Instituto Superior de Contabilidade e Administração

Politécnico de Coimbra

Instituto Superior de Contabilidade e Administração

Politécnico de Coimbra

Diogo André Andrade Pimenta

Desenvolvimento de componentes de configuração e treino de algoritmos de *Machine Learning* no Sistema de Recomendação de Medicamentos Anticancerígenos (SiReMA)

Coimbra, novembro de 2022

Desenvolvimento de componentes de configuração e treino de algoritmos de *Machine Learning* no Sistema de Recomendação de Medicamentos Anrticancerígenos (SiReMA)

Diogo André Andrade Pimenta

ISCAC | 2022

Instituto Superior
de Contabilidade
e Administração
Politécnico de Coimbra
COIMBRA BUSINESS SCHOOL
ISCAC.pt

Diogo André Andrade Pimenta

Desenvolvimento de componentes de configuração e treino de algoritmos de *Machine Learning* no Sistema de Recomendação de Medicamentos Anticancerígenos (SiReMA)

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 Doutor Fernando Paulo dos Santos Rodrigues Belfo e coorientação do Doutor António Rui Trigo Ribeiro.

Coimbra, novembro de 2022

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: secretariado@iscac.pt; www.iscac.pt

RESUMO

O cancro é uma doença caraterizada por mutações no *DNA* que ocorre a partir de uma multiplicação descontrolada das células, alterando as células normais em células cancerígenas. Estas linhas celulares cancerígenas são manipuladas artificialmente, *in vitro*, e proliferam indefinidamente mantendo as caraterísticas do tecido de origem, o que as torna uma ferramenta importante para testar a sensibilidade a medicamentos e assim adotar estratégias eficazes de tratamento, potencializando a adoção de uma medicina personalizada.

O presente projeto pretende contribuir para a medicina personalizada, e tem como objetivo principal o desenvolvimento de componentes de configuração e treino de algoritmos de *Machine Learning* num sistema de recomendação de medicamentos anticancerígenos, com base na similaridade das imagens *DNA microarray*. Para tal, usámos a base de dados *Genomics Drug Sensitivity Cancer* que é o maior recurso público disponível sobre sensibilidade a medicamentos em células cancerígenas e a base de dados da *ArrayExpress* que contém imagens *DNA* em formato *microarray* das linhas celulares cancerígenas. Validámos a premissa de que linhas celulares cancerígenas com semelhanças no *DNA* partilham de tratamentos semelhantes, e a partir desta validação prosseguimos então para a construção de dois algoritmos de recomendação, um personalizado e outro não personalizado, capazes de sugerir um conjunto de medicamentos mais eficazes para um determinado paciente, com base na expressão genómica do seu *DNA*. Os resultados obtidos, apesar de preliminares revelam-se promissores, constituindo uma ferramenta útil para o apoio à decisão clínica e ainda para adoção de novas estratégias para otimizar a terapia medicamentosa com base na informação genómica de cada paciente.

Por fim, foi ainda desenvolvida uma *API*, com o objetivo de disponibilizar os algoritmos construídos no âmbito deste projeto para que os mesmos possam ser utilizados por diferentes aplicações e serem dessa forma disponibilizados ao utilizador final.

Palavras-chave: *DNA*; genómica; *machine learning*; sistema de recomendação, similaridade; linha celular.

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

ABSTRACT

Cancer is a disease characterized by mutations in DNA that occur from uncontrolled cell multiplication, changing normal cells into cancer cells. These cancer cell lines are artificially manipulated, in vitro, and proliferate indefinitely maintaining the characteristics of the tissue of origin, which makes them an important tool to test drug sensitivity and thus adopt effective treatment strategies, enhancing the adoption of personalized medicine.

The present project aims to contribute to personalized medicine, and its main objective is to develop components for configuring and training Machine Learning algorithms in an anticancer drug recommendation system based on the similarity of DNA microarray images. To do this, we used the Genomics Drug Sensitivity Cancer database which is the largest public available resource on drug sensitivity in cancer cells and the ArrayExpress database which contains microarray format DNA images of cancer cell lines. We validated the premise that cancer cell lines with similarities in DNA share similar treatment, and from this validation we then proceeded to build two recommendation algorithms, personalized and non-personalized, capable of suggesting a set of most effective drugs for a given patient based on the genomic expression of their DNA. The results obtained, despite being preliminary, are promising, constituting a useful tool for clinical decision support and also for the adoption of new strategies to optimize drug therapy based on each patient's genomic information.

Finally, an API was also developed, with the goal of making available the algorithms built within this project so that they can be used by different applications and thus made available to the end user.

Keywords: DNA; genomics; machine learning; recommendation system, similarity; cell lines

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

ÍNDICE GERAL

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

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

Lista de abreviaturas, acrónimos e siglas

AED..... Análise Exploratória de Dados

API ..... *Application Programming Interface*

CRISP-DM ..... *Cross Industry Standard Process for Data Mining*

DNA ..... *Deoxyribonucleic Acid*

DP ..... *Deep Learning*

GDSC ..... *Genomics of Drug Sensitivity in Cancer*

IA ..... Inteligência Artificial

IC50 ..... *Half-Maximal Inhibitory Concentration*

IHME ..... *Institute for Health Metrics and Evaluation*

ML ..... *Machine Learning*

NHDRI ..... *National Human Genome Research Institute*

SR ..... Sistema de Recomendação

TCGA ..... *The Cancer Genome Atlas*

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

1 INTRODUÇÃO

Poderá a tecnologia de *DNA microarray* ser uma mais-valia no tratamento do cancro? Como é que um algoritmo de *machine learning* poderá ajudar a determinar qual o melhor tratamento para um determinado paciente com cancro?

O presente estudo conjuga tecnologias como o *DNA microarray* com algoritmos de recomendação a fim de desenvolver um sistema capaz de propor um conjunto de medicamentos eficazes para um determinado paciente com base na expressão genómica do *DNA* das suas células.

1.1 Motivação e Relevância do Estudo

O cancro é uma doença que causa milhões de mortes anualmente. De acordo com os dados do Instituto de Métricas e Avaliação da Saúde (IHME), em 2019, foram registados a nível mundial 23,6 milhões de novos casos e 10 milhões de mortes (Kocarnik et al., 2021). A dimensão desta doença e o seu impacto na saúde pública impulsiona grandes investimentos na medicina e investigação, procurando alcançar melhorias em termos de prevenção, diagnóstico e tratamento.

Nas últimas décadas temos assistido a avanços significativos em termos de diagnóstico e tratamento do cancro, o que contribuiu para o aumento da taxa de sobrevivência (Jiménez-Santos et al., 2022). Muitos esforços têm sido efetuados neste sentido, entre eles a publicação do Projeto Genoma Humano em 2003 e o projeto Atlas do Genoma do Cancro em 2006. Estes esforços proporcionaram progressos na compreensão do cancro, enquanto doença genómica (Verma, 2012; Weinstein et al., 2013).

1.2 Contextualização

Na área oncológica, as linhas celulares cancerígenas têm despertado interesse nos investigadores, sendo alvo de inúmeras pesquisas, dado o papel importante que desempenham no desenvolvimento terapêutico e descoberta de medicamentos. Uma vez que, através das linhas celulares é possível analisar as variações genéticas na resposta aos medicamentos, tal ajuda a verificar a resposta do paciente ao tratamento, potencializando uma terapia mais segura e eficaz (Goodspeed et al., 2016; Hynds et al., 2018; Wheeler et

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

al., 2012). Esta capacidade de prever a resposta dos pacientes a determinados medicamentos, baseada em informações moleculares, tais como dados de expressão genómica constitui uma oportunidade para a implementação de uma medicina personalizada (Costello et al., 2014; Suphavilai et al., 2018).

Os investigadores necessitam de ferramentas para analisar as alterações genéticas clinicamente relevantes para o processo de diagnóstico e consequente tratamento da doença. Tecnologias como os *microarrays* de *DNA*, permitem analisar a expressão de centenas de genes em simultâneo, capaz de identificar os genes relacionados a doenças, e constitui uma das ferramentas mais promissoras no ramo da bioinformática (Pokhriyal et al., 2019). As ferramentas de bioinformática são assim fundamentais para os avanços na biologia molecular, pois sem os recursos computacionais existentes, não seria possível obtermos o conhecimento que temos hoje desta área (Pimentel & Bueno, 2016).

O projeto aqui apresentado foi desenvolvido no âmbito do Mestrado em Análise de Dados e Sistemas de Apoio à Decisão. Este está integrado numa linha de investigação comum a outro projeto denominado "Desenvolvimento de componentes de execução e validação humana de modelos de *Machine Learning* no Sistema de Recomendação de Medicamentos Anticancerígenos (SiReMA)", no qual se dedicou ao desenvolvimento das componentes de execução, validação e ainda a interface gráfica do sistema capaz de disponibilizar os resultados obtidos ao utilizador final.

1.3 Objetivos da Investigação

O principal objetivo do presente projeto consiste no desenvolvimento de componentes de configuração e treino de algoritmos de *machine learning* no âmbito de um sistema de recomendação de medicamentos anticancerígenos, com base na expressão genómica de um paciente. Iremos providenciar um instrumento de apoio à decisão médica, capaz de tornar as deliberações de um médico mais eficazes, robustas e sustentadas.

A validação da premissa de que linhas celulares similares têm tratamentos similares, será um dos principais objetivos deste estudo. É através desta premissa que todo o sistema de recomendação será alicerçado, daí a importância da mesma. Os nossos objetivos passam

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

pelo desenvolvimento de dois algoritmos de recomendação robustos e eficazes, capazes de gerarem recomendações precisas e adequadas ao contexto em que estão inseridas.

Por fim, importa ressalvar que este projeto, de uma forma geral, tem como objetivo contribuir de forma positiva para os estudos do tratamento do cancro e da medicina personalizada, pois apesar de termos perfeita noção de que isto se trata de um tema extremamente complexo, sensível e profundo, iremos dar o nosso melhor de forma a, providenciar e auxiliar estes temas com os conhecimentos obtidos a partir deste projeto.

1.4 Estrutura do Documento

O presente documento encontra-se estruturado em sete capítulos.

No presente capítulo descreve-se de forma sucinta a contextualização, esclarecem-se os objetivos que nos levaram a esta investigação, bem como, a motivação deste estudo e a sua relevância, e ainda apresentamos a estrutura do documento.

No segundo capítulo, apresentamos a revisão da literatura, onde são abordados diversos temas diretamente relacionados com o projeto, nomeadamente, genómica (linhas celulares e bioinformática), medicina personalizada, inteligência artificial, *machine learning*, sistemas de recomendação e, por último, alguns estudos relacionados com este projeto existentes na literatura.

O terceiro capítulo dedica-se à compreensão do tema e à preparação dos dados. Nele são apresentadas as bases de dados, bem como, as informações relativas a estas, de forma a obtermos um conhecimento e uma compreensão dos dados que iremos utilizar para desenvolver este projeto. São ainda divulgadas todas as etapas relativas à limpeza e transformação de dados e ainda a análise exploratória a que os dados foram sujeitos de forma a obter novos conhecimentos e *insights*, da mesma. Ainda neste capítulo, foi validada a premissa, que nos permitiu seguir com o projeto de forma a desenvolver os algoritmos de recomendação. Por fim, foram descritas quais as tecnologias utilizadas para desenvolver todo o projeto aqui apresentado.

O quarto capítulo está centrado na implementação dos modelos. Este capítulo apresenta os algoritmos de recomendação utilizados neste projeto. Numa primeira fase, apresenta-se como será executado o fluxo do processamento das imagens, esquematizando-se o

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

funcionamento do cálculo da similaridade entres as diversas imagens *DNA microarray*. Já na segunda fase, será exposto o desenvolvimento dos dois algoritmos de recomendação, assim como o seu comportamento em relação aos dados. Este é um capítulo importante pois encontram-se aqui as bases teórico-práticas do projeto, que serão, posteriormente, aplicadas em contexto real.

O quinto capítulo foca-se em aspetos de avaliação. Neste está apresentado um exemplo específico de funcionamento dos algoritmos de recomendação, em que descrevemos todo o processo do sistema, assim como, as respetivas métricas de avaliação. Por último, selecionamos uma amostra de cinco órgãos, a que aplicamos os algoritmos desenvolvidos e verificamos o seu comportamento face às métricas de avaliação, de forma a, termos uma perceção mais consistente da eficácia dos algoritmos desenvolvidos neste projeto.

No sexto capítulo, descreve-se a forma como os modelos desenvolvidos neste projeto vão ser disponibilizados. Ainda neste tópico é abordada a questão da arquitetura do sistema SiReMA, apresentada num diagrama de pacotes. Por fim, é apresentada uma proposta de uma possível manutenção dos modelos.

No sétimo e último capítulo, apresentamos a conclusão, designadamente as contribuições para uma medicina personalizada mais eficiente. Identificamos ainda quais as principais limitações que este projeto possui, nomeadamente, a nível dos dados, assim como, diferentes abordagens que podem ser realizadas no âmbito do tema deste projeto.

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 DA LITERATURA

Neste capítulo, iremos abordar e rever a literatura quanto a diversos temas diretamente relacionados com o projeto aqui em desenvolvimento, tais como a genómica, medicina personalizada, inteligência artificial, *machine learning* e sistemas de recomendação. A abordagem a estas temáticas assim como o conhecimento adquirido neste capítulo, é fundamental para conseguirmos desenvolver este projeto de forma lógica, coerente e fundamentada.

2.1 Genómica

A genómica é a área da biologia que se dedica ao estudo do genoma completo de um organismo, sendo o genoma o conjunto do *DNA* de um ser vivo, incluindo todos os seus genes, presentes no núcleo das células (National Cancer Institute, 2022b; Neves, 2010). Estudar o genoma possibilita o entendimento da forma como os genes interagem entre si, ajudando a perceber como é que algumas doenças se desenvolvem, como é o caso do cancro (National Cancer Institute, 2022a). Importa referir que a genómica difere da genética. De acordo com a definição apresentada pelo *National Human Genome Research Institute* (NHGRI) a genética dedica-se ao estudo de genes individuais (um único gene) e de que forma é que certas caraterísticas são transmitidas de geração em geração, ou seja, há um foco na hereditariedade, enquanto a genómica dedica-se ao estudo de todos os genes, as suas funções, as interações entre si e com fatores ambientais identificando a sua influência em termos de desenvolvimento no organismo (Genome, 2018; World Health Organization, 2020).

