	0	0.0010892	0.2595889	0.6415850	0.6355490	0.6477758	0.5855993	0.6408674	0.6089138
A12_Breas	0.0045483	0.0031882	0.0019060	0.0078681	0.3007484	0.5884475	0.5602615	0.0036007	0.1860500	0.0019667	0.0086781	0.0010892	0	0.2372785	0.6276722	0.6214089	0.6340976	0.5696414	0.6269274	0.5937894
A12_Skin	0.2829998	0.2755262	0.2668201	0.2975023	0.0087338	0.2592131	0.2201818	0.2779362	0.0059143	0.2672883	0.3005444	0.2595889	0.2372785	0	0.3180641	0.3083307	0.3281766	0.2328550	0.3169003	0.2669256
B02_Skin	0.6556021	0.6511867	0.6459740	0.6640238	0.2546494	0.0075321	0.0206325	0.6526164	0.3691494	0.6462563	0.6657672	0.6415850	0.6276722	0.3180641	0	0.0002095	0.0002278	0.0156787	3.0049232	0.0057007
B09_Skin	0.6497983	0.6453094	0.6400104	0.6583609	0.2448639	0.0052472	0.0167792	0.6467629	0.3596803	0.6402974	0.6601335	0.6355490	0.6214089	0.3083307	0.0002095	0	0.0008738	0.0123245	0.0001623	0.0037357
D07_Skin	0.6615536	0.6572139	0.6520902	0.6698302	0.2648711	0.0103430	0.0250484	0.6586191	0.3789541	0.6523677	0.6715434	0.6477758	0.6340976	0.3281766	0.0002278	0.0008738	0	0.0195856	0.0002832	0.0081825
G02_Skin	0.6017163	0.5966353	0.5906418	0.6114170	0.1710071	0.0015216	0.0003568	0.5982801	0.2850042	0.5909663	0.6134266	0.5855993	0.5696414	0.2328550	0.0156787	0.0123245	0.0195856	0	0.0152567	0.0025369
H06_Skin	0.6549122	0.6504880	0.6452650	0.6633507	0.2534766	0.0072372	0.0201513	0.6519205	0.3680189	0.6455479	0.6650975	0.6408674	0.6269274	0.3169003	3.0049232	0.0001623	0.0002832	0.0152567	0	0.0054439
H10_Skin	0.6241717	0.6193634	0.6136893	0.6333478	0.2038706	0.0001296	0.0047838	0.6209201	0.3190121	0.6139965	0.6352481	0.6089138	0.5937894	0.2669256	0.0057007	0.0037357	0.0081825	0.0025365	0.0054439	0

Matriz de distâncias de similaridade de Sorensen

	A01.5500_	A01_Breas	A03_Breas	A05_Breas	A05_Skin	A06_Skin	A07_Skin	A08_Breas	A09.3978_	A09_Breas	A10_Breas	A11_Breas	A12_Breas	A12_Skin	B02_Skin	B09_Skin	D07_Skin	G02_Skin	H06_Skin	H10_Skin
A01.5500_	0	0.0055173	0.0119422	0.0107204	0.3417673	0.5376143	0.5167014	0.0037384	0.2642011	0.0115966	0.0129728	0.0172795	0.0337782	0.2996876	0.5678463	0.5629221	0.5729391	0.5235924	0.5672588	0.5416501
A01_Breas	0.0055173	0	0.0064253	0.0162368	0.3368853	0.5336800	0.5126455	0.0017788	0.2590615	0.0060797	0.0184888	0.0117633	0.0282661	0.2946575	0.5640963	0.5591414	0.5692211	0.5195760	0.5635051	0.5377398
A03_Breas	0.0119422	0.0064253	0	0.0226598	0.3311767	0.5290688	0.5078931	0.0082041	0.2530573	0.0034956	0.0249112	0.0053383	0.0218447	0.2887789	0.5596996	0.5547089	0.5648617	0.5148695	0.5591041	0.5331565
A05_Breas	0.0107204	0.0162368	0.0226598	0	0.3512010	0.5451926	0.5245164	0.0144583	0.2741452	0.0223144	0.0022526	0.0279948	0.0444826	0.3094140	0.5750660	0.5702015	0.5800965	0.5313304	0.5744856	0.5491816
A05_Skin	0.3417673	0.3368853	0.3311767	0.3512010	0	0.2399318	0.2124511	0.3384613	0.0852652	0.3314844	0.3531743	0.3264155	0.3115861	0.0468814	0.2805198	0.2738380	0.2874598	0.2214534	0.2797211	0.2452905
A06_Skin	0.5376143	0.5336800	0.5290688	0.5451926	0.2399318	0	0.0289566	0.5349510	0.3186776	0.5293177	0.5467738	0.5252139	0.5131548	0.2836229	0.0435169	0.0362905	0.0510488	0.0195153	0.0426518	0.0056937
A07_Skin	0.5167014	0.5126455	0.5078931	0.5245164	0.2124511	0.0289566	0	0.5139557	0.2924193	0.5081495	0.5261474	0.5039210	0.4915015	0.2567751	0.0723824	0.0651787	0.0798874	0.0094467	0.0715201	0.0346447
A08_Breas	0.0037384	0.0017788	0.0082041	0.0144583	0.3384613	0.5349510	0.5139557	0	0.2607202	0.0078585	0.0167105	0.0135419	0.0300435	0.2962811	0.5653079	0.5603629	0.5704224	0.5208735	0.5647179	0.5390030
A09.3978_	0.2642011	0.2590615	0.2530573	0.2741452	0.0852652	0.3186776	0.2924193	0.2607202	0	0.2533808	0.2762273	0.2480540	0.2324978	0.0385378	0.3572403	0.3509098	0.3638080	0.3010344	0.3564840	0.3237838
A09_Breas	0.0115966	0.0060797	0.0034956	0.0223144	0.3314844	0.5293177	0.5081495	0.0078585	0.2533808	0	0.0245658	0.0056839	0.0221902	0.2890957	0.5599369	0.5549481	0.5650970	0.5151235	0.5593416	0.5334039
A10_Breas	0.0129728	0.0184888	0.0249112	0.0022526	0.3531743	0.5467738	0.5261474	0.0167105	0.2762273	0.0245658	0	0.0302456	0.0467306	0.3114496	0.5765718	0.5717198	0.5815892	0.5329452	0.5759929	0.5507529
A11_Breas	0.0172795	0.0117633	0.0053383	0.0279948	0.3264155	0.5252139	0.5039210	0.0135419	0.2480540	0.0056839	0.0302456	0	0.0165083	0.2838782	0.5560225	0.5510022	0.5612157	0.5109355	0.5554235	0.5293247
A12_Breas	0.0337782	0.0282661	0.0218447	0.0444826	0.3115861	0.5131548	0.4915015	0.0300435	0.2324978	0.0221902	0.0467306	0.0165083	0	0.2686287	0.5445123	0.5394003	0.5498011	0.4986330	0.5439022	0.5173370
A12_Skin	0.2996876	0.2946575	0.2887789	0.3094140	0.0468814	0.2836229	0.2567751	0.2962811	0.0385378	0.2890957	0.3114496	0.2838782	0.2686287	0	0.3231514	0.3166542	0.3298954	0.2655776	0.3223750	0.2888503
B02_Skin	0.5678463	0.5640963	0.5596996	0.5750660	0.2805198	0.0435169	0.0723824	0.5653079	0.3572403	0.5599369	0.5765718	0.5560225	0.5445123	0.3231514	0	0.0072378	0.0075486	0.0629787	0.0008667	0.0378325
B09_Skin	0.5629221	0.5591414	0.5547089	0.5702015	0.2738380	0.0362905	0.0651787	0.5603629	0.3509098	0.5549481	0.5717198	0.5510022	0.5394003	0.3166542	0.0072378	0	0.0147857	0.0557663	0.0063711	0.0306030
D07_Skin	0.5729391	0.5692211	0.5648617	0.5800965	0.2874598	0.0510488	0.0798874	0.5704224	0.3638080	0.5650970	0.5815892	0.5612157	0.5498011	0.3298954	0.0075486	0.0147857	0	0.0704939	0.0084153	0.0453682
G02_Skin	0.5235924	0.5195760	0.5148695	0.5313304	0.2214534	0.0195153	0.0094467	0.5208735	0.3010344	0.5151235	0.5329452	0.5109355	0.4986330	0.2655776	0.0629787	0.0557663	0.0704939	0	0.0621154	0.0252062
H06_Skin	0.5672588	0.5635051	0.5591041	0.5744856	0.2797211	0.0426518	0.0715201	0.5647179	0.3564840	0.5593416	0.5759929	0.5554235	0.5439022	0.3223750	0.0008667	0.0063711	0.0084153	0.0621154	0	0.0369670
H10_Skin	0.5416501	0.5377398	0.5331565	0.5491816	0.2452905	0.0056937	0.0346447	0.5390030	0.3237838	0.5334039	0.5507529	0.5293247	0.5173370	0.2888503	0.0378325	0.0306030	0.0453682	0.0252062	0.0369670	0

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Matriz de distâncias de similaridade de Czekanowski

	A01.5500_	A01_Breas	A03_Breas	A05_Breas	A05_Skin	A06_Skin	A07_Skin	A08_Breas	A09.3978_	A09_Breas	A10_Breas	A11_Breas	A12_Breas	A12_Skin	B02_Skin	B09_Skin	D07_Skin	G02_Skin	H06_Skin	H10_Skin
A01.5500_	0	0.0055173	0.0119422	0.0107204	0.3417673	0.5376143	0.5167014	0.0037384	0.2642011	0.0115966	0.0129728	0.0172795	0.0337782	0.2996876	0.5678463	0.5629221	0.5729391	0.5235924	0.5672588	0.5416501
A01_Breas	0.0055173	0	0.0064253	0.0162368	0.3368853	0.5336800	0.5126455	0.0017788	0.2590615	0.0060797	0.0184888	0.0117633	0.0282661	0.2946575	0.5640963	0.5591414	0.5692211	0.5195760	0.5635051	0.5377398
A03_Breas	0.0119422	0.0064253	0	0.0226598	0.3311767	0.5290688	0.5078931	0.0082041	0.2530573	0.0003456	0.0249112	0.0053383	0.0218447	0.2887789	0.5596996	0.5547089	0.5648617	0.5148695	0.5591041	0.5331565
A05_Breas	0.0107204	0.0162368	0.0226598	0	0.3512010	0.5451926	0.5245164	0.0144583	0.2741452	0.0223144	0.0022526	0.0279948	0.0444826	0.3094140	0.5750660	0.5702015	0.5800965	0.5313304	0.5744856	0.5491816
A05_Skin	0.3417673	0.3368853	0.3311767	0.3512010	0	0.2399318	0.2124511	0.3384613	0.0852652	0.3314844	0.3531743	0.3264155	0.3115861	0.0468814	0.2805198	0.2738380	0.2874598	0.2214534	0.2797211	0.2452905
A06_Skin	0.5376143	0.5336800	0.5290688	0.5451926	0.2399318	0	0.0289566	0.5349510	0.3186776	0.5293177	0.5467738	0.5252139	0.5131548	0.2836229	0.0435169	0.0362905	0.0510488	0.0195153	0.0426518	0.0056937
A07_Skin	0.5167014	0.5126455	0.5078931	0.5245164	0.2124511	0.0289566	0	0.5139557	0.2924193	0.5081495	0.5261474	0.5039210	0.4915015	0.2567751	0.0723824	0.0651787	0.0798874	0.0094467	0.0715201	0.0346447
A08_Breas	0.0037384	0.0017788	0.0082041	0.0144583	0.3384613	0.5349510	0.5139557	0	0.2607202	0.0078585	0.0167105	0.0135419	0.0300435	0.2962811	0.5653079	0.5603629	0.5704224	0.5208735	0.5647179	0.5390030
A09.3978_	0.2642011	0.2590615	0.2530573	0.2741452	0.0852652	0.3186776	0.2924193	0.2607202	0	0.2533808	0.2762273	0.2480540	0.2324978	0.0385378	0.3572403	0.3509098	0.3638080	0.3010344	0.3564840	0.3237838
A09_Breas	0.0115966	0.0060797	0.0003456	0.0223144	0.3314844	0.5293177	0.5081495	0.0078585	0.2533808	0	0.0245658	0.0056839	0.0221902	0.2890957	0.5599369	0.5549481	0.5650970	0.5151235	0.5593416	0.5334039
A10_Breas	0.0129728	0.0184888	0.0249112	0.0022526	0.3531743	0.5467738	0.5261474	0.0167105	0.2762273	0.0245658	0	0.0302456	0.0467306	0.3114496	0.5765718	0.5717198	0.5815892	0.5329452	0.5759929	0.5507529
A11_Breas	0.0172795	0.0117633	0.0053383	0.0279948	0.3264155	0.5252139	0.5039210	0.0135419	0.2480540	0.0056839	0.0302456	0	0.0165083	0.2838782	0.5560225	0.5510022	0.5612157	0.5109355	0.5554235	0.5293247
A12_Breas	0.0337782	0.0282661	0.0218447	0.0444826	0.3115861	0.5131548	0.4915015	0.0300435	0.2324978	0.0221902	0.0467306	0.0165083	0	0.2686287	0.5445123	0.5394003	0.5498011	0.4986330	0.5439022	0.5173370
A12_Skin	0.2996876	0.2946575	0.2887789	0.3094140	0.0468814	0.2836229	0.2567751	0.2962811	0.0385378	0.2890957	0.3114496	0.2838782	0.2686287	0	0.3231514	0.3166542	0.3298954	0.2655776	0.3223750	0.2888503
B02_Skin	0.5678463	0.5640963	0.5596996	0.5750660	0.2805198	0.0435169	0.0723824	0.5653079	0.3572403	0.5599369	0.5765718	0.5560225	0.5445123	0.3231514	0	0.0072378	0.0075486	0.0629787	0.0088667	0.0378325
B09_Skin	0.5629221	0.5591414	0.5547089	0.5702015	0.2738380	0.0362905	0.0651787	0.5603629	0.3509098	0.5549481	0.5717198	0.5510022	0.5394003	0.3166542	0.0072378	0	0.0147857	0.0557663	0.0063711	0.0306030
D07_Skin	0.5729391	0.5692211	0.5648617	0.5800965	0.2874598	0.0510488	0.0798874	0.5704224	0.3638080	0.5650970	0.5815892	0.5612157	0.5498011	0.3298954	0.0075486	0.0147857	0	0.0704939	0.0084153	0.0453682
G02_Skin	0.5235924	0.5195760	0.5148695	0.5313304	0.2214534	0.0195153	0.0094467	0.5208735	0.3010344	0.5151235	0.5329452	0.5109355	0.4986330	0.2655776	0.0629787	0.0557663	0.0704939	0	0.0621154	0.0252062
H06_Skin	0.5672588	0.5635051	0.5591041	0.5744856	0.2797211	0.0426518	0.0715201	0.5647179	0.3564840	0.5593416	0.5759929	0.5554235	0.5439022	0.3223750	0.0008667	0.0063711	0.0084153	0.0621154	0	0.0369670
H10_Skin	0.5416501	0.5377398	0.5331565	0.5491816	0.2452905	0.0056937	0.0346447	0.5390030	0.3237838	0.5334039	0.5507529	0.5293247	0.5173370	0.2888503	0.0378325	0.0306030	0.0453682	0.0252062	0.0369670	0

Matriz de distâncias de similaridade de Minkowski

	A01.5500_	A01_Breas	A03_Breas	A05_Breas	A05_Skin	A06_Skin	A07_Skin	A08_Breas	A09.3978_	A09_Breas	A10_Breas	A11_Breas	A12_Breas	A12_Skin	B02_Skin	B09_Skin	D07_Skin	G02_Skin	H06_Skin	H10_Skin
A01.5500_	0	0.1342641	0.2887709	0.2651661	6.2326918	8.5555063	8.3360795	0.0911375	5.1137662	0.2805101	0.3216096	0.4156368	0.7995244	5.6422481	8.8623649	8.8131926	8.9128964	8.4090481	8.8565140	8.5971659
A01_Breas	0.1342641	0	0.1545067	0.3994303	6.0984276	8.4212421	8.2018154	0.0431266	4.9795020	0.1462460	0.4558737	0.2813726	0.6652602	5.5079839	8.7281007	8.6789285	8.7786323	8.2747839	8.7222498	8.4629017
A03_Breas	0.2887709	0.1545067	0	0.5539370	5.9439209	8.2667354	8.0473086	0.1976333	4.8249952	0.0082607	0.6103805	0.1268658	0.5107535	5.3534772	8.5735939	8.5244217	8.6241255	8.1202772	8.5677430	8.3083950
A05_Breas	0.2651661	0.3994303	0.5539370	0	6.4978580	8.8206725	8.6012457	0.3563037	5.3789323	0.5456763	0.0564434	0.6808029	1.0646906	5.9074143	9.1275310	9.0783588	9.1780626	8.6742142	9.1216801	8.8623321
A05_Skin	6.2326918	6.0984276	5.9439209	6.4978580	0	2.3228144	2.1033877	6.1415543	1.1189256	5.9521816	6.5543014	5.8170550	5.4331673	0.5904437	2.6296730	2.5805008	2.6802046	2.1763562	2.6238221	2.3644741
A06_Skin	8.5555063	8.4212421	8.2667354	8.8206725	2.3228144	0	0.2194267	8.4643688	3.4417401	8.2749961	8.8771159	8.1398695	7.7559818	2.9132582	0.3068585	0.2576863	0.3573901	0.1464582	0.3010076	0.0416596
A07_Skin	8.3360795	8.2018154	8.0473086	8.6012457	2.1033877	0.2194267	0	8.2449420	3.2223133	8.0555694	8.6576892	7.9204427	7.5365551	2.6938314	0.5262853	0.4771130	0.5768168	0.0729685	0.5204344	0.2610863
A08_Breas	0.0911375	0.0431266	0.1976333	0.3563037	6.1415543	8.4643688	8.2449420	0	5.0226286	0.1893726	0.4127471	0.3244992	0.7083869	5.5511105	8.7712273	8.7220551	8.8217589	8.3179105	8.7653764	8.5060284
A09.3978_	5.1137662	4.9795020	4.8249952	5.3789323	1.1189256	3.4417401	3.2223133	5.0226286	0	4.8332560	5.4353758	4.6981294	4.3142417	0.5284819	3.7485986	3.6994264	3.7991302	3.2952819	3.7427478	3.4833997
A09_Breas	0.2805101	0.1462460	0.0082607	0.5456763	5.9521816	8.2749961	8.0555694	0.1893726	4.8332560	0	0.6021198	0.1351266	0.5190142	5.3617379	8.5818547	8.5326824	8.6323862	8.1285379	8.5760038	8.3166557
A10_Breas	0.3216096	0.4558737	0.6103805	0.0564434	6.5543014	8.8771159	8.6576892	0.4127471	5.4353758	0.6021198	0	0.7372464	1.1211340	5.9638577	9.1839745	9.1348023	9.2345060	8.7306577	9.1781236	8.9187755
A11_Breas	0.4156368	0.2813726	0.1268658	0.6808029	5.8170550	8.1398695	7.9204427	0.3244992	4.6981294	0.1351266	0.7372464	0	0.3838876	5.2266113	8.4467281	8.3975558	8.4972596	7.9934113	8.4408772	8.1815293
A12_Breas	0.7995244	0.6652602	0.5107535	1.0646906	5.4331673	7.7559818	7.5365551	0.7083869	4.3142417	0.5190142	1.1211340	0.3838876	0	4.8427236	8.0628404	8.0136682	8.1133720	7.6095236	8.0569895	7.7976414
A12_Skin	5.6422481	5.5079839	5.3534772	5.9074143	0.5904437	2.9132582	2.6938314	5.5511105	0.5284819	5.3617379	5.9638577	5.2266113	4.8427236	0	3.2201167	3.1709445	3.2706483	2.7667999	3.2142658	2.9549178
B02_Skin	8.8623649	8.7281007	8.5735939	9.1275310	2.6296730	0.3068585	0.5262853	8.7712273	3.7485986	8.5818547	9.1839745	8.4467281	8.0628404	3.2201167	0	0.0491722	0.0505315	0.4533167	0.0058508	0.2651989
B09_Skin	8.8131926	8.6789285	8.5244217	9.0783588	2.5805008	0.2576863	0.4771130	8.7220551	3.6994264	8.5326824	9.1348023	8.3975558	8.0136682	3.1709445	0.0491722	0	0.0997037	0.4041445	0.0433213	0.2160267
D07_Skin	8.9128964	8.7786323	8.6241255	9.1780626	2.6802046	0.3573901	0.5768168	8.8217589	3.7991302	8.6323862	9.2345060	8.4972596	8.1133720	3.2706483	0.0505315	0.0997037	0	0.5038483	0.0563824	0.3157305
G02_Skin	8.4090481	8.2747839	8.1202772	8.6742142	2.1763562	0.1464582	0.0729685	8.3179105	3.2952819	8.1285379	8.7306577	7.9934113	7.6095236	2.7667999	0.4533167	0.4041445	0.5038483	0	0.4474658	0.1881178
H06_Skin	8.8565140	8.7222498	8.5677430	9.1216801	2.6238221	0.3010076	0.5204344	8.7653764	3.7427478	8.5760038	9.1781236	8.4408772	8.0569895	3.2142658	0.0058508	0.0433213	0.0563824	0.4474658	0	0.2593480
H10_Skin	8.5971659	8.4629017	8.3083950	8.8623321	2.3644741	0.0416596	0.2610863	8.5060284	3.4833997	8.3166557	8.9187755	8.1815291	7.7976414	2.9549178	0.2651989	0.2160267	0.3157305	0.1881178	0.2593480	0