O cancro é uma doença que surge a partir de mutações genéticas (alterações no *DNA*) nas células, onde se verifica uma alteração no comportamento das células causando um crescimento anormal, incontrolável e maligno. Estas alterações transformam as células normais em células cancerígenas (Silva et al., 2020).

2.1.1 Linhas Celulares Cancerígenas

As linhas celulares cancerígenas advêm de amostras de tecido cancerígeno extraídas do paciente e são depois armazenadas e isoladas em laboratórios, tornando-as imortais (Richter et al., 2021). Estas são manipuladas artificialmente, *in vitro*, em ambiente

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

laboratorial, capazes de se proliferarem indefinidamente, de forma ilimitada, sem perder as suas caraterísticas (Gonçalves & Sobral, 2020). Estas apresentam ainda uma facilidade de cultivo e manuseamento tornando-as ferramentas biológicas relevantes para testar a resposta dos pacientes a medicamentos (Capes-Davis et al., 2019; Carter & Shieh, 2015; Hynds et al., 2018). Nos últimos tempos, as linhas celulares cancerígenas têm sido recorrentemente utilizadas em estudos na área da farmacogenómica, que se dedica a estudar a forma como os genes/organismo reagem a um determinado medicamento (Goodspeed et al., 2016; Huo et al., 2020). Relativamente aos medicamentos contra o cancro, usualmente é utilizada uma métrica de avaliação da eficácia de um medicamento num recetor (por exemplo linha celular), designado *Half-Maximal Inhibitory Concentration* (IC50) (González-Larraza et al., 2020). Assim, o IC50 representa a concentração do composto (inibidor) necessária para a inibição de um processo biológico ou bioquímico pela metade, pelo que quanto mais baixo for mais eficiente é o medicamento, neste sentido, o IC50 é uma informação extremamente útil para a classificação de compostos ou medicamentos (Caldwell et al., 2012).

As linhas celulares são utilizadas para analisar os níveis de expressão de milhares de genes, o que possibilita a identificação de genes afetados pelo cancro, através da comparação da expressão génica das células cancerígenas com a das células normais (Kashyap et al., 2015; Neves, 2010).

Existem diversas ferramentas disponíveis para realizar a análise da expressão genética em células, no entanto, a tecnologia de *microarrays* de *DNA* é uma das ferramentas mais promissoras, no sentido em que esta, segundo Moreau et al. (2002) “permite medir os níveis de expressão de milhares de genes simultaneamente” numa única experiência (Moreau et al., 2002). Na Figura 2-1 é possível verificarmos um exemplo de uma imagem de *microarray DNA*.

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

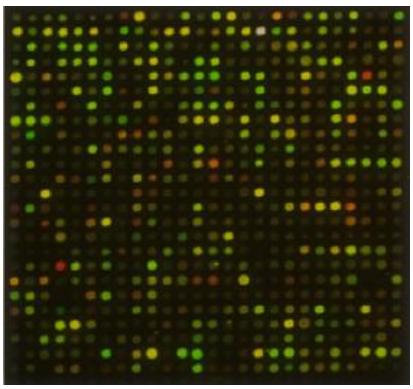

Figura 2-1 Imagem microarray DNA

Fonte: Neves (2010)

Esta tecnologia produz uma grande quantidade de dados, sob a forma de imagens, que se caraterizam por uma elevada complexidade em termos de processamento constituindo assim um desafio, requerendo técnicas computacionais para tornar os dados interpretáveis (Loewe & Nelson, 2011; Varella-Garcia, 2004).

2.1.2 Bioinformática

A bioinformática é uma área multidisciplinar que resulta da interação direta de duas disciplinas, a biologia e a informática, recorrendo ainda a outras disciplinas como a matemática e a estatística, e envolve o uso de tecnologia para analisar, interpretar e processar dados biológicos (Abdurakhmonov, 2016). Esta nova área da ciência faz uso de ferramentas computacionais para estudar problemas/fenómenos biológicos, procurando responder a questões essenciais no que concerne às ciências da vida e auxiliando no campo das ciências "ómicas" (Baxevanis & Ouellette, 2001; Hagen, 2000).

Os avanços nos últimos tempos da tecnologia e das investigações no ramo da biologia, impulsionaram o crescimento exponencial de dados biológicos. No entanto, este grande volume de dados constitui um desafio, uma vez que se torna necessário realizar uma gestão eficiente destes de modo a extrair conhecimento útil, requerendo assim o uso de técnicas de *machine learning* (Larrañaga et al., 2006).

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

2.2 Medicina Personalizada

Tradicionalmente, os pacientes com uma determinada patologia, com caraterísticas semelhantes, recebem o mesmo tratamento. No entanto, com os avanços no conhecimento constatou-se que o facto dos pacientes terem a mesma patologia não significa que estes devam receber o mesmo tratamento, pois é reconhecido que grande parte dos medicamentos apresentam variabilidade em termos de eficácia e toxicidade entre os pacientes. Com isto, surgiu a medicina personalizada, que possibilita o desenvolvimento de tratamento personalizado para cada paciente (Evans & Johnson, 2001; Gomes, 2019; Verma, 2012). Estudos como o sequenciamento de *DNA* permitiram avanços significativos no campo da medicina personalizada.

Partindo do princípio de que as pessoas possuem caraterísticas únicas e diferenciadas em termos moleculares, fisiológicos, ambientais e comportamentais, tal implica que necessitem de um tratamento adaptado a estas caraterísticas (Goetz & Schork, 2018).

A medicina personalizada é uma prática emergente na medicina que tem como base o desenvolvimento de um tratamento personalizado para um paciente a partir do uso de informações genéticas deste (Hoeben et al., 2021; Soares, 2020). Assim, esta informação genética do paciente ajuda os médicos a prescrever o medicamento certo na dose certa para a pessoa certa, ao invés da tradicional abordagem "*one fits all*", potencializando um tratamento adaptado ao paciente, garantindo assim um melhor atendimento e maior segurança, permitindo aos médicos uma tomada de decisão clínica mais eficaz (*Personalized Medicine*, 2022; Vogenberg et al., 2010).

Na Figura 2-2 é possível visualizarmos as diferenças entre a abordagem tradicional e a abordagem personalizada.

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

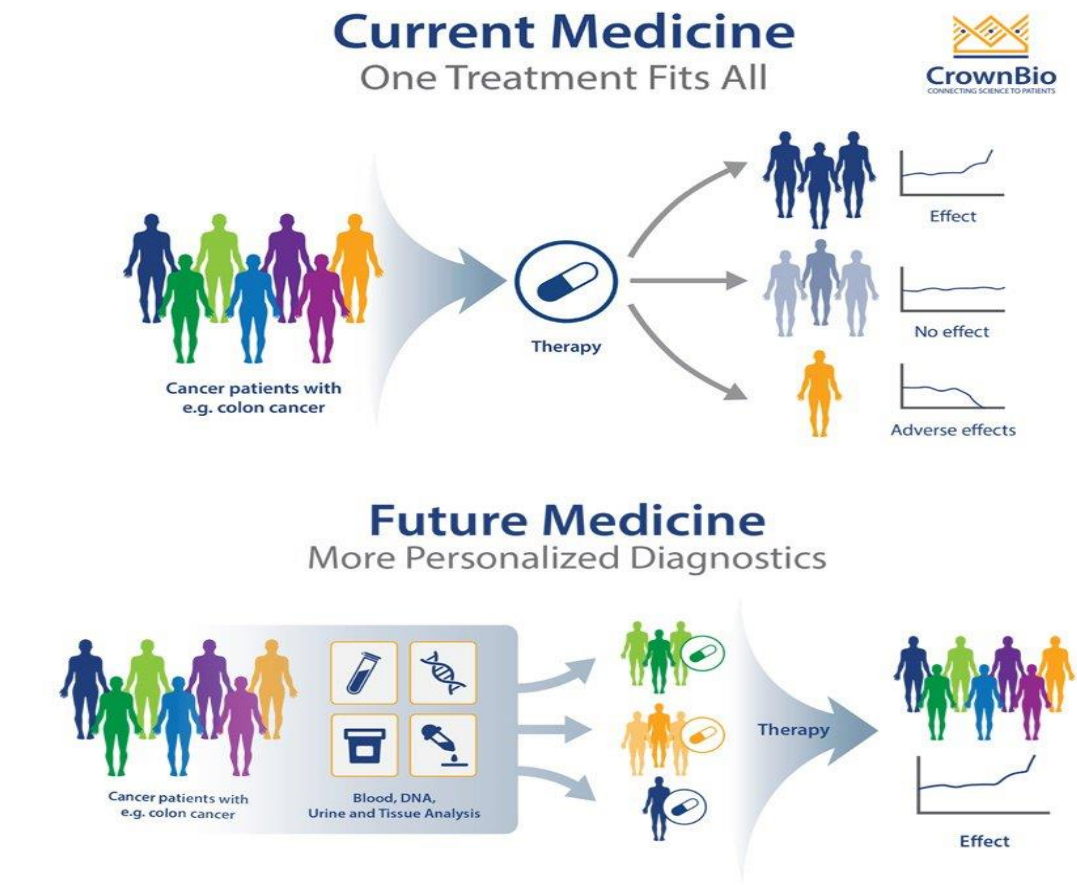

Figura 2-2 Medicina personalizada

Fonte: Barbeau (2018)

Conforme é possível visualizar na Figura 2-2, a abordagem tradicional está assente numa terapia universal em que o mesmo medicamento pode ser aplicado a vários pacientes, tendo como consequência diversos efeitos, tais como, positivos, adversos ou até mesmo neutros. Já a abordagem personalizada tem como critérios informações como *DNA* ou análises clínicas de cada paciente de modo a prescrever a terapia adequada a cada um destes, minimizando assim o risco de efeitos adversos ou ausência de efeitos.

2.3 Metodologia CRISP-DM

O *Cross Industry Standard Process for Data Mining (CRISP-DM)* é a metodologia mais utilizada nos projetos de ciências de dados para a realização da mineração 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

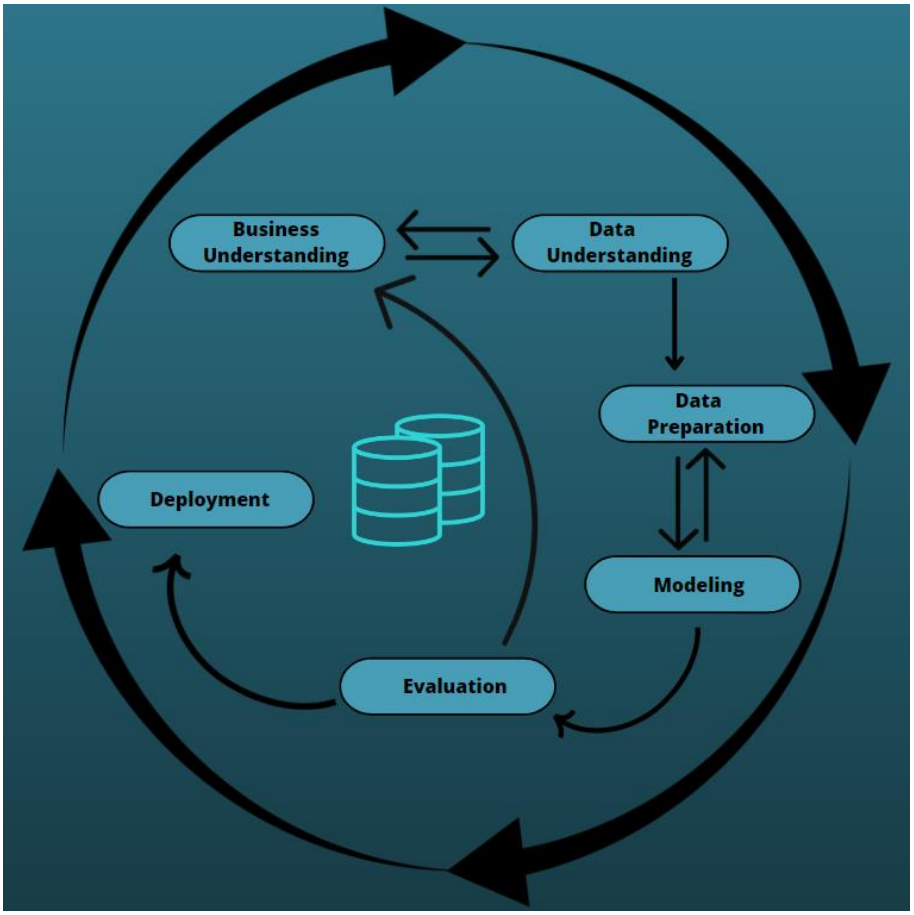

Figura 2-3 CRISP-DM

Adaptado de: Saltz (2021)

Esta metodologia tem por base uma sequência não rígida de seis fases (Colpani, 2018): compreensão do negócio, compreensão dos dados, preparação dos dados, modelação, avaliação e implementação. Na Figura 2-3 estão ilustradas as fases da metodologia CRISP-DM, onde é visível a iteratividade entre estas. Na prática, embora haja uma ordem natural, estas tarefas não seguem uma sequência rígida e podem ser executadas numa ordem diferente conforme necessário, sendo possível nalgumas fases regressar à fase anterior ou, no caso da fase de avaliação, regressar à fase inicial, caso se verifique a necessidade de repetir algumas tarefas.

A primeira fase consiste na compreensão do negócio e tem por base compreender as necessidades/requisitos e objetivos do projeto em termos de negócio. A segunda fase, compreensão dos dados, carateriza-se pela recolha, descrição e exploração dos dados para uma familiarização com estes. A terceira fase, preparação dos dados, é uma das tarefas

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

mais exigentes e extensa do trabalho e consiste na preparação dos dados, em termos de seleção, limpeza, construção, integração e formatação a fim de serem utilizados na modelação (Saltz, 2021). A modelação compreende desenvolver e encontrar o melhor modelo através da seleção e aplicação de algoritmos aos dados. Segue-se a fase de avaliação que se concentra em avaliar os resultados obtidos do modelo (Schafer et al., 2018). Na última fase deste processo, é implementado o modelo.

2.4 Inteligência Artificial

Os avanços tecnológicos, a crescente digitalização e a implementação de sistemas de informação na saúde, impulsionaram a geração de um grande volume de dados. Não obstante, é necessário haver uma gestão eficiente desta informação para que se traduza em *insights* valiosos permitindo uma tomada de decisão mais precisa (Barboza, 2019).