Matriz de distâncias de similaridade de Pearson

	A01.5500_	A01_Breas	A03_Breas	A05_Breas	A05_Skin	A06_Skin	A07_Skin	A08_Breas	A09.3978_	A09_Breas	A10_Breas	A11_Breas	A12_Breas	A12_Skin	B02_Skin	B09_Skin	D07_Skin	G02_Skin	H06_Skin	H10_Skin
A01.5500_	0	0.0014897	0.0069805	0.0056251	6.4722732	19.894921	17.824444	0.0006839	3.6723711	0.0065823	0.0082375	0.0146165	0.0559012	4.8290235	23.290150	22.701406	23.914846	18.483811	23.219123	20.319220
A01_Breas	0.0014897	0	0.0019983	0.0127637	6.1964262	19.275387	17.254893	0.0001531	3.4820632	0.0017891	0.0165511	0.0066985	0.0387027	4.6019329	22.589807	22.014987	23.199765	17.898275	22.520459	19.689514
A03_Breas	0.0069805	0.0019983	0	0.0245480	5.8864244	18.574574	16.610917	0.0032164	3.2693289	5.7084599	0.0296715	0.0013617	0.0228129	4.3473726	21.797107	21.238120	22.390304	17.236122	21.729666	18.977136
A05_Breas	0.0056251	0.0127637	0.0245480	0	7.0347061	21.147264	18.976451	0.0104543	4.0630951	0.0249087	0.0002537	0.0392157	0.0991300	5.2935840	24.704704	24.088009	25.358987	19.667907	24.630309	21.591975
A05_Skin	6.4722732	6.1964262	5.8864244	7.0347061	0	1.4664925	1.1348304	3.1060720	0.1758195	2.9636933	3.4213052	2.8630198	2.5814543	0.0528824	2.0505791	1.9462302	2.1625480	1.2381060	2.0379287	1.5369705
A06_Skin	19.894921	19.275387	18.574574	21.147264	1.4664925	0	0.0123501	5.8998923	1.6634930	5.7281781	6.2759977	5.6059975	5.2605583	1.2873982	0.0279221	0.0194074	0.0384516	0.0056069	0.0268210	0.0004771
A07_Skin	17.824444	17.254893	16.610917	18.976451	1.1348304	0.0123501	0	5.5979645	1.4581438	5.4284194	5.9695689	5.3078291	4.9671129	1.1007679	0.0821325	0.0665317	0.1001626	0.0013917	0.0801774	0.0187397
A08_Breas	0.0006839	0.0001531	0.0032164	0.0104543	3.1060720	5.8998923	5.5979645	0	3.5426395	0.0029999	0.0135677	0.0089093	0.0438833	4.6742798	22.813597	22.234321	23.428271	18.085326	22.743711	19.890700
A09.3978_	3.6723711	3.4820632	3.2693289	4.0630951	0.1758195	1.6634930	1.4581438	3.5426395	0	1.9541617	2.3528711	1.8675316	1.6276732	0.0423657	4.1668793	3.9999542	4.3450846	2.8384615	4.1466876	3.3358233
A09_Breas	0.0065823	0.0017891	5.7084599	0.0249087	2.9636933	5.7281781	5.4284194	0.0029999	1.9541617	0	0.0288738	0.0015448	0.0235568	4.3607995	21.839130	21.279302	22.433219	17.271208	21.771589	19.014892
A10_Breas	0.0082375	0.0165511	0.0296715	0.0002537	3.4213052	6.2759977	5.9695689	0.0135677	2.3528711	0.0288738	0	0.0459878	0.1099191	5.3952243	25.011191	24.388468	25.671852	19.924700	24.936069	21.867890
A11_Breas	0.0146165	0.0066985	0.0013617	0.0392157	2.8630198	5.6059975	5.3078291	0.0089093	1.8675316	0.0015448	0.0459878	0	0.0128874	4.1437673	21.156803	20.610666	21.736401	16.701758	21.090912	18.402014
A12_Breas	0.0559012	0.0387027	0.0228129	0.0991300	2.5814543	5.2605583	4.9671129	0.0438833	1.6276732	0.0235568	0.1099191	0.0128874	0	3.5574133	19.277431	18.769337	19.816759	15.136058	19.216124	16.715637
A12_Skin	4.8290235	4.6019329	4.3473726	5.2935840	0.0528824	1.2873982	1.1007679	4.6742798	0.0423657	4.3607995	5.3952243	4.1437673	3.5574133	0	3.0747957	2.9387559	3.2203091	2.0010290	3.0583260	2.4004199
B02_Skin	23.290150	22.589807	21.797107	24.704704	2.0505791	0.0279221	0.0821325	22.813597	4.1668793	21.839130	25.011191	21.156803	19.277431	3.0747957	0	0.0007066	0.0007686	0.0537157	1.0133551	0.0193348
B09_Skin	22.701406	22.014987	21.238120	24.088009	1.9462302	0.0194074	0.0665317	22.234321	3.9999542	21.279302	24.388468	20.610666	18.769337	2.9387559	0.0007066	0	0.0029926	0.0426944	0.0005555	0.0128295
D07_Skin	23.914846	23.199765	22.390304	25.358987	2.1625480	0.0384516	0.1001626	23.428271	4.3450846	22.433219	25.671852	21.736401	19.816759	3.2203091	0.0007686	0.0029926	0	0.0663586	0.0009410	0.0274049
G02_Skin	18.483811	17.898275	17.236122	19.667907	1.2381060	0.0056069	0.0013917	18.085326	2.8384615	17.271208	19.924700	16.701758	15.136058	2.0010290	0.0537157	0.0426944	0.0663586	0	0.0592707	0.0097287
H06_Skin	23.219123	22.520459	21.729666	24.630309	2.0379287	0.0268210	0.0801774	22.743711	4.1466876	21.771589	24.936069	21.090912	19.216124	3.0583260	1.0133551	0.0005555	0.0009410	0.0592707	0	0.0184910
H10_Skin	20.319220	19.689514	18.977136	21.591979	1.5369705	0.0004771	0.0187397	19.890700	3.3358233	19.014892	21.867890	18.402014	16.715637	2.4004199	0.0193348	0.0128295	0.0274049	0.0097287	0.0184910	0

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Matriz de distâncias de similaridade de Intersection

	A01.5500_	A01_Breas	A03_Breas	A05_Breas	A05_Skin	A06_Skin	A07_Skin	A08_Breas	A09.3978_	A09_Breas	A10_Breas	A11_Breas	A12_Breas	A12_Skin	B02_Skin	B09_Skin	D07_Skin	G02_Skin	H06_Skin	H10_Skin
A01.5500_	12.234670	12.100406	11.945899	12.234670	6.0019789	3.6791644	3.8985912	12.143533	7.1209046	11.954160	12.234670	11.819034	11.435146	6.5924227	3.3723059	3.4214781	3.3217743	3.8256227	3.3781568	3.6375048
A01_Breas	12.100406	12.100406	11.945899	12.100406	6.0019789	3.6791644	3.8985912	12.100406	7.1209046	11.954160	12.100406	11.819034	11.435146	6.5924227	3.3723059	3.4214781	3.3217743	3.8256227	3.3781568	3.6375048
A03_Breas	11.945899	11.945899	11.945899	11.945899	6.0019789	3.6791644	3.8985912	11.945899	7.1209046	11.945899	11.945899	11.819034	11.435146	6.5924227	3.3723059	3.4214781	3.3217743	3.8256227	3.3781568	3.6375048
A05_Breas	12.234670	12.100406	11.945899	12.499837	6.0019789	3.6791644	3.8985912	12.143533	7.1209046	11.954160	12.499837	11.819034	11.435146	6.5924227	3.3723059	3.4214781	3.3217743	3.8256227	3.3781568	3.6375048
A05_Skin	6.0019789	6.0019789	6.0019789	6.0019789	6.0019789	3.6791644	3.8985912	6.0019789	6.0019789	6.0019789	6.0019789	6.0019789	6.0019789	3.3723059	3.4214781	3.3217743	3.8256227	3.3781568	3.6375048	
A06_Skin	3.6791644	3.6791644	3.6791644	3.6791644	3.6791644	3.6791644	3.6791644	3.6791644	3.6791644	3.6791644	3.6791644	3.6791644	3.6791644	3.3723059	3.4214781	3.3217743	3.6791644	3.3781568	3.6375048	
A07_Skin	3.8985912	3.8985912	3.8985912	3.8985912	3.8985912	3.6791644	3.8985912	3.8985912	3.8985912	3.8985912	3.8985912	3.8985912	3.8985912	3.3723059	3.4214781	3.3217743	3.8256227	3.3781568	3.6375048	
A08_Breas	12.143533	12.100406	11.945899	12.143533	6.0019789	3.6791644	3.8985912	12.143533	7.1209046	11.954160	12.143533	11.819034	11.435146	6.5924227	3.3723059	3.4214781	3.3217743	3.8256227	3.3781568	3.6375048
A09.3978_	7.1209046	7.1209046	7.1209046	7.1209046	6.0019789	3.6791644	3.8985912	7.1209046	7.1209046	7.1209046	7.1209046	7.1209046	7.1209046	6.5924227	3.3723059	3.4214781	3.3217743	3.8256227	3.3781568	3.6375048
A09_Breas	11.954160	11.954160	11.945899	11.954160	6.0019789	3.6791644	3.8985912	11.954160	7.1209046	11.954160	11.954160	11.819034	11.435146	6.5924227	3.3723059	3.4214781	3.3217743	3.8256227	3.3781568	3.6375048
A10_Breas	12.234670	12.100406	11.945899	12.499837	6.0019789	3.6791644	3.8985912	12.143533	7.1209046	11.954160	12.556280	11.819034	11.435146	6.5924227	3.3723059	3.4214781	3.3217743	3.8256227	3.3781568	3.6375048
A11_Breas	11.819034	11.819034	11.819034	11.819034	6.0019789	3.6791644	3.8985912	11.819034	7.1209046	11.819034	11.819034	11.819034	11.435146	6.5924227	3.3723059	3.4214781	3.3217743	3.8256227	3.3781568	3.6375048
A12_Breas	11.435146	11.435146	11.435146	11.435146	6.0019789	3.6791644	3.8985912	11.435146	7.1209046	11.435146	11.435146	11.435146	11.435146	6.5924227	3.3723059	3.4214781	3.3217743	3.8256227	3.3781568	3.6375048
A12_Skin	6.5924227	6.5924227	6.5924227	6.5924227	6.0019789	3.6791644	3.8985912	6.5924227	6.5924227	6.5924227	6.5924227	6.5924227	6.5924227	6.5924227	3.3723059	3.4214781	3.3217743	3.8256227	3.3781568	3.6375048
B02_Skin	3.3723059	3.3723059	3.3723059	3.3723059	3.3723059	3.3723059	3.3723059	3.3723059	3.3723059	3.3723059	3.3723059	3.3723059	3.3723059	3.3723059	3.3723059	3.3723059	3.3723059	3.3723059	3.3723059	3.3723059
B09_Skin	3.4214781	3.4214781	3.4214781	3.4214781	3.4214781	3.4214781	3.4214781	3.4214781	3.4214781	3.4214781	3.4214781	3.4214781	3.4214781	3.4214781	3.3723059	3.4214781	3.3217743	3.4214781	3.3781568	3.4214781
D07_Skin	3.3217743	3.3217743	3.3217743	3.3217743	3.3217743	3.3217743	3.3217743	3.3217743	3.3217743	3.3217743	3.3217743	3.3217743	3.3217743	3.3217743	3.3217743	3.3217743	3.3217743	3.3217743	3.3217743	3.3217743
G02_Skin	3.8256227	3.8256227	3.8256227	3.8256227	3.8256227	3.6791644	3.8256227	3.8256227	3.8256227	3.8256227	3.8256227	3.8256227	3.8256227	3.8256227	3.3723059	3.4214781	3.3217743	3.8256227	3.3781568	3.6375048
H06_Skin	3.3781568	3.3781568	3.3781568	3.3781568	3.3781568	3.3781568	3.3781568	3.3781568	3.3781568	3.3781568	3.3781568	3.3781568	3.3781568	3.3781568	3.3723059	3.3781568	3.3217743	3.3781568	3.3781568	3.3781568
H10_Skin	3.6375048	3.6375048	3.6375048	3.6375048	3.6375048	3.6375048	3.6375048	3.6375048	3.6375048	3.6375048	3.6375048	3.6375048	3.6375048	3.6375048	3.3723059	3.4214781	3.3217743	3.6375048	3.3781568	3.6375048

Matriz de distâncias de Manhattan

	A01.5500_	A01_Breas	A03_Breas	A05_Breas	A05_Skin	A06_Skin	A07_Skin	A08_Breas	A09.3978_	A09_Breas	A10_Breas	A11_Breas	A12_Breas	A12_Skin	B02_Skin	B09_Skin	D07_Skin	G02_Skin	H06_Skin	H10_Skin
A01.5500_	0	0.1342641	0.2887709	0.2651661	6.2326918	8.5555063	8.3360795	0.0911375	5.1137662	0.2805101	0.3216096	0.4156368	0.7995244	5.6422481	8.8623649	8.8131926	8.9128964	8.4090481	8.8565140	8.5971659
A01_Breas	0.1342641	0	0.1545067	0.3994303	6.0984276	8.4212421	8.2018154	0.0431266	4.9795020	0.1462460	0.4558737	0.2813726	0.6652602	5.5079839	8.7281007	8.6789285	8.7786323	8.2747839	8.7222498	8.4629017
A03_Breas	0.2887709	0.1545067	0	0.5539370	5.9439209	8.2667354	8.0473086	0.1976333	4.8249952	0.0082607	0.6103805	0.1268658	0.5107535	5.3534772	8.5735939	8.5244217	8.6241255	8.1202772	8.5677430	8.3083950
A05_Breas	0.2651661	0.3994303	0.5539370	0	6.4978580	8.8206725	8.6012457	0.3563037	5.3789323	0.5456763	0.0564434	0.6808029	1.0646906	5.9074143	9.1275310	9.0783588	9.1780626	8.6742142	9.1216801	8.8623321
A05_Skin	6.2326918	6.0984276	5.9439209	6.4978580	0	2.3228144	2.1033877	6.1415543	1.1189256	5.9521816	6.5543014	5.8170550	5.4331673	0.5904437	2.6296730	2.5805008	2.6802046	2.1763562	2.6238221	2.3644741
A06_Skin	8.5555063	8.4212421	8.2667354	8.8206725	2.3228144	0	0.2194267	8.4643688	3.4417401	8.2749961	8.8771159	8.1398695	7.7559818	2.9132582	0.3068585	0.2576863	0.3573901	0.1464582	0.3010076	0.0416596
A07_Skin	8.3360795	8.2018154	8.0473086	8.6012457	2.1033877	0.2194267	0	8.2449420	3.2223133	8.0555694	8.6576892	7.9204427	7.5365551	2.6938314	0.5262853	0.4771130	0.5768168	0.0729685	0.5204344	0.2610863
A08_Breas	0.0911375	0.0431266	0.1976333	0.3563037	6.1415543	8.4643688	8.2449420	0	5.0226286	0.1893726	0.4127471	0.3244992	0.7083869	5.5511105	8.7712273	8.7220551	8.8217589	8.3179105	8.7653764	8.5060284
A09.3978_	5.1137662	4.9795020	4.8249952	5.3789323	1.1189256	3.4417401	3.2223133	5.0226286	0	4.8332560	5.4353758	4.6981294	4.3142417	0.5284819	3.7485986	3.6994264	3.7991302	3.2952819	3.7427478	3.4833997
A09_Breas	0.2805101	0.1462460	0.0082607	0.5456763	5.9521816	8.2749961	8.0555694	0.1893726	4.8332560	0	0.6021198	0.1351266	0.5190142	5.3617379	8.5818547	8.5326824	8.6323862	8.1285379	8.5760038	8.3166557
A10_Breas	0.3216096	0.4558737	0.6103805	0.0564434	6.5543014	8.8771159	8.6576892	0.4127471	5.4353758	0.6021198	0	0.7372464	1.1211340	5.9638577	9.1839745	9.1348023	9.2345060	8.7306577	9.1781236	8.9187755
A11_Breas	0.4156368	0.2813726	0.1268658	0.6808029	5.8170550	8.1398695	7.9204427	0.3244992	4.6981294	0.1351266	0.7372464	0	0.3838876	5.2266113	8.4467281	8.3975558	8.4972596	7.9934113	8.4408772	8.1815291
A12_Breas	0.7995244	0.6652602	0.5107535	1.0646906	5.4331673	7.7559818	7.5365551	0.7083869	4.3142417	0.5190142	1.1211340	0.3838876	0	4.8427236	8.0628404	8.0136682	8.1133720	7.6095236	8.0569895	7.7976414
A12_Skin	5.6422481	5.5079839	5.3534772	5.9074143	0.5904437	2.9132582	2.6938314	5.5511105	0.5284819	5.3617379	5.9638577	5.2266113	4.8427236	0	3.2201167	3.1709445	3.2706483	2.7667999	3.2142658	2.9549178
B02_Skin	8.8623649	8.7281007	8.5735939	9.1275310	2.6296730	0.3068585	0.5262853	8.7712273	3.7485986	8.5818547	9.1839745	8.4467281	8.0628404	3.2201167	0	0.0491722	0.0505315	0.4533167	0.0058508	0.2651989
B09_Skin	8.8131926	8.6789285	8.5244217	9.0783588	2.5805008	0.2576863	0.4771130	8.7220551	3.6994264	8.5326824	9.1348023	8.3975558	8.0136682	3.1709445	0.0491722	0	0.0997037	0.4041445	0.0433213	0.2160267
D07_Skin	8.9128964	8.7786323	8.6241255	9.1780626	2.6802046	0.3573901	0.5768168	8.8217589	3.7991302	8.6323862	9.2345060	8.4972596	8.1133720	3.2706483	0.0505315	0.0997037	0	0.5038483	0.0563824	0.3157305
G02_Skin	8.4090481	8.2747839	8.1202772	8.6742142	2.1763562	0.1464582	0.0729685	8.3179105	3.2952819	8.1285379	8.7306577	7.9934113	7.6095236	2.7667999	0.4533167	0.4041445	0.5038483	0	0.4474658	0.1881178
H06_Skin	8.8565140	8.7222498	8.5677430	9.1216801	2.6238221	0.3010076	0.5204344	8.7653764	3.7427478	8.5760038	9.1781236	8.4408772	8.0569895	3.2142658	0.0058508	0.0433213	0.0563824	0.4474658	0	0.2593480
H10_Skin	8.5971659	8.4629017	8.3083950	8.8623321	2.3644741	0.0416596	0.2610863	8.5060284	3.4833997	8.3166557	8.9187755	8.1815291	7.7976414	2.9549178	0.2651989	0.2160267	0.3157305	0.1881178	0.2593480	0

Matriz de distâncias de similaridade de Tanimoto

	A01.5500_	A01_Breas	A03_Breas	A05_Breas	A05_Skin	A06_Skin	A07_Skin	A08_Breas	A09.3978_	A09_Breas	A10_Breas	A11_Breas	A12_Breas	A12_Skin	B02_Skin	B09_Skin	D07_Skin	G02_Skin	H06_Skin	H10_Skin
A01.5500_	0	0.0109740	0.0236026	0.0212135	0.5094286	0.6992837	0.6813489	0.0074491	0.4179733	0.0229274	0.0256134	0.0339720	0.0653490	0.4611687	0.7243648	0.7203457	0.7284949	0.6873129	0.7238865	0.7026887
A01_Breas	0.0109740	0	0.0127687	0.0319548	0.5039853	0.6959470	0.6778132	0.0035514	0.4115152	0.0120860	0.0363064	0.0232531	0.0549783	0.4551899	0.7213063	0.7172427	0.7254824	0.6838434	0.7208228	0.6993898
A03_Breas	0.0236026	0.0127687	0	0.0443155	0.4975699	0.6920144	0.6736460	0.0162747	0.4039038	0.0006910	0.0486115	0.0106200	0.0427555	0.4481434	0.7177018	0.7135855	0.7219318	0.6797543	0.7172120	0.6955018
A05_Breas	0.0212135	0.0319548	0.0443155	0	0.5198354	0.7056630	0.6881086	0.0285046	0.4303202	0.0436546	0.0044952	0.0544649	0.0851763	0.4725993	0.7302120	0.7262781	0.7342545	0.6939461	0.7297439	0.7089958
A05_Skin	0.5094286	0.5039853	0.4975699	0.5198354	0	0.3870081	0.3504490	0.5057468	0.1571325	0.4979171	0.5219938	0.4921768	0.4751288	0.0895639	0.4381343	0.4299416	0.4465534	0.3626064	0.4371595	0.3939490
A06_Skin	0.6992837	0.6959470	0.6920144	0.7056630	0.3870081	0	0.0562836	0.6970268	0.4833290	0.6922272	0.7069861	0.6887085	0.6782582	0.4419101	0.0834044	0.0700393	0.0971389	0.0382834	0.0818141	0.0113231
A07_Skin	0.6813489	0.6778132	0.6736460	0.6881086	0.3504490	0.0562836	0	0.6789574	0.4525146	0.6738716	0.6895106	0.6701429	0.6590694	0.4086254	0.1349937	0.1223808	0.1479552	0.0187166	0.1334929	0.0669694
A08_Breas	0.0074491	0.0035514	0.0162747	0.0285046	0.5057468	0.6970268	0.6789574	0	0.4136052	0.0155945	0.0328717	0.0267219	0.0583344	0.4571248	0.7222961	0.7182468	0.7264573	0.6849662	0.7218143	0.7004574
A09.3978_	0.4179733	0.4115152	0.4039038	0.4303202	0.1571325	0.4833290	0.4525146	0.4136052	0	0.4043157	0.4328810	0.3975053	0.3772791	0.0742155	0.5264216	0.5195163	0.5335179	0.4627616	0.5256000	0.4891793
A09_Breas	0.0229274	0.0120860	0.0006910	0.0436546	0.4979171	0.6922272	0.6738716	0.0155945	0.4043157	0	0.0479536	0.0113037	0.0434170	0.4485248	0.7178968	0.7137834	0.7221239	0.6799756	0.7174074	0.6957122
A10_Breas	0.0256134	0.0363064	0.0486115	0.0044952	0.5219938	0.7069861	0.6895106	0.0328717	0.4328810	0.0479536	0	0.0587153	0.0892887	0.4749700	0.7314247	0.7275086	0.7354491	0.6953219	0.7309587	0.7103039
A11_Breas	0.0339720	0.0232531	0.0106200	0.0544649	0.4921768	0.6887085	0.6701429	0.0267219	0.3975053	0.0113037	0.0587153	0	0.0324804	0.4422198	0.7146716	0.7105111	0.7189470	0.6763168	0.7141765	0.6922333
A12_Breas	0.0653490	0.0549783	0.0427555	0.0851763	0.4751288	0.6782582	0.6590694	0.0583344	0.3772791	0.0434170	0.0892887	0.0324804	0	0.4234946	0.7050928	0.7007927	0.7095118	0.6654504	0.7045812	0.6819013
A12_Skin	0.4611687	0.4551899	0.4481434	0.4725993	0.0895639	0.4419101	0.4086254	0.4571248	0.0742155	0.4485248	0.4749700	0.4422198	0.4234946	0	0.4884572	0.4809983	0.4961223	0.4196939	0.4875697	0.4482294
B02_Skin	0.7243648	0.7213063	0.7177018	0.7302120	0.4381343	0.0834044	0.1349937	0.7222961	0.5264216	0.7178968	0.7314247	0.7146716	0.7050928	0.4884572	0	0.0143716	0.0149842	0.1184948	0.0017319	0.0729068
B09_Skin	0.7203457	0.7172427	0.7135855	0.7262781	0.4299416	0.0700393	0.1223808	0.7182468	0.5195163	0.7137834	0.7275086	0.7105111	0.7007927	0.4809983	0.0143716	0	0.0291405	0.1056415	0.0126615	0.0593887
D07_Skin	0.7284949	0.7254824	0.7219318	0.7342545	0.4465534	0.0971389	0.1479552	0.7264573	0.5335179	0.7221239	0.7354491	0.7189470	0.7095118	0.4961223	0.0149842	0.0291405	0	0.1317036	0.0166902	0.0867986
G02_Skin	0.6873129	0.6838434	0.6797543	0.6939461	0.3626064	0.0382834	0.0187166	0.6849662	0.4627616	0.6799756	0.6953219	0.6763168	0.6654504	0.4196939	0.1184948	0.1056415	0.1317036	0	0.1169655	0.0491731
H06_Skin	0.7238865	0.7208228	0.7172120	0.7297439	0.4371595	0.0818141	0.1334929	0.7218143	0.5256000	0.7174074	0.7309587	0.7141765	0.7045812	0.4875697	0.0017319	0.0126615	0.0166902	0.1169655	0	0.0712983
H10_Skin	0.7026887	0.6993898	0.6955018	0.7089958	0.3939490	0.0113231	0.0669694	0.7004574	0.4891793	0.6957122	0.7103039	0.6922333	0.6819013	0.4482294	0.0729068	0.0593887	0.0867986	0.0491731	0.0712983	0