Tecnologias como a inteligência artificial e o *machine learning* têm ganho elevada popularidade nos últimos anos, sendo ferramentas úteis para lidar com grandes quantidades de dados biológicos extraindo informação/conhecimento útil desses dados de forma eficiente (Hanif et al., 2019; Larrañaga et al., 2006).

A Inteligência Artificial (IA) engloba os conceitos de *Machine Learning* (ML) e *Deep Learning* (DL), conforme se encontra ilustrado na Figura 2-4. De uma forma geral, a IA, o ML e o DL são técnicas utilizadas para tentar simular o comportamento humano, de forma a, serem capazes de automatizar e otimizar processos, tarefas e ações, que até então eram realizadas por humanos (Shinde & Shah, 2018).

A IA é uma tecnologia do ramo da ciência da computação que surgiu por volta da década de 1950 tendo como um dos pioneiros, John McCarthy, que definiu a IA como: “a ciência e a engenharia de fazer máquinas inteligentes, especialmente programas de computador inteligentes” (McCarthy, 2007). Ou seja, o objetivo da aplicação da IA passa por dotar uma máquina da capacidade de simulação do comportamento do humano e consequentemente apoiar na tomada de decisões e resolução problemas, tornando-a assim numa máquina inteligente (Shinde & Shah, 2018).

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

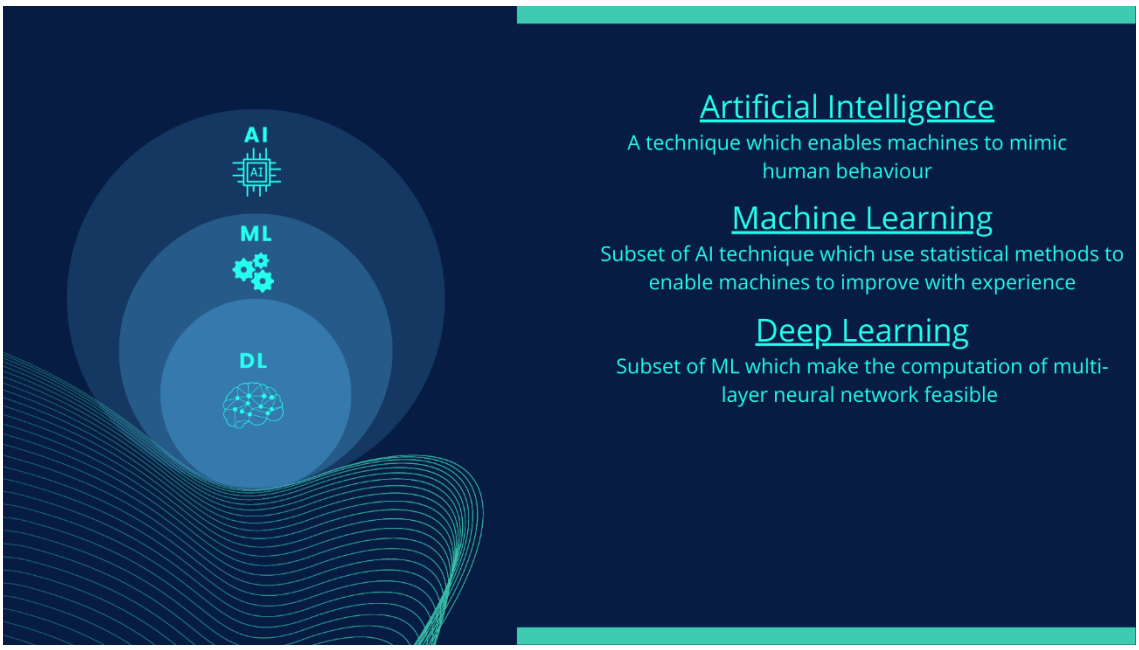

Figura 2-4 Dimensionamento da Inteligência Artificial

Adaptado de: Shrestha (2021)

2.5 *Machine Learning*

De uma forma muito genérica, os humanos conseguem aprender através da observação do mundo real, e, por conseguinte, da deteção e análise de padrões existentes nesse mesmo mundo. Embora o ML utilize o mesmo “raciocínio" usado pelos humanos, lida com a informação de forma mais eficiente. Este ramo da IA consiste na aprendizagem dos computadores por meio da aplicação de algoritmos aos dados, capacitando-os da identificação/reconhecimento de padrões para construir modelos de classificação, regressão, previsão, entre outros (Quazi, 2022; Vasconcelos & Barão, 2017). Desta forma, tendo em conta o resultado desejado, diversas abordagens de aprendizagem podem ser aplicadas, entre elas a supervisionada, não supervisionada e de reforço.

No processo de construção de um modelo de ML, frequentemente é realizada uma divisão da base de dados em dois conjuntos, treino e teste, de forma aleatória e normalmente é utilizada a regra 70/30 onde 70% dos dados são utilizados no conjunto de treino e 30% no conjunto de teste (Mahesh, 2020). A primeira fase passa por treinar o modelo, ou seja, efetuar o processo de aprendizagem que consiste em utilizar o conjunto de treino e aplicar algoritmos a estes dados para a construção do modelo (Jayanthi & Mahesh, 2018). Podem

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

ser aplicados diversas técnicas, entre elas, a indução de árvores de decisão, a regressão linear, a regressão logística ou a classificação de *Naive Bayes* (Mahesh, 2020). Posteriormente, depois de construído o modelo, segue-se a fase de teste que passa por utilizar o conjunto de teste, o qual é composto por dados que não foram utilizados no treino. Estes dados são usados no modelo de forma a avaliar o comportamento deste com esses dados, não conhecidos, e calcular o seu desempenho (Castiglioni et al., 2021).

Um exemplo de aplicação do ML está no desenvolvimento de sistemas de recomendação, onde o sistema efetua uma aprendizagem baseando-se nos dados históricos, produzindo uma recomendação com base nesses dados. O subcapítulo "Sistemas de Recomendação" detalha este tema.

2.6 Sistemas de Recomendação

Com os avanços tecnológicos, existe imensa informação disponível na Internet. Contudo, esta informação excessiva dificulta muitas vezes os utilizadores encontrarem o conteúdo de que realmente necessitam (Li, 2021). Neste sentido, foram desenvolvidos sistemas de recomendação que se caraterizam pela enorme utilidade que têm para os utilizadores, pois permitem, apresentar informação filtrada e fornecer recomendações aos utilizadores por exemplo em *websites* de *e-commerce* ou plataformas de filmes, melhorando a navegação e experiência destes na Internet (Adak & Ucar, 2021).

Os sistemas de recomendação constituem uma importante ferramenta estratégica para as empresas pois através destes, as empresas conseguem ajustar os produtos/serviços que são pertinentes aos clientes, identificando as suas preferências, resultando na oferta de um produto adequado e personalizado ao perfil de cada cliente (Oliveira, 2014).

Os sistemas de recomendação são mecanismos baseados em ML e IA, no qual aos dados históricos, informação passada sobre as preferências e pesquisas do cliente, é aplicado um algoritmo que tem como resultado final uma recomendação (Mana & Sasipraba, 2021). Estes sistemas podem ser categorizados em abordagens personalizadas ou não personalizadas. A abordagem personalizada pode ser filtragem colaborativa, filtragem baseada em conteúdo ou abordagem híbrida (Feng et al., 2020). A título exemplificativo, apresentamos o caso da Netflix que tem na sua génese a utilização de algoritmos de

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

recomendação. A Netflix é uma plataforma de *streaming* que oferece uma variedade de séries, filmes e documentários. Esta diversidade de conteúdo implica muita informação para os utilizadores e por esta razão a empresa aplica algoritmos de recomendação objetivando analisar as preferências dos consumidores, de forma a oferecer um produto ajustado ao desejo e gosto do cliente (Choy, 2021). Assim, quanto mais informação a empresa tiver sobre o cliente, em termos dos seus gostos e hábitos, mais facilmente esta consegue recomendar o conteúdo direcionado para aquele cliente. Estes sistemas demonstraram elevada eficiência uma vez que as recomendações geradas pelo algoritmo representam cerca de 80% dos conteúdos transmitidos (Chong, 2020).

Na Figura 2-5 apresentamos o exemplo de uma captura de ecrã do sistema de recomendação da Netflix, onde na primeira linha "melhores escolhas para Josué" está representada a abordagem de filtragem colaborativa, que consiste na recomendação de conteúdo que outros utilizadores com perfil semelhante assistiram. Na segunda linha, designada por "tendências atuais", está representada a abordagem não personalizada, esta é uma abordagem que não tem em conta as preferências do utilizador, fornecendo recomendações com base nas tendências mais populares do momento (Poriya et al., 2014). Ainda a abordagem baseada no conteúdo está representada na terceira linha designada, "porque viu Narcos", onde é recomendado conteúdo que compartilha caraterísticas comuns com um conteúdo que o utilizador viu recentemente.

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

Figura 2-5 Exemplo de captura de tela do sistema de recomendação da Netflix

Fonte: Naujoks (2019)

2.7 Trabalhos Relacionados

As linhas celulares representam um importante biomarcador molecular pois permitem compreender a biologia do cancro e auxiliam na previsão da resposta aos medicamentos. Desde a criação de bases de dados de sensibilidade a medicamentos com base nas linhas celulares, tais como a GDSC e CCLE, que estudos em larga escala têm sido realizados, incorporando a aplicação de algoritmos de ML o que tem permitido obter conhecimentos valiosos. Partindo do princípio de que nem todos os pacientes respondem da mesma forma aos medicamentos, a análise da expressão genómica contida nas linhas celulares poderá ser um fator determinante indicativo sobre a resposta dos pacientes aos medicamentos. Dedicamos esta secção do projeto a mostrar alguns trabalhos que têm sido realizados neste âmbito.

Na literatura existente, diversas abordagens têm sido desenvolvidas, no entanto, a generalidade tem-se dedicado à previsão da resposta aos medicamentos (que é quantificado pelo IC50), pelo que estas abordagens são mais direcionadas para o desenvolvimento/triagem de medicamentos. Uma diferente abordagem, ainda pouco

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

explorada na literatura, passa por construir um sistema de recomendação de medicamentos anticancerígenos para uma linha celular, permitindo assim adotar uma medicina personalizada.

Menden et al. (2013) utilizaram a base de dados da GDSC e propuseram uma abordagem para determinar esta questão da previsão da sensibilidade aos medicamentos, mas baseado nas caraterísticas genómicas das linhas celulares e nas propriedades químicas dos medicamentos. Os autores recorreram a algoritmos de ML (redes neurais e *random forest*), e os resultados obtidos foram significativos, onde a avaliação em termos preditivos do modelo desenvolvido, atingiu um coeficiente de determinação de 0,68, o que demonstra que o modelo consegue prever os valores de IC50 com uma precisão de 68% para novas linhas celulares (Menden et al., 2013).

Sharma & Rani (2018) usaram os dados da GDSC e da base de dados CCLE, e apresentaram uma abordagem para previsão da resposta a medicamentos, que utiliza a técnica de fatoração de matrizes kernelizada, que é comumente usada nos sistemas de recomendação, pois ajuda a encontrar similaridade entre itens e utilizadores (Sharma & Rani, 2018). Comparativamente com as demais técnicas, e utilizando a métrica do erro quadrado médio, esta técnica foi a que obteve o melhor valor.

Brandão (2020) desenvolveu um sistema de recomendação para medicamentos anticancerígenos. O seu trabalho focou-se na utilização de técnicas de processamento de imagens, tais como *wavelets*, para calcular a similaridade das linhas celulares, com o intuito de perceber se esta técnica era capaz de gerar similaridades mais corretas e assim gerar melhores resultados (Brandão, 2020).

Zhang et al. (2015) apresentaram uma proposta para previsão de resposta a medicamentos, através da aplicação de informações de similaridade entre linhas celulares e medicamentos. Utilizando o conjunto de dados CCLE e Cancer Genome Project (CGP), os autores desenvolveram um modelo ponderado de rede de camada dupla, integrando uma rede de similaridade de linha celular, baseada nas similaridades dos perfis de expressão génicas das linhas celulares e ainda uma rede de similaridade de medicamento, baseada na similaridade das caraterísticas químicas dos medicamentos. Foi obtido um coeficiente de correlação de Pearson de 0,6 o que permitiu aos autores concluir que

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

"linhas celulares com perfis de expressão génica similares exibem respostas similares ao mesmo medicamento" (Zhang et al., 2015). Ainda em termos de avaliação da capacidade preditiva do modelo, o modelo de camada dupla comparativamente com a *random forest*, a regressão vetorial de suporte e a rede elástica, foi o que apresentou o melhor desempenho.

Importa referir que algumas propostas, em termos de desenvolvimento de sistemas de recomendação, baseiam-se na suposição de que linhas celulares semelhantes exibem respostas semelhantes a medicamentos. Este projeto procurou aprofundar e testar a premissa de que efetivamente linhas celulares partilham de tratamento similar, avaliando se existe uma correlação entre estas variáveis.

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

3 COMPREENSÃO DO TEMA E PREPARAÇÃO DOS DADOS

Este capítulo tem como finalidade, contextualizar os dados presentes neste projeto, de forma a adquirirmos um bom entendimento do tema aqui exposto. Ainda é realizada a preparação dos dados, que compreende desde a limpeza e tratamento até à análise exploratória destes, de forma a conseguirmos reunir o máximo de informações úteis e relevantes. Por fim, é testada a premissa base deste projeto.

3.1 Entendimento do tema e dos dados

Os dados utilizados neste estudo foram extraídos de duas plataformas distintas, pelo que procedemos ao cruzamento da informação que consta em ambas. A plataforma *Genomics of Drug Sensitivity in Cancer (GDSC)* disponibiliza conjuntos de dados referentes à sensibilidade de várias linhas celulares cancerígenas a diversos medicamentos anticancerígenos (*Genomics of Drug Sensitivity in Cancer*, 2022). A plataforma *ArrayExpress* disponibiliza imagens *DNA microarray* de linhas celulares cancerígenas (*ArrayExpress - Functional Genomics Data*, 2022).

3.1.1 Base de dados *Genomics of Drug Sensitivity in Cancer*

A plataforma *GDSC* disponibiliza uma base de dados, que compreende dados de 2010 a 2015, e contém 310 904 registos, descrevendo a resposta de 987 linhas celulares a 345 medicamentos. Estes registos dizem respeito à testagem de centenas de medicamentos anticancerígenos em várias linhas celulares cancerígenas.

Esta base de dados é composta por diversas variáveis. Para este estudo, iremo-nos focar exclusivamente em três variáveis, sendo estas:

- *CELL_LINE_NAME:* referente aos diferentes nomes de linhas celulares disponíveis na base de dados;
- *DRUG_NAME:* referente aos nomes dos medicamentos testados nas diversas linhas celulares;
- *LN_IC50:* referente ao logaritmo da concentração inibitória média de um medicamento testado numa determinada linha celular;

Na Tabela 3-1, podemos visualizar uma amostra da base 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

Tabela 3-1 Amostra da base de dados GDSC

CELL_LINE_NAME	DRUG_NAME	LN_IC50
42-MG-BA	PICTILISIB	0,280776
42-MG-BA	ZIBOTENTAN	5,495883
A2780	DOXORUBICIN	-4,246344
A2780	EPOTHILONE B	-6,753907

3.1.2 Base de dados *ArrayExpress*

A base de dados *ArrayExpress* disponibiliza 1018 imagens *DNA microarray* que representam 1018 linhas celulares cancerígenas, sendo composta por três variáveis:

- *CELL_LINE:* referente ao nome da linha celular;
- *ORGAN:* referente ao nome do órgão a que pertence a linha celular;
- *ASSAY_FILE:* referente ao nome do ficheiro da imagem *DNA microarray* da linha celular correspondente;

Na Tabela 3-2, podemos visualizar uma amostra da base de dados.

Tabela 3-2 Amostra da base de dados ArrayExpress

CELL_LINE	ORGAN	ASSAY_FILE
A2058	SKIN	5500994157493061613625_A06.CEL
AU565	BREAST	5500994157493061613625_C06.CEL
A2780	OVARY	5500994172383112813929_B03.CEL
CAPAN-1	PANCREAS	5500994158987071513209_F06.CEL

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

3.2 Preparação dos dados

A preparação dos dados é uma das primeiras tarefas a ser realizada e compreende a limpeza e tratamento de dados. Esta tarefa é de extrema importância pois envolve a eliminação de dados inválidos, imprecisos, inconsistentes ou pouco informativos, resultando na obtenção de um conjunto de dados com uma melhor qualidade informativa e consequentemente, proporcionando resultados mais precisos, evitando enviesamentos. Para além de uma melhor precisão dos dados, esta tarefa é também muito útil para a familiarização com os mesmos, visto que, ao executar esta tarefa inevitavelmente adquirimos um maior domínio em relação a estes.

3.2.1 Limpeza e tratamento de dados (*GDSC*)

Posto isto, o primeiro passo a efetuar na base de dados *GDSC*, consistiu na eliminação das linhas celulares testadas com menos de 100 medicamentos, pois consideramos que estas linhas eram pouco informativas em termos de medicamentos testados relativamente à totalidade dos medicamentos disponíveis.

O segundo passo consistiu na remoção dos medicamentos testados em duplicado na mesma linha celular, pois o objetivo é que cada linha celular seja testada com os vários medicamentos disponíveis apenas uma só vez. O fato de aparecerem dois medicamentos iguais na mesma linha celular acontece porque, por exemplo, pode ter ocorrido um erro na primeira testagem desse medicamento na linha celular o que levou a uma segunda tentativa de testagem. Tendo em conta este facto, a opção tomada nestas situações foi de eliminar a primeira ocorrência, ficando a base de dados apenas com as segundas ocorrências dos casos em que existem medicamentos duplicados.

O passo seguinte passou pela transformação dos nomes das linhas celulares, mais especificamente, colocá-los todos em letras maiúsculas. Por fim, foram corrigidos os nomes de duas linhas celulares que estavam incorretos.

Após esta transformação ficamos então com 977 linhas celulares distintas e 345 medicamentos.

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

3.2.2 Limpeza e tratamento de dados (*ArrayExpress)*

A etapa da limpeza e tratamento de dados na presente base de dados, passou essencialmente por quatro fases distintas, sendo estas: a alteração dos nomes das linhas celulares para letra maiúscula; a correção de alguns nomes de linhas celulares que estavam escritos de forma incorreta; a eliminação de linhas celulares duplicadas e consequentemente as imagens duplicadas e por fim a eliminação de uma linha celular que tinha associada uma imagem que estava corrompida (formato incorreto).

Após esta etapa, ficamos com 1013 linhas celulares distintas, ou seja, 1013 imagens que estão associadas a cada uma destas linhas celulares.

3.2.3 Transformação de dados (*GDSC*)

De forma geral, o processo da transformação de dados, consiste em modificar os dados brutos/originais em formatos mais adequados. Dentro da transformação de dados existem diversas atividades. No presente projeto foram desenvolvidas atividades como a normalização dos dados e a geração de hierarquia de conceitos.

Em relação à normalização dos dados, como apresentado anteriormente, os valores da coluna *LN_IC50* estão num formato de logaritmo, pelo que, houve a necessidade de normalizar estes valores, de forma que, estes se encontrem dentro de um intervalo entre 0 e 1. Para tal, foi utilizada a seguinte função logística, utilizada por Menden et al (2013):

$$norm(y) = \frac{1}{1 + \exp(y)^{-0.1}}$$

Onde: $y = LN_{IC50}$

Por fim, a atividade referente à geração de hierarquia de conceitos, que, de uma forma geral, consiste na conversão para um nível superior de um determinado atributo, foi realizada através da criação de uma nova coluna, designada Órgão. Para tal, foi determinado que cada linha celular pertence a um órgão e a partir daí foi efetuada uma associação entre as várias linhas celulares e os órgãos a que pertencem. Para visualizarmos melhor esta associação podemos observar a Figura 3-1:

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

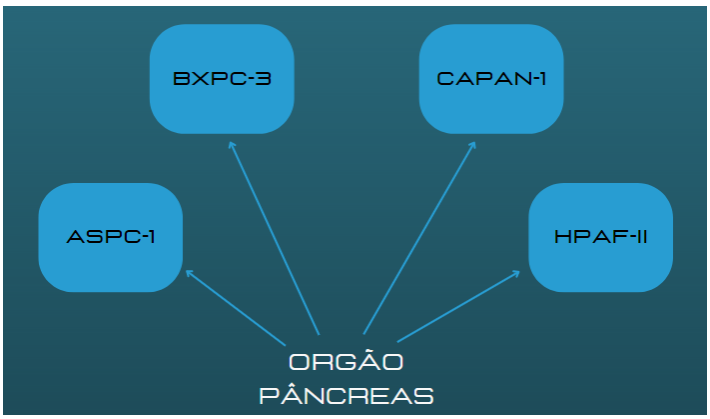

Figura 3-1 Demonstração das linhas celulares de um órgão

Após todas estas alterações à base de dados original, o formato final da base de dados encontra-se representado na Tabela 3-3.

Tabela 3-3 Amostra da base de dados GDSC final

CELL_LINE_NAME	DRUG_NAME	IC50	ORGAN
42-MG-BA	PICTILISIB	0,507018	BRAIN
42-MG-BA	ZIBOTENTAN	0,634040	BRAIN
A2780	DOXORUBICIN	0,395408	OVARY
A2780	EPOTHILONE B	0,337290	OVARY

3.2.4 Cruzamento de dados (linhas celulares)

Depois de termos as duas bases de dados limpas e tratadas, o passo seguinte foi uniformizar as informações presentes em cada uma destas, de forma a selecionar apenas os dados que são comuns entre estas e que serão utilizados efetivamente, nomeadamente, a nível das linhas celulares, pois facilmente reparamos que enquanto na base de dados do *ArrayExpress* estão presentes 1013 linhas celulares distintas, na base de dados *GDSC* estão presentes apenas 977. Logo, existe a necessidade de igualar o número de linhas celulares de forma que, fiquemos apenas com as linhas celulares que são comuns entre as bases de dados. Após esta etapa, validamos que existem 932 linhas celulares comuns entre ambas as bases de dados, logo, serão apenas nestas que nos iremos focar.

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

3.2.5 Exploração dos Dados

A realização de uma análise exploratória de dados à base de dados *GDSC* torna-se quase imprescindível, pois esta é uma tarefa que nos permite analisar a base de dados, e consequentemente, sintetizar as suas principais características. É igualmente importante perceber que outra das vantagens em aplicar esta análise está no facto de que, através desta, podemos ainda identificar erros, interpretar os padrões existentes, detetar anomalias, como *outliers*, e ainda observar as diferentes relações entre as variáveis.

Os resultados obtidos com a análise exploratória de dados são apresentados em quatro *dashboards* principais, de forma a tornar a apresentação destes mais intuitiva.

Figura 3-2 Dashboard 1 - Linhas Celulares

Ao observarmos as informações presentes no *dashboard* da Figura 3-2, podemos constatar que existem 932 linhas celulares únicas presentes na base de dados. Em relação à frequência destas, facilmente percebemos que temos linhas celulares, tais como, *SK-MEL-1*, *A253*, *KCL-22*, *TE-12*, *KNS-42* que estão entre as mais frequentes, apresentando uma frequência de 345 cada uma. O valor da frequência indica que estas linhas foram testadas com todos os medicamentos disponíveis, ou seja, com cada um dos 345 medicamentos disponíveis. Já no caso das linhas celulares menos frequentes tais como, *U-266*, *COR-L311*, *NCI-H1650*, *OCI-LY7*, *IGROV-1*, estas têm uma frequência de 119, 118, 117, 117, 115, respetivamente, significando que foram testadas com apenas cerca de

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

120 dos 345 medicamentos disponíveis. Tal como podemos verificar nem todas as linhas celulares foram testadas com todos os medicamentos disponíveis por isso é interessante apresentarmos a média da frequência das linhas celulares, que neste caso é de cerca de 300. Isto é, em média, cada linha celular foi testada com 300 dos 345 medicamentos disponíveis. Percentualmente podemos aferir que em média cada linha celular foi testada com aproximadamente 87% dos medicamentos disponíveis.

Por fim, apresentamos uma comparação entre as 10 linhas celulares que têm os valores de IC50 associados mais baixos (melhores) e as 10 linhas celulares com valores de IC50 mais elevados (piores). Esta estatística é importante, pois permite-nos ter uma visão geral, do grau de discrepância que existe entre as linhas celulares com valores de IC50 mais elevados e mais baixos. Como exemplo, podemos observar a linha celular com o IC50 mais baixo que é a *ME-180* com um valor de IC50 de 0.2577, em comparação com a linha celular com o IC50 mais elevado que é a linha celular *SW1116* com um valor de IC50 de 0.7749. Isto significa que, o valor do IC50 mais alto é praticamente 300% superior ao valor do IC50 mais baixo, ou seja, as diferenças entre as linhas celulares com IC50 mais baixos e com IC50 mais elevados, são muito significativas, o que indica que existem diferenças consideráveis na eficácia entre os diversos medicamentos.

Figura 3-3 Dashboard 2 – Medicamentos

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

O *dashboard* da Figura 3-3 reflete informações gerais da variável medicamentos. Neste caso, existem 345 medicamentos disponíveis. De forma análoga à análise anterior, podemos verificar que os cinco medicamentos mais frequentes são o *Avagacestat, o SB505124, o UNC0638, o Doxorubicin e o Gemcitabine*, com frequências de 930, 929, 928, 927 e 926, respetivamente. Os valores das frequências indicam-nos que o medicamento *Avagacestat,* foi testado em 930 linhas celulares, das 932 existentes, ou seja, este medicamento foi testado em praticamente todas as linhas celulares. Já os medicamentos *Dasatinib, Tozasertib, Bortezomib, JW-7-52-1 e Rapamycin*, são aqueles que menos vezes foram testados, com valores de 377, 377, 377, 368 e 342, respetivamente, o que uma vez mais nos indica que estes medicamentos apenas foram testados em menos de 380 linhas celulares. Em termos percentuais, isto significa que estes cinco medicamentos foram testados em menos de 40% das linhas celulares. Já em termos médios, temos que a frequência média toma o valor de aproximadamente 812, o que sugere que em média cada medicamento é testado em 812 linhas celulares.

Uma outra perspetiva que podemos analisar, são os 20 medicamentos com melhores médias de IC50. Como exemplo, podemos observar o medicamento *Bortezomib* que tem uma média de IC50 de 0.3637, o que sugere que este medicamento teve uma boa eficácia nas linhas celulares onde foi testado. No entanto, como podemos observar, este foi utilizado em apenas 377 linhas celulares. Já o medicamento *Epothilone B*, conta com uma média de IC50 a rondar os 0.3795, que é igualmente um valor satisfatório, contudo este medicamento foi aplicado em 833 linhas celulares, o que pode sugerir que foi eficaz num número superior de linhas celulares, em relação ao medicamento *Bortezomib*. O gráfico no canto inferior direito apresentado na *dashboard* permite-nos também perceber que os medicamentos com melhores médias de IC50 (mais baixas) têm frequências altas, ou seja, são testados em muitas linhas celulares.

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

Figura 3-4 Dashboard 3 - IC50

A *dashboard* da Figura 3-4 ilustra informações detalhadas sobre a variável IC50. Ao analisarmos esta variável verificamos que existem 280 141 valores associados, o que está correto, pois inicialmente havia 310 904 dados, porém, as tarefas da limpeza e tratamento de dados implicaram que alguns dados fossem eliminados.

Já o valor médio do IC50 toma o valor de 0.55, tendo associado um desvio padrão de 0.06, o que reforça que o valor da média está correto. Uma das razões para o valor médio do IC50 ser elevado prende-se com o facto de grande parte dos medicamentos testados nas linhas celulares terem valores de IC50 superiores a 0.5. Isto é, só uma pequena parte dos medicamentos testados em cada linha celular é que têm valores de IC50 baixos, ou seja, são eficazes, enquanto a outra parte são medicamentos com valores de IC50 mais elevados, o que implica a média apresentar um valor de IC50 de aproximadamente 0.55.

Em relação ao valor mínimo e máximo temos os valores de 0.2577 e 0.7749, respetivamente.