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Matriz de distâncias de similaridade de Euclideana

	A01.5500_	A01_Breas	A03_Breas	A05_Breas	A05_Skin	A06_Skin	A07_Skin	A08_Breas	A09.3978_	A09_Breas	A10_Breas	A11_Breas	A12_Breas	A12_Skin	B02_Skin	B09_Skin	D07_Skin	G02_Skin	H06_Skin	H10_Skin
A01.5500_	0	0.1342641	0.2887709	0.2651661	6.2326918	8.5555063	8.3360795	0.0911375	5.1137662	0.2805101	0.3216096	0.4156368	0.7995244	5.6422481	8.8623649	8.8131926	8.9128964	8.4090481	8.8565140	8.5971659
A01_Breas	0.1342641	0	0.1545067	0.3994303	6.0984276	8.4212421	8.2018154	0.0431266	4.9795020	0.1462460	0.4558737	0.2813726	0.6652602	5.5079839	8.7281007	8.6789285	8.7786323	8.2747839	8.7222498	8.4629017
A03_Breas	0.2887709	0.1545067	0	0.5539370	5.9439209	8.2667354	8.0473086	0.1976333	4.8249952	0.0082607	0.6103805	0.1268658	0.5107535	5.3534772	8.5735939	8.5244217	8.6241255	8.1202772	8.5677430	8.3083950
A05_Breas	0.2651661	0.3994303	0.5539370	0	6.4978580	8.8206725	8.6012457	0.3563037	5.3789323	0.5456763	0.0564434	0.6808029	1.0646906	5.9074143	9.1275310	9.0783588	9.1780626	8.6742142	9.1216801	8.8623321
A05_Skin	6.2326918	6.0984276	5.9439209	6.4978580	0	2.3228144	2.1033877	6.1415543	1.1189256	5.9521816	6.5543014	5.8170550	5.4331673	0.5904437	2.6296730	2.5805008	2.6802046	2.1763562	2.6238221	2.3644741
A06_Skin	8.5555063	8.4212421	8.2667354	8.8206725	2.3228144	0	0.2194267	8.4643688	3.4417401	8.2749961	8.8771159	8.1398695	7.7559818	2.9132582	0.3068585	0.2576863	0.3573901	0.1464582	0.3010076	0.0416596
A07_Skin	8.3360795	8.2018154	8.0473086	8.6012457	2.1033877	0.2194267	0	8.2449420	3.2223133	8.0555694	8.6576892	7.9204427	7.5365551	2.6938314	0.5262853	0.4771130	0.5768168	0.0729685	0.5204344	0.2610863
A08_Breas	0.0911375	0.0431266	0.1976333	0.3563037	6.1415543	8.4643688	8.2449420	0	5.0226286	0.1893726	0.4127471	0.3244992	0.7083869	5.5511105	8.7712273	8.7220551	8.8217589	8.3179105	8.7653764	8.5060284
A09.3978_	5.1137662	4.9795020	4.8249952	5.3789323	1.1189256	3.4417401	3.2223133	5.0226286	0	4.8332560	5.4353758	4.6981294	4.3142417	0.5284819	3.7485986	3.6994264	3.7991302	3.2952819	3.7427478	3.4833997
A09_Breas	0.2805101	0.1462460	0.0082607	0.5456763	5.9521816	8.2749961	8.0555694	0.1893726	4.8332560	0	0.6021198	0.1351266	0.5190142	5.3617379	8.5818547	8.5326824	8.6323862	8.1285379	8.5760038	8.3166557
A10_Breas	0.3216096	0.4558737	0.6103805	0.0564434	6.5543014	8.8771159	8.6576892	0.4127471	5.4353758	0.6021198	0	0.7372464	1.1211340	5.9638577	9.1839745	9.1348023	9.2345060	8.7306577	9.1781236	8.9187755
A11_Breas	0.4156368	0.2813726	0.1268658	0.6808029	5.8170550	8.1398695	7.9204427	0.3244992	4.6981294	0.1351266	0.7372464	0	0.3838876	5.2266113	8.4467281	8.3975558	8.4972596	7.9934113	8.4408772	8.1815291
A12_Breas	0.7995244	0.6652602	0.5107535	1.0646906	5.4331673	7.7559818	7.5365551	0.7083869	4.3142417	0.5190142	1.1211340	0.3838876	0	4.8427236	8.0628404	8.0136682	8.1133720	7.6095236	8.0569895	7.7976414
A12_Skin	5.6422481	5.5079839	5.3534772	5.9074143	0.5904437	2.9132582	2.6938314	5.5511105	0.5284819	5.3617379	5.9638577	5.2266113	4.8427236	0	3.2201167	3.1709445	3.2706483	2.7667999	3.2142658	2.9549178
B02_Skin	8.8623649	8.7281007	8.5735939	9.1275310	2.6296730	0.3068585	0.5262853	8.7712273	3.7485986	8.5818547	9.1839745	8.4467281	8.0628404	3.2201167	0	0.0491722	0.0505315	0.4533167	0.0058508	0.2651989
B09_Skin	8.8131926	8.6789285	8.5244217	9.0783588	2.5805008	0.2576863	0.4771130	8.7220551	3.6994264	8.5326824	9.1348023	8.3975558	8.0136682	3.1709445	0.0491722	0	0.0997037	0.4041445	0.0433213	0.2160267
D07_Skin	8.9128964	8.7786323	8.6241255	9.1780626	2.6802046	0.3573901	0.5768168	8.8217589	3.7991302	8.6323862	9.2345060	8.4972596	8.1133720	3.2706483	0.0505315	0.0997037	0	0.5038483	0.0563824	0.3157305
G02_Skin	8.4090481	8.2747839	8.1202772	8.6742142	2.1763562	0.1464582	0.0729685	8.3179105	3.2952819	8.1285379	8.7306577	7.9934113	7.6095236	2.7667999	0.4533167	0.4041445	0.5038483	0	0.4474658	0.1881178
H06_Skin	8.8565140	8.7222498	8.5677430	9.1216801	2.6238221	0.3010076	0.5204344	8.7653764	3.7427478	8.5760038	9.1781236	8.4408772	8.0569895	3.2142658	0.0058508	0.0433213	0.0563824	0.4474658	0	0.2593480
H10_Skin	8.5971659	8.4629017	8.3083950	8.8623321	2.3644741	0.0416596	0.2610863	8.5060284	3.4833997	8.3166557	8.9187755	8.1815291	7.7976414	2.9549178	0.2651989	0.2160267	0.3157305	0.1881178	0.2593480	0

Gene Sonda ID - 11716887_a_at

Matriz de distâncias de similaridade de Dice

	A01.5500_	A01_Breas	A03_Breas	A05_Breas	A05_Skin	A06_Skin	A07_Skin	A08_Breas	A09.3978_	A09_Breas	A10_Breas	A11_Breas	A12_Breas	A12_Skin	B02_Skin	B09_Skin	D07_Skin	G02_Skin	H06_Skin	H10_Skin
A01.5500_	0	0.0056555	0.0018923	0.0033881	0.4433516	0.4904744	0.3854717	0.0005867	0.4692408	0.0011977	0.0001616	0.0003495	0.0011647	0.4886633	0.4416079	0.4853829	0.4840370	0.3836616	0.3735128	0.4516413
A01_Breas	0.0056555	0	0.0139764	0.0002920	0.3929398	0.4424007	0.3330694	0.0098407	0.4200404	0.0016587	0.0039143	0.0087861	0.0016980	0.4404891	0.3911213	0.4370286	0.4356096	0.3312147	0.3208358	0.4015973
A03_Breas	0.0018923	0.0139764	0	0.0102795	0.4714071	0.5169943	0.4150120	0.0003726	0.4964856	0.0060781	0.0031557	0.0006165	0.0060039	0.5152471	0.4697150	0.5120813	0.5107821	0.4132403	0.4032968	0.4794463
A05_Breas	0.0033881	0.0002920	0.0102795	0	0.4045520	0.4535262	0.3450576	0.0067708	0.4314037	0.0005597	0.0020735	0.0058980	0.0005827	0.4516360	0.4027485	0.4482138	0.4468102	0.3432102	0.3328675	0.4131348
A05_Skin	0.4433516	0.3929398	0.4714071	0.4045520	0	0.0054406	0.0074019	0.4590697	0.0016079	0.4204614	0.4350027	0.4555047	0.4207828	0.0050214	7.0964977	0.0043066	0.0040298	0.0078584	0.0106540	0.0001620
A06_Skin	0.4904744	0.4424007	0.5169943	0.4535262	0.0054406	0	0.0251353	0.5053513	0.0011415	0.4687170	0.4825517	0.5019814	0.4690233	8.5122577	0.0058378	6.6996445	0.0001069	0.0259519	0.0307361	0.0037337
A07_Skin	0.3854717	0.3330694	0.4150120	0.3450576	0.0074019	0.0251353	0	0.4019906	0.0157735	0.3615641	0.3767308	0.3982369	0.3618986	0.0242466	0.0069546	0.0226802	0.0220535	6.9615262	0.0003019	0.0097273
A08_Breas	0.0005867	0.0098407	0.0003726	0.0067708	0.4590697	0.5053513	0.4019906	0	0.4845159	0.0034538	0.0013632	3.0582896	0.0033977	0.5035754	0.4573542	0.5003580	0.4990378	0.4002010	0.3901608	0.4672229
A09.3978_	0.4692408	0.4200404	0.4964856	0.4314037	0.0016079	0.0011415	0.0157735	0.4845159	0	0.4469420	0.4611152	0.4810539	0.4472556	0.0009531	0.0018281	0.0006558	0.0005502	0.0164296	0.0203295	0.0007501
A09_Breas	0.0011977	0.0016587	0.0060781	0.0005597	0.4204614	0.4687170	0.3615641	0.0034538	0.4469420	0	0.0004799	0.0028367	2.3167828	0.4668579	0.4186807	0.4634913	0.4621103	0.3597298	0.3494533	0.4289320
A10_Breas	0.0001616	0.0039143	0.0031557	0.0020735	0.4350027	0.4825517	0.3767308	0.0013632	0.4611152	0.0004799	0	0.0009859	0.0004590	0.4807227	0.4332450	0.4774100	0.4760510	0.3749111	0.3647114	0.4433609
A11_Breas	0.0003495	0.0087861	0.0006165	0.0058980	0.4555047	0.5019814	0.3982369	3.0582896	0.4810539	0.0028367	0.0009859	0	0.0027858	0.5001974	0.4537826	0.4969655	0.4956392	0.3964424	0.3863761	0.4636897
A12_Breas	0.0011647	0.0016980	0.0060039	0.0005827	0.4207828	0.4690233	0.3618986	0.0033977	0.4472556	2.3167828	0.0004590	0.0027858	0	0.4671648	0.4190027	0.4637994	0.4624189	0.3600645	0.3497896	0.4292510
A12_Skin	0.4886633	0.4404891	0.5152471	0.4516360	0.0050214	8.5122577	0.0242466	0.5035754	0.0009531	0.4668579	0.4807227	0.5001974	0.4671648	0	0.0054035	2.7748826	5.5097183	0.0250497	0.0297594	0.0033871
B02_Skin	0.4416079	0.3911213	0.4697150	0.4027485	7.0964977	0.0058378	0.0069546	0.4573542	0.0018281	0.4186807	0.4332450	0.4537826	0.4190027	0.0054035	0	0.0046613	0.0043732	0.0073975	0.0101181	0.0002368
B09_Skin	0.4853829	0.4370286	0.5120813	0.4482138	0.0043066	6.6996445	0.0226802	0.5003580	0.0006558	0.4634913	0.4774100	0.4969655	0.4637994	2.7748826	0.0046613	0	4.6446176	0.0234586	0.0280330	0.0028039
D07_Skin	0.4840370	0.4356096	0.5107821	0.4468102	0.0040298	0.0001069	0.0220535	0.4990378	0.0005502	0.4621103	0.4760510	0.4956392	0.4624189	5.5097183	0.0043732	4.6446176	0	0.0228218	0.0273405	0.0025810
G02_Skin	0.3836616	0.3312147	0.4132403	0.3432102	0.0078584	0.0259519	6.9615262	0.4002010	0.0164296	0.3597298	0.3749111	0.3964424	0.3600645	0.0250497	0.0073975	0.0234586	0.0228218	0	0.0002172	0.0102481
H06_Skin	0.3735128	0.3208358	0.4032968	0.3328675	0.0106540	0.0307361	0.0003019	0.3901608	0.0203295	0.3494533	0.3647114	0.3863761	0.3497896	0.0297594	0.0101181	0.0280330	0.0273405	0.0002172	0	0.0133995
H10_Skin	0.4516413	0.4015973	0.4794463	0.4131348	0.0001620	0.0037337	0.0097273	0.4672229	0.0007501	0.4289320	0.4433609	0.4636897	0.4292510	0.0033871	0.0002368	0.0028039	0.0025810	0.0102481	0.0133995	0

Matriz de distâncias de similaridade de Jaccard

	A01.5500_	A01_Breas	A03_Breas	A05_Breas	A05_Skin	A06_Skin	A07_Skin	A08_Breas	A09.3978_	A09_Breas	A10_Breas	A11_Breas	A12_Breas	A12_Skin	B02_Skin	B09_Skin	D07_Skin	G02_Skin	H06_Skin	H10_Skin
A01.5500_	0	0.0112475	0.0037774	0.0067533	0.6143362	0.6581454	0.5564483	0.0011727	0.6387527	0.0023926	0.0003232	0.0006988	0.0023267	0.6565129	0.6126602	0.6535458	0.6523247	0.5545599	0.5438796	0.6222492
A01_Breas	0.0112475	0	0.0275675	0.0005839	0.5641878	0.6134228	0.4997031	0.0194897	0.5915894	0.0033119	0.0077981	0.0174193	0.0033904	0.6115827	0.5623109	0.6082393	0.6068635	0.4976127	0.4858073	0.5730566
A03_Breas	0.0037774	0.0275675	0	0.0203499	0.6407568	0.6816035	0.5865844	0.0007449	0.6635354	0.0120829	0.0062915	0.0012323	0.0119361	0.6800832	0.6391919	0.6773198	0.6761823	0.5848125	0.5747848	0.6481429
A05_Breas	0.0067533	0.0005839	0.0203499	0	0.5760584	0.6240358	0.5130748	0.0134507	0.6027702	0.0011188	0.0041385	0.0117269	0.0011647	0.6222441	0.5742277	0.6189884	0.6176487	0.5110298	0.4994758	0.5847069
A05_Skin	0.6143362	0.5641878	0.6407568	0.5760584	0	0.0108223	0.0146950	0.6292636	0.0032106	0.5920068	0.6062744	0.6259062	0.5923253	0.0099926	1.4192894	0.0085762	0.0080272	0.0155942	0.0210834	0.0003239
A06_Skin	0.6581454	0.6134228	0.6816035	0.6240358	0.0108223	0	0.0490381	0.6714065	0.0022804	0.6382673	0.6509746	0.6684256	0.6385512	1.7024370	0.0116078	0.0001339	0.0002138	0.0505909	0.0596391	0.0074396
A07_Skin	0.5564483	0.4997031	0.5865844	0.5130748	0.0146950	0.0490381	0	0.5734569	0.0310571	0.5311012	0.5472831	0.5696272	0.5314619	0.0473453	0.0138131	0.0443544	0.0431553	1.3922955	0.0006037	0.0192673
A08_Breas	0.0011727	0.0194897	0.0007449	0.0134507	0.6292636	0.6714065	0.5734569	0	0.6527594	0.0068839	0.0027227	0.6163922	0.0067724	0.6698372	0.6276500	0.6669848	0.6658108	0.5716336	0.5613175	0.6368806
A09.3978_	0.6387527	0.5915894	0.6635354	0.6027702	0.0032106	0.0022804	0.0310571	0.6527594	0	0.6177746	0.6311826	0.6496102	0.6180741	0.0019044	0.0036496	0.0013109	0.0010998	0.0323281	0.0398488	0.0014991
A09_Breas	0.0023926	0.0033119	0.0120829	0.0011188	0.5920068	0.6382673	0.5311012	0.0068839	0.6177746	0	0.0009593	0.0056573	4.6335647	0.6365414	0.5902396	0.6334049	0.6321141	0.5291195	0.5179184	0.6003532
A10_Breas	0.0003232	0.0077981	0.0062915	0.0041385	0.6062744	0.6509746	0.5472831	0.0027227	0.6311826	0.0009593	0	0.0019699	0.0009177	0.6493082	0.6045652	0.6462797	0.6450332	0.5453605	0.5344887	0.6143452
A11_Breas	0.0006988	0.0174193	0.0012323	0.0117269	0.6259062	0.6684256	0.5696272	0.6163922	0.6496102	0.0056573	0.0019699	0	0.0055562	0.6668421	0.6242784	0.6639638	0.6627791	0.5677891	0.5573900	0.6335901
A12_Breas	0.0023267	0.0033904	0.0119361	0.0011647	0.5923253	0.6385512	0.5314619	0.0067724	0.6180741	4.6335647	0.0009177	0.0055562	0	0.6368267	0.5905594	0.6336926	0.6324028	0.5294816	0.5182876	0.6006657
A12_Skin	0.6565129	0.6115827	0.6800832	0.6222441	0.0099926	1.7024370	0.0473453	0.6698372	0.0019044	0.6365414	0.6493082	0.6668421	0.6368267	0	0.0107489	5.5496113	0.0001101	0.0488751	0.0577988	0.0067515
B02_Skin	0.6126602	0.5623109	0.6391919	0.5742277	1.4192894	0.0116078	0.0138131	0.6276500	0.0036496	0.5902396	0.6045652	0.6242784	0.5905594	0.0107489	0	0.0092793	0.0087084	0.0146865	0.0200335	0.0004736
B09_Skin	0.6535458	0.6082393	0.6773198	0.6189884	0.0085762	0.0001339	0.0443544	0.6669848	0.0013109	0.6334049	0.6462797	0.6639638	0.6336926	5.5496113	0.0092793	0	9.2891921	0.0458419	0.0545373	0.0055921
D07_Skin	0.6523247	0.6068635	0.6761823	0.6176487	0.0080272	0.0002138	0.0431553	0.6658108	0.0010998	0.6321141	0.6450332	0.6627791	0.6324028	0.0001101	0.0087084	9.2891921	0	0.0446253	0.0532259	0.0051488
G02_Skin	0.5545599	0.4976127	0.5848125	0.5110298	0.0155942	0.0505909	1.3922955	0.5716336	0.0323281	0.5291195	0.5453605	0.5677891	0.5294816	0.0488751	0.0146865	0.0458419	0.0446253	0	0.0004343	0.0202883
H06_Skin	0.5438796	0.4858073	0.5747848	0.4994758	0.0210834	0.0596391	0.0006037	0.5613175	0.0398488	0.5179184	0.5344887	0.5573900	0.5182876	0.0577988	0.0200335	0.0545373	0.0532259	0.0004343	0	0.0264448
H10_Skin	0.6222492	0.5730566	0.6481429	0.5847069	0.0003239	0.0074396	0.0192673	0.6368806	0.0014991	0.6003532	0.6143452	0.6335901	0.6006657	0.0067515	0.0004736	0.0055921	0.0051488	0.0202883	0.0264448	0

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Matriz de distâncias de similaridade de Sorensen

	A01.5500_	A01_Breas	A03_Breas	A05_Breas	A05_Skin	A06_Skin	A07_Skin	A08_Breas	A09.3978_	A09_Breas	A10_Breas	A11_Breas	A12_Breas	A12_Skin	B02_Skin	B09_Skin	D07_Skin	G02_Skin	H06_Skin	H10_Skin
A01.5500_	0	0.0532523	0.0307742	0.0411939	0.5336775	0.5700172	0.4886224	0.0171304	0.5536616	0.0244791	0.0089902	0.0132208	0.0241390	0.5686228	0.5323289	0.5660969	0.5650605	0.4872009	0.4792117	0.5400837
A01_Breas	0.0532523	0	0.0838891	0.0120848	0.4944780	0.5329421	0.4470011	0.0703186	0.5156115	0.0288107	0.0442832	0.0664264	0.0291507	0.5314634	0.4930535	0.5287853	0.5276867	0.4455070	0.4371141	0.5012475
A03_Breas	0.0307742	0.0838891	0	0.0718771	0.5553313	0.5904341	0.5117022	0.0136510	0.5746448	0.0552118	0.0397535	0.0175605	0.0548725	0.5890886	0.5540271	0.5866511	0.5856507	0.5103237	0.5025743	0.5615250
A05_Breas	0.0411939	0.0120848	0.0718771	0	0.5035538	0.5415392	0.4566194	0.0582833	0.5244286	0.0167316	0.0322156	0.0543852	0.0170719	0.5400795	0.5021464	0.5374359	0.5363512	0.4551414	0.4468386	0.5102416
A05_Skin	0.5336775	0.4944780	0.5553313	0.5035538	0	0.0522276	0.0609483	0.5458180	0.0283655	0.5159385	0.5272167	0.5430666	0.5161882	0.0501699	0.0018836	0.0464537	0.0449330	0.0628069	0.0731815	0.0900003
A06_Skin	0.5700172	0.5329421	0.5904341	0.5415392	0.0522276	0	0.1128168	0.5814698	0.0238974	0.5532579	0.5639167	0.5788756	0.5534940	0.0020630	0.0541059	0.0057878	0.0073117	0.1146584	0.1249316	0.0432475
A07_Skin	0.4886224	0.4470011	0.5117022	0.4566194	0.0609483	0.1128168	0	0.5015547	0.0891597	0.4697621	0.4817484	0.4986222	0.4700273	0.1107796	0.0590714	0.1070989	0.1055922	0.0018656	0.0122879	0.0699104
A08_Breas	0.0171304	0.0703186	0.0136510	0.0582833	0.5458180	0.5814698	0.5015547	0	0.5654293	0.0415922	0.0261167	0.0039104	0.0412524	0.5801026	0.5444941	0.5776259	0.5766095	0.5001570	0.4923008	0.5521061
A09.3978_	0.5536616	0.5156115	0.5746448	0.5244286	0.0283655	0.0238974	0.0891597	0.5654293	0	0.5364531	0.5473961	0.5627632	0.5366955	0.0218355	0.0302476	0.0181121	0.0165886	0.0910103	0.1013367	0.0193700
A09_Breas	0.0244791	0.0288107	0.0552118	0.0167316	0.5159385	0.5532579	0.4697621	0.0415922	0.5364531	0	0.0154923	0.0376878	0.0003403	0.5518247	0.5145549	0.5492287	0.5481636	0.4683068	0.4601301	0.5225125
A10_Breas	0.0089902	0.0442832	0.0397535	0.0322156	0.5272167	0.5639167	0.4817484	0.0261167	0.5473961	0.0154923	0	0.0222085	0.0151520	0.5625081	0.5258553	0.5599565	0.5589095	0.4803144	0.4722560	0.5336847
A11_Breas	0.0132208	0.0664264	0.0175605	0.0543852	0.5430666	0.5788756	0.4986222	0.0039104	0.5627632	0.0376878	0.0222085	0	0.0373480	0.5775022	0.5417371	0.5750143	0.5739933	0.4972190	0.4893324	0.5493818
A12_Breas	0.0241390	0.0291507	0.0548725	0.0170719	0.5161882	0.5534940	0.4700273	0.0412524	0.5366955	0.0003403	0.0151520	0.0373480	0	0.5520614	0.5148051	0.5494664	0.5484017	0.4685725	0.4603984	0.5227599
A12_Skin	0.5686228	0.5314634	0.5890886	0.5400795	0.0501699	0.0020630	0.1107796	0.5801026	0.0218355	0.5518247	0.5625081	0.5775022	0.5520614	0	0.0520487	0.0037248	0.0052487	0.1126220	0.1229003	0.0411881
B02_Skin	0.5323289	0.4930535	0.5540271	0.5021464	0.0018836	0.0541059	0.0590714	0.5444941	0.0302476	0.5145549	0.5258553	0.5417371	0.5148051	0.0520487	0	0.0483332	0.0468127	0.0609304	0.0713077	0.0108838
B09_Skin	0.5660969	0.5287853	0.5866511	0.5374359	0.0464537	0.0057878	0.1070989	0.5776259	0.0181121	0.5492287	0.5599565	0.5750143	0.5494664	0.0037248	0.0483332	0	0.0015239	0.1089428	0.1192300	0.0374690
D07_Skin	0.5650605	0.5276867	0.5856507	0.5363512	0.0449330	0.0073117	0.1055922	0.5766095	0.0165886	0.5481636	0.5589095	0.5739933	0.5484017	0.0052487	0.0468127	0.0015239	0	0.1074367	0.1177274	0.0359472
G02_Skin	0.4872009	0.4455070	0.5103237	0.4551414	0.0628069	0.1146584	0.0018656	0.5001570	0.0910103	0.4683068	0.4803144	0.4972190	0.4685725	0.1126220	0.0609304	0.1089428	0.1074367	0	0.0104225	0.0717667
H06_Skin	0.4792117	0.4371141	0.5025743	0.4468386	0.0731815	0.1249316	0.0122879	0.4923008	0.1013367	0.4601301	0.4722560	0.4893324	0.4603984	0.1229003	0.0713077	0.1192300	0.1177274	0.0104225	0	0.0821278
H10_Skin	0.5400837	0.5012475	0.5615250	0.5102416	0.0090003	0.0432475	0.0699104	0.5521061	0.0193700	0.5225125	0.5336847	0.5493818	0.5227599	0.0411881	0.0108838	0.0374690	0.0359472	0.0717667	0.0821278	0

Matriz de distâncias de Czekanowski

	A01.5500_	A01_Breas	A03_Breas	A05_Breas	A05_Skin	A06_Skin	A07_Skin	A08_Breas	A09.3978_	A09_Breas	A10_Breas	A11_Breas	A12_Breas	A12_Skin	B02_Skin	B09_Skin	D07_Skin	G02_Skin	H06_Skin	H10_Skin
A01.5500_	0	0.0532523	0.0307742	0.0411939	0.5336775	0.5700172	0.4886224	0.0171304	0.5536616	0.0244791	0.0089902	0.0132208	0.0241390	0.5686228	0.5323289	0.5660969	0.5650605	0.4872009	0.4792117	0.5400837
A01_Breas	0.0532523	0	0.0838891	0.0120848	0.4944780	0.5329421	0.4470011	0.0703186	0.5156115	0.0288107	0.0442832	0.0664264	0.0291507	0.5314634	0.4930535	0.5287853	0.5276867	0.4455070	0.4371141	0.5012475
A03_Breas	0.0307742	0.0838891	0	0.0718771	0.5553313	0.5904341	0.5117022	0.0136510	0.5746448	0.0552118	0.0397535	0.0175605	0.0548725	0.5890886	0.5540271	0.5866511	0.5856507	0.5103237	0.5025743	0.5615250
A05_Breas	0.0411939	0.0120848	0.0718771	0	0.5035538	0.5415392	0.4566194	0.0582833	0.5244286	0.0167316	0.0322156	0.0543852	0.0170719	0.5400795	0.5021464	0.5374359	0.5363512	0.4551414	0.4468386	0.5102416
A05_Skin	0.5336775	0.4944780	0.5553313	0.5035538	0	0.0522276	0.0609483	0.5458180	0.0283655	0.5159385	0.5272167	0.5430666	0.5161882	0.0501699	0.0018836	0.0464537	0.0449330	0.0628069	0.0731815	0.0900003
A06_Skin	0.5700172	0.5329421	0.5904341	0.5415392	0.0522276	0	0.1128168	0.5814698	0.0238974	0.5532579	0.5639167	0.5788756	0.5534940	0.0020630	0.0541059	0.0057878	0.0073117	0.1146584	0.1249316	0.0432475
A07_Skin	0.4886224	0.4470011	0.5117022	0.4566194	0.0609483	0.1128168	0	0.5015547	0.0891597	0.4697621	0.4817484	0.4986222	0.4700273	0.1107796	0.0590714	0.1070989	0.1055922	0.0018656	0.0122879	0.0699104
A08_Breas	0.0171304	0.0703186	0.0136510	0.0582833	0.5458180	0.5814698	0.5015547	0	0.5654293	0.0415922	0.0261167	0.0039104	0.0412524	0.5801026	0.5444941	0.5776259	0.5766095	0.5001570	0.4923008	0.5521061
A09.3978_	0.5536616	0.5156115	0.5746448	0.5244286	0.0283655	0.0238974	0.0891597	0.5654293	0	0.5364531	0.5473961	0.5627632	0.5366955	0.0218355	0.0302476	0.0181121	0.0165886	0.0910103	0.1013367	0.0193700
A09_Breas	0.0244791	0.0288107	0.0552118	0.0167316	0.5159385	0.5532579	0.4697621	0.0415922	0.5364531	0	0.0154923	0.0376878	0.0003403	0.5518247	0.5145549	0.5492287	0.5481636	0.4683068	0.4601301	0.5225125
A10_Breas	0.0089902	0.0442832	0.0397535	0.0322156	0.5272167	0.5639167	0.4817484	0.0261167	0.5473961	0.0154923	0	0.0222085	0.0151520	0.5625081	0.5258553	0.5599565	0.5589095	0.4803144	0.4722560	0.5336847
A11_Breas	0.0132208	0.0664264	0.0175605	0.0543852	0.5430666	0.5788756	0.4986222	0.0039104	0.5627632	0.0376878	0.0222085	0	0.0373480	0.5775022	0.5417371	0.5750143	0.5739933	0.4972190	0.4893324	0.5493818
A12_Breas	0.0241390	0.0291507	0.0548725	0.0170719	0.5161882	0.5534940	0.4700273	0.0412524	0.5366955	0.0003403	0.0151520	0.0373480	0	0.5520614	0.5148051	0.5494664	0.5484017	0.4685725	0.4603984	0.5227599
A12_Skin	0.5686228	0.5314634	0.5890886	0.5400795	0.0501699	0.0020630	0.1107796	0.5801026	0.0218355	0.5518247	0.5625081	0.5775022	0.5520614	0	0.0520487	0.0037248	0.0052487	0.1126220	0.1229003	0.0411881
B02_Skin	0.5323289	0.4930535	0.5540271	0.5021464	0.0018836	0.0541059	0.0590714	0.5444941	0.0302476	0.5145549	0.5258553	0.5417371	0.5148051	0.0520487	0	0.0483332	0.0468127	0.0609304	0.0713077	0.0108838
B09_Skin	0.5660969	0.5287853	0.5866511	0.5374359	0.0464537	0.0057878	0.1070989	0.5776259	0.0181121	0.5492287	0.5599565	0.5750143	0.5494664	0.0037248	0.0483332	0	0.0015239	0.1089428	0.1192300	0.0374690
D07_Skin	0.5650605	0.5276867	0.5856507	0.5363512	0.0449330	0.0073117	0.1055922	0.5766095	0.0165886	0.5481636	0.5589095	0.5739933	0.5484017	0.0052487	0.0468127	0.0015239	0	0.1074367	0.1177274	0.0359472
G02_Skin	0.4872009	0.4455070	0.5103237	0.4551414	0.0628069	0.1146584	0.0018656	0.5001570	0.0910103	0.4683068	0.4803144	0.4972190	0.4685725	0.1126220	0.0609304	0.1089428	0.1074367	0	0.0104225	0.0717667
H06_Skin	0.4792117	0.4371141	0.5025743	0.4468386	0.0731815	0.1249316	0.0122879	0.4923008	0.1013367	0.4601301	0.4722560	0.4893324	0.4603984	0.1229003	0.0713077	0.1192300	0.1177274	0.0104225	0	0.0821278
H10_Skin	0.5400837	0.5012475	0.5615250	0.5102416	0.0090003	0.0432475	0.0699104	0.5521061	0.0193700	0.5225125	0.5336847	0.5493818	0.5227599	0.0411881	0.0108838	0.0374690	0.0359472	0.0717667	0.0821278	0

Matriz de distâncias de similaridade de Minkowski

	A01.5500_	A01_Breas	A03_Breas	A05_Breas	A05_Skin	A06_Skin	A07_Skin	A08_Breas	A09.3978_	A09_Breas	A10_Breas	A11_Breas	A12_Breas	A12_Skin	B02_Skin	B09_Skin	D07_Skin	G02_Skin	H06_Skin	H10_Skin
A01.5500_	0	1.0446562	0.6560396	0.8174649	7.1897167	7.5015408	6.7819686	0.3601139	7.3630024	0.4936974	0.1840997	0.2768263	0.4869990	7.4898426	7.1778605	7.4685986	7.4598615	6.7687016	6.6936658	7.2457555
A01_Breas	1.0446562	0	1.7006958	0.2271912	6.1450605	6.4568846	5.7373124	1.4047701	6.3183462	0.5509587	0.8605564	1.3214825	0.5576571	6.4451864	6.1332043	6.4239424	6.4152053	5.7240454	5.6490096	6.2010993
A03_Breas	0.6560396	1.7006958	0	1.4735046	7.8457564	8.1575805	7.4380083	0.2959257	8.0190421	1.1497371	0.8401394	0.3792133	1.1430387	8.1458823	7.8339002	8.1246383	8.1159012	7.4247413	7.3497055	7.9017952
A05_Breas	0.8174649	0.2271912	1.4735046	0	6.3722518	6.6840758	5.9645037	1.1775789	6.5455374	0.3237674	0.6333652	1.0942913	0.3304659	6.6723776	6.3603955	6.6511336	6.6423966	5.9512367	5.8762008	6.4282906
A05_Skin	7.1897167	6.1450605	7.8457564	6.3722518	0	0.3118240	0.4077480	7.5498307	0.1732856	6.6960192	7.0056170	7.4665431	6.7027177	0.3001258	0.0118562	0.2788818	0.2701448	0.4210150	0.4960509	0.0560388
A06_Skin	7.5015408	6.4568846	8.1575805	6.6840758	0.3118240	0	0.7195721	7.8616548	0.1385383	7.0078433	7.3174411	7.7783672	7.0145418	0.0116981	0.3236802	0.0329422	0.0416792	0.7328391	0.8078749	0.2557852
A07_Skin	6.7819686	5.7373124	7.4380083	5.9645037	0.4077480	0.7195721	0	7.1420826	0.5810337	6.2882711	6.5978689	7.0587950	6.2949696	0.7078739	0.3958918	0.6866299	0.6778929	0.0132670	0.0883028	0.4637868
A08_Breas	0.3601139	1.4047701	0.2959257	1.1775789	7.5498307	7.8616548	7.1420826	0	7.7231164	0.8538114	0.5442137	0.0832876	0.8471130	7.8499566	7.5379745	7.8287125	7.8199755	7.1288156	7.0537798	7.6058695
A09.3978_	7.3630024	6.3183462	8.0190421	6.5455374	0.1732856	0.1385383	0.5810337	7.7231164	0	6.8693049	7.1789027	7.6398288	6.8760034	0.1268401	0.1851418	0.1055961	0.0968591	0.5943007	0.6693365	0.1172468
A09_Breas	0.4936974	0.5509587	1.1497371	0.3237674	6.6960192	7.0078433	6.2882711	0.8538114	6.8693049	0	0.3095977	0.7705238	0.0066984	6.9961451	6.6841630	6.9749011	6.9661640	6.2750041	6.1999683	6.7520580
A10_Breas	0.1840997	0.8605564	0.8401394	0.6333652	7.0056170	7.3174411	6.5978689	0.5442137	7.1789027	0.3095977	0	0.4609260	0.3028992	7.3057429	6.9937608	7.2844988	7.2757618	6.5846019	6.5095661	7.0616558
A11_Breas	0.2768263	1.3214825	0.3792133	1.0942913	7.4665431	7.7783672	7.0587950	0.0832876	7.6398288	0.7705238	0.4609260	0	0.7638253	7.7666690	7.4546869	7.7454249	7.7366879	7.0455280	6.9704922	7.5225819
A12_Breas	0.4869990	0.5576571	1.1430387	0.3304659	6.7027177	7.0145418	6.2949696	0.8471130	6.8760034	0.0066984	0.3028992	0.7638253	0	7.0028436	6.6908615	6.9815995	6.9728625	6.2817026	6.2066668	6.7587565
A12_Skin	7.4898426	6.4451864	8.1458823	6.6723776	0.3001258	0.0116981	0.7078739	7.8499566	0.1268401	6.9961451	7.3057429	7.7666690	7.0028436	0	0.3119820	0.0212440	0.0299810	0.7211409	0.7961767	0.2440870
B02_Skin	7.1778605	6.1332043	7.8339002	6.3603955	0.0118562	0.3236802	0.3958918	7.5379745	0.1851418	6.6841630	6.9937608	7.4546869	6.6908615	0.3119820	0	0.2907380	0.2820010	0.4091588	0.4841947	0.0678950
B09_Skin	7.4685986	6.4239424	8.1246383	6.6511336	0.2788818	0.0329422	0.6866299	7.8287125	0.1055961	6.9749011	7.2844988	7.7454249	6.9815995	0.0212440	0.2907380	0	0.0087370	0.6998969	0.7749327	0.2228430
D07_Skin	7.4598615	6.4152053	8.1159012	6.6423966	0.2701448	0.0416792	0.6778929	7.8199755	0.0968591	6.9661640	7.2757618	7.7366879	6.9728625	0.0299810	0.2820010	0.0087370	0	0.6911599	0.7661957	0.2141060
G02_Skin	6.7687016	5.7240454	7.4247413	5.9512367	0.4210150	0.7328391	0.0132670	7.1288156	0.5943007	6.2750041	6.5846019	7.0455280	6.2817026	0.7211409	0.4091588	0.6998969	0.6911599	0	0.0750358	0.4770538
H06_Skin	6.6936658	5.6490096	7.3497055	5.8762008	0.4960509	0.8078749	0.0883028	7.0537798	0.6693365	6.1999683	6.5095661	6.9704922	6.2066668	0.7961767	0.4841947	0.7749327	0.7661957	0.0750358	0	0.5520897
H10_Skin	7.2457555	6.2010993	7.9017952	6.4282906	0.0560388	0.2557852	0.4637868	7.6058695	0.1172468	6.7520580	7.0616558	7.5225819	6.7587565	0.2440870	0.0678950	0.2228430	0.2141060	0.4770538	0.5520897	0

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Matriz de distâncias de similaridade de Pearson

	A01.5500_	A01_Breas	A03_Breas	A05_Breas	A05_Skin	A06_Skin	A07_Skin	A08_Breas	A09.3978_	A09_Breas	A10_Breas	A11_Breas	A12_Breas	A12_Skin	B02_Skin	B09_Skin	D07_Skin	G02_Skin	H06_Skin	H10_Skin
A01.5500_	0	0.1175189	0.0391728	0.0702428	16.456380	19.889201	12.960374	0.0121300	18.266917	0.0247771	0.0033402	0.0072242	0.0240929	19.745577	16.340474	19.488001	19.383266	12.861635	12.318569	17.017508
A01_Breas	0.1175189	0	0.2632556	0.0054255	12.021622	14.735417	9.2751936	0.1845834	13.451236	0.0308580	0.0729845	0.1646272	0.0315913	14.621617	11.930248	14.417579	14.334631	9.1979620	8.7735803	12.464245
A03_Breas	0.0391728	0.2632556	0	0.2282269	19.596585	23.520098	15.589037	0.0081912	21.667078	0.1343775	0.0695624	0.0135564	0.1327259	23.356121	19.463937	23.062009	22.942404	15.475622	14.851561	20.238585
A05_Breas	0.0702428	0.0054255	0.2282269	0	12.926966	15.790618	10.024312	0.1297066	14.435970	0.0106560	0.0395348	0.1128872	0.0110939	15.670602	12.830479	15.455406	15.367917	9.9425988	9.4934812	13.394287
A05_Skin	16.456380	12.021622	19.596585	12.926966	0	0.0343665	0.0468478	5.3315896	0.0101176	4.5578818	4.8368756	5.2555484	4.5638978	0.0317052	4.4582701	0.0271725	0.0254190	0.0497600	0.0676526	0.0010179
A06_Skin	19.889201	14.735417	23.520098	15.790618	0.0343665	0	0.1458997	5.7810966	0.0064668	4.9922743	5.2770427	5.7036881	4.9984191	4.8168381	0.0332282	0.0003791	0.0006050	0.1507657	0.1794405	0.0212070
A07_Skin	12.960374	9.2751936	15.589037	10.024312	0.0468478	0.1458997	0	4.7712484	0.1137517	4.0196866	4.2902194	4.6972106	4.0255139	0.1763747	0.0497081	0.1647155	0.1600617	4.9411894	0.0021437	0.0697213
A08_Breas	0.0121300	0.1845834	0.0081912	0.1297066	5.3315896	5.7810966	4.7712484	0	20.097430	0.0741060	0.0291884	0.0006539	0.0728981	21.689971	18.021210	21.412618	21.299833	14.266591	13.679682	18.751082
A09.3978_	18.266917	13.451236	21.667078	14.435970	0.0101176	0.0064668	0.1137517	20.097430	0	4.7968404	5.0791176	5.5023237	4.8029295	0.0056628	0.0108713	0.0038957	0.0032677	0.0991512	0.1231746	0.0044558
A09_Breas	0.0247771	0.0308580	0.1343775	0.0106560	4.5578818	4.9922743	4.0196866	0.0741060	4.7968404	0	0.0094464	0.0559694	4.5581200	17.228286	14.169962	16.996722	16.902572	11.053848	10.568446	14.777499
A10_Breas	0.0033402	0.0729845	0.0695624	0.0395348	4.8368756	5.2770427	4.2902194	0.0291884	5.0791176	0.0094464	0	0.0200281	0.0093203	18.786818	15.513013	18.539091	18.438363	12.171511	11.650278	16.163734
A11_Breas	0.0072242	0.1646272	0.0135564	0.1128872	5.2555484	5.7036881	4.6972106	0.0006539	5.5023237	0.0559694	0.0200281	0	0.0592682	21.232154	17.625174	20.959435	20.848536	13.935179	13.358543	18.342666
A12_Breas	0.0240929	0.0315913	0.1327259	0.0110939	4.5638978	4.9984191	4.0255139	0.0728981	4.8029295	4.5581200	0.0093203	0.0592682	0	17.261292	14.198376	17.029384	16.935094	11.077460	10.591295	14.806834
A12_Skin	19.745577	14.621617	23.356121	15.670602	0.0317052	4.8168381	0.1763747	21.689971	0.0056628	17.228286	18.786818	21.232154	17.261292	0	0.0308698	0.0001576	0.0003130	0.1459909	0.1742814	0.0193115
B02_Skin	16.340474	11.930248	19.463937	12.830479	4.4582701	0.0332282	0.0497081	18.021210	0.0108713	14.169962	15.513013	17.625174	14.198376	0.0308698	0	0.0295320	0.0276991	0.0469969	0.0644573	0.0014941
B09_Skin	19.488001	14.417579	23.062009	15.455406	0.0271725	0.0003791	0.1647155	21.412618	0.0038957	16.996722	18.539091	20.959435	17.029384	0.0001576	0.0295320	0	2.6588443	0.1375161	0.1651050	0.0160963
D07_Skin	19.383266	14.334631	22.942404	15.367917	0.0254190	0.0006050	0.1600617	21.299833	0.0032677	16.902572	18.438363	20.848536	16.935094	0.0003130	0.0276991	2.6588443	0	0.1341042	0.1614030	0.0148588
G02_Skin	12.861635	9.1979620	15.475622	9.9425988	0.0497600	0.1507657	4.9411894	14.266591	0.0991512	11.053848	12.171511	13.935179	11.077460	0.1459909	0.0469969	0.1375161	0.1341042	0	0.0015479	0.0737672
H06_Skin	12.318569	8.7735803	14.851561	9.4934812	0.0676526	0.1794405	0.0021437	13.679682	0.1231746	10.568446	11.650278	13.358543	10.591295	0.1742814	0.0644573	0.1651050	0.1614030	0.0015479	0	0.0987979
H10_Skin	17.017508	12.464245	20.238585	13.394287	0.0010179	0.0212070	0.0697213	18.751082	0.0044558	14.777499	16.163734	18.342666	14.806834	0.0193115	0.0014941	0.0160963	0.0148588	0.0737672	0.0987979	0