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

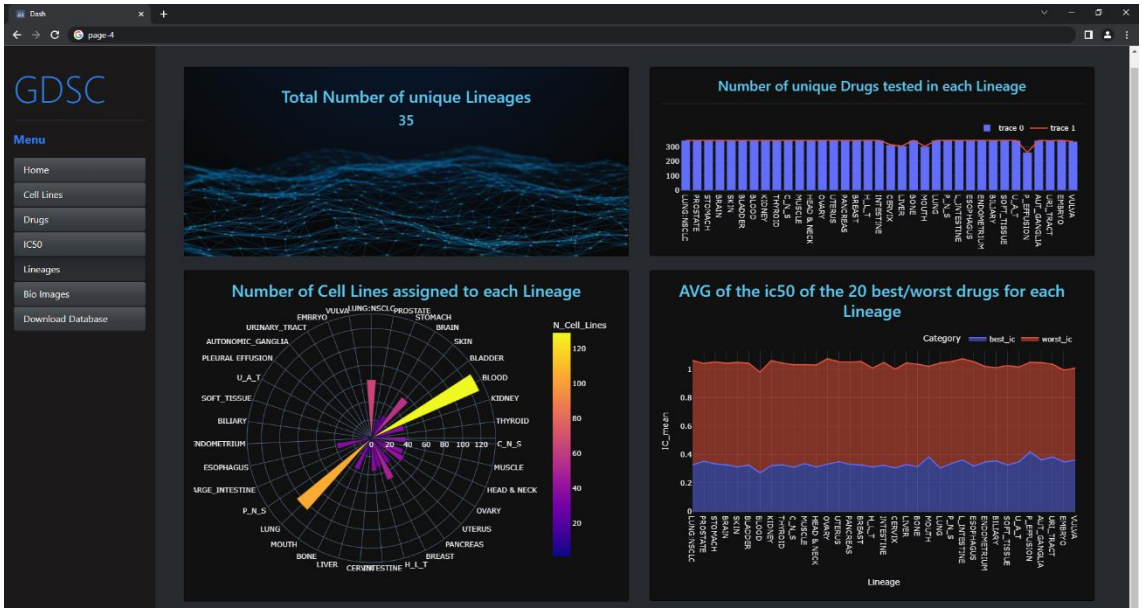

Figura 3-5 Dashboard 4 – Órgão

De acordo com a informação presente no *dashboard* da Figura 3-5, verificamos que na base de dados estão representados 35 órgãos. Além de analisarmos o valor do número de órgãos presentes na base de dados, podemos considerar também a sua frequência, assim como o número de linhas celulares que estes têm associadas. Uma das conclusões que podemos extrair desta análise é que existem efetivamente órgãos que estão mais bem representados a nível informativo (isto é, têm uma maior quantidade de dados associados) do que outros, o que pode influenciar o comportamento dos algoritmos de recomendação.

Uma outra estatística presente na *dashboard* está relacionada com o facto de, praticamente todos os órgãos terem sido testados com todos os medicamentos disponíveis, à exceção do mieloma múltiplo que apenas foi testado com 262 medicamentos. Uma das justificações para este resultado, prende-se com o facto deste cancro ter associado apenas uma linha celular.

Por fim, temos o gráfico de área que visa comparar a média do IC50 relativa aos 20 melhores e piores medicamentos de cada parte do corpo. Desta forma, é possível observarmos através das áreas, o grau de proporcionalidade entre as médias dos melhores e piores IC50. A título de exemplo, podemos observar que para o sangue a média do IC50 dos 20 melhores medicamentos é de aproximadamente, 0.27, enquanto a média do IC50 dos 20 piores medicamentos é de 0.70, logo a área relativa aos melhores IC50 será menor

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

que a área correspondente aos 20 piores IC50. De forma análoga para as outras partes do corpo, verificamos, que a área dos melhores IC50 é praticamente sempre menor que a área dos 20 piores IC50.

3.3 Teste da premissa

O teste da premissa tem como objetivo validar se linhas celulares similares têm tratamentos similares. Para tal, é necessário desenvolver uma métrica para a similaridade das imagens das linhas celulares assim como uma métrica para a similaridade dos tratamentos. Por fim, foi então necessário um criar um *dataset* que contenha estas 2 métricas para depois podermos correlacioná-las.

3.3.1 Similaridade das imagens

A análise à base de dados de imagens, inicia-se com o cálculo da similaridade entre todas estas imagens (todos os pares possíveis), que resultou em 868 624 interações, ou seja, foram obtidas 868 624 similaridades correspondentes a cada uma destas combinações (932*932). Importa referir ainda, que por questões de eficiência, desenhamos o algoritmo de forma que este ignorasse todas as combinações repetidas e inversas (por exemplo, a similaridade da imagem A com a imagem A, que não faz sentido registar, e a similaridade da imagem B com a imagem A, uma vez que já obtivemos a similaridade de A com B), restando 433 846 resultados de similaridade. Este resultado advém do seguinte cálculo:

$$\frac{[(932 * 932) - 932]}{2}$$

3.3.2 Similaridade dos tratamentos

A similaridade dos tratamentos consiste em perceber o grau de similaridade entre os 30 melhores medicamentos de duas linhas celulares. Isto é, vamos perceber quantos medicamentos em comum é que existem entre duas linhas celulares. Importa ainda referir que optamos por escolher o valor de 30 medicamentos, pois tal como evidenciado na análise exploratória de dados, verificamos que, quando os medicamentos estão ordenados por ordem crescente, através do IC50, em média, os primeiros 30 medicamentos são efetivamente os mais eficazes.

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

No que respeita à similaridade dos tratamentos, o objetivo passa por desenvolver uma métrica que seja capaz de retornar um valor para a similaridade do tratamento. Para tal, utilizamos o algoritmo *Nearest Neighbors*, em que o parâmetro escolhido relativo à métrica da distância foi o *cosine*, pois esta tem a vantagem de ter em conta o número de medicamentos comuns entre ambas as linhas celulares de forma robusta.

Uma vez que o próximo passo passou por calcular a correlação entre estas duas novas variáveis (similaridade das imagens e similaridade dos tratamentos), os valores destas similaridades foram incorporados numa base de dados, conforme é possível observar na Tabela 3-4.

Tabela 3-4 Amostra da nova base de dados

CELL_LINE_X	ORGAN_X	CELL_LINE_Y	ORGAN_Y	IMAGE_SIMILARITY	TREATMENT_SIMILARITY
CAL-51	BREAST	A2058	SHIN	0.716233	0.666973
JIMT-1	BREAST	TT	THYROID	0.588508	0.507093
PL4	PANCREAS	TT	THYROID	0.543423	0.628828
CAL-51	BREAST	MDA-MB-157	BREAST	0.582037	0.615457

Numa primeira abordagem testámos a correlação entre todos os dados das similaridades (os 433 846), e o resultado obtido, quantificado pelo coeficiente de *Pearson*, foi de aproximadamente 0,08, o que é uma correlação demasiado baixa, invalidando a premissa.

No entanto, decidimos não desistir nesta fase e tentar encontrar alguma solução que permitisse provar tal premissa, pois como verificamos anteriormente na revisão de literatura alguns atores conseguiram validar esta premissa, apesar de utilizarem abordagens/base de dados diferentes. Posto isto, submetemos os dados a diversas técnicas estatísticas até que detetamos que a técnica de agrupamento seria uma boa estratégia para estes dados. A partir daí, a estratégia adotada passou então por dividir em subconjuntos a base de dados inicial com base no tipo de órgão. Tal como visto na AED, existem 35 órgãos diferentes, pelo que houve a necessidade de, antes de executar o algoritmo, filtrar a base de dados por órgão. Com esta abordagem, conseguimos obter correlações mais altas. A título de exemplo, o órgão Pâncreas teve uma correlação de 0.56.

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

4 MODELO

Este capítulo descreve todas as fases do processo de desenvolvimento dos algoritmos de recomendação de medicamentos anticancerígenos, desde o cálculo das similaridades entre imagens *DNA microarray* até à recomendação de um conjunto de medicamentos.

4.1 Algoritmo de Recomendação

Os algoritmos de recomendação desenvolvidos neste projeto têm como objetivo central gerar recomendações dos melhores medicamentos para uma determinada linha celular, isto é, pretendemos construir um modelo capaz de recomendar um conjunto de possíveis melhores tratamentos, para um determinado paciente. Com isto, facilmente percebemos que o sistema é constituído essencialmente por duas fases distintas, sendo estas, cálculo da similaridade do paciente alvo, e posteriormente, recomendação dos melhores tratamentos.

O propósito principal da execução da primeira fase é de testar a similaridade da linha celular do paciente alvo, com todas as outras linhas celulares disponíveis na base de dados e assim, identificar quais as linhas celulares que têm um maior grau de similaridade em relação a esta.

A fase seguinte, prende-se com a recomendação de um conjunto de medicamentos para a linha celular alvo, com base nas similaridades testadas na fase 1, pois tal como podemos verificar na etapa acima apresentada, foi provado que linhas celulares similares têm tratamentos similares.

4.1.1 Fase 1 – Similaridade entre linhas celulares

O principal objetivo desta fase é de testar a similaridade de uma nova linha celular, que entra no sistema, com todas as outras que estão presentes na base de dados, pertencentes ao mesmo órgão, de forma a determinar as linhas celulares mais similares com esta.

Desta forma, conseguimos então determinar quais as linhas celulares que partilham de uma expressão genómica mais idêntica com a linha celular alvo. Como forma de exemplificar o funcionamento desta fase na prática, apresentamos na Figura 4-1 um esquema que demonstra o fluxo de processos envolvidos nesta fase.

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

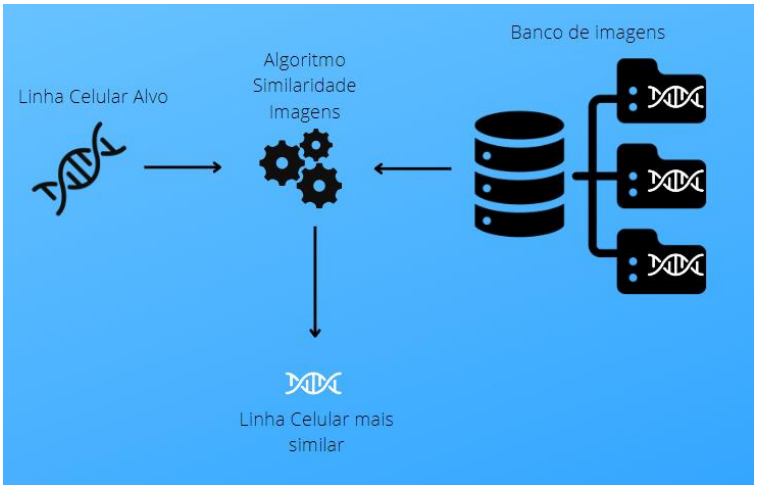

Figura 4-1 Fluxo do processo de similaridade de imagens

4.1.2 Fase 2 – Construção de um algoritmo de recomendação

A próxima fase consistiu no desenvolvimento do algoritmo de recomendação, cujo principal objetivo é o de atribuir um conjunto de possíveis medicamentos mais eficientes para uma determinada linha celular alvo, alicerçado nas similaridades mais expressivas, das linhas celulares, testadas na fase 1.

De forma a tornar mais eficiente e robusta a fase da recomendação de medicamentos anticancerígenos, decidimos desenvolver não um, mas dois algoritmos de recomendação, diferentes, mas com um propósito comum, o de recomendar os medicamentos mais eficazes para determinada linha celular.

4.1.2.1 Algoritmo de recomendação – collaborative filtering

Uma vez determinadas quais as linhas celulares mais similares com a linha celular alvo, e sabendo que linhas celulares similares têm tratamentos similares, podemos então atribuir os medicamentos mais eficazes, segundo o IC50, das linhas celulares mais similares, à linha celular alvo. Esta recomendação tem por base a utilização da técnica de recomendação denominada *collaborative filtering*, na qual consoante a similaridade das imagens de *DNA*, são recomendados os medicamentos mais eficazes.

O esquema da Figura 4-2 ilustra visualmente o funcionamento deste algoritmo de recomendação.

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

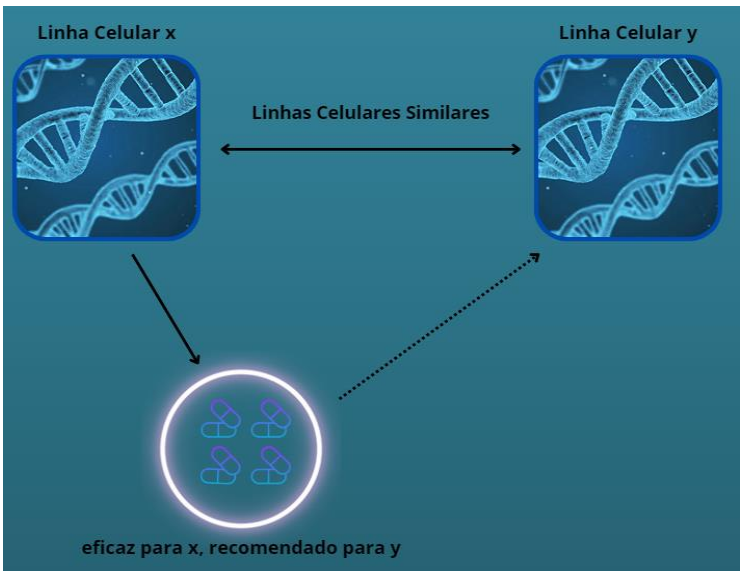

Figura 4-2 Esquema do sistema de recomendação collaborative filtering

4.1.2.2 Algoritmo de recomendação – non personalized

Este algoritmo tem como função principal, analisar os dados presentes na base de dados, com o objetivo de definir, para um determinado órgão, a seguinte informação: Quais os medicamentos mais frequentes e eficazes para aquele órgão.

Com isto, podemos retribuir um conjunto de medicamentos eficazes que são comuns a várias linhas celulares daquele órgão. Através desta abordagem, conseguimos determinar quais são os medicamentos que melhor resultam para o tratamento de um determinado órgão, o que pode ser benéfico pois, como a linha celular alvo pertence também àquele órgão, então provavelmente, alguns dos medicamentos mais frequentes e eficazes naquele órgão também servirão para aquela linha celular. O esquema da Figura 4-3 ilustra visualmente o funcionamento deste algoritmo de recomendação.