Matriz de distâncias de Intersection

	A01.5500_	A01_Breas	A03_Breas	A05_Breas	A05_Skin	A06_Skin	A07_Skin	A08_Breas	A09.3978_	A09_Breas	A10_Breas	A11_Breas	A12_Breas	A12_Skin	B02_Skin	B09_Skin	D07_Skin	G02_Skin	H06_Skin	H10_Skin
A01.5500_	10.330870	9.2862145	10.330870	9.5134058	3.1411540	2.8293299	3.5489021	10.330870	2.9678683	9.8371732	10.146771	10.330870	9.8438717	2.8410281	3.1530102	2.8622721	2.8710091	3.5621691	3.6372049	3.0851152
A01_Breas	9.2862145	9.2862145	9.2862145	9.2862145	3.1411540	2.8293299	3.5489021	9.2862145	2.9678683	9.2862145	9.2862145	9.2862145	9.2862145	2.8410281	3.1530102	2.8622721	2.8710091	3.5621691	3.6372049	3.0851152
A03_Breas	10.330870	9.2862145	10.986910	9.5134058	3.1411540	2.8293299	3.5489021	10.690984	2.9678683	9.8371732	10.146771	10.607697	9.8438717	2.8410281	3.1530102	2.8622721	2.8710091	3.5621691	3.6372049	3.0851152
A05_Breas	9.5134058	9.2862145	9.5134058	9.5134058	3.1411540	2.8293299	3.5489021	9.5134058	2.9678683	9.5134058	9.5134058	9.5134058	9.5134058	2.8410281	3.1530102	2.8622721	2.8710091	3.5621691	3.6372049	3.0851152
A05_Skin	3.1411540	3.1411540	3.1411540	3.1411540	3.1411540	2.8293299	3.1411540	3.1411540	2.9678683	3.1411540	3.1411540	3.1411540	3.1411540	2.8410281	3.1411540	2.8622721	2.8710091	3.1411540	3.1411540	3.0851152
A06_Skin	2.8293299	2.8293299	2.8293299	2.8293299	2.8293299	2.8293299	2.8293299	2.8293299	2.8293299	2.8293299	2.8293299	2.8293299	2.8293299	2.8293299	2.8293299	2.8293299	2.8293299	2.8293299	2.8293299	2.8293299
A07_Skin	3.5489021	3.5489021	3.5489021	3.5489021	3.1411540	2.8293299	3.5489021	3.5489021	2.9678683	3.5489021	3.5489021	3.5489021	3.5489021	2.8410281	3.1530102	2.8622721	2.8710091	3.5489021	3.5489021	3.0851152
A08_Breas	10.330870	9.2862145	10.690984	9.5134058	3.1411540	2.8293299	3.5489021	10.690984	2.9678683	9.8371732	10.146771	10.607697	9.8438717	2.8410281	3.1530102	2.8622721	2.8710091	3.5621691	3.6372049	3.0851152
A09.3978_	2.9678683	2.9678683	2.9678683	2.9678683	2.9678683	2.8293299	2.9678683	2.9678683	2.9678683	2.9678683	2.9678683	2.9678683	2.9678683	2.8410281	2.9678683	2.8622721	2.8710091	2.9678683	2.9678683	2.9678683
A09_Breas	9.8371732	9.2862145	9.8371732	9.5134058	3.1411540	2.8293299	3.5489021	9.8371732	2.9678683	9.8371732	9.8371732	9.8371732	9.8371732	2.8410281	3.1530102	2.8622721	2.8710091	3.5621691	3.6372049	3.0851152
A10_Breas	10.146771	9.2862145	10.146771	9.5134058	3.1411540	2.8293299	3.5489021	10.146771	2.9678683	9.8371732	10.146771	10.146771	9.8438717	2.8410281	3.1530102	2.8622721	2.8710091	3.5621691	3.6372049	3.0851152
A11_Breas	10.330870	9.2862145	10.607697	9.5134058	3.1411540	2.8293299	3.5489021	10.607697	2.9678683	9.8371732	10.146771	10.607697	9.8438717	2.8410281	3.1530102	2.8622721	2.8710091	3.5621691	3.6372049	3.0851152
A12_Breas	9.8438717	9.2862145	9.8438717	9.5134058	3.1411540	2.8293299	3.5489021	9.8438717	2.9678683	9.8371732	9.8438717	9.8438717	9.8438717	2.8410281	3.1530102	2.8622721	2.8710091	3.5621691	3.6372049	3.0851152
A12_Skin	2.8410281	2.8410281	2.8410281	2.8410281	2.8410281	2.8293299	2.8410281	2.8410281	2.8410281	2.8410281	2.8410281	2.8410281	2.8410281	2.8410281	2.8410281	2.8410281	2.8410281	2.8410281	2.8410281	2.8410281
B02_Skin	3.1530102	3.1530102	3.1530102	3.1530102	3.1411540	2.8293299	3.1530102	3.1530102	2.9678683	3.1530102	3.1530102	3.1530102	3.1530102	2.8410281	3.1530102	2.8622721	2.8710091	3.1530102	3.1530102	3.0851152
B09_Skin	2.8622721	2.8622721	2.8622721	2.8622721	2.8622721	2.8293299	2.8622721	2.8622721	2.8622721	2.8622721	2.8622721	2.8622721	2.8622721	2.8410281	2.8622721	2.8622721	2.8622721	2.8622721	2.8622721	2.8622721
D07_Skin	2.8710091	2.8710091	2.8710091	2.8710091	2.8710091	2.8293299	2.8710091	2.8710091	2.8710091	2.8710091	2.8710091	2.8710091	2.8710091	2.8410281	2.8710091	2.8622721	2.8710091	2.8710091	2.8710091	2.8710091
G02_Skin	3.5621691	3.5621691	3.5621691	3.5621691	3.1411540	2.8293299	3.5489021	3.5621691	2.9678683	3.5621691	3.5621691	3.5621691	3.5621691	2.8410281	3.1530102	2.8622721	2.8710091	3.5621691	3.5621691	3.0851152
H06_Skin	3.6372049	3.6372049	3.6372049	3.6372049	3.1411540	2.8293299	3.5489021	3.6372049	2.9678683	3.6372049	3.6372049	3.6372049	3.6372049	2.8410281	3.1530102	2.8622721	2.8710091	3.5621691	3.6372049	3.0851152
H10_Skin	3.0851152	3.0851152	3.0851152	3.0851152	3.0851152	2.8293299	3.0851152	3.0851152	2.9678683	3.0851152	3.0851152	3.0851152	3.0851152	2.8410281	3.0851152	2.8622721	2.8710091	3.0851152	3.0851152	3.0851152

Matriz de distâncias de similaridade de Manhattan

	A01.5500_	A01_Breas	A03_Breas	A05_Breas	A05_Skin	A06_Skin	A07_Skin	A08_Breas	A09.3978_	A09_Breas	A10_Breas	A11_Breas	A12_Breas	A12_Skin	B02_Skin	B09_Skin	D07_Skin	G02_Skin	H06_Skin	H10_Skin
A01.5500_	0	1.0446562	0.6560396	0.8174649	7.1897167	7.5015408	6.7819686	0.3601139	7.3630024	0.4936974	0.1840997	0.2768263	0.4869990	7.4898426	7.1778605	7.4685986	7.4598615	6.7687016	6.6936658	7.2457555
A01_Breas	1.0446562	0	1.7006958	0.2271912	6.1450605	6.4568846	5.7373124	1.4047701	6.3183462	0.5509587	0.8605564	1.3214825	0.5576571	6.4451864	6.1332043	6.4239424	6.4152053	5.7240454	5.6490096	6.2010993
A03_Breas	0.6560396	1.7006958	0	1.4735046	7.8457564	8.1575805	7.4380083	0.2959257	8.0190421	1.1497371	0.8401394	0.3792133	1.1430387	8.1458823	7.8339002	8.1246383	8.1159012	7.4247413	7.3497055	7.9017952
A05_Breas	0.8174649	0.2271912	1.4735046	0	6.3722518	6.6840758	5.9645037	1.1775789	6.5455374	0.3237674	0.6333652	1.0942913	0.3304659	6.6723776	6.3603955	6.6511336	6.6423966	5.9512367	5.8762008	6.4282906
A05_Skin	7.1897167	6.1450605	7.8457564	6.3722518	0	0.3118240	0.4077480	7.5498307	0.1732856	6.6960192	7.0056170	7.4665431	6.7027177	0.3001258	0.0118562	0.2788818	0.2701448	0.4210150	0.4960509	0.0560388
A06_Skin	7.5015408	6.4568846	8.1575805	6.6840758	0.3118240	0	0.7195721	7.8616548	0.1385383	7.0078433	7.3174411	7.7783672	7.0145418	0.0116981	0.3236802	0.0329422	0.0416792	0.7328391	0.8078749	0.2557852
A07_Skin	6.7819686	5.7373124	7.4380083	5.9645037	0.4077480	0.7195721	0	7.1420826	0.5810337	6.2882711	6.5978689	7.0587950	6.2949696	0.7078739	0.3958918	0.6866299	0.6778929	0.0132670	0.0883028	0.4637868
A08_Breas	0.3601139	1.4047701	0.2959257	1.1775789	7.5498307	7.8616548	7.1420826	0	7.7231164	0.8538114	0.5442137	0.0832876	0.8471130	7.8499566	7.5379745	7.8287125	7.8199755	7.1288156	7.0537798	7.6058695
A09.3978_	7.3630024	6.3183462	8.0190421	6.5455374	0.1732856	0.1385383	0.5810337	7.7231164	0	6.8693049	7.1789027	7.6398288	6.8760034	0.1268401	0.1851418	0.1055961	0.0968591	0.5943007	0.6693365	0.1172468
A09_Breas	0.4936974	0.5509587	1.1497371	0.3237674	6.6960192	7.0078433	6.2882711	0.8538114	6.8693049	0	0.3095977	0.7705238	0.0066984	6.9961451	6.6841630	6.9749011	6.9661640	6.2750041	6.1999683	6.7520580
A10_Breas	0.1840997	0.8605564	0.8401394	0.6333652	7.0056170	7.3174411	6.5978689	0.5442137	7.1789027	0.3095977	0	0.4609260	0.3028992	7.3057429	6.9937608	7.2844988	7.2757618	6.5846019	6.5095661	7.0616558
A11_Breas	0.2768263	1.3214825	0.3792133	1.0942913	7.4665431	7.7783672	7.0587950	0.0832876	7.6398288	0.7705238	0.4609260	0	0.7638253	7.7666690	7.4546869	7.7454249	7.7366879	7.0455280	6.9704922	7.5225819
A12_Breas	0.4869990	0.5576571	1.1430387	0.3304659	6.7027177	7.0145418	6.2949696	0.8471130	6.8760034	0.0066984	0.3028992	0.7638253	0	7.0028436	6.6908615	6.9815995	6.9728625	6.2817026	6.2066668	6.7587565
A12_Skin	7.4898426	6.4451864	8.1458823	6.6723776	0.3001258	0.0116981	0.7078739	7.8499566	0.1268401	6.9961451	7.3057429	7.7666690	7.0028436	0	0.3119820	0.0212440	0.0299810	0.7211409	0.7961767	0.2440870
B02_Skin	7.1778605	6.1332043	7.8339002	6.3603955	0.0118562	0.3236802	0.3958918	7.5379745	0.1851418	6.6841630	6.9937608	7.4546869	6.6908615	0.3119820	0	0.2907380	0.2820010	0.4091588	0.4841947	0.0678950
B09_Skin	7.4685986	6.4239424	8.1246383	6.6511336	0.2788818	0.0329422	0.6866299	7.8287125	0.1055961	6.9749011	7.2844988	7.7454249	6.9815995	0.0212440	0.2907380	0	0.0087370	0.6998969	0.7749327	0.2228430
D07_Skin	7.4598615	6.4152053	8.1159012	6.6423966	0.2701448	0.0416792	0.6778929	7.8199755	0.0968591	6.9661640	7.2757618	7.7366879	6.9728625	0.0299810	0.2820010	0.0087370	0	0.6911599	0.7661957	0.2141060
G02_Skin	6.7687016	5.7240454	7.4247413	5.9512367	0.4210150	0.7328391	0.0132670	7.1288156	0.5943007	6.2750041	6.5846019	7.0455280	6.2817026	0.7211409	0.4091588	0.6998969	0.6911599	0	0.0750358	0.4770538
H06_Skin	6.6936658	5.6490096	7.3497055	5.8762008	0.4960509	0.8078749	0.0883028	7.0537798	0.6693365	6.1999683	6.5095661	6.9704922	6.2066668	0.7961767	0.4841947	0.7749327	0.7661957	0.0750358	0	0.5520897
H10_Skin	7.2457555	6.2010993	7.9017952	6.4282906	0.0560388	0.2557852	0.4637868	7.6058695	0.1172468	6.7520580	7.0616558	7.5225819	6.7587565	0.2440870	0.0678950	0.2228430	0.2141060	0.4770538	0.5520897	0

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Matriz de distâncias de similaridade de Tanimoto

	A01.5500_	A01_Breas	A03_Breas	A05_Breas	A05_Skin	A06_Skin	A07_Skin	A08_Breas	A09.3978_	A09_Breas	A10_Breas	A11_Breas	A12_Breas	A12_Skin	B02_Skin	B09_Skin	D07_Skin	G02_Skin	H06_Skin	H10_Skin
A01.5500_	0	0.1011198	0.0597110	0.0791283	0.6959448	0.7261286	0.6564759	0.0336838	0.7127184	0.0477885	0.0178203	0.0260967	0.0471401	0.7249962	0.6947972	0.7229398	0.7220941	0.6551917	0.6479285	0.7013692
A01_Breas	0.1011198	0	0.1547929	0.0238811	0.6617400	0.6953193	0.6178311	0.1313976	0.6804006	0.0560078	0.0848108	0.1245777	0.0566501	0.6940596	0.6604633	0.6917719	0.6908310	0.6164024	0.6083221	0.6677747
A03_Breas	0.0597110	0.1547929	0	0.1341145	0.7141003	0.7424817	0.6769881	0.0269343	0.7298723	0.1046460	0.0764673	0.0345150	0.1040364	0.7414170	0.7130212	0.7394834	0.7386882	0.6757806	0.6689510	0.7192008
A05_Breas	0.0791283	0.0238811	0.1341145	0	0.6698181	0.7025954	0.6269577	0.1101469	0.6880330	0.0329126	0.0624203	0.1031601	0.0335707	0.7013658	0.6685718	0.6991327	0.6982143	0.6255632	0.6176758	0.6757086
A05_Skin	0.6959448	0.6617400	0.7141003	0.6698181	0	0.0992705	0.1148941	0.7061866	0.0551662	0.6806853	0.6904282	0.7038797	0.6809025	0.0955463	0.0037602	0.0887832	0.0860017	0.1181906	0.1363824	0.0178401
A06_Skin	0.7261286	0.6953193	0.7424817	0.7025954	0.0992705	0	0.2027590	0.7353536	0.0466794	0.7123838	0.7211595	0.7332757	0.7125795	0.0041175	0.1026575	0.0115091	0.0145172	0.2057283	0.2221142	0.0829094
A07_Skin	0.6564759	0.6178311	0.6769881	0.6269577	0.1148941	0.2027590	0	0.6680472	0.1637221	0.6392355	0.6502432	0.6654408	0.6394810	0.1994628	0.1115533	0.1934767	0.1910148	0.0037244	0.0242776	0.1306846
A08_Breas	0.0336838	0.1313976	0.0269343	0.1101469	0.7061866	0.7353536	0.6680472	0	0.7223952	0.0798627	0.0509039	0.0077904	0.0792361	0.7342594	0.7050776	0.7322723	0.7314551	0.6668062	0.6597876	0.7114283
A09.3978_	0.7127184	0.6804006	0.7298723	0.6880330	0.0551662	0.0466794	0.1637221	0.7223952	0	0.6983006	0.7075061	0.7202155	0.6985059	0.0427378	0.0587190	0.0355798	0.0326359	0.1668367	0.1840249	0.0380040
A09_Breas	0.0477885	0.0560078	0.1046460	0.0329126	0.6806853	0.7123838	0.6392355	0.0798627	0.6983006	0	0.0305119	0.0726381	0.0006804	0.7111946	0.6794800	0.7090350	0.7081469	0.6378869	0.6302591	0.6863819
A10_Breas	0.0178203	0.0848108	0.0764673	0.0624203	0.6904282	0.7211595	0.6502432	0.0509039	0.7075061	0.0305119	0	0.0434520	0.0298517	0.7200066	0.6892597	0.7179130	0.7170519	0.6489356	0.6415406	0.6959510
A11_Breas	0.0260967	0.1245777	0.0345150	0.1031601	0.7038797	0.7332757	0.6654408	0.0077904	0.7202155	0.0726381	0.0434520	0	0.0720067	0.7321729	0.7027620	0.7301702	0.7293466	0.6641901	0.6571164	0.7091625
A12_Breas	0.0471401	0.0566501	0.1040364	0.0335707	0.6809025	0.7125795	0.6394810	0.0792361	0.6985059	0.0006804	0.0298517	0.0720067	0	0.7113911	0.6796981	0.7092330	0.7083455	0.6381333	0.6305107	0.6865953
A12_Skin	0.7249962	0.6940596	0.7414170	0.7013658	0.0955463	0.0041175	0.1994628	0.7342594	0.0427378	0.7111946	0.7200066	0.7321729	0.7113911	0	0.0989473	0.0074220	0.0104426	0.2024443	0.2188979	0.0791176
B02_Skin	0.6947972	0.6604633	0.7130212	0.6685718	0.0037602	0.1026575	0.1115533	0.7050776	0.0587190	0.6794800	0.6892597	0.7027620	0.6796981	0.0989473	0	0.0922096	0.0894386	0.1148622	0.1331227	0.0215333
B09_Skin	0.7229398	0.6917719	0.7394834	0.6991327	0.0887832	0.0115091	0.1934767	0.7322723	0.0355798	0.7090350	0.7179130	0.7301702	0.7092330	0.0074220	0.0922096	0	0.0030431	0.1964805	0.2130572	0.0722316
D07_Skin	0.7220941	0.6908310	0.7386882	0.6982143	0.0860017	0.0145172	0.1910148	0.7314551	0.0326359	0.7081469	0.7170519	0.7293466	0.7083455	0.0104426	0.0894386	0.0030431	0	0.1940278	0.2106550	0.0693996
G02_Skin	0.6551917	0.6164024	0.6757806	0.6255632	0.1181906	0.2057283	0.0037244	0.6668062	0.1668367	0.6378869	0.6489356	0.6641901	0.6381333	0.2024443	0.1148622	0.1964805	0.1940278	0	0.0206300	0.1339223
H06_Skin	0.6479285	0.6083221	0.6689510	0.6176758	0.1363824	0.2221142	0.0242776	0.6597876	0.1840249	0.6302591	0.6415406	0.6571164	0.6305107	0.2188979	0.1331227	0.2130572	0.2106550	0.0206300	0	0.1517895
H10_Skin	0.7013692	0.6677747	0.7192008	0.6757086	0.0178401	0.0829094	0.1306846	0.7114283	0.0380040	0.6863819	0.6959510	0.7091625	0.6865953	0.0791176	0.0215333	0.0722316	0.0693996	0.1339223	0.1517895	0

Matriz de distâncias de Euclideana

	A01.5500_	A01_Breas	A03_Breas	A05_Breas	A05_Skin	A06_Skin	A07_Skin	A08_Breas	A09.3978_	A09_Breas	A10_Breas	A11_Breas	A12_Breas	A12_Skin	B02_Skin	B09_Skin	D07_Skin	G02_Skin	H06_Skin	H10_Skin
A01.5500_	0	1.0446562	0.6560396	0.8174649	7.1897167	7.5015408	6.7819686	0.3601139	7.3630024	0.4936974	0.1840997	0.2768263	0.4869990	7.4898426	7.1778605	7.4685986	7.4598615	6.7687016	6.6936658	7.2457555
A01_Breas	1.0446562	0	1.7006958	0.2271912	6.1450605	6.4568846	5.7373124	1.4047701	6.3183462	0.5509587	0.8605564	1.3214825	0.5576571	6.4451864	6.1332043	6.4239424	6.4152053	5.7240454	5.6490096	6.2010993
A03_Breas	0.6560396	1.7006958	0	1.4735046	7.8457564	8.1575805	7.4380083	0.2959257	8.0190421	1.1497371	0.8401394	0.3792133	1.1430387	8.1458823	7.8339002	8.1246383	8.1159012	7.4247413	7.3497055	7.9017952
A05_Breas	0.8174649	0.2271912	1.4735046	0	6.3722518	6.6840758	5.9645037	1.1775789	6.5455374	0.3237674	0.6333652	1.0942913	0.3304659	6.6723776	6.3603955	6.6511336	6.6423966	5.9512367	5.8762008	6.4282906
A05_Skin	7.1897167	6.1450605	7.8457564	6.3722518	0	0.3118240	0.4077480	7.5498307	0.1732856	6.6960192	7.0056170	7.4665431	6.7027177	0.3001258	0.0118562	0.2788818	0.2701448	0.4210150	0.4960509	0.0560388
A06_Skin	7.5015408	6.4568846	8.1575805	6.6840758	0.3118240	0	0.7195721	7.8616548	0.1385383	7.0078433	7.3174411	7.7783672	7.0145418	0.0116981	0.3236802	0.0329422	0.0416792	0.7328391	0.8078749	0.2557852
A07_Skin	6.7819686	5.7373124	7.4380083	5.9645037	0.4077480	0.7195721	0	7.1420826	0.5810337	6.2882711	6.5978689	7.0587950	6.2949696	0.7078739	0.3958918	0.6866299	0.6778929	0.0132670	0.0883028	0.4637868
A08_Breas	0.3601139	1.4047701	0.2959257	1.1775789	7.5498307	7.8616548	7.1420826	0	7.7231164	0.8538114	0.5442137	0.0832876	0.8471130	7.8499566	7.5379745	7.8287125	7.8199755	7.1288156	7.0537798	7.6058695
A09.3978_	7.3630024	6.3183462	8.0190421	6.5455374	0.1732856	0.1385383	0.5810337	7.7231164	0	6.8693049	7.1789027	7.6398288	6.8760034	0.1268401	0.1851418	0.1055961	0.0968591	0.5943007	0.6693365	0.1172468
A09_Breas	0.4936974	0.5509587	1.1497371	0.3237674	6.6960192	7.0078433	6.2882711	0.8538114	6.8693049	0	0.3095977	0.7705238	0.0066984	6.9961451	6.6841630	6.9749011	6.9661640	6.2750041	6.1999683	6.7520580
A10_Breas	0.1840997	0.8605564	0.8401394	0.6333652	7.0056170	7.3174411	6.5978689	0.5442137	7.1789027	0.3095977	0	0.4609260	0.3028992	7.3057429	6.9937608	7.2844988	7.2757618	6.5846019	6.5095661	7.0616558
A11_Breas	0.2768263	1.3214825	0.3792133	1.0942913	7.4665431	7.7783672	7.0587950	0.0832876	7.6398288	0.7705238	0.4609260	0	0.7638253	7.7666690	7.4546869	7.7454249	7.7366879	7.0455280	6.9704922	7.5225819
A12_Breas	0.4869990	0.5576571	1.1430387	0.3304659	6.7027177	7.0145418	6.2949696	0.8471130	6.8760034	0.0066984	0.3028992	0.7638253	0	7.0028436	6.6908615	6.9815995	6.9728625	6.2817026	6.2066668	6.7587565
A12_Skin	7.4898426	6.4451864	8.1458823	6.6723776	0.3001258	0.0116981	0.7078739	7.8499566	0.1268401	6.9961451	7.3057429	7.7666690	7.0028436	0	0.3119820	0.0212440	0.0299810	0.7211409	0.7961767	0.2440870
B02_Skin	7.1778605	6.1332043	7.8339002	6.3603955	0.0118562	0.3236802	0.3958918	7.5379745	0.1851418	6.6841630	6.9937608	7.4546869	6.6908615	0.3119820	0	0.2907380	0.2820010	0.4091588	0.4841947	0.0678950
B09_Skin	7.4685986	6.4239424	8.1246383	6.6511336	0.2788818	0.0329422	0.6866299	7.8287125	0.1055961	6.9749011	7.2844988	7.7454249	6.9815995	0.0212440	0.2907380	0	0.0087370	0.6998969	0.7749327	0.2228430
D07_Skin	7.4598615	6.4152053	8.1159012	6.6423966	0.2701448	0.0416792	0.6778929	7.8199755	0.0968591	6.9661640	7.2757618	7.7366879	6.9728625	0.0299810	0.2820010	0.0087370	0	0.6911599	0.7661957	0.2141060
G02_Skin	6.7687016	5.7240454	7.4247413	5.9512367	0.4210150	0.7328391	0.0132670	7.1288156	0.5943007	6.2750041	6.5846019	7.0455280	6.2817026	0.7211409	0.4091588	0.6998969	0.6911599	0	0.0750358	0.4770538
H06_Skin	6.6936658	5.6490096	7.3497055	5.8762008	0.4960509	0.8078749	0.0883028	7.0537798	0.6693365	6.1999683	6.5095661	6.9704922	6.2066668	0.7961767	0.4841947	0.7749327	0.7661957	0.0750358	0	0.5520897
H10_Skin	7.2457555	6.2010993	7.9017952	6.4282906	0.0560388	0.2557852	0.4637868	7.6058695	0.1172468	6.7520580	7.0616558	7.5225819	6.7587565	0.2440870	0.0678950	0.2228430	0.2141060	0.4770538	0.5520897	0

Gene Sonda com ID - 11724186_a_at

Matriz de distâncias de similaridade de Dice

	A01.5500_	A01_Breas	A03_Breas	A05_Breas	A05_Skin	A06_Skin	A07_Skin	A08_Breas	A09.3978_	A09_Breas	A10_Breas	A11_Breas	A12_Breas	A12_Skin	B02_Skin	B09_Skin	D07_Skin	G02_Skin	H06_Skin	H10_Skin
A01.5500_	0	0.0056555	0.0018923	0.0033881	0.4433516	0.4904744	0.3854717	0.0005867	0.4692408	0.0011977	0.0001616	0.0003495	0.0011647	0.4886633	0.4416079	0.4853829	0.4840370	0.3836616	0.3735128	0.4516413
A01_Breas	0.0056555	0	0.0139764	0.0002920	0.3929398	0.4424007	0.3330694	0.0098407	0.4200404	0.0016587	0.0039143	0.0087861	0.0016980	0.4404891	0.3911213	0.4370286	0.4356096	0.3312147	0.3208358	0.4015973
A03_Breas	0.0018923	0.0139764	0	0.0102795	0.4714071	0.5169943	0.4150120	0.0003726	0.4964856	0.0060781	0.0031557	0.0006165	0.0060039	0.5152471	0.4697150	0.5120813	0.5107821	0.4132403	0.4032968	0.4794463
A05_Breas	0.0033881	0.0002920	0.0102795	0	0.4045520	0.4535262	0.3450576	0.0067708	0.4314037	0.0005597	0.0020735	0.0058980	0.0005827	0.4516360	0.4027485	0.4482138	0.4468102	0.3432102	0.3328675	0.4131348
A05_Skin	0.4433516	0.3929398	0.4714071	0.4045520	0	0.0054406	0.0074019	0.4590697	0.0016079	0.4204614	0.4350027	0.4555047	0.4207828	0.0050214	0.7964977	0.0043066	0.0040298	0.0078584	0.0106540	0.0001620
A06_Skin	0.4904744	0.4424007	0.5169943	0.4535262	0.0054406	0	0.0251353	0.5053513	0.0011415	0.4687170	0.4825517	0.5019814	0.4690233	8.5122577	0.0058378	6.6996445	0.0001069	0.0259519	0.0307361	0.0037337
A07_Skin	0.3854717	0.3330694	0.4150120	0.3450576	0.0074019	0.0251353	0	0.4019906	0.0157735	0.3615641	0.3767308	0.3982369	0.3618986	0.0242466	0.0069546	0.0226802	0.0220535	0.4990378	0.3901608	0.0097273
A08_Breas	0.0005867	0.0098407	0.0003726	0.0067708	0.4590697	0.5053513	0.4019906	0	0.4845159	0.0034538	0.0013632	3.0582896	0.0033977	0.5035754	0.4573542	0.5003580	0.4990378	0.4002010	0.3901608	0.4672229
A09.3978_	0.4692408	0.4200404	0.4964856	0.4314037	0.0016079	0.0011415	0.0157735	0.4845159	0	0.4469420	0.4611152	0.4810539	0.4472556	0.0009531	0.0018281	0.0006558	0.0005502	0.0164296	0.0203295	0.0007501
A09_Breas	0.0011977	0.0016587	0.0060781	0.0005597	0.4204614	0.4687170	0.3615641	0.0034538	0.4469420	0	0.0004799	0.0028367	2.3167828	0.4668579	0.4186807	0.4634913	0.4621103	0.3597298	0.3494533	0.4289320
A10_Breas	0.0001616	0.0039143	0.0031557	0.0020735	0.4350027	0.4825517	0.3767308	0.0013632	0.4611152	0.0004799	0	0.0009859	0.0004590	0.4807227	0.4332450	0.4774100	0.4760510	0.3749111	0.3647114	0.4433609
A11_Breas	0.0003495	0.0087861	0.0006165	0.0058980	0.4555047	0.5019814	0.3982369	0.3582896	0.4810539	0.0028367	0.0009859	0	0.0027858	0.5001974	0.4537826	0.4969655	0.4956392	0.3964424	0.3863761	0.4636897
A12_Breas	0.0011647	0.0016980	0.0060039	0.0005827	0.4207828	0.4690233	0.3618986	0.0033977	0.4472556	2.3167828	0.0004590	0.0027858	0	0.4671648	0.4190027	0.4637994	0.4624189	0.3600645	0.3497896	0.4292510
A12_Skin	0.4886633	0.4404891	0.5152471	0.4516360	0.0050214	8.5122577	0.0242466	0.5035754	0.0009531	0.4668579	0.4807227	0.5001974	0.4671648	0	0.0054035	2.7748826	5.5097183	0.0250497	0.0297594	0.0033871
B02_Skin	0.4416079	0.3911213	0.4697150	0.4027485	7.0964977	0.0058378	0.0069546	0.4573542	0.0018281	0.4186807	0.4332450	0.4537826	0.4190027	0.0054035	0	0.0046613	0.0043732	0.0073975	0.0101181	0.0002368
B09_Skin	0.4853829	0.4370286	0.5120813	0.4482138	0.0043066	6.6996445	0.0226802	0.5003580	0.0006558	0.4634913	0.4774100	0.4969655	0.4637994	2.7748826	0.0046613	0	4.6446176	0.0234586	0.0280330	0.0028039
D07_Skin	0.4840370	0.4356096	0.5107821	0.4468102	0.0040298	0.0001069	0.0220535	0.4990378	0.0005502	0.4621103	0.4760510	0.4956392	0.4624189	5.5097183	0.0043732	4.6446176	0	0.0228218	0.0273405	0.0025810
G02_Skin	0.3836616	0.3312147	0.4132403	0.3432102	0.0078584	0.0259519	6.9615262	0.4002010	0.0164296	0.3597298	0.3749111	0.3964424	0.3600645	0.0250497	0.0073975	0.0234586	0.0228218	0	0.0002172	0.0102481
H06_Skin	0.3735128	0.3208358	0.4032968	0.3328675	0.0106540	0.0307361	0.0003019	0.3901608	0.0203295	0.3494533	0.3647114	0.3863761	0.3497896	0.0297594	0.0101181	0.0280330	0.0273405	0.0002172	0	0.0133995
H10_Skin	0.4516413	0.4015973	0.4794463	0.4131348	0.0001620	0.0037337	0.0097273	0.4672229	0.0007501	0.4289320	0.4433609	0.4636897	0.4292510	0.0033871	0.0002368	0.0028039	0.0025810	0.0102481	0.0133995	0

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Matriz de distâncias de similaridade de Jaccard

	A01.5500_	A01_Breas	A03_Breas	A05_Breas	A05_Skin	A06_Skin	A07_Skin	A08_Breas	A09.3978_	A09_Breas	A10_Breas	A11_Breas	A12_Breas	A12_Skin	B02_Skin	B09_Skin	D07_Skin	G02_Skin	H06_Skin	H10_Skin
A01.5500_	0	0.0055693	2.1146316	0.0062567	0.0008799	0.0012718	0.0240754	0.0034871	0.0219969	0.0012787	0.0001537	0.0010226	0.0090043	0.0003102	3.3090626	0.0013836	0.0086813	0.0184689	0.0007490	0.0052210
A01_Breas	0.0055693	0	0.0049088	2.0318615	0.0020341	0.0015294	0.0066964	0.0002455	0.0055980	0.0015218	0.0038824	0.0018310	0.0281879	0.0032624	0.0064534	0.0014118	0.0003506	0.0038609	0.0103451	5.7069742
A03_Breas	2.1146316	0.0049088	0	0.0055559	0.0006285	0.0009654	0.0227139	0.0029673	0.0206926	0.0009714	6.0843718	0.0007499	0.0098863	0.0001694	0.0001071	0.0010631	0.0078548	0.0172698	0.0010214	0.0045818
A05_Breas	0.0062567	2.0318615	0.0055559	0	0.0024597	0.0019014	0.0059851	0.0004071	0.0049483	0.0018929	0.0044610	0.0022359	0.0296604	0.0037950	0.0071908	0.0017701	0.0002021	0.0033236	0.0112682	4.7560923
A05_Skin	0.0008799	0.0020341	0.0006285	0.0024597	0	3.6088796	0.0159503	0.0008678	0.0142489	3.7275011	0.0002984	5.3700410	0.0153949	0.0001454	0.0012537	5.6917645	0.0040643	0.0114119	0.0032453	0.0018249
A06_Skin	0.0012718	0.0015294	0.0009654	0.0019014	3.6088796	0	0.0144991	0.0005502	0.0128761	9.5915214	0.0005418	1.3617065	0.0168872	0.0003263	0.0017142	2.3626121	0.0033379	0.0101825	0.0039621	0.0013486
A07_Skin	0.0240754	0.0066964	0.0227139	0.0059851	0.0159503	0.0144991	0	0.0094735	5.0009496	0.0144760	0.0204894	0.0153826	0.0600949	0.0190621	0.0258306	0.0141385	0.0039994	0.0003934	0.0329341	0.0070893
A08_Breas	0.0034871	0.0002455	0.0029673	0.0004071	0.0008678	0.0005502	0.0094735	0	0.0081623	0.0005456	0.0021809	0.0007367	0.0233388	0.0017220	0.0041956	0.0004805	0.0011821	0.0060378	0.0074353	0.0001764
A09.3978_	0.0219969	0.0055980	0.0206926	0.0049483	0.0142489	0.0128761	5.0009496	0.0081623	0	0.0128543	0.0185663	0.0137114	0.0569454	0.0172053	0.0236810	0.0125356	0.0031588	0.0001629	0.0305243	0.0059582
A09_Breas	0.0012787	0.0015218	0.0009714	0.0018929	3.7275011	9.5915214	0.0144760	0.0005456	0.0128543	0	0.0005464	1.4349438	0.0169121	0.0003298	0.0017223	2.2711327	0.0033266	0.0101630	0.0039744	0.0013415
A10_Breas	0.0001537	0.0038824	6.0843718	0.0044610	0.0002984	0.0005418	0.0204894	0.0021809	0.0185663	0.0005464	0	0.0003837	0.0114739	2.7215529	0.0003293	0.0006157	0.0065454	0.0153235	0.0015798	0.0035919
A11_Breas	0.0010226	0.0018310	0.0007499	0.0022359	5.3700410	1.3617065	0.0153826	0.0007367	0.0137114	1.4349438	0.0003837	0	0.0159627	0.0002066	0.0014228	2.7323297	0.0037757	0.0109296	0.0035134	0.0016327
A12_Breas	0.0090043	0.0281879	0.0098863	0.0296604	0.0153949	0.0168872	0.0600949	0.0233388	0.0569454	0.0169121	0.0114739	0.0159627	0	0.0125994	0.0079576	0.0172796	0.0345288	0.0514408	0.0045921	0.0274215
A12_Skin	0.0003102	0.0032624	0.0001694	0.0037950	0.0001454	0.0003263	0.0190621	0.0017220	0.0172053	0.0003298	2.7215529	0.0002066	0.0125994	0	0.0005458	0.0003842	0.0057352	0.0140840	0.0020206	0.0029963
B02_Skin	3.3090626	0.0064534	0.0001071	0.0071908	0.0012537	0.0017142	0.0258306	0.0041956	0.0236810	0.0017223	0.0003293	0.0014228	0.0079576	0.0005458	0	0.0018435	0.0097720	0.0200227	0.0004674	0.0060785
B09_Skin	0.0013836	0.0014118	0.0010631	0.0017701	5.6917645	2.3626121	0.0141385	0.0004805	0.0125356	2.2711327	0.0006157	2.7323297	0.0172796	0.0003842	0.0018435	0	0.0031634	0.0098788	0.0041569	0.0012383
D07_Skin	0.0086813	0.0003506	0.0078548	0.0002021	0.0040643	0.0033379	0.0039994	0.0011821	0.0031588	0.0033266	0.0065454	0.0037757	0.0345288	0.0057352	0.0097720	0.0031634	0	0.0018906	0.0144292	0.0004457
G02_Skin	0.0184689	0.0038609	0.0172698	0.0033236	0.0114119	0.0101825	0.0003934	0.0060378	0.0001629	0.0101630	0.0153235	0.0109296	0.0514408	0.0140840	0.0200227	0.0098788	0.0018906	0	0.0263874	0.0041617
H06_Skin	0.0007490	0.0103451	0.0010214	0.0112682	0.0032453	0.0039621	0.0329341	0.0074353	0.0305243	0.0039744	0.0015798	0.0035134	0.0045921	0.0020206	0.0004674	0.0041569	0.0144292	0.0263874	0	0.0098714
H10_Skin	0.0052210	5.7069742	0.0045818	4.7560923	0.0018249	0.0013486	0.0070893	0.0001764	0.0059582	0.0013415	0.0035919	0.0016327	0.0274215	0.0029963	0.0060785	0.0012383	0.0004457	0.0041617	0.0098714	0

Matriz de distâncias de Sorensen

	A01.5500_	A01_Breas	A03_Breas	A05_Breas	A05_Skin	A06_Skin	A07_Skin	A08_Breas	A09.3978_	A09_Breas	A10_Breas	A11_Breas	A12_Breas	A12_Skin	B02_Skin	B09_Skin	D07_Skin	G02_Skin	H06_Skin	H10_Skin
A01.5500_	0	0.0373923	0.0022992	0.0396428	0.0148368	0.0178397	0.0782914	0.0295645	0.0747761	0.0178887	0.0061994	0.0159952	0.0476066	0.0088077	0.0028762	0.0186080	0.0467393	0.0684258	0.0136877	0.0361994
A01_Breas	0.0373923	0	0.0350960	0.0022538	0.0225680	0.0195656	0.0410191	0.0078364	0.0374886	0.0195166	0.0312001	0.0214099	0.0848479	0.0285940	0.0402642	0.0187973	0.0093634	0.0311131	0.0510539	0.0011944
A03_Breas	0.0022992	0.0350960	0	0.0373469	0.0125379	0.0155411	0.0760058	0.0272671	0.0724893	0.0155900	0.0039002	0.0136964	0.0499004	0.0065085	0.0051754	0.0163094	0.0444448	0.0661370	0.0159865	0.0339030
A05_Breas	0.0396428	0.0022538	0.0373469	0	0.0248205	0.0218184	0.0387689	0.0100900	0.0352378	0.0217695	0.0334516	0.0236626	0.0870851	0.0308458	0.0425142	0.0210502	0.0071097	0.0288613	0.0533016	0.0034482
A05_Skin	0.0148368	0.0225680	0.0125379	0.0248205	0	0.0030037	0.0635284	0.0147341	0.0600059	0.0030527	0.0086381	0.0011586	0.0623994	0.0060298	0.0177123	0.0037722	0.0319246	0.0536435	0.0285188	0.0213741
A06_Skin	0.0178397	0.0195656	0.0155411	0.0218184	0.0030037	0	0.0605362	0.0117309	0.0570124	4.8968157	0.0116416	0.0018450	0.0653909	0.0090334	0.0207149	0.0007685	0.0289237	0.0506479	0.0315198	0.0183715
A07_Skin	0.0782914	0.0410191	0.0760058	0.0387689	0.0635284	0.0605362	0	0.0488399	0.0035359	0.0604874	0.0721270	0.0623743	0.1254306	0.0695316	0.0811494	0.0597704	0.0316679	0.0099186	0.0918807	0.0422115
A08_Breas	0.0295645	0.0078364	0.0272671	0.0100900	0.0147341	0.0117309	0.0488399	0	0.0453118	0.0116820	0.0233694	0.0135757	0.0770627	0.0207622	0.0324380	0.0109625	0.0171985	0.0389401	0.0432348	0.0066420
A09.3978_	0.0747761	0.0374886	0.0724893	0.0352378	0.0600059	0.0570124	0.0035359	0.0453118	0	0.0569636	0.0686085	0.0588513	0.1219487	0.0660119	0.0776357	0.0562464	0.0281351	0.0063829	0.0883735	0.0386814
A09_Breas	0.0178887	0.0195166	0.0155900	0.0217695	0.0030527	4.8968157	0.0604874	0.0116820	0.0569636	0	0.0116905	0.0018940	0.0654396	0.0090824	0.0207639	0.0007195	0.0288747	0.0505991	0.0315687	0.0183226
A10_Breas	0.0061994	0.0312001	0.0039002	0.0334516	0.0086381	0.0116416	0.0721270	0.0233694	0.0686085	0.0116905	0	0.0097967	0.0537902	0.0026084	0.0090755	0.0124100	0.0405517	0.0622528	0.0198855	0.0300067
A11_Breas	0.0159952	0.0214099	0.0136964	0.0236626	0.0011586	0.0018450	0.0623743	0.0135757	0.0588513	0.0018940	0.0097967	0	0.0635535	0.0071884	0.0188706	0.0026136	0.0307671	0.0524881	0.0296765	0.0202159
A12_Breas	0.0476066	0.0848479	0.0499004	0.0870851	0.0623994	0.0653909	0.1254306	0.0770627	0.1219487	0.0654396	0.0537902	0.0635535	0	0.0563907	0.0447365	0.0661561	0.0941365	0.1156558	0.0339410	0.0836619
A12_Skin	0.0088077	0.0285940	0.0065085	0.0308458	0.0060298	0.0090334	0.0695316	0.0207622	0.0660119	0.0090824	0.0026084	0.0071884	0.0563907	0	0.0116836	0.0098019	0.0379472	0.0596541	0.0224928	0.0274004
B02_Skin	0.0028762	0.0402642	0.0051754	0.0425142	0.0177123	0.0207149	0.0811494	0.0324380	0.0776357	0.0207639	0.0090755	0.0188706	0.0447365	0.0116836	0	0.0214831	0.0496089	0.0712881	0.0108119	0.0390716
B09_Skin	0.0186080	0.0187973	0.0163094	0.0210502	0.0037722	0.0007685	0.0597704	0.0109625	0.0562464	0.0007195	0.0124100	0.0026136	0.0661561	0.0098019	0.0214831	0	0.0281558	0.0498813	0.0322876	0.0176032
D07_Skin	0.0467393	0.0093634	0.0444448	0.0071097	0.0319246	0.0289237	0.0316679	0.0171985	0.0281351	0.0288747	0.0405517	0.0307671	0.0941365	0.0379472	0.0496089	0.0281558	0	0.0217561	0.0603885	0.0105577
G02_Skin	0.0684258	0.0311131	0.0661370	0.0288613	0.0536435	0.0506479	0.0099186	0.0389401	0.0063829	0.0505991	0.0622528	0.0524881	0.1156558	0.0596541	0.0712881	0.0498813	0.0217561	0	0.0820368	0.0323064
H06_Skin	0.0136877	0.0510539	0.0159865	0.0533016	0.0285188	0.0315198	0.0918807	0.0432348	0.0883735	0.0315687	0.0198855	0.0296765	0.0339410	0.0224928	0.0108119	0.0322876	0.0603885	0.0820368	0	0.0498625
H10_Skin	0.0361994	0.0011944	0.0339030	0.0034482	0.0213741	0.0183715	0.0422115	0.0066420	0.0386814	0.0183226	0.0300067	0.0202159	0.0836619	0.0274004	0.0390716	0.0176032	0.0105577	0.0323064	0.0498625	0