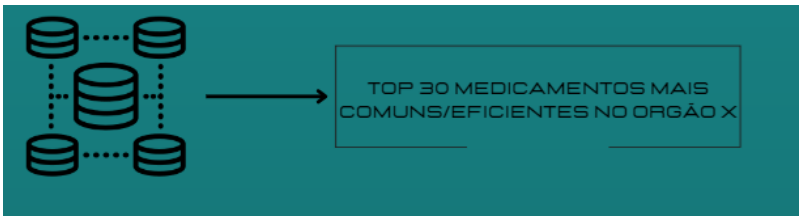

Figura 4-3 Esquema sistema de recomendação non personalized

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

4.2 Tecnologias utilizadas

O *Visual Studio Code* foi o ambiente de desenvolvimento utilizado em todo o processo de construção do presente projeto. A escolha deste ambiente de desenvolvimento prende-se com o facto de suportar um grande conjunto de linguagens / tecnologias, tais como, *Python*, *HTML*, *CSS*, *JavaScript*, entre outras. Para além de permitir a utilização de uma variedade de linguagens diferentes, este inclui ajuda na área da depuração, assim como, controlo de versões GIT incorporado. Estas funcionalidades são muito úteis, principalmente no desenvolvimento de aplicações, tal como é o caso da linha de investigação no qual este projeto está inserido.

Figura 4-4 Tecnologias utilizadas

Em relação às bibliotecas disponíveis no *Python*, foram utilizadas neste projeto:

- Pandas: para manipulação e análise das bases de dados;
- *Numpy*: computação matemática, matrizes;
- *OS*: manipulação dos diretórios do sistema operativo;
- *Biopython*: manipulação de dados biológicos;
- *SciKit-image*: processamento de imagens, similaridade;
- *Multiprocessing*: utilizado para paralelizar código
- tqdm: barras de progresso;
- *Scipy*: calcular estatísticas, teste de hipóteses, correlações;
- *Sklearn:* ML, similaridades de tratamentos;
- *Flask:* desenvolvimento da *API.*
- *Dash:* criação dos *dashboards* interativos para uma melhor visualização dos resultados da análise exploratória 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

5 AVALIAÇÃO

Os algoritmos desenvolvidos ao longo deste projeto, tiveram como principal objetivo, a criação de um sistema de recomendação de medicamentos anticancerígenos capaz de gerar um conjunto de medicamentos eficazes para uma determinada linha celular cancerígena alvo.

No entanto, torna-se imprescindível testar a eficácia e a segurança das recomendações de medicamentos geradas pelos algoritmos, o que nos leva à fase da avaliação destes mesmos algoritmos.

Assim, foram utilizadas duas métricas de avaliação diferentes, de modo a percebermos o quão eficazes estes algoritmos são.

5.1 Algoritmo de recomendação - *collaborative filtering*: exemplo pâncreas

A título exemplificativo, ao executarmos o algoritmo de recomendação baseado na *collaborative filtering*, (tendo como linha celular alvo a SW1990) este recomenda os medicamentos mais eficazes da linha celular BXPC-3 à linha celular SW1990, pois a linha celular BXPC-3 é a mais similar em termos de imagem *DNA*.

Na Tabela 5-1 é possível visualizar os 30 medicamentos mais eficazes reais, para a linha celular SW1990, ou seja, a nossa linha celular alvo.

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

Tabela 5-1 Lista dos 30 melhores medicamentos reais da linha celular SW1990

LINHA CELULAR ALVO - SW1990	
DRUG_NAME	RANK
THAPSIGARGIN	1
ELESCLOMOL	2
DOCETAXEL	3
LUMINESPIB	4
SN-38	5
EPOTHILONE B	6
VINBLASTINE	7
VINORELBINE	8
DACTOLISIB	9
AZD4877	10
GEMCITABINE	11
ARRY-520	12
DACINOSTAT	13
BRYOSTATIN 1	14
TANESPIMYCIN	15
NSC319726	16
TRAMETINIB	17
TEMSIROLIMUS	18
PANOBINOSTAT	19
DOXORUBICIN	20
OMIPALISIB	21
AZD7762	22
OBATOCLAX MESYLATE	23
PEVONEDISTAT	24
RTRAIL	25
PD0325901	26
PLX_6522	27
TRICHOSTATIN A	28
MITOMYCIN-C	29
FLAVOPIRIDOL	30

Na Tabela 5-2 podemos observar a lista de medicamentos recomendados, pelo algoritmo, para a linha celular SW1990.

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

Tabela 5-2 Lista dos 30 medicamentos recomendados para a linha celular SW1990 (collaborative filtering)

SISTEMA DE RECOMENDAÇÃO - COLLABORATIVE FILTERING	
LINHA CELULAR MAIS SIMILAR - BXPC-3	
DRUG_NAME	RANK
SN-38	1
EPOTHILONE B	2
SEPANTRONIUM BROMIDE	3
DOCETAXEL	4
VINORELBINE	5
VINBLASTINE	6
THAPSIGARGIN	7
BRYOSTATIN 1	8
AZD4877	9
TRAMETINIB	10
GEMCITABINE	11
DACINOSTAT	12
AZD5582	13
TH2-2-102-1	14
LUMINESPIB	15
RTRAIL	16
PLH_6522	17
OMIPALISIB	18
SNX-2112	19
DACTOLISIB	20
PANOBINOSTAT	21
ELESCLOMOL	22
OBATOCLAX MESYLATE	23
JQ1	24
ARRY-520	25
IAP_5620	26
TANESPIMYCIN	27
DOXORUBICIN	28
IAP_7638	29
NSC319726	30

Observando a Tabela 5-2, de uma forma geral, conseguimos validar que a maioria dos medicamentos recomendados pelo algoritmo estão em concordância com os reais (da Tabela 5-1), obtendo assim uma taxa de acerto elevada.

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

De forma a avaliar os resultados, em primeiro lugar, utilizamos uma métrica de avaliação denominada *HitRate*. Esta métrica calcula o número de acertos sobre o número total de possibilidades através da seguinte fórmula:

$$HitRate = \frac{n1 \cap n2}{n}$$

Aplicando a fórmula aos resultados apresentados na tabela temos o seguinte *Hit-Rate*:

$$HitRate = \frac{23}{30} = 0.7666$$

A taxa de 0.7666, significa que o algoritmo desenvolvido acertou em 23 dos 30 medicamentos possíveis, isto é, o algoritmo teve uma precisão de aproximadamente 77%, para esta linha celular em específico.

Agora utilizamos uma outra métrica de avaliação, denominada *cosine similarity*, que pode ser representada pela seguinte fórmula:

$$cosine\ similarity = Sc(A,B) = \frac{A * B}{||A|| * ||B||} = \frac{\sum_{i=1}^{n} Ai * Bi}{\sqrt{\sum_{i=1}^{n} Ai^2} * \sqrt{\sum_{i=1}^{n} Bi^2}}$$

Aplicando a fórmula aos dados apresentados nas tabelas acima representadas, temos o seguinte resultado:

$$Sc(A,B) = 0.6832$$

O valor de 0.6832 indica-nos que o grau de similaridade entre os medicamentos recomendados para a linha celular SW1990 e os medicamentos reais desta mesma linha, é de aproximadamente 68%.

A principal diferença entre esta última métrica de avaliação e a métrica do *HitRate*, está na fórmula matemática utilizada para obter o resultado da similaridade. Isto é, esta métrica permite-nos perceber em quantos medicamentos o algoritmo de recomendação acertou de forma mais completa e robusta.

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

5.2 Algoritmo de recomendação – *non-personalized*: exemplo pâncreas

Para o algoritmo de recomendação *non personalized*, apresentamos na Tabela 5-3 com os medicamentos mais frequentes para um determinado órgão, que neste caso é o pâncreas.

Tabela 5-3 Lista dos 30 medicamentos recomendados (non personalized)

SISTEMA DE RECOMENDAÇÃO - NON PERSONALIZED	
DRUG_NAME	FREQUENCY
VINBLASTINE	22
DOCETAXEL	22
SEPANTRONIUM BROMIDE	22
OMIPALISIB	22
BRYOSTATIN 1	21
EPOTHILONE B	21
DACINOSTAT	21
SN-38	21
ELESCLOMOL	20
AZD4877	20
VINORELBINE	20
LUMINESPIB	19
THAPSIGARGIN	19
PLX_6522	19
PANOBINOSTAT	19
DACTOLISIB	18
DOXORUBICIN	18
TEMSIROLIMUS	18
GEMCITABINE	18
PD0325901	17
ISPINESIB MESYLATE	17
TH2-2-102-1	16
NSC319726	16
ARRY-520	16
TANESPIMYCIN	15
TRAMETINIB	14
SNX-2112	12
OBATOCLAX MESYLATE	10
DAPORINAD	10
RTRAIL	8

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

Observando a Tabela 5-3, de uma forma geral, conseguimos validar que a maioria dos medicamentos recomendados pelo algoritmo estão em concordância com os reais (da Tabela 5-1), obtendo assim uma taxa de acerto elevada.

Uma vez mais, de forma a medir o desempenho deste algoritmo, utilizamos a métrica do *HitRate*.

$$HitRate = \frac{25}{30} = 0.8333$$

O facto deste algoritmo acertar em 25 dos 30 medicamentos possíveis, fez com que este tivesse uma precisão de aproximadamente 83%, para esta linha celular em específico.

Este resultado tem um significado diferente do anterior, pois estes 83% significam que, 25 dos medicamentos mais frequentes no tratamento do cancro do pâncreas, são comuns com os melhores medicamentos reais da linha celular SW1990.

Aplicando agora a fórmula do *cosine similarity*, temos o seguinte resultado:

$$Sc(A, B) = 0.7591$$

Este resultado, novamente, traduz-se num resultado positivo.

5.3 Resultados para diferentes órgãos

Depois de analisarmos o resultado dos algoritmos numa determinada linha celular alvo, podemos agora analisar os resultados globais para os órgãos.

Para avaliar os resultados para o órgão pâncreas temos de dividir o *dataset* em treino e teste. O órgão pâncreas tem 32 linhas celulares associadas, pelo que 22 linhas celulares são afetas ao treino e as restantes 10 afetas ao teste, tendo assim uma proporção de aproximadamente 2/3 e 1/3.

Para avaliar os algoritmos, utilizamos novamente a métrica do *HitRate* e do *cosine similarity*, no entanto, agora efetuamos as médias de ambas as métricas de avaliação para cada uma das linhas celulares presentes na base de teste, de forma a perceber qual é a precisão dos algoritmos para o órgão pâncreas.

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

Ao observarmos os resultados abaixo apresentados, verificamos que para o pâncreas, o algoritmo de recomendação *non personalized* obteve uma classificação ligeiramente maior que o algoritmo de recomendação *collaborative filtering*. Podemos também verificar que, tanto num algoritmo como no outro, a métrica de avaliação *cosine similarity* obtém sempre um valor ligeiramente inferior ao da métrica *HitRate*, o que acaba por ser normal, pois tal como abordado anteriormente, o *cosine similarity* é uma métrica de avaliação mais completa e robusta. No entanto, de uma forma geral, podemos concluir que, tanto a métrica do *HitRate* como a *cosine similarity*, têm valores muito próximos, o que evidencia coerência e conformidade dos resultados exibidos.

Contudo, importa ainda analisar os resultados destas métricas para os outros órgãos presentes na base de dados, de forma a determinar a real eficiência destes dois algoritmos, num contexto mais abrangente. Para tal, aplicamos a lógica relativa ao órgão pâncreas, para os restantes órgãos, com a finalidade de, apresentarmos os resultados obtidos para cada um destes.

Objetivando uma melhor visualização dos dados obtidos, apresentamos uma tabela onde constam as métricas de avaliação para uma amostra de 5 órgãos distintos:

Tabela 5-4 Métricas de avaliação vários órgãos

ORGÃOS	MÉTRICAS DE AVALIAÇÃO			
	COLLABORATIVE FILTERING		NON-PERSONALIZED	
	HIT RATE	COSINE SIMILARITY	HIT RATE	COSINE SIMILARITY
PANCREAS	69.00	68.32	77.33	75.91
BREAST	56.25	53.87	66.04	64.31
THYROID	74.66	72.54	76.66	74.64
BRAIN	75.00	73.76	80.42	80.34
UTERUS	68.89	66.31	75.56	72.32

De uma forma geral, os resultados obtidos a partir das métricas de avaliação utilizadas, são promissores, no sentido em que, temos taxas de eficácia a rondar os 70%, em ambos os algoritmos de recomendação. Podemos ainda observar que os valores relativos ao *cosine similarity* são quase sempre ligeiramente inferiores aos valores do *HitRate*.

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

Em relação aos resultados da avaliação auferidos para ambos os algoritmos utilizados, conferimos que estes possuem valores muito idênticos, o que é favorável a este estudo, pois de certa maneira, prova que existe coerência e consistência nas análises realizadas a estes dados.

Por fim, podemos reparar que existem órgãos que têm melhores taxas de eficácia do que outros, o que é perfeitamente expetável, pois nem todos os órgãos têm a mesma quantidade e heterogeneidade de dados, isto é, existem órgãos que têm associadas mais linhas celulares e consequentemente mais dados/informações do que outros, o que tem uma influência direta no comportamento e na eficácia dos algoritmos, daí existirem órgãos com classificações mais altas do que outros.

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

6 DISPONIBILIZAÇÃO

Este capítulo é dedicado à explicação deste projeto como um todo, uma vez que este projeto está inserido numa linha de investigação comum entre dois projetos, tendo como meta final, o desenvolvimento de um Sistema de Apoio à Decisão Clínica. Posto isto, podemos então dividir este sistema em duas partes distintas, sendo estas a construção de componentes referentes à interface gráfica do utilizador e a construção de componentes de configuração e treino dos modelos usados no sistema de recomendação de medicamentos anticancerígenos. Deste modo, temos de conseguir integrar estas duas componentes do sistema, com o propósito, de conceber um sistema completo, original e totalmente funcional.

6.1 *Application Programming Interface*

Nesta última fase, importa agora arquitetar uma estratégia, que seja capaz de disponibilizar todos os modelos/algoritmos desenvolvidos neste projeto, que vão desde o cálculo das similaridades das imagens *DNA microarray*, até aos algoritmos de recomendação, de forma que o trabalho desenvolvido na mesma linha de investigação que este, consiga obter todos os resultados que necessita, para depois poder trabalhá-los e providenciá-los através de um sistema de apoio à decisão clínica. No fundo, o que é pretendido nesta fase é desenvolver uma aplicação completa, em que todo o processo de análise de dados se funde com a disponibilização gráfica dos resultados gerados.

Para tal, foi utilizada uma tecnologia denominada por interface de programação de aplicações, do inglês *Application Programming Interface*, normalmente conhecida pela sigla *API*, em que todos os processos desenvolvidos no presente projeto, assim como os seus resultados, são facultados num determinado servidor. Depois disso, a interface gráfica que está a ser desenvolvida pelo outro projeto, terá de conectar-se com o servidor *API*, de modo a providenciar os resultados gerados pelos modelos de recomendação, graficamente mais acessível, percetível e compreensível perante o utilizador final. Deste modo, os utilizadores, terão ao seu dispor um conjunto de informações otimizadas e coerentes capazes de ajudar no processo de tomada de decisão.

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

Importa ainda referir que a *API* foi construída a partir de um *framework*, denominado *Flask*, o que se tornou muito útil, uma vez que os modelos de recomendação foram desenvolvidos através da linguagem *Python*, tornando a integração destes modelos na *API* mais produtiva.

Para conseguirmos visualizar melhor o funcionamento da *API* aqui desenvolvida, é apresentado na Figura 6-1 um esquema onde são representados todos os processos presentes na *API*.

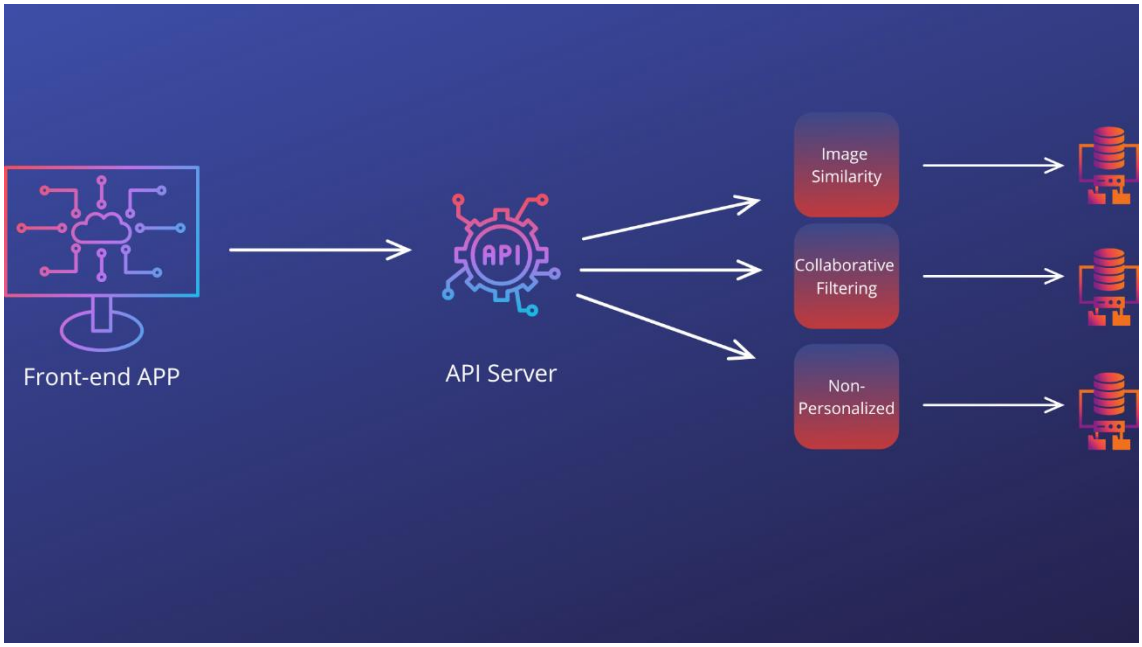

Figura 6-1 Arquitetura do SiReMA

6.2 Diagrama de pacotes do SiReMA

Nesta secção apresentamos o diagrama de pacotes do SiReMA, e através deste, conseguimos visualizar toda a arquitetura da aplicação, dividida em agrupamentos lógicos (pacotes), demonstrando assim os fluxos e as relações entre estes. Importa realçar que este diagrama é importante para este projeto no sentido em que indica de que forma os serviços fornecidos pela *API* serão disponibilizados ao outro projeto da mesma linha de investigação.

Na Figura 6-2 está representado o diagrama de pacotes do SiReMA, e em primeiro lugar, importa evidenciar a razão pela qual foram utilizadas diferentes cores para representar os pacotes presentes no diagrama. Sendo este um projeto que está inserido numa linha de

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

investigação conjunta com um outro projeto, é essencial definir em que áreas é que cada um se foca e atua. Nesse sentido, as diferentes cores utilizadas representam a divisão das tarefas desenvolvidas em ambos os projetos, mais especificamente, a cor verde determina quais as tarefas efetivadas pelo presente projeto, nomeadamente, o desenvolvimento dos modelos de recomendação, enquanto a cor azul, representa todas as tarefas relacionadas com o outro projeto, responsável pela criação da interface gráfica do utilizador e que integra os modelos de recomendação aqui desenvolvidos.

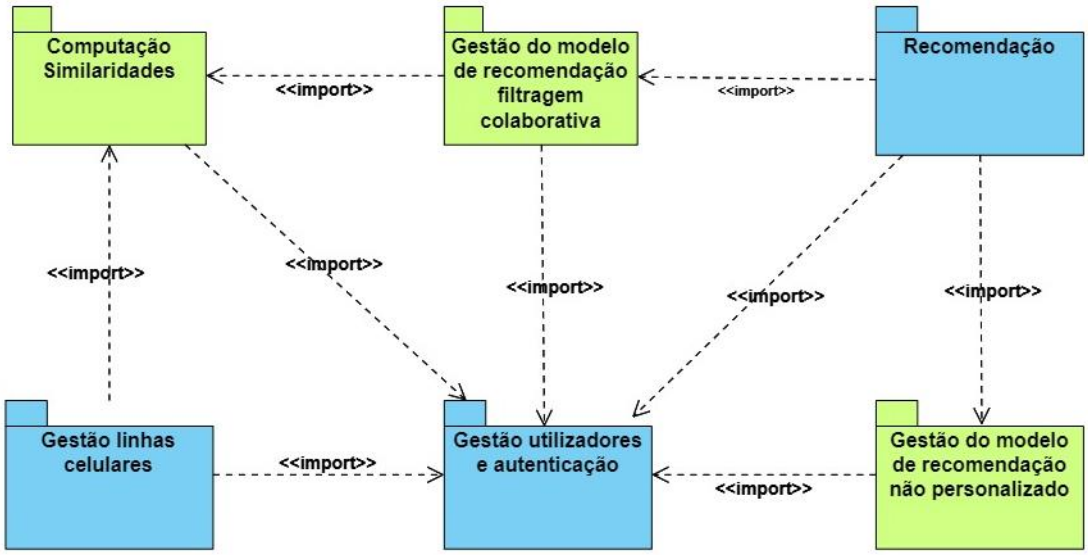

Figura 6-2 Diagrama de pacotes

Ao analisar a Figura 6-2, facilmente observamos os fluxos dos processos presentes na mesma, desde a realização do *login*, até a apresentação de um conjunto de medicamentos recomendados para um dado paciente. Podemos também visualizar, de que forma é que os modelos de recomendação foram integrados na aplicação, bem como, expomos a relação que esta aplicação tem com as bases de dados utilizadas para desenvolver os modelos.

6.3 Manutenção do modelo

Antes de mais importa relembrar que os modelos de recomendação aqui desenvolvidos, foram desenhados e otimizados para lidar com as variáveis presentes nas bases de dados selecionadas para este projeto. Em segundo lugar, é necessário ter em consideração que estas bases de dados, estão em constante progresso e evolução, isto é, ao longo do tempo

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

vão sendo inseridos novos dados, novas informações, bem como, novas técnicas, o que é positivo, pois desta maneira o conhecimento que se adquire através destas bases de dados torna-se mais consistente, confiável, sustentado e fundamentado.

Posto isto, então facilmente percebermos que à medida que as bases de dados vão sofrendo alterações, os modelos também terão de voltar a ser treinados e avaliados, de forma a, percebermos quais são os reais impactos causados pela evolução das bases de dados.

Deste modo, consideramos que esta é uma fase extremamente importante, pois permite minimizar o risco destes modelos, bem como da aplicação desenvolvida, se tornar obsoleta e ineficaz.

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

7 CONCLUSÃO

A realização do presente projeto em conjunto com outro projeto, da mesma linha de investigação, culminou na conceção de um sistema de apoio à decisão clínica. Os projetos apesar de terem um objetivo final comum, conceber um sistema, focaram-se em desenvolver diferentes componentes deste, onde o projeto aqui apresentado dedicou-se ao desenvolvimento das componentes de configuração e treino, sendo assim responsável pelos algoritmos de recomendação e consequente criação do modelo de recomendação, enquanto o projeto parceiro dedicou-se ao desenvolvimento das componentes de execução, validação e ainda a interface gráfica do sistema capaz de disponibilizar os resultados obtidos ao utilizador final. Os algoritmos desenvolvidos neste projeto, nomeadamente o cálculo das similaridades e as recomendações serão integrados no sistema final através de uma *API*, onde o outro projeto teve de se conectar com o servidor *API* para ter acesso aos resultados gerados pelos algoritmos criados e assim apresentá-los na interface gráfica.

Para o desenvolvimento do sistema de recomendação, os algoritmos desenvolvidos objetivaram a geração de um conjunto de medicamentos anticancerígenos mais eficazes para um determinado paciente. Estes algoritmos para além de gerarem recomendações visam otimizar e automatizar os processos inerentes à atividade clínica, proporcionando um serviço mais eficiente.

7.1 Principais contributos

Este sistema constitui um instrumento de apoio a decisão clínica no sentido em que os resultados gerados, neste caso as recomendações, têm como objetivo primordial dotar o médico de um conjunto de informações importantes, auxiliando assim o processo de tomada de decisão, tornando este processo mais sustentado e preciso. O projeto aqui desenvolvido pretende contribuir de forma significativa para a implementação de uma medicina personalizada.

Neste sentido foram desenvolvidos dois algoritmos que tiveram por base abordagens diferentes, *collaborative filtering* e *non personalized.* O *collaborative filtering* teve como objetivo gerar um conjunto de medicamentos eficazes com base na expressão genómica

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

presente nas imagens *microarray DNA* do paciente. Enquanto o *non personalized* teve como objetivo gerar um conjunto de medicamentos eficazes, tendo por base os medicamentos mais eficazes para aquele órgão.

Desta forma, facilmente percebemos que este último atua no sentido de complementar e reforçar os resultados gerados pelo primeiro algoritmo, tornando assim a decisão do medico mais robusta e eficaz.

Ambos os projetos que foram desenvolvidos nesta linha de investigação, mais especificamente, o presente projeto e o outro que teve como propósito desenvolver uma interface gráfica, têm os modelos, algoritmos, códigos entre outras informações relativas a todo o projeto, disponibilizadas no *github*, através do *link*: https://github.com/MonicaTeles/SiReMA.

Importa ainda referir que no âmbito deste projeto foi desenvolvido e publicado um artigo científico cujo propósito foi de tentar provar que linhas celulares similares partilham de tratamentos similares. O estudo desta premissa teve particular importância neste projeto pois, o algoritmo *collaborative filtering*, está alicerçado nesta premissa.

7.2 Limitações

Uma das principais limitações que enfrentamos neste projeto está relacionada com a quantidade/qualidade dos dados disponíveis. Por exemplo, a base de dados GDSC originalmente tinha 310 904 registos, no entanto, após o processo de preparação dos dados (limpeza e tratamento) restaram apenas 280 141 registos, o que se traduz numa perda de aproximadamente 10% de informação. Como sabemos, os algoritmos de recomendação necessitam de uma grande quantidade de dados para se tornarem precisos e confiáveis caso contrário podemos estar perante um enviesamento dos resultados.

Um outro ponto tão ou mais importante que a quantidade de dados está na sua heterogeneidade. A título de exemplo, como é possível visualizar na Figura 3-6 *dashboard* 4 anteriormente apresentada, existem partes do corpo que têm muitas linhas celulares associadas, ou seja, têm muita informação, enquanto existem outras partes do corpo que têm muito poucas linhas celulares associadas. Estas diferenças têm um impacto direto na viabilidade e eficiência dos algoritmos.

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

Por fim, uma outra limitação com o qual nos deparamos está nos valores em falta, nomeadamente no que diz respeito à variável IC50. Isto porque existem 345 medicamentos e seria expectável que cada um destes fosse testado em todas as linhas celulares, no entanto tal não se verificou, ficando medicamentos por testar em algumas linhas celulares.

7.3 Trabalhos futuros

Uma das áreas que pode ser interessante explorar mais aprofundadamente no âmbito deste projeto, relaciona-se com o tipo de imagens que podemos analisar. Neste projeto utilizamos imagens *DNA microarray*, no entanto, existe uma panóplia de outro tipo de imagens que podemos utilizar, como por exemplo, radiografias, tomografias, ressonância magnética, entre outras.

Uma outra recomendação para possíveis trabalhos futuros, vincula-se com a previsão da resposta das linhas celulares aos medicamentos utilizados, ou seja, tem por base a construção de algoritmos capazes de prever o IC50 de uma determinada linha celular quando testada com um determinado medicamento / composto.

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

REFERÊNCIAS

ArrayExpress - Functional Genomics Data. (2022). https://www.ebi.ac.uk/biostudies/arrayexpress

Abdurakhmonov, I. Y. (2016). Bioinformatics: Basics, Development, and Future. *Bioinformatics - Updated Features and Applications*. https://doi.org/10.5772/63817

Adak, M. F., & Ucar, M. (2021). A Book Recommendation System Using Decision Tree-based Fuzzy Logic for E-Commerce Sites. *HORA 2021 - 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Proceedings*. https://doi.org/10.1109/HORA52670.2021.9461319

Barbeau, J. (2018, June 5). *PDX and Personalized Medicine*. Crown Bioscience. https://blog.crownbio.com/pdx-personalized-medicine

Barboza, D. C. (2019). Artificial Intelligence and HR : The New Wave of Technology. *Technology Journal of Advances in Social Science and Humanities*, *5*(4), 715–720. https://doi.org/10.15520/jassh5

Baxevanis, A. D., & Ouellette, B. F. F. (2001). Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. In P. John Wiley & Sons (Ed.), *Briefings in Bioinformatics* (Vol. 2). https://doi.org/10.1093/bib/2.4.407

Brandão, L. C. P. (2020). *Wavelet-Based Cancer Drug Recommender System*. Polytechnic Institute of Coimbra.

Caldwell, G. W., Yan, Z., Lang, W., & Masucci, J. A. (2012). The IC(50) concept revisited. *Current Topics in Medicinal Chemistry*, *12*(11), 1282–1290. https://doi.org/10.2174/156802612800672844

Capes-Davis, A., Bairoch, A., Barrett, T., Burnett, E. C., Dirks, W. G., Hall, E. M., Healy, L., Kniss, D. A., Korch, C., Liu, Y., Neve, R. M., Nims, R. W., Parodi, B., Schweppe, R. E., Storts, D. R., & Tian, F. (2019). Cell Lines as Biological Models: Practical Steps for More Reliable Research. *Chemical Research in Toxicology*, *32*(9), 1733–1736. https://doi.org/10.1021/ACS.CHEMRESTOX.9B00215/ASSET/IMAGES/MEDIU

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

M/TX-2019-002152_0001.GIF

Carter, M., & Shieh, J. (2015). Cell Culture Techniques. *Guide to Research Techniques in Neuroscience*, 295–310. https://doi.org/10.1016/B978-0-12-800511-8.00014-9

Castiglioni, I., Rundo, L., Codari, M., Di Leo, G., Salvatore, C., Interlenghi, M., Gallivanone, F., Cozzi, A., D'Amico, N. C., & Sardanelli, F. (2021). AI applications to medical images: From machine learning to deep learning. *Physica Medica*, *83*, 9–24. https://doi.org/10.1016/J.EJMP.2021.02.006

Cauduro, A. (2018). *Deep Learning: o motor dos negócios na era da inteligência artificial*. Medium. https://medium.com/huia/inteligência-artificial-uma-corrida-desleal-80bfa53075ed

Chong, D. (2020). *Deep Dive into Netflix's Recommender System Towards Data Science*. Towards Data Science. https://towardsdatascience.com/deep-dive-into-netflixs-recommender-system-341806ae3b48

Costello, J. C., Heiser, L. M., Georgii, E., Gönen, M., Menden, M. P., Wang, N. J., Bansal, M., Ammad-Ud-Din, M., Hintsanen, P., Khan, S. A., Mpindi, J. P., Kallioniemi, O., Honkela, A., Aittokallio, T., Wennerberg, K., Collins, J. J., Gallahan, D., Singer, D., Saez-Rodriguez, J., Stolovitzky, G. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology*, *32*(12), 1202–1212. https://doi.org/10.1038/nbt.2877

Das, D., Sahoo, L., & Datta, S. (2017). A Survey on Recommendation System. *International Journal of Computer Applications*, *160*(7), 6–10. https://doi.org/10.5120/IJCA2017913081

Evans, W. E., & Johnson, J. A. (2001). Pharmacogenomics: the inherited basis for interindividual differences in drug response. *Annual Review of Genomics and Human Genetics*, *2*, 9–39. https://doi.org/10.1146/ANNUREV.GENOM.2.1.9

Feng, C., Khan, M., Rahman, A. U., & Ahmad, A. (2020). News Recommendation Systems-Accomplishments, Challenges Future Directions. *IEEE Access*, *8*, 16702–16725. https://doi.org/10.1109/ACCESS.2020.2967792

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

Genome, N. H. (2018). *Genetics vs. Genomics Fact Sheet*. https://www.genome.gov/about-genomics/fact-sheets/Genetics-vs-Genomics

Genomics of Drug Sensitivity in Cancer. (2022). https://www.cancerrxgene.org/

Goetz, L. H., & Schork, N. J. (2018). Personalized medicine: motivation, challenges, and progress. *Fertility and Sterility*, *109*(6), 952–963. https://doi.org/10.1016/J.FERTNSTERT.2018.05.006

Gomes, C. P. (2019). *Dispositivos Médicos e Medicina Personalizada*. Universidade de Lisboa.

Gonçalves, J. C. R., & Sobral, M. V. (2020). *Cultivo de células da teoria à bancada* (Editora UFPB (Ed.).

González-Larraza, P. G., López-Goerne, T. M., Padilla-Godínez, F. J., González-López, M. A., Hamdan-Partida, A., & Gómez, E. (2020). IC50 Evaluation of Platinum Nanocatalystsfor Cancer Treatment in Fibroblast, HeLa, and DU-145 Cell Lines. *ACS Omega*, *5*(39), 25381. https://doi.org/10.1021/ACSOMEGA.0C03759

Goodspeed, A., Heiser, L. M., Gray, J. W., & Costello, J. C. (2016). Tumor-Derived Cell Lines as Molecular Models of Cancer Pharmacogenomics. *Molecular Cancer Research: MCR*, *14*(1), 3. https://doi.org/10.1158/1541-7786.MCR-15-0189

Hagen, J. B. (2000). The origins of bioinformatics. *Macmillan Magazines Ltd*, *1*. https://bcb.unl.edu/yyin/teach/PBB2015/The origins of bioinformatics.pdf

Hanif, W., Amin Afzal, M., Ansar, S., Saleem, M., Ikram, A., Afzal, S., Amjad Fateh Khan, S., Ahmad Larra, S., Noor, H., & Saf, K. (2019). Artificial intelligence in bioinformatics . *Biomedical Letters*, *5*(1), 1–7.

Hoeben, A., Joosten, E. A. J., & van den Beuken-Van Everdingen, M. H. J. (2021). Personalized Medicine: Recent Progress in Cancer Therapy. *Cancers*, *13*(2), 1–3. https://doi.org/10.3390/CANCERS13020242

Huo, K. G., D'Arcangelo, E., & Tsao, M. S. (2020). Patient-derived cell line, xenograft and organoid models in lung cancer therapy. *Translational Lung Cancer Research*, *9*(5), 2214–2232. https://doi.org/10.21037/TLCR-20-154

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

Hynds, R. E., Vladimirou, E., & Janes, S. M. (2018). The secret lives of cancer cell lines. *DMM Disease Models and Mechanisms*, *11*(11). https://doi.org/10.1242/DMM.037366

Jayanthi, K., & Mahesh, C. (2018). A Study on machine learning methods and applications in genetics and genomics. *International Journal of Engineering and Technology(UAE)*, *7*(1.7 Special Issue 7), 201–204. https://doi.org/10.14419/IJET.V7I1.7.10653

Jiménez-Santos, M.J., García-Martín, S., Fustero-Torre, C., Di Domenico, T., Gómez-López, G., & Al-Shahrour, F. (2022). Bioinformatics roadmap for therapy selection in cancer genomics. *Molecular Oncology*. https://doi.org/10.1002/1878-0261.13286

Kashyap, H., Ahmed, H. A., Hoque, N., Roy, S., & Bhattacharyya, D. K. (2015). *Big Data Analytics in Bioinformatics: A Machine Learning Perspective* (Vol. 13, Issue 9). http://arxiv.org/abs/1506.05101

Kocarnik, J. M., Compton, K., Dean, F. E., Fu, W., Gaw, B. L., Harvey, J. D., Henrikson, H. J., Lu, D., Pennini, A., Xu, R., Ababneh, E., Abbasi-Kangevari, M., Abbastabar, H., Abd-Elsalam, S. M., Abdoli, A., Abedi, A., Abidi, H., Abolhassani, H., Adedeji, I. A., Force, L. M. (2021). Cancer Incidence, Mortality, Years of Life Lost, Years Lived with Disability, and Disability-Adjusted Life Years for 29 Cancer Groups from 2010 to 2019: A Systematic Analysis for the Global Burden of Disease Study 2019. *JAMA Oncology*. https://doi.org/10.1001/jamaoncol.2021.6987

Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J. A., Armañanzas, R., Santafé, G., Pérez, A., & Robles, V. (2006). Machine learning in bioinformatics. *Briefings in Bioinformatics*, *7*(1), 86–112. https://doi.org/10.1093/bib/bbk007

Li, X. (2021). Research on the Application of Collaborative Filtering Algorithm in Mobile E-Commerce Recommendation System. *Proceedings of IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers, IPEC 2021*, 924–926. https://doi.org/10.1109/IPEC51340.2021.9421092

Loewe, R. P., & Nelson, P. J. (2011). Microarray bioinformatics. *Methods in Molecular*

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

Biology (Clifton, N.J.), *671*, 295–320. https://doi.org/10.1007/978-1-59745-551-0_18

Mahesh, B. (2020). Machine Learning Algorithms - A Review. *International Journal of Science and Research* , *9*(1). https://doi.org/10.21275/ART20203995

Mana, S. C., & Sasipraba, T. (2021). A machine learning based implementation of product and service recommendation models. *Proceedings of the 7th International Conference on Electrical Energy Systems, ICEES 2021*, 543–547. https://doi.org/10.1109/ICEES51510.2021.9383732

Menden, M. P., Iorio, F., Garnett, M., McDermott, U., Benes, C. H., Ballester, P. J., & Saez-Rodriguez, J. (2013). Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. *PLoS ONE*, *8*(4). https://doi.org/10.1371/journal.pone.0061318

McCarthy, J. (2007). *What Is Artificial Intelligence?* http://www-formal.stanford.edu/jmc/

Moreau, Y., De Smet, F., Thijs, G., Marchal, K., & De Moor, B. (2002). *Functional Bioinformatics of Microarray Data: From Expression to Regulation*. https://doi.org/10.1109/JPROC.2002.804681

National Cancer Institute. (2022-a). *Definition of genome*. Retrieved August 26, 2022, from https://www.cancer.gov/publications/dictionaries/cancer-terms/def/genome

National Cancer Institute. (2022-b). *Definition of genomics*. Retrieved August 26, 2022, from https://www.cancer.gov/publications/dictionaries/cancer-terms/def/genomics

Naujoks, J. (2019). *Recommendations in time & context* . Medium. https://medium.com/swlh/recommendations-in-time-context-93b32f73d98d

Neves, C. E. (2010). *Experimentos de microarrays e teoria da resposta ao item*. Universidade de São Paulo.

Oliveira, C. G. (2014). *PREFREC: Uma Metodologia para Desenvolvimento de Sistemas de Recomendação utilizando Algoritmos de Mineração de Preferências*. Universidade Federal de Uberlândia Faculdade de Ciência da Computação.

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

Personalized Medicine. (2022, September 22). https://www.genome.gov/genetics-glossary/Personalized-Medicine

Pimentel, A. L., & Bueno, P. S. A. B. (2016). *Ferramentas de bioinformática na caracterização de alvos de medicamentos*.

Pokhriyal, M., Ratta, B., Yadav, B. S., Pokhriyal, M., Ratta, B., & Yadav, B. S. (2019). Bioinformatics and Microarray-Based Technologies to Viral Genome Sequence Analysis. *Microbial Genomics in Sustainable Agroecosystems*, 115–129. https://doi.org/10.1007/978-981-13-8739-5_6

Poriya, A., Bhagat, T., Patel, N., & Sharma, R. (2014). Non-Personalized Recommender Systems and User-based Collaborative Recommender Systems. *International Journal of Applied Information Systems*, *6*. www.ijais.org

Quazi, S. (2022). Artificial intelligence and machine learning in precision and genomic medicine. *Medical Oncology*, *39*(8), 1–18. https://doi.org/10.1007/S12032-022-01711-1

Richter, M., Piwocka, O., Musielak, M., Piotrowski, I., Suchorska, W. M., & Trzeciak, T. (2021). From Donor to the Lab: A Fascinating Journey of Primary Cell Lines. *Frontiers in Cell and Developmental Biology*, *9*, 1869. *https://doi.org/10.3389/FCELL.2021.711381/BIBTEX*

Saltz, J. S. (2021). CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps. *2021 IEEE International Conference on Big Data*. https://doi.org/10.1109/BIGDATA52589.2021.9671634

Sharma, A., & Rani, R. (2018). KSRMF: Kernelized similarity based regularized matrix factorization framework for predicting anti-cancer drug responses. *Journal of Intelligent and Fuzzy Systems*, *35*(2), 1779–1790. https://doi.org/10.3233/JIFS-169713

Shinde, P. P., & Shah, S. (2018). A Review of Machine Learning and Deep Learning Applications. *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, 1–6. https://doi.org/10.1109 / ICCUBEA.2018.8697857

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

Shrestha, P. (2021). *Application of Machine Learning and Deep Learning Techniques for Nepal stock market price prediction* [London Metropolitan University]. https://www.researchgate.net/publication/353634645_Application_of_Machine_Learning_and_Deep_Learning_Techniques_for_Nepal_stock_market_price_prediction

Silva, A. R. da, Fernandes, A. S., Fiorin, V. P., Zeppenfeld, T., França, G. B., Tatsch, J. S. ., Sagrillo, M. R., & Simão, E. M. (2020). Análise diferencial de genes em linhagens de células de leucemia. *Scientia Plena*. https://www.scientiaplena.org.br/sp/article/view/5003/2284

Soares, B. F. (2020). Medicina Personalizada. *Revista de Ciência Elementar*, *8*(4). https://doi.org/10.24927/RCE2020.053

Suphavilai, C., Bertrand, D., & Nagarajan, N. (2018). Predicting Cancer Drug Response using a Recommender System. *Bioinformatics (Oxford, England)*, *34*(22), 3907–3914. https://doi.org/10.1093/BIOINFORMATICS/BTY452

Varella-Garcia, M. (2004). *Análise Genômica: do laboratório à prática oncológica Genome Analysis: from laboratory to oncology practice*. *11*(1), 40–43.

Vasconcelos, J. B. de, & Barão, A. (2017). *Ciência dos Dados nas Organizações: Aplicações em Python* (1st ed.).

Verma, M. (2012). Personalized Medicine and Cancer. *Journal of Personalized Medicine*, *2*(1), 1. https://doi.org/10.3390/JPM2010001

Vogenberg, F. R., Barash, C. I., & Pursel, M. (2010). Personalized Medicine: Part 1: Evolution and Development into Theranostics. *Pharmacy and Therapeutics*, *35*(10), 560. /pmc/articles/PMC2957753/

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., & Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics* , *45*(10), 1113–1120. https://doi.org/10.1038/ng.2764

Wheeler, H. E., Maitland, M. L., Dolan, M. E., Cox, N. J., & Ratain, M. J. (2012). Cancer

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

pharmacogenomics: strategies and challenges. *Nature Reviews Genetics 2012 14:1*, *14*(1), 23–34. https://doi.org/10.1038/nrg3352

World Health Organization. (2020). Genomics. https://www.who.int/news-room/questions-and-answers/item/genomics

Zhang, N., Wang, H., Fang, Y., Wang, J., Zheng, X., & Shirley Liu, X. (2015). Predicting Anticancer Drug Responses Using a Dual-Layer Integrated Cell Line-Drug Network Model. *PLoS Comput Biol*, *11*(9), 1004498. https://doi.org/10.1371/journal.pcbi.1004498

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