Matriz de distâncias de similaridade de Czekanowski

	A01.5500_	A01_Breas	A03_Breas	A05_Breas	A05_Skin	A06_Skin	A07_Skin	A08_Breas	A09.3978_	A09_Breas	A10_Breas	A11_Breas	A12_Breas	A12_Skin	B02_Skin	B09_Skin	D07_Skin	G02_Skin	H06_Skin	H10_Skin
A01.5500_	0	0.0373923	0.0022992	0.0396428	0.0148368	0.0178397	0.0782914	0.0295645	0.0747761	0.0178887	0.0061994	0.0159952	0.0476066	0.0088077	0.0028762	0.0186080	0.0467393	0.0684258	0.0136877	0.0361994
A01_Breas	0.0373923	0	0.0350960	0.0022538	0.0225680	0.0195656	0.0410191	0.0078364	0.0374886	0.0195166	0.0312001	0.0214099	0.0848479	0.0285940	0.0402642	0.0187973	0.0093634	0.0311131	0.0510539	0.0011944
A03_Breas	0.0022992	0.0350960	0	0.0373469	0.0125379	0.0155411	0.0760058	0.0272671	0.0724893	0.0155900	0.0039002	0.0136964	0.0499004	0.0065085	0.0051754	0.0163094	0.0444448	0.0661370	0.0159865	0.0339030
A05_Breas	0.0396428	0.0022538	0.0373469	0	0.0248205	0.0218184	0.0387689	0.0100900	0.0352378	0.0217695	0.0334516	0.0236626	0.0870851	0.0308458	0.0425142	0.0210502	0.0071097	0.0288613	0.0533016	0.0034482
A05_Skin	0.0148368	0.0225680	0.0125379	0.0248205	0	0.0030037	0.0635284	0.0147341	0.0600059	0.0030527	0.0086381	0.0011586	0.0623994	0.0060298	0.0177123	0.0037722	0.0319246	0.0536435	0.0285188	0.0213741
A06_Skin	0.0178397	0.0195656	0.0155411	0.0218184	0.0030037	0	0.0605362	0.0117309	0.0570124	4.8968157	0.0116416	0.0018450	0.0653909	0.0090334	0.0207149	0.0007685	0.0289237	0.0506479	0.0315198	0.0183715
A07_Skin	0.0782914	0.0410191	0.0760058	0.0387689	0.0635284	0.0605362	0	0.0488399	0.0035359	0.0604874	0.0721270	0.0623743	0.1254306	0.0695316	0.0811494	0.0597704	0.0316679	0.0099186	0.0918807	0.0422115
A08_Breas	0.0295645	0.0078364	0.0272671	0.0100900	0.0147341	0.0117309	0.0488399	0	0.0453118	0.0116820	0.0233694	0.0135757	0.0770627	0.0207622	0.0324380	0.0109625	0.0171985	0.0389401	0.0432348	0.0066420
A09.3978_	0.0747761	0.0374886	0.0724893	0.0352378	0.0600059	0.0570124	0.0035359	0.0453118	0	0.0569636	0.0686085	0.0588513	0.1219487	0.0660119	0.0776357	0.0562464	0.0281351	0.0063829	0.0883735	0.0386814
A09_Breas	0.0178887	0.0195166	0.0155900	0.0217695	0.0030527	4.8968157	0.0604874	0.0116820	0.0569636	0	0.0116905	0.0018940	0.0654396	0.0090824	0.0207639	0.0007195	0.0288747	0.0505991	0.0315687	0.0183226
A10_Breas	0.0061994	0.0312001	0.0039002	0.0334516	0.0086381	0.0116416	0.0721270	0.0233694	0.0686085	0.0116905	0	0.0097967	0.0537902	0.0026084	0.0090755	0.0124100	0.0405517	0.0622528	0.0198855	0.0300067
A11_Breas	0.0159952	0.0214099	0.0136964	0.0236626	0.0011586	0.0018450	0.0623743	0.0135757	0.0588513	0.0018940	0.0097967	0	0.0635535	0.0071884	0.0188706	0.0026136	0.0307671	0.0524881	0.0296765	0.0202159
A12_Breas	0.0476066	0.0848479	0.0499004	0.0870851	0.0623994	0.0653909	0.1254306	0.0770627	0.1219487	0.0654396	0.0537902	0.0635535	0	0.0563907	0.0447365	0.0661561	0.0941365	0.1156558	0.0339410	0.0836619
A12_Skin	0.0088077	0.0285940	0.0065085	0.0308458	0.0060298	0.0090334	0.0695316	0.0207622	0.0660119	0.0090824	0.0026084	0.0071884	0.0563907	0	0.0116836	0.0098019	0.0379472	0.0596541	0.0224928	0.0274004
B02_Skin	0.0028762	0.0402642	0.0051754	0.0425142	0.0177123	0.0207149	0.0811494	0.0324380	0.0776357	0.0207639	0.0090755	0.0188706	0.0447365	0.0116836	0	0.0214831	0.0496089	0.0712881	0.0108119	0.0390716
B09_Skin	0.0186080	0.0187973	0.0163094	0.0210502	0.0037722	0.0007685	0.0597704	0.0109625	0.0562464	0.0007195	0.0124100	0.0026136	0.0661561	0.0098019	0.0214831	0	0.0281558	0.0498813	0.0322876	0.0176032
D07_Skin	0.0467393	0.0093634	0.0444448	0.0071097	0.0319246	0.0289237	0.0316679	0.0171985	0.0281351	0.0288747	0.0405517	0.0307671	0.0941365	0.0379472	0.0496089	0.0281558	0	0.0217561	0.0603885	0.0105577
G02_Skin	0.0684258	0.0311131	0.0661370	0.0288613	0.0536435	0.0506479	0.0099186	0.0389401	0.0063829	0.0505991	0.0622528	0.0524881	0.1156558	0.0596541	0.0712881	0.0498813	0.0217561	0	0.0820368	0.0323064
H06_Skin	0.0136877	0.0510539	0.0159865	0.0533016	0.0285188	0.0315198	0.0918807	0.0432348	0.0883735	0.0315687	0.0198855	0.0296765	0.0339410	0.0224928	0.0108119	0.0322876	0.0603885	0.0820368	0	0.0498625
H10_Skin	0.0361994	0.0011944	0.0339030	0.0034482	0.0213741	0.0183715	0.0422115	0.0066420	0.0386814	0.0183226	0.0300067	0.0202159	0.0836619	0.0274004	0.0390716	0.0176032	0.0105577	0.0323064	0.0498625	0

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Matriz de distâncias de similaridade de Minkowski

	A01.5500_	A01_Breas	A03_Breas	A05_Breas	A05_Skin	A06_Skin	A07_Skin	A08_Breas	A09.3978_	A09_Breas	A10_Breas	A11_Breas	A12_Breas	A12_Skin	B02_Skin	B09_Skin	D07_Skin	G02_Skin	H06_Skin	H10_Skin
A01.5500_	0	0.5975106	0.0380276	0.6321009	0.2423543	0.2905466	1.2036054	0.4760185	1.1533241	0.2913298	0.1021347	0.2609784	0.8286241	0.1447311	0.0478172	0.3028307	0.7402022	1.0616518	0.2300512	0.5791153
A01_Breas	0.5975106	0	0.5594829	0.0345903	0.3551562	0.3069639	0.6060948	0.1214920	0.5558135	0.3061807	0.4953758	0.3365321	1.4261348	0.4527794	0.6453278	0.2946799	0.1426916	0.4641412	0.8275618	0.0183952
A03_Breas	0.0380276	0.5594829	0	0.5940733	0.2043267	0.2525189	1.1655777	0.4379909	1.1152965	0.2533022	0.0641071	0.2229507	0.8666518	0.1067034	0.0858449	0.2648030	0.7021745	1.0236242	0.2680788	0.5410877
A05_Breas	0.6321009	0.0345903	0.5940733	0	0.3897466	0.3415543	0.5715044	0.1560823	0.5212231	0.3407710	0.5299662	0.3711225	1.4607251	0.4873698	0.6799182	0.3292702	0.1081012	0.4295509	0.8621521	0.0529855
A05_Skin	0.2423543	0.3551562	0.2043267	0.3897466	0	0.0481922	0.9612510	0.2336642	0.9109697	0.0489755	0.1402196	0.0186240	1.0709785	0.0976232	0.2901716	0.0604763	0.4978478	0.8192975	0.4724055	0.3367610
A06_Skin	0.2905466	0.3069639	0.2525189	0.3415543	0.0481922	0	0.9130587	0.1854719	0.8627775	0.0007832	0.1884118	0.0295682	1.1191708	0.1458155	0.3383639	0.0122840	0.4496556	0.7711052	0.5205978	0.2885687
A07_Skin	1.2036054	0.6060948	1.1655777	0.5715044	0.9612510	0.9130587	0	0.7275868	0.0502812	0.9122755	1.1014706	0.9426269	2.0322296	1.0588742	1.2514227	0.9007747	0.4634031	0.1419535	1.4336566	0.6244900
A08_Breas	0.4760185	0.1214920	0.4379909	0.1560823	0.2336642	0.1854719	0.7275868	0	0.6773055	0.1846886	0.3738838	0.2150401	1.3046427	0.3312874	0.5238358	0.1731878	0.2641836	0.5856333	0.7060698	0.1030968
A09.3978_	1.1533241	0.5558135	1.1152965	0.5212231	0.9109697	0.8627775	0.0502812	0.6773055	0	0.8619942	1.0511894	0.8923457	1.9819483	1.0085930	1.2011414	0.8504934	0.4131219	0.0916722	1.3833753	0.5742087
A09_Breas	0.2913298	0.3061807	0.2533022	0.3407710	0.0489755	0.0007832	0.9122755	0.1846886	0.8619942	0	0.1891951	0.0303514	1.1199540	0.1465987	0.3391471	0.0115008	0.4488723	0.7703220	0.5213811	0.2877854
A10_Breas	0.1021347	0.4953758	0.0641071	0.5299662	0.1402196	0.1884118	1.1014706	0.3738838	1.0511894	0.1891951	0	0.1588436	0.9307589	0.0425963	0.1499520	0.2006959	0.6380674	0.9595171	0.3321859	0.4769806
A11_Breas	0.2609784	0.3365321	0.2229507	0.3711225	0.0186240	0.0295682	0.9426269	0.2150401	0.8923457	0.0303514	0.1588436	0	1.0896026	0.1162473	0.3087957	0.0418522	0.4792238	0.8006734	0.4910296	0.3181369
A12_Breas	0.8286241	1.4261348	0.8666518	1.4607251	1.0709785	1.1191708	2.0322296	1.3046427	1.9819483	1.1199540	0.9307589	1.0896026	0	0.9733553	0.7808069	1.1314548	1.5688264	1.8902760	0.5985729	1.4077395
A12_Skin	0.1447311	0.4527794	0.1067034	0.4873698	0.0976232	0.1458155	1.0588742	0.3312874	1.0085930	0.1465987	0.0425963	0.1162473	0.9733553	0	0.1925484	0.1580995	0.5954711	0.9169207	0.3747823	0.4343842
B02_Skin	0.0478172	0.6453278	0.0858449	0.6799182	0.2901716	0.3383639	1.2514227	0.5238358	1.2011414	0.3391471	0.1499520	0.3087957	0.7808069	0.1925484	0	0.3506479	0.7880195	1.1094691	0.1822339	0.6269326
B09_Skin	0.3028307	0.2946799	0.2648030	0.3292702	0.0604763	0.0122840	0.9007747	0.1731878	0.8504934	0.0115008	0.2006959	0.0418522	1.1314548	0.1580995	0.3506479	0	0.4373715	0.7588211	0.5328819	0.2762846
D07_Skin	0.7402022	0.1426916	0.7021745	0.1081012	0.4978478	0.4496556	0.4634031	0.2641836	0.4131219	0.4488723	0.6380674	0.4792238	1.5688264	0.5954711	0.7880195	0.4373715	0	0.3214496	0.9702534	0.1610868
G02_Skin	1.0616518	0.4641412	1.0236242	0.4295509	0.8192975	0.7711052	0.1419535	0.5856333	0.0916722	0.7703220	0.9595171	0.8006734	1.8902760	0.9169207	1.1094691	0.7588211	0.3214496	0	1.2917031	0.4825365
H06_Skin	0.2300512	0.8275618	0.2680788	0.8621521	0.4724055	0.5205978	1.4336566	0.7060698	1.3833753	0.5213811	0.3321859	0.4910296	0.5985729	0.3747823	0.1822339	0.5328819	0.9702534	1.2917031	0	0.8091666
H10_Skin	0.5791153	0.0183952	0.5410877	0.0529855	0.3367610	0.2885687	0.6244900	0.1030968	0.5742087	0.2877854	0.4769806	0.3181369	1.4077395	0.4343842	0.6269326	0.2762846	0.1610868	0.4825365	0.8091666	0

Matriz de distâncias de similaridade de Pearson

	A01.5500_	A01_Breas	A03_Breas	A05_Breas	A05_Skin	A06_Skin	A07_Skin	A08_Breas	A09.3978_	A09_Breas	A10_Breas	A11_Breas	A12_Breas	A12_Skin	B02_Skin	B09_Skin	D07_Skin	G02_Skin	H06_Skin	H10_Skin
A01.5500_	0	0.0464204	0.0001752	0.0521853	0.0072998	0.0105548	0.2044724	0.0290040	0.1864223	0.0106128	0.0012742	0.0084845	0.0753107	0.0025721	0.0002742	0.0114838	0.0725857	0.1559607	0.0062127	0.0435021
A01_Breas	0.0464204	0	0.0379397	0.0001562	0.0156765	0.0117813	0.0518498	0.0018893	0.0432965	0.0117224	0.0299763	0.0141081	0.2230813	0.0251737	0.0499558	0.0108740	0.0026974	0.0298092	0.0803961	4.3892540
A03_Breas	0.0001752	0.0379397	0	0.0460951	0.0051887	0.0079727	0.1917560	0.0245550	0.1743315	0.0080230	0.0005020	0.0061920	0.0823818	0.0013980	0.0008840	0.0087808	0.0653192	0.1449880	0.0084364	0.0379765
A05_Breas	0.0521853	0.0001562	0.0460951	0	0.0188788	0.0145861	0.0461005	0.0031183	0.0380752	0.0145207	0.0343087	0.0171574	0.2340340	0.0291669	0.0554547	0.0135766	0.0015481	0.0255317	0.0872573	0.0003641
A05_Skin	0.0072998	0.0156765	0.0051887	0.0188788	0	0.0002903	0.1304187	0.0069886	0.1163062	0.0002999	0.0024017	4.3208374	0.1258066	0.0011702	0.0101003	0.0004579	0.0328355	0.0928825	0.0261977	0.0147103
A06_Skin	0.0105548	0.0117813	0.0079727	0.0145861	0.0002903	0	0.1176695	0.0044031	0.1043260	7.6712561	0.0043363	0.0001089	0.1373835	0.0026108	0.0137338	1.8896110	0.0267861	0.0822769	0.0318155	0.0108013
A07_Skin	0.2044724	0.0518498	0.1917560	0.0461005	0.1304187	0.1176695	0	0.0677611	0.0003543	0.1040676	0.1482022	0.1106873	0.4529889	0.1376775	0.1878597	0.1016063	0.0284491	0.0027883	0.2412817	0.0505861
A08_Breas	0.0290040	0.0018893	0.0245550	0.0031183	0.0069886	0.0044031	0.0677611	0	0.0642931	0.0042652	0.0170758	0.0057604	0.1866918	0.0134767	0.0329166	0.0037559	0.0092461	0.0474572	0.0585233	0.0013787
A09.3978_	0.1864223	0.0432965	0.1743315	0.0380752	0.1163062	0.1043260	0.0003543	0.0642931	0	0.0929121	0.1349804	0.0991938	0.4308505	0.1249126	0.1730668	0.0905796	0.0226103	0.0011628	0.2246540	0.0427680
A09_Breas	0.0106128	0.0117224	0.0080230	0.0145207	0.0002999	7.6712561	0.1040676	0.0042652	0.0929121	0	0.0043724	0.0001147	0.1375759	0.0026389	0.0137975	1.6563239	0.0266929	0.0821099	0.0319113	0.0107428
A10_Breas	0.0012742	0.0299763	0.0005020	0.0343087	0.0024017	0.0043363	0.1482022	0.0170758	0.1349804	0.0043724	0	0.0031431	0.0950203	0.0002228	0.0026973	0.0050438	0.0539366	0.1273961	0.0129537	0.0295108
A11_Breas	0.0084845	0.0141081	0.0061920	0.0171574	4.3208374	0.0001089	0.1106873	0.0057604	0.0991938	0.0001147	0.0031431	0	0.1302201	0.0016593	0.0114384	0.0002193	0.0304247	0.0887078	0.0283041	0.0131282
A12_Breas	0.0753107	0.2230813	0.0823818	0.2340340	0.1258066	0.1373835	0.4529889	0.1866918	0.4308505	0.1375759	0.0950203	0.1302201	0	0.1163368	0.0731329	0.1603109	0.3260623	0.4944261	0.0420599	0.2570542
A12_Skin	0.0025721	0.0251737	0.0013980	0.0291669	0.0011702	0.0026108	0.1376775	0.0134767	0.1249126	0.0026389	0.0002228	0.0016593	0.1163368	0	0.0044473	0.0031300	0.0469755	0.1163361	0.0164889	0.0244753
B02_Skin	0.0002742	0.0499558	0.0008840	0.0554547	0.0101003	0.0137338	0.1878597	0.0329166	0.1730668	0.0137975	0.0026973	0.0114384	0.0731329	0.0044473	0	0.0153968	0.0822668	0.1703261	0.0038984	0.0509825
B09_Skin	0.0114838	0.0108740	0.0087808	0.0135766	0.0004579	1.8896110	0.1016063	0.0037559	0.0905796	1.6563239	0.0050438	0.0002193	0.1603109	0.0031300	0.0153968	0	0.0253426	0.0796764	0.0333346	0.0099013
D07_Skin	0.0725857	0.0026974	0.0653192	0.0015481	0.0328355	0.0267861	0.0284491	0.0092461	0.0226103	0.0266929	0.0539366	0.0304247	0.3260623	0.0469755	0.0822668	0.0253426	0	0.0142980	0.1105107	0.0033658
G02_Skin	0.1559607	0.0298092	0.1449880	0.0255317	0.0928825	0.0822769	0.0027883	0.0474572	0.0011628	0.0821099	0.1273961	0.0887078	0.4944261	0.1163361	0.1703261	0.0796764	0.0142980	0	0.1958662	0.0302023
H06_Skin	0.0062127	0.0803961	0.0084364	0.0872573	0.0261977	0.0318155	0.2412817	0.0585233	0.2246540	0.0319113	0.0129537	0.0283041	0.0420599	0.0164889	0.0038984	0.0333346	0.1105107	0.1958662	0	0.0849290
H10_Skin	0.0435021	4.3892540	0.0379765	0.0003641	0.0147103	0.0108013	0.0505861	0.0013787	0.0427680	0.0107428	0.0295108	0.0131282	0.2570542	0.0244753	0.0509825	0.0099013	0.0033658	0.0302023	0.0849290	0

Matriz de distâncias de similaridade de Intersection

	A01.5500_	A01_Breas	A03_Breas	A05_Breas	A05_Skin	A06_Skin	A07_Skin	A08_Breas	A09.3978_	A09_Breas	A10_Breas	A11_Breas	A12_Breas	A12_Skin	B02_Skin	B09_Skin	D07_Skin	G02_Skin	H06_Skin	H10_Skin
A01.5500_	8.2885019	7.6909913	8.2504743	7.6564009	8.0461475	7.9979553	7.0848965	7.8124833	7.1351777	7.9971720	8.1863671	8.0275235	8.2885019	8.1437708	8.2885019	7.9856712	7.5482997	7.2268500	8.2885019	7.7093865
A01_Breas	7.6909913	7.6909913	7.6909913	7.6564009	7.6909913	7.6909913	7.0848965	7.6909913	7.1351777	7.6909913	7.6909913	7.6909913	7.6909913	7.6909913	7.6909913	7.6909913	7.5482997	7.2268500	7.6909913	7.6909913
A03_Breas	8.2504743	7.6909913	8.2504743	7.6564009	8.0461475	7.9979553	7.0848965	7.8124833	7.1351777	7.9971720	8.1863671	8.0275235	8.2504743	8.1437708	8.2504743	7.9856712	7.5482997	7.2268500	8.2504743	7.7093865
A05_Breas	7.6564009	7.6564009	7.6564009	7.6564009	7.6564009	7.6564009	7.0848965	7.6564009	7.1351777	7.6564009	7.6564009	7.6564009	7.6564009	7.6564009	7.6564009	7.5482997	7.2268500	7.6564009	7.6564009	7.6564009
A05_Skin	8.0461475	7.6909913	8.0461475	7.6564009	8.0461475	7.9979553	7.0848965	7.8124833	7.1351777	7.9971720	8.0461475	8.0275235	8.0461475	8.0461475	8.0461475	7.9856712	7.5482997	7.2268500	8.0461475	7.7093865
A06_Skin	7.9979553	7.6909913	7.9979553	7.6564009	7.9979553	7.9979553	7.0848965	7.8124833	7.1351777	7.9971720	7.9979553	7.9979553	7.9979553	7.9979553	7.9979553	7.9856712	7.5482997	7.2268500	7.9979553	7.7093865
A07_Skin	7.0848965	7.0848965	7.0848965	7.0848965	7.0848965	7.0848965	7.0848965	7.0848965	7.0848965	7.0848965	7.0848965	7.0848965	7.0848965	7.0848965	7.0848965	7.0848965	7.0848965	7.0848965	7.0848965	7.0848965
A08_Breas	7.8124833	7.6909913	7.8124833	7.6564009	7.8124833	7.8124833	7.0848965	7.8124833	7.1351777	7.8124833	7.8124833	7.8124833	7.8124833	7.8124833	7.8124833	7.8124833	7.5482997	7.2268500	7.8124833	7.7093865
A09.3978_	7.1351777	7.1351777	7.1351777	7.1351777	7.1351777	7.1351777	7.0848965	7.1351777	7.1351777	7.1351777	7.1351777	7.1351777	7.1351777	7.1351777	7.1351777	7.1351777	7.1351777	7.1351777	7.1351777	7.1351777
A09_Breas	7.9971720	7.6909913	7.9971720	7.6564009	7.9971720	7.9971720	7.0848965	7.8124833	7.1351777	7.9971720	7.9971720	7.9971720	7.9971720	7.9971720	7.9971720	7.9856712	7.5482997	7.2268500	7.9971720	7.7093865
A10_Breas	8.1863671	7.6909913	8.1863671	7.6564009	8.0461475	7.9979553	7.0848965	7.8124833	7.1351777	7.9971720	8.1863671	8.0275235	8.1863671	8.1437708	8.1863671	7.9856712	7.5482997	7.2268500	8.1863671	7.7093865
A11_Breas	8.0275235	7.6909913	8.0275235	7.6564009	8.0275235	7.9979553	7.0848965	7.8124833	7.1351777	7.9971720	8.0275235	8.0275235	8.0275235	8.0275235	8.0275235	7.9856712	7.5482997	7.2268500	8.0275235	7.7093865
A12_Breas	8.2885019	7.6909913	8.2504743	7.6564009	8.0461475	7.9979553	7.0848965	7.8124833	7.1351777	7.9971720	8.1863671	8.0275235	9.1171261	8.1437708	8.3363192	7.9856712	7.5482997	7.2268500	8.5185531	7.7093865
A12_Skin	8.1437708	7.6909913	8.1437708	7.6564009	8.0461475	7.9979553	7.0848965	7.8124833	7.1351777	7.9971720	8.1437708	8.0275235	8.1437708	8.1437708	8.1437708	7.9856712	7.5482997	7.2268500	8.1437708	7.7093865
B02_Skin	8.2885019	7.6909913	8.2504743	7.6564009	8.0461475	7.9979553	7.0848965	7.8124833	7.1351777	7.9971720	8.1863671	8.0275235	8.3363192	8.1437708	8.3363192	7.9856712	7.5482997	7.2268500	8.3363192	7.7093865
B09_Skin	7.9856712	7.6909913	7.9856712	7.6564009	7.9856712	7.9856712	7.0848965	7.8124833	7.1351777	7.9856712	7.9856712	7.9856712	7.9856712	7.9856712	7.9856712	7.9856712	7.5482997	7.2268500	7.9856712	7.7093865
D07_Skin	7.5482997	7.5482997	7.5482997	7.5482997	7.5482997	7.5482997	7.0848965	7.5482997	7.1351777	7.5482997	7.5482997	7.5482997	7.5482997	7.5482997	7.5482997	7.5482997	7.5482997	7.2268500	7.5482997	7.5482997
G02_Skin	7.2268500	7.2268500	7.2268500	7.2268500	7.2268500	7.2268500	7.0848965	7.2268500	7.1351777	7.2268500	7.2268500	7.2268500	7.2268500	7.2268500	7.2268500	7.2268500	7.2268500	7.2268500	7.2268500	7.2268500
H06_Skin	8.2885019	7.6909913	8.2504743	7.6564009	8.0461475	7.9979553	7.0848965	7.8124833	7.1351777	7.9971720	8.1863671	8.0275235	8.5185531	8.1437708	8.3363192	7.9856712	7.5482997	7.2268500	8.5185531	7.7093865
H10_Skin	7.7093865	7.6909913	7.7093865	7.6564009	7.7093865	7.7093865	7.0848965	7.7093865	7.1351777	7.7093865	7.7093865	7.7093865	7.7093865	7.7093865	7.7093865	7.7093865	7.5482997	7.2268500	7.7093865	7.7093865

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Matriz de distâncias de similaridade de Manhattan

	A01.5500_	A01_Breas	A03_Breas	A05_Breas	A05_Skin	A06_Skin	A07_Skin	A08_Breas	A09.3978_	A09_Breas	A10_Breas	A11_Breas	A12_Breas	A12_Skin	B02_Skin	B09_Skin	D07_Skin	G02_Skin	H06_Skin	H10_Skin
A01.5500_	0	0.5975106	0.0380276	0.6321009	0.2423543	0.2905466	1.2036054	0.4760185	1.1533241	0.2913298	0.1021347	0.2609784	0.8286241	0.1447311	0.0478172	0.3028307	0.7402022	1.0616518	0.2300512	0.5791153
A01_Breas	0.5975106	0	0.5594829	0.0345903	0.3551562	0.3069639	0.6060948	0.1214920	0.5558135	0.3061807	0.4953758	0.3365321	1.4261348	0.4527794	0.6453278	0.2946799	0.1426916	0.4641412	0.8275618	0.0183952
A03_Breas	0.0380276	0.5594829	0	0.5940733	0.2043267	0.2525189	1.1655777	0.4379909	1.1152965	0.2533022	0.0641071	0.2229507	0.8666518	0.1067034	0.0858449	0.2648030	0.7021745	1.0236242	0.2680788	0.5410877
A05_Breas	0.6321009	0.0345903	0.5940733	0	0.3897466	0.3415543	0.5715044	0.1560823	0.5212231	0.3407710	0.5299662	0.3711225	1.4607251	0.4873698	0.6799182	0.3292702	0.1081012	0.4295509	0.8621521	0.0529855
A05_Skin	0.2423543	0.3551562	0.2043267	0.3897466	0	0.0481922	0.9612510	0.2336642	0.9109697	0.0489755	0.1402196	0.0186240	1.0709785	0.0976232	0.2901716	0.0604763	0.4978478	0.8192975	0.4724055	0.3367610
A06_Skin	0.2905466	0.3069639	0.2525189	0.3415543	0.0481922	0	0.9130587	0.1854719	0.8627775	0.0007832	0.1884118	0.0295682	1.1191708	0.1458155	0.3383639	0.0122840	0.4496556	0.7711052	0.5205978	0.2885687
A07_Skin	1.2036054	0.6060948	1.1655777	0.5715044	0.9612510	0.9130587	0	0.7275868	0.0502812	0.9122755	1.1014706	0.9426269	2.0322296	1.0588742	1.2514227	0.9007747	0.4634031	0.1419535	1.4336566	0.6244900
A08_Breas	0.4760185	0.1214920	0.4379909	0.1560823	0.2336642	0.1854719	0.7275868	0	0.6773055	0.1846886	0.3738838	0.2150401	1.3046427	0.3312874	0.5238358	0.1731878	0.2641836	0.5856333	0.7060698	0.1030968
A09.3978_	1.1533241	0.5558135	1.1152965	0.5212231	0.9109697	0.8627775	0.0502812	0.6773055	0	0.8619942	1.0511894	0.8923457	1.9819483	1.0085930	1.2011414	0.8504934	0.4131219	0.0916722	1.3833753	0.5742087
A09_Breas	0.2913298	0.3061807	0.2533022	0.3407710	0.0489755	0.0007832	0.9122755	0.1846886	0.8619942	0	0.1891951	0.0303514	1.1199540	0.1465987	0.3391471	0.0115008	0.4488723	0.7703220	0.5213811	0.2877854
A10_Breas	0.1021347	0.4953758	0.0641071	0.5299662	0.1402196	0.1884118	1.1014706	0.3738838	1.0511894	0.1891951	0	0.1588436	0.9307589	0.0425963	0.1499520	0.2006959	0.6380674	0.9595171	0.3321859	0.4769806
A11_Breas	0.2609784	0.3365321	0.2229507	0.3711225	0.0186240	0.0295682	0.9426269	0.2150401	0.8923457	0.0303514	0.1588436	0	1.0896026	0.1162473	0.3087957	0.0418522	0.4792238	0.8006734	0.4910296	0.3181369
A12_Breas	0.8286241	1.4261348	0.8666518	1.4607251	1.0709785	1.1191708	2.0322296	1.3046427	1.9819483	1.1199540	0.9307589	1.0896026	0	0.9733553	0.7808069	1.1314548	1.5688264	1.8902760	0.5985729	1.4077395
A12_Skin	0.1447311	0.4527794	0.1067034	0.4873698	0.0976232	0.1458155	1.0588742	0.3312874	1.0085930	0.1465987	0.0425963	0.1162473	0.9733553	0	0.1925484	0.1580995	0.5954711	0.9169207	0.3747823	0.4343842
B02_Skin	0.0478172	0.6453278	0.0858449	0.6799182	0.2901716	0.3383639	1.2514227	0.5238358	1.2011414	0.3391471	0.1499520	0.3087957	0.7808069	0.1925484	0	0.3506479	0.7880195	1.1094691	0.1822339	0.6269326
B09_Skin	0.3028307	0.2946799	0.2648030	0.3292702	0.0604763	0.0122840	0.9007747	0.1731878	0.8504934	0.0115008	0.2006959	0.0418522	1.1314548	0.1580995	0.3506479	0	0.4373715	0.7588211	0.5328819	0.2762846
D07_Skin	0.7402022	0.1426916	0.7021745	0.1081012	0.4978478	0.4496556	0.4634031	0.2641836	0.4131219	0.4488723	0.6380674	0.4792238	1.5688264	0.5954711	0.7880195	0.4373715	0	0.3214496	0.9702534	0.1610868
G02_Skin	1.0616518	0.4641412	1.0236242	0.4295509	0.8192975	0.7711052	0.1419535	0.5856333	0.0916722	0.7703220	0.9595171	0.8006734	1.8902760	0.9169207	1.1094691	0.7588211	0.3214496	0	1.2917031	0.4825365
H06_Skin	0.2300512	0.8275618	0.2680788	0.8621521	0.4724055	0.5205978	1.4336566	0.7060698	1.3833753	0.5213811	0.3321859	0.4910296	0.5985729	0.3747823	0.1822339	0.5328819	0.9702534	1.2917031	0	0.8091666
H10_Skin	0.5791153	0.0183952	0.5410877	0.0529855	0.3367610	0.2885687	0.6244900	0.1030968	0.5742087	0.2877854	0.4769806	0.3181369	1.4077395	0.4343842	0.6269326	0.2762846	0.1610868	0.4825365	0.8091666	0

Matriz de distâncias de similaridade de Tanimoto

	A01.5500_	A01_Breas	A03_Breas	A05_Breas	A05_Skin	A06_Skin	A07_Skin	A08_Breas	A09.3978_	A09_Breas	A10_Breas	A11_Breas	A12_Breas	A12_Skin	B02_Skin	B09_Skin	D07_Skin	G02_Skin	H06_Skin	H10_Skin
A01.5500_	0	0.0720890	0.0045879	0.0762623	0.0292398	0.0350541	0.1452138	0.0574311	0.1391474	0.0351486	0.0123224	0.0314868	0.0908865	0.0174616	0.0057360	0.0365362	0.0893047	0.1280873	0.0270059	0.0698697
A01_Breas	0.0720890	0	0.0678122	0.0044975	0.0441399	0.0383803	0.0788058	0.0155510	0.0722681	0.0382861	0.0605122	0.0419222	0.1564237	0.0555982	0.0774116	0.0369010	0.0185530	0.0603486	0.0971481	0.0023860
A03_Breas	0.0045879	0.0678122	0	0.0720047	0.0247654	0.0306066	0.1412740	0.0530867	0.1351796	0.0307015	0.0077701	0.0270227	0.0950575	0.0129330	0.0102977	0.0320954	0.0851071	0.1240685	0.0314700	0.0655826
A05_Breas	0.0762623	0.0044975	0.0720047	0	0.0484389	0.0427052	0.0746440	0.0199785	0.0680767	0.0426114	0.0647376	0.0462312	0.1602177	0.0598457	0.0815609	0.0412326	0.0141190	0.0561035	0.1012087	0.0068728
A05_Skin	0.0292398	0.0441399	0.0247654	0.0484389	0	0.0059894	0.1194672	0.0290405	0.1132181	0.0060868	0.0171284	0.0023146	0.1174688	0.0119874	0.0348081	0.0075161	0.0618740	0.1018248	0.0554560	0.0418536
A06_Skin	0.0350541	0.0383803	0.0306066	0.0427052	0.0059894	0	0.1141615	0.0231899	0.1078747	9.7931518	0.0230153	0.0036833	0.1227547	0.0179051	0.0405891	0.0015359	0.0562213	0.0964127	0.0611134	0.0360803
A07_Skin	0.1452138	0.0788058	0.1412740	0.0746440	0.1194672	0.1141615	0	0.0931313	0.0070469	0.1140747	0.1345493	0.1174243	0.2229024	0.1300226	0.1501169	0.1127988	0.0613917	0.0196425	0.1682981	0.0810038
A08_Breas	0.0574311	0.0155510	0.0530867	0.0199785	0.0290405	0.0231899	0.0931313	0	0.0866952	0.0230942	0.0456715	0.0267878	0.1430980	0.0406798	0.0628377	0.0216873	0.0338155	0.0749612	0.0828861	0.0131964
A09.3978_	0.1391474	0.0722681	0.1351796	0.0680767	0.1132181	0.1078747	0.0070469	0.0866952	0	0.1077873	0.1284073	0.1111607	0.2173873	0.1238484	0.1440853	0.1065024	0.0547304	0.0126849	0.1623955	0.0744817
A09_Breas	0.0351486	0.0382861	0.0307015	0.0426114	0.0060868	9.7931518	0.1140747	0.0230942	0.1077873	0	0.0231110	0.0037809	0.1228406	0.0180013	0.0406830	0.0014381	0.0561288	0.0963243	0.0612053	0.0359859
A10_Breas	0.0123224	0.0605122	0.0077701	0.0647376	0.0171284	0.0230153	0.1345493	0.0456715	0.1284073	0.0231110	0	0.0194034	0.1020890	0.0052033	0.0179877	0.0245158	0.0779426	0.1172091	0.0389955	0.0582652
A11_Breas	0.0314868	0.0419222	0.0270227	0.0462312	0.0023146	0.0036833	0.1174243	0.0267878	0.1111607	0.0037809	0.0194034	0	0.1195116	0.0142743	0.0370422	0.0052135	0.0596975	0.0997410	0.0576423	0.0396307
A12_Breas	0.0908865	0.1564237	0.0950575	0.1602177	0.1174688	0.1227547	0.2229024	0.1430980	0.2173873	0.1228406	0.1020890	0.1195116	0	0.1067611	0.0856417	0.1241021	0.1720746	0.2073324	0.0656536	0.1544060
A12_Skin	0.0174616	0.0555982	0.0129330	0.0598457	0.0119874	0.0179051	0.1300226	0.0406798	0.1238484	0.0180013	0.0052033	0.0142743	0.1067611	0	0.0230975	0.0194135	0.0731198	0.1125916	0.0439960	0.0533394
B02_Skin	0.0057360	0.0774116	0.0102977	0.0815609	0.0348081	0.0405891	0.1501169	0.0628377	0.1440853	0.0406830	0.0179877	0.0370422	0.0856417	0.0230975	0	0.0420626	0.0945284	0.1330886	0.0213925	0.0752049
B09_Skin	0.0365362	0.0369010	0.0320954	0.0412326	0.0075161	0.0015359	0.1127988	0.0216873	0.1065024	0.0014381	0.0245158	0.0052135	0.1241021	0.0194135	0.0420626	0	0.0547695	0.0950228	0.0625554	0.0345975
D07_Skin	0.0893047	0.0185530	0.0851071	0.0141190	0.0618740	0.0562213	0.0613917	0.0338155	0.0547304	0.0561288	0.0779426	0.0596975	0.1720746	0.0731198	0.0945284	0.0547695	0	0.0425857	0.1138988	0.0208945
G02_Skin	0.1280873	0.0603486	0.1240685	0.0561035	0.1018248	0.0964127	0.0196425	0.0749612	0.0126849	0.0963243	0.1172091	0.0997410	0.2073324	0.1125916	0.1330886	0.0950228	0.0425857	0	0.1516340	0.0625907
H06_Skin	0.0270059	0.0971481	0.0314700	0.1012087	0.0554560	0.0611134	0.1682981	0.0828861	0.1623955	0.0612053	0.0389955	0.0576423	0.0656536	0.0439960	0.0213925	0.0625554	0.1138988	0.1516340	0	0.0949887
H10_Skin	0.0698697	0.0023860	0.0655826	0.0068728	0.0418536	0.0360803	0.0810038	0.0131964	0.0744817	0.0359859	0.0582652	0.0396307	0.1544060	0.0533394	0.0752049	0.0345975	0.0208949	0.0625907	0.0949887	0

Matriz de distâncias de similaridade de Euclideana

	A01.5500_	A01_Breas	A03_Breas	A05_Breas	A05_Skin	A06_Skin	A07_Skin	A08_Breas	A09.3978_	A09_Breas	A10_Breas	A11_Breas	A12_Breas	A12_Skin	B02_Skin	B09_Skin	D07_Skin	G02_Skin	H06_Skin	H10_Skin
A01.5500_	0	0.5975106	0.0380276	0.6321009	0.2423543	0.2905466	1.2036054	0.4760185	1.1533241	0.2913298	0.1021347	0.2609784	0.8286241	0.1447311	0.0478172	0.3028307	0.7402022	1.0616518	0.2300512	0.5791153
A01_Breas	0.5975106	0	0.5594829	0.0345903	0.3551562	0.3069639	0.6060948	0.1214920	0.5558135	0.3061807	0.4953758	0.3365321	1.4261348	0.4527794	0.6453278	0.2946799	0.1426916	0.4641412	0.8275618	0.0183952
A03_Breas	0.0380276	0.5594829	0	0.5940733	0.2043267	0.2525189	1.1655777	0.4379909	1.1152965	0.2533022	0.0641071	0.2229507	0.8666518	0.1067034	0.0858449	0.2648030	0.7021745	1.0236242	0.2680788	0.5410877
A05_Breas	0.6321009	0.0345903	0.5940733	0	0.3897466	0.3415543	0.5715044	0.1560823	0.5212231	0.3407710	0.5299662	0.3711225	1.4607251	0.4873698	0.6799182	0.3292702	0.1081012	0.4295509	0.8621521	0.0529855
A05_Skin	0.2423543	0.3551562	0.2043267	0.3897466	0	0.0481922	0.9612510	0.2336642	0.9109697	0.0489755	0.1402196	0.0186240	1.0709785	0.0976232	0.2901716	0.0604763	0.4978478	0.8192975	0.4724055	0.3367610
A06_Skin	0.2905466	0.3069639	0.2525189	0.3415543	0.0481922	0	0.9130587	0.1854719	0.8627775	0.0007832	0.1884118	0.0295682	1.1191708	0.1458155	0.3383639	0.0122840	0.4496556	0.7711052	0.5205978	0.2885687
A07_Skin	1.2036054	0.6060948	1.1655777	0.5715044	0.9612510	0.9130587	0	0.7275868	0.0502812	0.9122755	1.1014706	0.9426269	2.0322296	1.0588742	1.2514227	0.9007747	0.4634031	0.1419535	1.4336566	0.6244900
A08_Breas	0.4760185	0.1214920	0.4379909	0.1560823	0.2336642	0.1854719	0.7275868	0	0.6773055	0.1846886	0.3738838	0.2150401	1.3046427	0.3312874	0.5238358	0.1731878	0.2641836	0.5856333	0.7060698	0.1030968
A09.3978_	1.1533241	0.5558135	1.1152965	0.5212231	0.9109697	0.8627775	0.0502812	0.6773055	0	0.8619942	1.0511894	0.8923457	1.9819483	1.0085930	1.2011414	0.8504934	0.4131219	0.0916722	1.3833753	0.5742087
A09_Breas	0.2913298	0.3061807	0.2533022	0.3407710	0.0489755	0.0007832	0.9122755	0.1846886	0.8619942	0	0.1891951	0.0303514	1.1199540	0.1465987	0.3391471	0.0115008	0.4488723	0.7703220	0.5213811	0.2877854
A10_Breas	0.1021347	0.4953758	0.0641071	0.5299662	0.1402196	0.1884118	1.1014706	0.3738838	1.0511894	0.1891951	0	0.1588436	0.9307589	0.0425963	0.1499520	0.2006959	0.6380674	0.9595171	0.3321859	0.4769806
A11_Breas	0.2609784	0.3365321	0.2229507	0.3711225	0.0186240	0.0295682	0.9426269	0.2150401	0.8923457	0.0303514	0.1588436	0	1.0896026	0.1162473	0.3087957	0.0418522	0.4792238	0.8006734	0.4910296	0.3181369
A12_Breas	0.8286241	1.4261348	0.8666518	1.4607251	1.0709785	1.1191708	2.0322296	1.3046427	1.9819483	1.1199540	0.9307589	1.0896026	0	0.9733553	0.7808069	1.1314548	1.5688264	1.8902760	0.5985729	1.4077395
A12_Skin	0.1447311	0.4527794	0.1067034	0.4873698	0.0976232	0.1458155	1.0588742	0.3312874	1.0085930	0.1465987	0.0425963	0.1162473	0.9733553	0	0.1925484	0.1580995	0.5954711	0.9169207	0.3747823	0.4343842
B02_Skin	0.0478172	0.6453278	0.0858449	0.6799182	0.2901716	0.3383639	1.2514227	0.5238358	1.2011414	0.3391471	0.1499520	0.3087957	0.7808069	0.1925484	0	0.3506479	0.7880195	1.1094691	0.1822339	0.6269326
B09_Skin	0.3028307	0.2946799	0.2648030	0.3292702	0.0604763	0.0122840	0.9007747	0.1731878	0.8504934	0.0115008	0.2006959	0.0418522	1.1314548	0.1580995	0.3506479	0	0.4373715	0.7588211	0.5328819	0.2762846
D07_Skin	0.7402022	0.1426916	0.7021745	0.1081012	0.4978478	0.4496556	0.4634031	0.2641836	0.4131219	0.4488723	0.6380674	0.4792238	1.5688264	0.5954711	0.7880195	0.4373715	0	0.3214496	0.9702534	0.1610868
G02_Skin	1.0616518	0.4641412	1.0236242	0.4295509	0.8192975	0.7711052	0.1419535	0.5856333	0.0916722	0.7703220	0.9595171	0.8006734	1.8902760	0.9169207	1.1094691	0.7588211	0.3214496	0	1.2917031	0.4825365
H06_Skin	0.2300512	0.8275618	0.2680788	0.8621521	0.4724055	0.5205978	1.4336566	0.7060698	1.3833753	0.5213811	0.3321859	0.4910296	0.5985729	0.3747823	0.1822339	0.5328819	0.9702534	1.2917031	0	0.8091666
H10_Skin	0.5791153	0.0183952	0.5410877	0.0529855	0.3367610	0.2885687	0.6244900	0.1030968	0.5742087	0.2877854	0.4769806	0.3181369	1.4077395	0.4343842	0.6269326	0.2762846	0.1610868	0.4825365	0.8091666	0

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt