UNIVERSIDADE FEDERAL DE PELOTAS

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA

Dissertação

Bambu: Desenvolvimento de uma ferramenta para QSAR baseada em aprendizado de máquina

Isadora Leitzke Guidotti

Pelotas, 2021

Isadora Leitzke Guidotti

Bambu: Desenvolvimento de uma ferramenta para QSAR baseada em aprendizado de máquina

Dissertação apresentada ao Programa de Pós-Graduação em Biotecnologia da Universidade Federal de Pelotas, como requisito parcial à obtenção do título de Mestre em Biotecnologia.

Orientador: Frederico Schmitt Kremer
Coorientador(a): Fabiana Kommling Seixas

Pelotas, 2021

Universidade Federal de Pelotas / Sistema de Bibliotecas
Catalogação na Publicação

G948b Guidotti, Isadora Leitzke

Bambu : desenvolvimento de uma ferramenta para QSAR baseada em aprendizado de máquina / Isadora Leitzke Guidotti ; Frederico Schmitt Kremer, orientador ; Fabiana Kommling Seixas, coorientador. — Pelotas, 2021.

66 f. : il.

Dissertação (Mestrado) — Programa de Pós-Graduação em Biotecnologia, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, 2021.

1. Bioinformática. 2. Quimioinformática. 3. Terapia alvo. 4. HTS. I. Kremer, Frederico Schmitt, orient. II. Seixas, Fabiana Kommling, coorient. III. Título.

CDD : 620.8

Elaborada por Ubirajara Buddin Cruz CRB: 10/901

BANCA EXAMINADORA

Prof. Drª. Lucielli Savegnago (Universidade Federal de Pelotas, Centro de Desenvolvimento Tecnológico

Prof. Drª. Karina dos Santos Machado (Universidade Federal de Rio Grande, Centro de Ciências Computacionais)

Suplente

Prof. Drª. Fabiana Kommling Seixas (Universidade Federal de Pelotas, Centro de Desenvolvimento Tecnológico)

Dedico para a minha família, amigos e colegas de profissão.

Agradecimentos

Em 2020 quando entrei no mestrado achei que seria meu ano, fui a primeira semana para o laboratório e depois veio o anúncio que estávamos em pandemia e que a universidade entraria em quarentena, seguimos nessa até o momento que estou escrevendo isso, em outubro de 2021. Acho que antes de agradecer a todos que me ajudaram nesse projeto, deveria agradecer ao universo por ter mantido todos vivos e salvos dessa doença, hoje vacinados. Sinto por todas as quase 600 mil pessoas que perderam a sua vida durante essa pandemia, a todas que sofrem pela perda de um ente querido ou por problemas financeiros ou outros, pois os problemas que a pandemia trouxe foram vários. Então meu primeiro obrigada vai para o universo por ter me dado o privilégio de todos que eu amo estarem comigo ainda hoje.

Meu segundo agradecimento vai com um imenso carinho para a Professora Fabiana Seixas que em 2020 me recebeu no seu laboratório e aceitou me orientar e foi a pessoa que me colocou no caminho que eu estou hoje e me fez novamente encontrar o Frederico. Obrigada Fabi, por ter acreditado no meu potencial e me dar um novo rumo.

Agradeço com todo o carinho, ao meu orientador, amigo e sócio Frederico por ter estado do meu lado, me passado tanto conhecimento acadêmico e de vida, se um dia eu chegar a ser 10% da pessoa, professor e orientador que tu é eu fico muito feliz. Obrigada por ter acreditado que eu era capaz de realizar tudo que conquistamos nesse último ano e em tão pouco tempo. É um prazer enorme ter sido sua primeira orientada.

Por último e não menos importante agradeço aos meus pais, irmãos e amigos que estiveram comigo durante esse período, que me deram suporte e a oportunidade de ter chegado aonde cheguei. E a Professora Karina que faz parte da minha banca e durante esse processo me acompanhou e contribuiu com esse projeto.

Obrigada!

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

“Science may set limits to knowledge, but should not set limits to imagination”

Bertrand Russell

Resumo

GUIDOTTI, Isadora L. **Bambu: Desenvolvimento de uma ferramenta para QSAR baseada em aprendizado de máquina.** 2021. 66f. Dissertação (Mestrado) - Programa de Pós-Graduação em Biotecnologia. Universidade Federal de Pelotas, Pelotas.

BioAssays Model Builder - BAMBU é uma ferramenta desenvolvida para auxiliar pesquisadores em pesquisas relacionadas ao desenvolvimento de novos fármacos. A identificação de moléculas com potencial farmacológico é tradicionalmente realizada por meio de descobertas de compostos naturais, para isso se usa a abordagem de drug discovery, dentro dela o *High Throughput Screening* (HTS). Outra abordagem usada para o desenvolvimento dessas ferramentas é o aprendizado de máquina (AM), uma subárea da inteligência artificial, que visa desenvolver algoritmos capazes de solucionar problemas para os quais não foram explicitamente programados. Pode ser dividida em duas áreas fundamentais: a aprendizagem supervisionada que visa aprender com os dados fornecidos e a aprendizagem não supervisionada que aprender com base nos padrões de amostras. Uma das metodologias empregadas no ML supervisionado no contexto de drug discovery é a *Quantitative Structure-Activity Relationship* (QSAR), um estudo quantitativo para ver interações entre moléculas orgânicas e estruturas químicas de forma tridimensional que visa estudar o ligante. A ferramenta bambu une os bancos de dados oriundos de estudos de HTS junto da metodologia QSAR que emprega o uso do AM supervisionado e usa algoritmos de AM para distinguir moléculas com potencial farmacológico daquelas que não possuem e pode ser usada para auxiliar na pesquisa de novas abordagens terapêuticas para várias doenças incluindo câncer e doenças neurodegenerativas. Para compor a ferramenta usamos modelos baseados em árvores de decisão, redes neurais e regressão linear. Como estratégia de balanceamento são usadas abordagens de undersampling, oversampling, tomek links e SMOTE. Para avaliar o funcionamento da ferramenta é usado métricas de classificação como precisão, recall, f1-score e acurácia. O uso dos modelos permite que a ferramenta consiga separar moléculas ativas de inativas e isso se comprova e atinge o objetivo do trabalho quando observamos os dados da precisão que é uma das métricas usadas para validar que nos diz se o modelo está conseguindo separar moléculas ativas de moléculas inativas. Também observamos que os modelos baseados em árvores de decisão e ensembles de árvores de decisão são os que obtiveram resultados mais satisfatórios.

Palavras-chave: Bioinformática, Quimioinformática, Terapia alvo, HTS.

Abstract

GUIDOTTI, Isadora L. **Bambu: Development of a machine learning-based QSAR.** 2021. 66f. Dissertação (mestrado) - Programa de Pós-Graduação em Biotecnologia. Universidade Federal de Pelotas, Pelotas.

BioAssays Model Builder - BAMBU is a tool developed to assist researchers in research related to the development of new drugs. The identification of molecules with pharmacological potential is traditionally performed through discoveries of natural compounds, for which the drug discovery approach is used, within it the High Throughput Screening (HTS). Another approach used to develop these tools is machine learning (ML), a sub-area of artificial intelligence, which aims to develop algorithms capable of solving problems for which they were not explicitly programmed. It can be divided into two fundamental areas, supervised learning that aims to learn from the data provided and unsupervised learning that learns from sample patterns. One of the methodologies used in supervised ML in the context of drug discovery is the Quantitative Structure-Activity Relationship (QSAR), a quantitative study to see interactions between organic molecules and chemical structures in a three-dimensional way that aims to study the ligand. The bamboo tool joins the databases from HTS studies with the QSAR methodology that employs the use of supervised ML and uses ML algorithms to distinguish molecules with pharmacological potential from those that do not and can be used to assist in the search for new approaches therapeutics for various diseases including cancer and neurodegenerative diseases. To compose the tool we used models based on decision trees, neural networks and linear regression. As a balancing strategy, undersampling, oversampling, tomek links and SMOTE approaches are used. To assess the functioning of the tool, ranking metrics are used such as precision, recall, f1-score and accuracy. The use of models allows the tool to be able to separate active from inactive molecules and this is proven and achieves the objective of the work when we observe the precision data, which is one of the metrics used to validate that tells us if the model is managing to separate active molecules from inactive molecules. We also observed that models based on decision trees and decision tree ensembles are the ones with the most satisfactory results.

Keywords: Bioinformatics, Chemoinformatics, Target Therapy, HTS.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

ADMET – Absorção, distribuição, metabolismo, excreção e toxicidade

AM – Aprendizado de Máquina

BAMBU - *BioAssays Model Builder* (Construtor de modelo BioAssays)

BCG – *Bacille Calmette-Guérin*

CADD – *Computer Aided Drug Discovery* (Descoberta de drogas auxiliada por computador)

CAR – Chimeric antigen receptors (Receptores de antígenos quiméricos)

CART – *Classification and regression trees* (Árvores de classificação e regressão)

ER – Receptor de estrogênio

HTS - *High throughput screen* (Triagem de alto rendimento)

INCA – Instituto nacional do câncer

ML – *Machine learning* (Aprendizado de máquina)

PR – Receptor de progesterona

QSAR – Quantitative structure-activity relationship (Relação quantitativa estrutura-atividade)

RBF – *Radial basis function* (Função de base radial)

SMOTE - *Synthetic minority oversampling technique* (Técnica de sobreamostragem de minoria sintética)

SVM - *Support vector machine* (Máquina de vetores de suporte)

TCR – T cells receptors (Células T receptoras)

TR-FET – *Time-resolved fluorescence energy transfer* (Transferência de energia de fluorescência resolvida no tempo)

TNM – Tumor; nódulo e metástase

UV – Ultravioleta

Sumário

1. INTRODUÇÃO GERAL

BioAssays Model Builder - BAMBU é uma ferramenta desenvolvida para auxiliar na pesquisa para o desenvolvimento de novos fármacos. A identificação de moléculas com potencial farmacológico é tradicionalmente realizada por meio de descobertas de compostos naturais. Os medicamentos descobertos, hoje auxiliam no tratamento de inúmeras doenças, inclusive câncer, mas isso sempre foi laborioso e custoso para os órgãos que financiam pesquisas para *drug discovery*. As grandes empresas farmacêuticas e centros de pesquisa costumam usar uma abordagem denominada *High Throughput Screening* - HTS (Triagem de alto rendimento), onde são testados milhares de moléculas para atividades biológicas de forma automatizada. A maior problemática envolvendo essa tecnologia é o custo de infraestrutura e das bibliotecas (SHINN et al., 2019).

Novas abordagens vêm sendo desenvolvidas, graças aos dados produzidos por HTS, programas que conseguem filtrar informações relevantes no meio dos dados disponibilizados nesses bancos estão sendo desenvolvidos, permitindo o estudo de compostos para o desenvolvimento de novas terapias capazes de analisar estruturas tridimensionais de moléculas alvo e de seus ligantes (PYZER-KNAPP et al., 2015). O uso dessas ferramentas de bioinformática auxiliam diminuem os custos dos estudos de câncer e se concentram na análise de componentes celulares, atividades enzimáticas, na resistência de novos compostos terapêuticos, biologia molecular e descoberta e desenvolvimento de novas drogas, além de vacinas para câncer (YANG, 2008).

Dentro da bioinformática é usado para o desenvolvimento de novas ferramentas o aprendizado de máquina - AM, que é uma sub-área da inteligência artificial e visa desenvolver algoritmos capazes de solucionar problemas para os quais não foram explicitamente programados. Pode ser dividida em duas áreas fundamentais, a aprendizagem supervisionada e a aprendizagem não supervisionada (BI et al., 2019). Uma das metodologias empregadas no AM supervisionado no contexto de *drug discovery* é a *Quantitative Structure-Activity Relationship* - QSAR (Relação quantitativa estrutura-atividade), um estudo

quantitativo para ver a relação entre a estrutura de moléculas orgânicas e sua atividade biológica, o que ajuda na descoberta de possíveis ligantes. Pode ser usada para *virtual screening*, e ser usada na predição da atividade química, inibição enzimática, inibição de crescimento celular entre outros (VERMA; KHEDKAR; COUTINHO, 2010).

Bambu une ambas as abordagens já que utiliza os bancos de dados oriundos dos dados de HTS e também o QSAR que emprega o AM. A ferramenta pode ser usada para auxiliar e pode ser vista como um filtro para selecionar moléculas com potencial farmacológico para diversas doenças, incluindo o câncer.

O Câncer é uma das principais causas de morte no mundo, sendo um grupo de doenças caracterizada pela divisão desorientada de células que sofreram mutações genéticas, e possuem a capacidade de invadir diferentes grupos celulares. Estas mutações são ocasionadas por alguns fatores de risco, que incluem tabagismo, alcoolismo, má alimentação e estresses repetitivos (LEWANDOWSKA et al, 2019).

O tratamento de câncer pode ser realizado com quimioterapia e radioterapia, muitas vezes ambos os tratamentos são recomendados para os pacientes com câncer, mas essas opções costumam ser danosas para a pessoa que o recebe. Por isso, as pesquisas para o desenvolvimento de novas terapias para o tratamento de câncer têm avançado de forma expressiva (HERRERO, 2015). Devido aos incessantes estudos na área, diferentes metodologias estão sendo desenvolvidas tornando as abordagens terapêuticas, mais eficazes e menos agressivas. Por exemplo, graças a biotecnologia, pacientes diagnosticados com câncer têm acesso hoje a diferentes tratamentos, como a imunoterapia (RILLEY, MITCHELL, 2019), que é uma abordagem que ajuda o sistema imunológico a combater as células cancerosas e também a terapia alvo (TSIMBERIDOU, 2015) que agem direcionando medicamentos quase que exclusivamente as células tumorais o que ajuda a manter o tecido saudável intacto.

2. REVISÃO BIBLIOGRÁFICA

2.1 Câncer

2.1.1 Epidemiologia do Câncer

O câncer é um grupo de doenças responsável por levar a óbito milhares de pessoas todos os anos, sendo a segunda maior causa de morte no mundo perdendo apenas para as doenças cardiovasculares (NAGAI; KIM, 2017). No ano de 2020 foi estimado mais de 19 milhões de novos casos de câncer para o ano em todo mundo, e estima-se que até 2040 esse número aumente em 47%, ultrapassando os 28 milhões de casos. Em relação às mortes no ano de 2020 ocorrem cerca de 10 milhões de mortes por câncer, sendo o câncer de mama a mais letal para as mulheres e o câncer de pulmão para ambos os sexos (SUNG et al., 2020). Os cânceres mais incidentes no ano de 2020 foram os cânceres de mama, pulmão, colorretal, próstata e estômago (Figura 1).

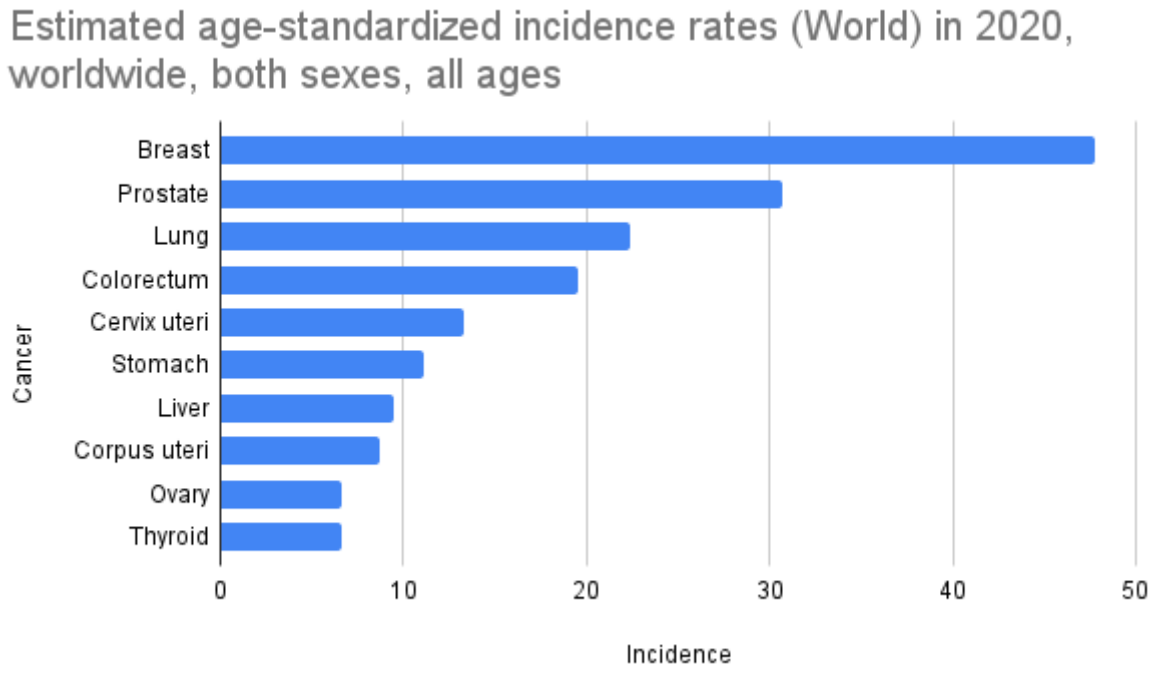


Figura 1. Cânceres mais incidentes no ano de 2020. ADAPTADO DE GLOBOCAN, 2020.

No Brasil, o Instituto Nacional do Câncer - INCA, no seu levantamento mostrou que o câncer de próstata é o mais incidente entre os homens ultrapassando os 65 mil casos no ano de 2020, em segundo lugar sendo o câncer de colorretal e de pulmão e para as mulheres o câncer de mama é o mais incidente ultrapassando o também mais de 65 mil casos diagnosticados no ano de 2020, para as mulheres em segundo lugar também vem o câncer de colorretal e em seguida o câncer de colo de útero. Em relação à mortalidade (Figura 2), o câncer de mama continua sendo o mais letal para as mulheres, enquanto o câncer de pulmão é o mais letal para os homens levando a óbito mais de 16 mil homens a óbito no Brasil em 2019 (“Estatísticas de câncer”, 2021).

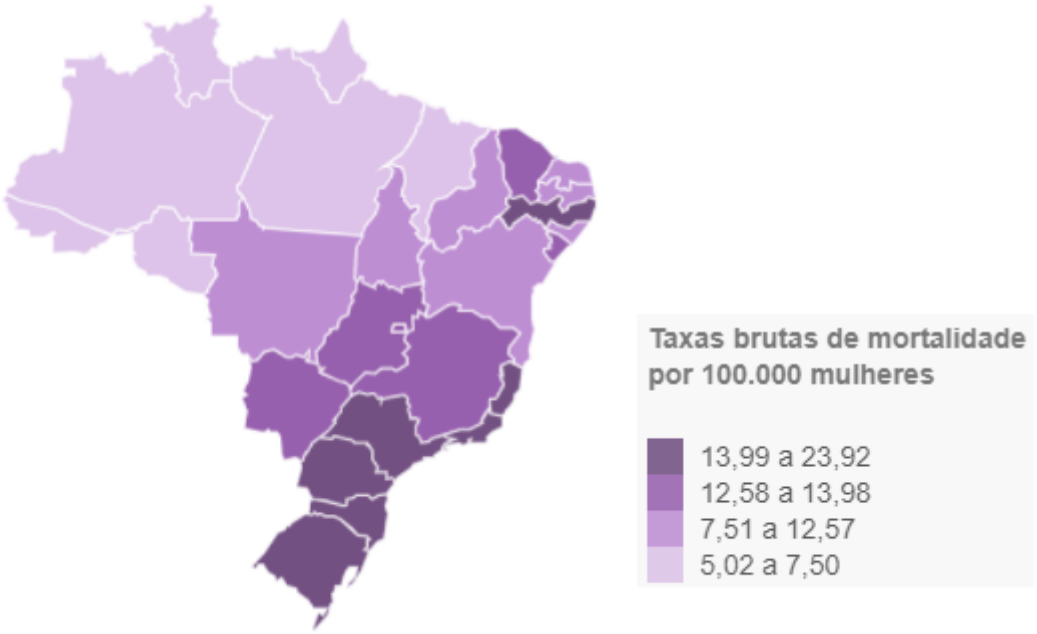


Figura 2. Estatísticas da mortalidade por câncer de mama no Brasil. Representação espacial das taxas brutas de mortalidade por câncer de mama, a cada 100.000 mulheres, pelas unidades da federação do Brasil, entre 2009 e 2019. Atlas Online Inca (https://www.inca.gov.br/app/mortalidade).

2.1.2 Câncer de mama

O câncer de mama é o mais comum entre as mulheres e o mais letal, geralmente atinge mulheres acima dos 50 anos de idade, mas pode acometer

mulheres mais jovens. Os fatores de risco mais atribuídos a essa patologia estão fatores genéticos, dieta, cigarro, álcool, uso de alguns tipos de anticoncepcionais à base de estrogênio e progesterona (ROJAS; STUCKEY, 2016). Em países da Europa e da América do Norte, onde ocorrem campanhas de incentivo à mamografia e a mesma é acessível houve redução no número de mortes por câncer de mama, além disso a triagem de rotina também é aderida pelas pacientes (AHMAD, 2019). Existem pelo menos cinco subtipos de câncer de mama, que possuem diferentes assinaturas genéticas e diferentes abordagens terapêuticas. Podem ser classificados de acordo com o tamanho, estado nodal e se há metástases, o sistema usado para realizar a classificação é o TNM (ESCRIG SOS; GÓMEZ QUILES; MAIOCCHI, 2019), onde, o T representa o tamanho do tumor, N se há nódulos e o M se há presença de metástase. Esse sistema é comumente usado em quase todos os tipos de câncer para classificar o tamanho e auxiliar no prognóstico. Molecularmente, o câncer de mama pode ser dividido em luminal A; luminal B; HER2-superexpresso; basal-*like* e normal-*like* (YEO; GUAN, 2017). Luminal A é o subtipo mais incidente, apresentando um grau baixo (chances menores de disseminação e crescimento lento) e prognóstico favorável normalmente são receptor de estrogênio (ER) positivo e/ou receptor de progesterona (PR) positivo, enquanto os receptores de fator de crescimento epidérmico (HER2) é negativo (TSANG; TSE, 2020). O luminal B, representa cerca de 24% dos casos incidentes e tende a ter um grau mais alto (chances maiores de disseminação e crescimento acelerado) e apresentar um prognóstico pior do que o luminal A, normalmente é caracterizado por ER/PR positiva e HER2 também (ADES et al., 2014). HER2-superexpresso, é incidente em aproximadamente 15% dos casos, é caracterizado pela superexpressão de HER2, sendo tumores de alto grau e que normalmente apresentam ER/PR negativo e tendem ser bem agressivos, porém respondem bem a terapias que usam o anti-HER2 (VODUC et al., 2010). O basal-*like* diferente dos outros não apresenta expressão em ER/PR e HER2, sendo conhecido como triplo negativo ou *claudin low*, é responsável por cerca de 19% dos casos e tende a ser o de mais alto grau, visto que afeta as células mamárias basais e células mioepiteliais normais, o que torna esse tipo de câncer de mama o mais agressivo e as pacientes com esse subtipo tendem a ter recidiva em até 5 anos

(BADVE et al., 2011). Por fim, o normal-like representa cerca de 2% dos casos, sendo ele o mais raro e apresenta poucas células malignas, podendo ser considerado como uma contaminação (RUSSNES et al., 2017).

De acordo com os subtipos, no luminal A são comuns mutações no gene PIK3CA, e algumas vezes podem ocorrer mutações em genes como MAP3K1, MAP2K4 que se apresentam em cerca de 20% dos casos incidentes. No luminal B, é frequente encontrar mutações nos genes TP53 e PIK3CA, esses dois genes em específicos são comuns de aparecer em subtipos como o basal-*like* e HER2-superexpresso, no caso do basal-*like* o gene TP53 é mais comum do que nos outros (KOBOLDT et al., 2012). Normalmente essas mutações ocorrem de forma esporádica, podendo ser causadas por fatores de risco como tabagismo e alcoolismo. Quando falamos no câncer de mama que aparece de forma hereditária, que são mutações progressivas e cumulativas, os genes mais envolvidos são o BRCA1 e BRCA2. BRCA1 é o mais frequente e o que apresenta o pior prognóstico, geralmente o paciente com essa mutação desenvolve o subtipo triplo negativo, enquanto o BRCA2 têm uma porcentagem menor de casos diagnosticados, ele continua também sendo mais agressivo do que os cânceres de mama que aparecem de forma esporádica (BARETTA et al., 2016).

O tratamento para o câncer de mama é como todo o tratamento oncológico. Depende do grau de estadiamento, definido pelo sistema TNM, o subtipo do tumor e de outras características do paciente. O tratamento pode ser realizado de forma local que é com cirurgia, radioterapia, ou pode ser usado o tratamento sistêmico com quimioterapia, terapia alvo, hormonioterapia. Ainda pode optar-se por fazer combinações entre essas terapias (AHMAD, 2019).

As terapias podem ser adjuvantes, que visam o tratamento após o tratamento cirúrgico, neoadjuvantes aquelas que antecedem o tratamento cirúrgico e paliativos que visam aumentar a sobrevida e melhorar a qualidade de vida do paciente (FISUSI; AKALA, 2019).

2.1.3 Melanoma

O câncer de pele pode ser dividido em câncer de pele não melanoma e câncer de pele melanoma, o que difere ambos os tipos é que o melanoma é maligno sendo o tipo mais agressivo de câncer de pele e o que mais leva a óbito. O melanoma se desenvolve nos melanócitos, células responsáveis pela produção de melanina (AMERICAN CANCER SOCIETY, 2019).

É estimado pela Globocan (Figura 3) que a Europa é o continente que mais teve incidência de casos de melanoma em 2020, ultrapassando os mais de 140 mil casos. Na América, em particular na América latina o continente teve menos de 20 mil casos, sendo que a maior parte dos casos de melanoma foram na América do norte, onde ultrapassou os 100 mil. O Brasil, segundo estimativas do INCA estimou que para 2020 haveria 8450 casos, destes 4200 em homens e 4250 em mulheres. Além disso, no Brasil, segundo o INCA houveram 1978 óbitos por causa do melanoma em 2019, sendo a maioria homens com 1259 mortes e 819 mulheres.

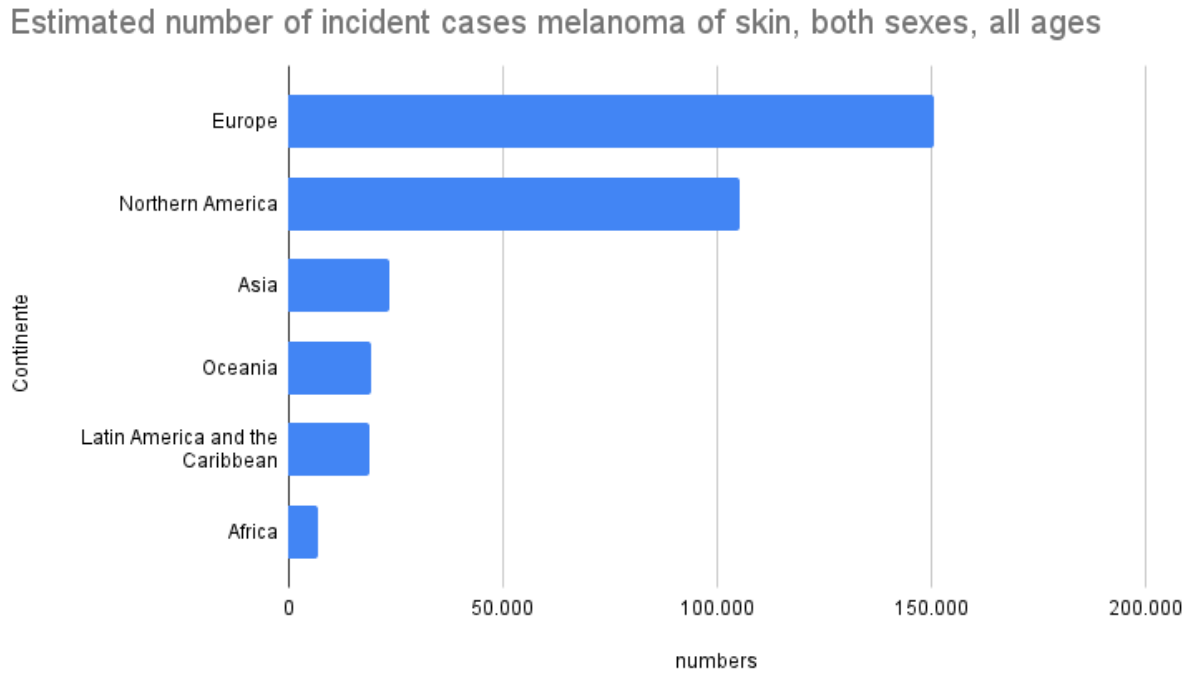


Figura 3. Número de casos de melanoma. Casos incidentes em cada continente no ano de 2020. ADAPTADO DE GLOBOCAN, 2021.

Normalmente o melanoma ocorre na pele mas pode acontecer também nos olhos, boca, região genital e anal. Além disso, as manchas de melanoma costumam ter uma cor marrom ou preta, isso porque as células de melanócito cancerosas seguem produzindo melanina (MISIR et al., 2016).

O sol é o principal fator de risco para o desenvolvimento do melanoma, em países nórdicos e na Austrália, o sol foi responsável por mais de 90% dos casos de melanoma nesses países. A exposição ao sol pode ser de curta duração, intermitente, onde a pessoa tem uma exposição ao sol contínua, como por exemplo ser agricultor e trabalhar horas exposto ao sol, sendo que a exposição curta, isso é, banhos de sol e recreações ao ar livre, são as que mais estão associadas ao desenvolvimento de melanoma (BERWICK et al., 2016). Os raios ultravioleta - UV são considerados carcinogênicos para os humanos, existem três tipos de raios UV que podem ser classificados em: raios UV-A representam 95% dos raios que adentram a camada de ozônio, os raios UV-B representam 5% dos raios solares que estão na terra, e os raios UV-C são aqueles que não ultrapassam a camada de ozônio. A exposição solar causa “impressões digitais”, isso é, dependendo do raio UV, genes específicos são atingidos. (EL GHISSASSI et al., 2009).

Outro fator de risco é a própria genética, pessoas com alterações nos genes CDKN2A representam cerca de 20%-40% dos melanomas familiares. Esse gene está presente no cromossomo 9p21 e possui 4 exons que codificam duas proteínas, a proteína p16INK4A é produzida por meio dos exons 1α, 2 e 3 enquanto a p14ARF é produzida pelos exons 1 , 2 e 3 (READ; WADT; HAYWARD, 2016). Quando a atividade da proteína p16INK4A é inibida por meio da CDK4 e CDK6 ela é mantida em um quadro hipofosforilado, impedindo a entrada na fase S do ciclo celular. A proteína p14ARF é um regulador de p53, então mutações que afetam essa proteína, permite que as células escapem da barreira de senescência. (GOLDSTEIN et al., 2006). Além dessa mutação, genes responsável por outras síndromes também podem colaborar para o desenvolvimento de melanoma, como BRCA2 responsável pelo câncer de mama, OCA2 gene do albinismo, MC1R gene associado ao cabelo ruivo e fenótipo sardento (O’NEILL; SCOGGINS, 2019). Um estudo feito pela *Cancer Genome Atlas Network* classificou os melanomas cutâneos em 4 subtipos

genômicos: mutante BRAF; mutante NRAS; mutante NF1 e tipo selvagem triplo que é caracterizado pela ausência de BRAF, K-RAS e NF1 (AKBANI et al., 2015).

Para a prevenção a Sociedade Brasileira de Dermatologia, criou um guia de recomendações para evitar o câncer de pele, algumas dessas recomendações é o uso de protetor solar, evitar sair no sol entre as 10h - 16h e fazer bronzeamento artificial, recomenda o uso de barreiras físicas como bonés, chapéus, guarda-chuvas, roupa comprida (SCHALKA; STEINER, 2012).

O melanoma pode ser encontrado em alguns subtipos como: melanoma lentigo maligno; melanoma lentiginoso acral; melanoma extensivo superficial e melanoma nodular (CLARK; ELDER; VAN HORN, 1986). O melanoma lentigo maligno normalmente está associado à exposição solar crônica, apresenta um padrão de crescimento radial e representa cerca de 79% - 83% dos casos de melanoma *in situ* e normalmente aparecem em pessoas idosas e principalmente homens (Figura 4) (DEWANE et al., 2019). O melanoma lentiginoso acral é raro em pessoas brancas, atingindo cerca de 1% - 2% de todos os melanomas, não apresenta relação com a exposição solar e geralmente é encontrado nas palmas das mãos e dos pés (Figura 5) (BAÑULS, 2018). O melanoma extensivo superficial é o mais comum e pode apresentar diferentes colorações (Figura 6) (LONGO; PELLACANI, 2016). Apresenta características predominantes como dispersão e proliferação pagetoide ao longo da junção dermoepidérmica, além de possuir grandes melanócitos epitelióides isolados ou em ninhos na derme (FORMAN et al., 2008). O melanoma nodular é o mais detectado como uma lesão (Figura 7), corresponde a 9% a 15% dos casos, possui um crescimento rápido e geralmente costuma ser diagnosticado quando já está em um estágio mais avançado (MENZIES et al., 2013).

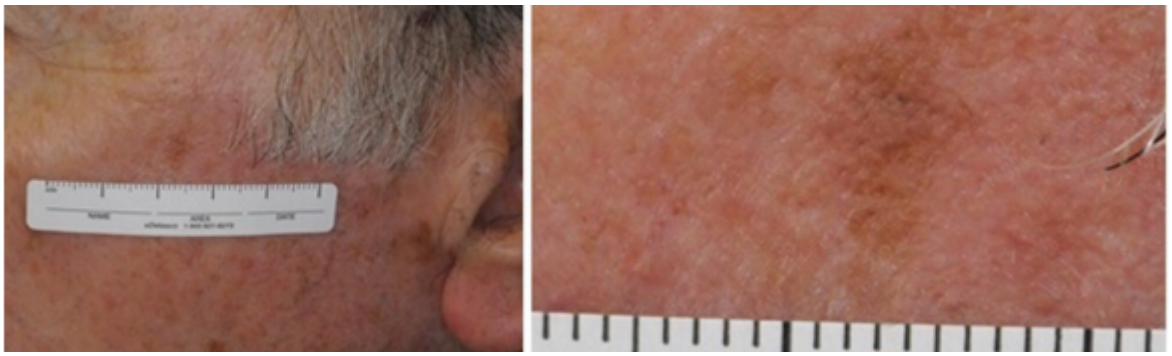


Figura 4. Melanoma lentigo maligno. ADAPTADO DE: DEWANE et al., 2019.

Figura 5. Melanoma lentiginoso acral. Melanoma lentiginoso acral na sola do pé. ADAPTADO DE: (BRAZEN et al., 2020).

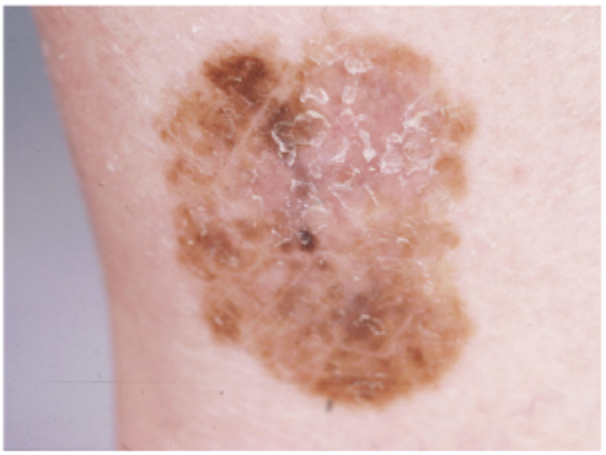


Figura 6. Melanoma extensivo superficial. ADAPTADO DE: (BRISTOW; BOWLING, 2009).

Figura 7. Melanoma nodular. ADAPTADO DE: (WEILER; THOMAS; DALLE, 2013).

O diagnóstico de melanoma consiste na anamnese, exame clínico e histopatológico. Um dos exames realizados é a dermatoscopia, um método não invasivo que examina as lesões pigmentadas e não pigmentadas. Com os últimos anos a dermatoscopia vem melhorando cada vez mais devido a implementação do *machine learning* que treina novos algoritmos que usam a regra do ABCDE para diagnosticar melanoma (WEBER et al., 2018). A regra do ABCDE é baseada em critérios que devem ser preenchidos para o diagnóstico de melanoma. O A representa assimetria, B a irregularidade das bordas, C a cor, D o diâmetro e E a evolução, se aquela mancha muda de cor ou formato, para ser benigno a pontuação deve variar entre 1.0 e 4.75, suspeita varia entre 4.75 e 5.45 e acima de 5.45 é altamente suspeito (SAIYED; HAMILTON; AUSTIN, 2017). Outro meio usado dessa vez para avaliar o prognóstico é a escala de Breslow e os níveis de Clark, exames histopatológicos. Breslow mede a espessura do tumor em milímetros, isso quer dizer que ele mede da mancha até as células cancerosas mais profundas e os níveis de Clark mede a anatomia, isto é avalia se a mancha está apenas na epiderme, ou se chega até a derme ou passa por todas as camadas anatômicas de pele (ELDER, 2011).

O tratamento convencional para melanoma pode ser cirúrgico, onde é retirada a mancha cancerosa e é feita a busca pelo linfonodo sentinela para evitar que ocorra metástases (FALK DELGADO; ZOMMORODI; FALK DELGADO, 2019). A quimioterapia pode ser administrada via oral, intralesional ou de forma sistêmica, o quimioterápico padrão ouro é o dacarbazina, mas outros como temolozomide que é

administrado de forma oral, cisplatina e vimblastina são bastante usados e podem ser combinados (YANG; CHAPMAN, 2009). A radioterapia é indicada para evitar o crescimento de novas células tumorais, normalmente as células tumorais de melanoma são resistentes à radioterapia (TAKAHASHI; NAGASAWA, 2020).

Existem outros tratamentos biotecnológicos como a imunoterapia e a terapia alvo. A imunoterapia faz uso de substâncias que auxiliam o sistema imunológico a combater as células tumorais. A imunoterapia intralesional usa BCG - *Bacilo Calmette-Guérin* é usado para tratar lesões avançadas, mas o uso dessa abordagem ocasiona em alguns efeitos colaterais como a sepse por BCG, febre e problemas cutâneos (ONITILO; WITTIG, 2020). A terapia alvo tem o objetivo de inibir apenas as células tumorais preservando o tecido saudável, para o melanoma a principal terapia alvo é relacionado ao gene BRAF, por ser uma mutação mais frequente nesse tipo de câncer (NAMIKAWA; YAMAZAKI, 2019). O primeiro da classe foi o vemurafenib, que é um inibidor da mutação V600 do BRAF. Outros inibidores de outros genes como NRAS e MEK (LUKE et al., 2017).

2.1.4 Terapia Alvo e Imunoterapia

A imunoterapia e a terapia alvo são terapias biotecnológicas. A terapia alvo visa atacar apenas as células cancerosas, ou seja ela age em uma via específica, ou em algum gene específico da célula alvo (TSIMBERIDOU, 2015) enquanto a imunoterapia visa estimular o próprio sistema imunológico a combater as células cancerosas (RILEY et al., 2019).

A primeira imunoterapia desenvolvida foi com a citocina IL-2 (Aldesleukin) (NOBLE; GOA, 1997) secretada pelas células T CD4+, desenvolvida para o tratamento de melanoma metastático e câncer renal (ROSENBERG, 2014). Outras imunoterapias usando o antígeno 4 de linfócito T citotóxico - CTLA-4 (Ipilimumab) (SONDAK et al., 2011), morte programada-1 - PD-1 (Nivolumab) (SCOTT, 2015) ou seu ligante - PD-L1 (Atezolizumab) (INMAN et al., 2017), usados para o tratamento de vários cânceres, como melanoma, câncer de pulmão, renal, cabeça e pescoço (VAREKI; GARRIGÓS, 2017) as quais pertencem a classe de desenvolvimento de

checkpoints inhibitors é a abordagem mais usada para o desenvolvimento de imunoterapias (PARDOLL, 2012). Outra abordagem para o desenvolvimento é a engenharia de células T que usa receptores de antígenos quiméricos (CAR) e células T receptoras (TCR), essa abordagem utiliza as células T que são coletadas do paciente e em seguidas modificadas geneticamente para expressar proteínas CARs (LIM; JUNE, 2017), usando essa abordagem foi desenvolvido o Tisagenlecleucel, usado para pacientes com leucemia ou linfoma não-Hodgkin (FRIGAULT et al., 2019). Ainda existem também as vacinas anti-câncer como a BCG que está sendo estudada para o tratamento de câncer de bexiga (REDELMAN-SIDI; GLICKMAN; BOCHNER, 2014) e melanoma metastático (BENITEZ et al., 2019).

A terapia alvo pode utilizar tanto moléculas de pequeno peso molecular (*small molecules*) quanto biofármacos (ex: peptídeos, anticorpos monoclonais), devendo, em ambos os casos, a atividade do mesmo ser seletiva para o alvo de interesse (diferente da quimioterapia tradicional). O futuro dos inibidores de moléculas pequenas é promissor, pois é capaz de contribuir para a criação de fármacos precisos, que evitam o uso de modelos clínicos para prever quais os fármacos são candidatos, minimizando o uso de animais visto que isso hoje pode ser feito através de ferramentas de bioinformática e os testes em animais ficam apenas para estudos clínicos (BEDARD et al., 2020). Existem algumas categorias de inibidores de moléculas pequenas, como as multiquisanes, nessa categoria encontramos fármacos como sorafenib e sunitinib (PARK et al., 2012), que são inibidores de VEGFR1, VEGFR2, KIT e PDGFR-α, imatinib que inibe o BCR-ABL que é uma proteína de fusão presente na leucemia mielóide crônica (DRUKER et al., 2001). Outra categoria é os inibidores seletivos de pequenas moléculas que inibem oncogenes que são moléculas pequenas seletivas com a capacidade de facilitar o desenvolvimento de cânceres, medicamentos como o Vemurafenib são usados para o tratamento de melanoma e tem como alvo o oncogene BRAF Val600Glu (HEAKAL; KESTER; SAVAGE, 2011). A biotecnologia pode atuar tanto na identificação de possíveis moléculas candidatas quanto no planejamento de sistemas de entrega direcionados (*drug delivery*) para possibilitar a terapia alvo.

2.2 Bioinformática

2.2.1 Bioinformática e Quimioinformática

A bioinformática é uma área interdisciplinar que engloba a computação, biologia, matemática, química, engenharias e outras áreas. Hoje está em ascendência e muitos pesquisadores usam abordagens de bioinformática, conhecidas como abordagens *in sillico*. A bioinformática desenvolve softwares e programas capazes de predizer e analisar rotas metabólicas, estruturas tridimensionais, interpretar dados biológicos, na criação de novos algoritmos para o desenvolvimento de testes de diagnóstico entre outras centenas de análises (YANG, 2008) .

A quimioinformática estuda as moléculas químicas, usando abordagens computacionais e é indispensável para os pesquisadores que atuam na descoberta de novas drogas. Com o grande acúmulo de dados de moléculas com o potencial farmacológico, abordagens de ML são interessantes para ajudar a filtrar e entender melhor o comportamento das moléculas (LO et al., 2018). A junção entre a bioinformática e a quimioinformática pode render uma melhor contribuição para que entendamos como organismos vivos funcionam, desenvolver ferramentas e abordagens que auxiliam em todos os aspectos da vida, incluindo a cura de várias doenças (GASTEIGER, 2016).

2.2.2 Descoberta de Novas Drogas

A bioinformática e a quimioinformática são amplamente usadas para a descoberta de drogas. Ambas as áreas possuem várias metodologias que auxiliam os cientistas na descoberta de novos compostos com potencial farmacológico. Uma técnica amplamente utilizada pela indústria farmacêutica é o HTS que gera bancos com grandes volumes de dados pois analisa inúmeras moléculas que são filtradas até conseguir separar moléculas ativas (com potencial farmacológico) de moléculas inativas (sem potencial farmacológico) (SHINN et al., 2019). O HTS usa o “funil computacional” (figura 8), que filtra essas milhares de moléculas, cada nível do funil

possui cálculos matemáticos com limites de erro que definem quais estruturas devem ser descartadas e a cada nível que as moléculas alcançam é mais complexo computacionalmente, e apenas as moléculas que tenham todas as características necessárias chegam ao último nível do funil são as moléculas mais promissoras (PYZER-KNAPP et al., 2015).

Pelo alto número de moléculas analisadas foram criados bancos de dados de HTS como PubChem (WANG et al., 2012) que é um banco de dados dentro do NCBI (SAYERS et al., 2011) que arquiva moléculas pequenas e siRNA, amplamente utilizadas por pesquisadores que estudam novos fármacos. Além do PubChem existe o chEMBL (BÜHLMANN; REYMOND, 2020) que também é um banco de dados de estruturas químicas amplamente usado para a descoberta de fármacos.

Para analisar os dados desses bancos de dados existem abordagens *Computer Aided Drug Discovery* - CADD (WARR, 2017) que são estratégias computacionais que podem ser usadas para a descoberta de drogas, como mineração de dados, estudos baseados em ligantes como o QSAR ou baseados em receptores como o docking molecular e a dinâmica molecular, além disso é uma forma viável de começar a pesquisa com moléculas com potencial farmacológico.

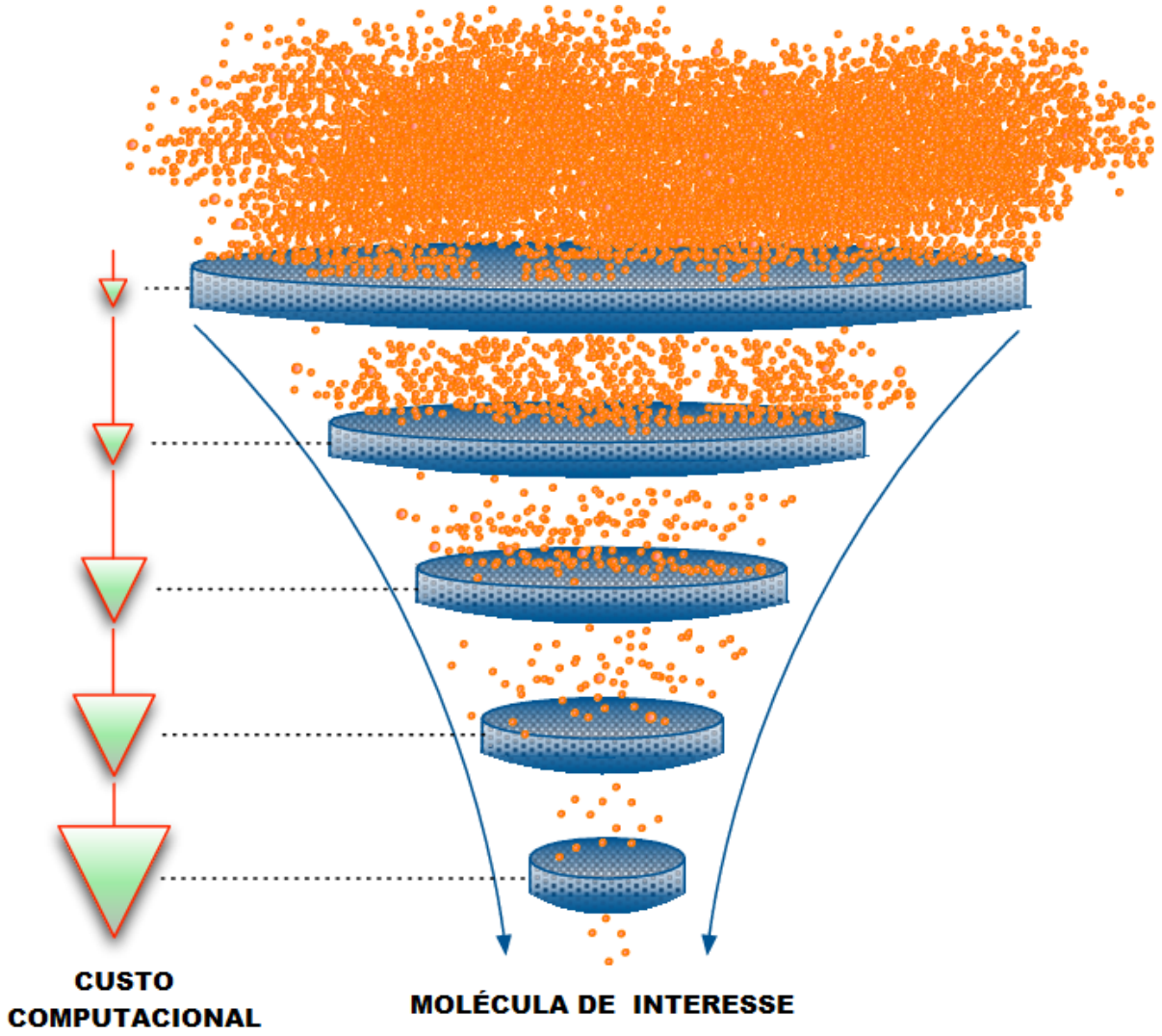


Figura 8. Esquema de High Throughput Screening - Funil computacional. No HTS milhares de moléculas são analisadas e vão passando por filtros até serem selecionadas aquelas que apresentam potencial farmacológico. ADAPTADO DE: PYZER-KNAPP et al., 2015.

2.2.3 Aprendizado de Máquina

Aprendizado de máquina é uma sub-área da inteligência artificial - IA que visa aprender através de dados fornecidos. Existem tipos de aprendizado como o aprendizado supervisionado, aprendizado não supervisionado, aprendizado semi-supervisionado e por reforço o objetivo do AM é criar algoritmos que sejam capazes de resolver problemas para os quais não foram explicitamente programados (BI et al., 2019).

O aprendizado supervisionado visa aprender a predição de “rótulos” (numéricos ou categóricos) com base em dados rotulados fornecidos (Figura 9a), esse tipo de aprendizado é usado para classificação, na medicina pode ser usado para ajudar os médicos a diagnosticarem doenças e estimarem o risco da doença, mas também é amplamente usado para negócios (DEO, 2015). Dentro do aprendizado supervisionado existe o QSAR que é uma metodologia usada para o descobrimento de drogas e é um método baseado no ligante, ou seja o estudo é baseado entre a estrutura do ligante e a sua atividade (AMBURE et al., 2020).

O aprendizado não supervisionado reconhece agrupamentos ou representações nas amostras (Figura 9b), ou seja, diferente do aprendizado supervisionado que recebe rótulos, o aprendizado não supervisionado analisa similaridade nos dados (PATEL et al., 2020). Aprendizado semi-supervisionado é menos usado e visa aprender com dados rotulados e não rotulados (VANDIST; STORMS; VAN DEN BUSSCHE, 2019).

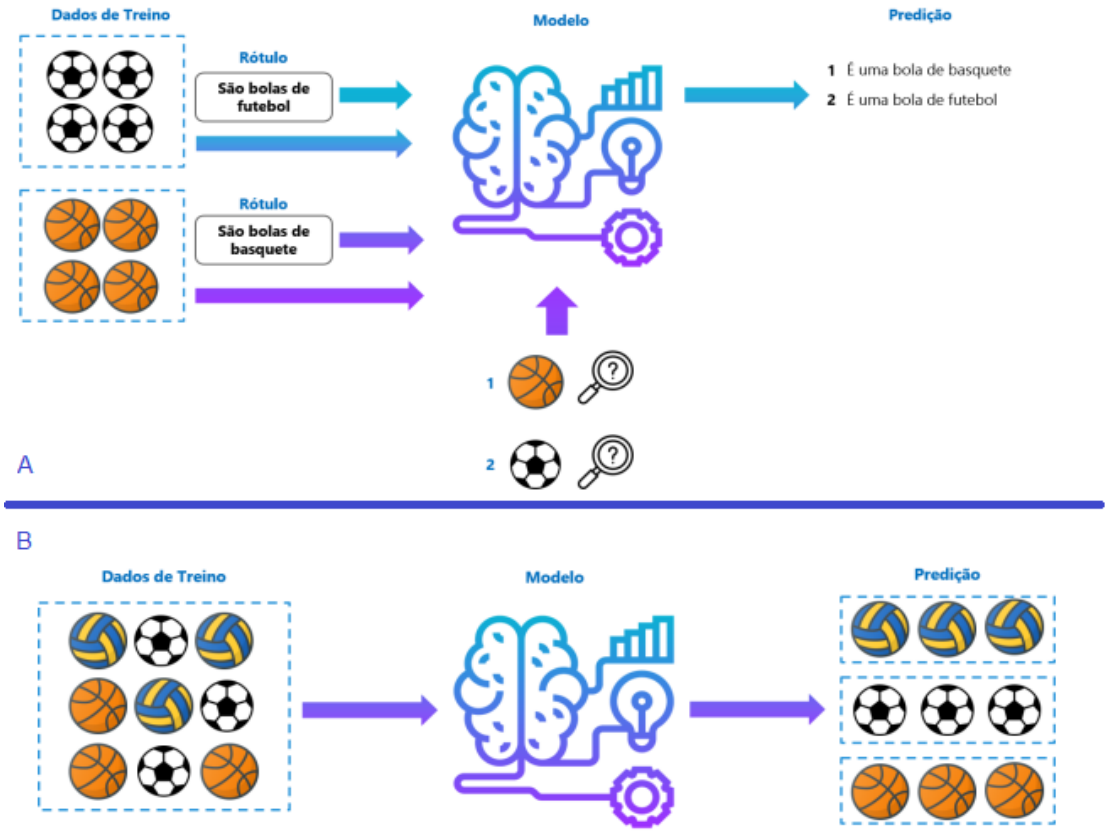


Figura 9. Diferenças entre aprendizado supervisionado e não supervisionado. A) aprendizado supervisionado que aprende com os dados fornecidos. B) aprendizado não supervisionado que reconhece padrões. ADAPTADO DE: CARRICONDE, 2020.

2.2.4 QSAR

QSAR é um método computacional quantitativo que visa estudar a relação entre a estrutura de uma molécula e a sua atividade (VERMA; KHEDKAR; COUTINHO, 2010). QSAR emprega o AM supervisionado, e visa estudar o ligante e como o mesmo se comporta a nível estrutural e na sua atividade (LIN; LI; LIN, 2020).

QSAR utiliza descritores moleculares para análises QSAR, esses descritores nada mais são do que as características topológicas, químicas e físicas de cada molécula que são usadas para compreender a função e o comportamento da molécula (GINI, 2018). As predições realizadas por QSAR podem utilizar modelos baseados em redes neurais, árvores de decisões e regressão linear, o padrão ouro é o *Random Forest* que é um *ensemble* de árvores de decisão, que costumam ser mais simples de entender e mais robustos (KWON et al., 2019).

Existem alguns tipos de QSAR como o 2D QSAR que permite a representação bidimensional dos descritores moleculares. O 3D QSAR permite muitas expressões QSAR sugerem que a seletividade biológica resulta de cada alvo formando interações altamente específicas como ligações de hidrogênio com um ligante (CHERKASOV et al., 2014).

O ADMET (absorção, distribuição, metabolismo, excreção e toxicidade) é uma parte importante quando se trata da descoberta de novos fármacos, e visa estudar comportamento de uma molécula química com potencial farmacológico (VAN DE WATERBEEMD; GIFFORD, 2003). O QSAR faz parte de uma das abordagens disponíveis para ADMET, assim como os modelos de ML, que podem ser usados para a criação de novas ferramentas para analisar dados disponíveis em bancos *Drug Bank* (https://go.drugbank.com/) ChEMBL (https://www.ebi.ac.uk/chembl/) e PubChem (https://pubchem.ncbi.nlm.nih.gov/) (GUAN et al., 2018).

2.2.5 *Decision Tree*

Decision Tree (Figura 10) em tradução livre “árvore de decisão” é similar a um fluxograma, sua estrutura é composta pela primeira variável que é denominada de raiz, as variáveis onde decisões são tomadas são chamadas de nós e toda a

saída de dados é chamada de ramos, ou seja é o trajeto percorrido e o resultado é chamado de nodo folha (SHALEV-SHWARTZ; BEN-DAVID, 2014).

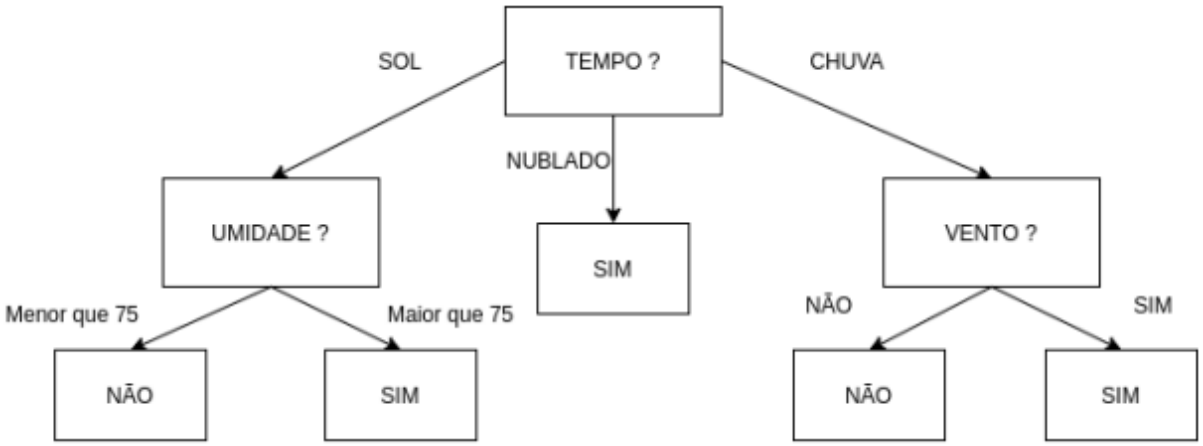


Figura 10. Modelo estrutural de uma árvore de decisão. ADAPTADO DE: CARRICONDE, 2020.

Árvores de decisões usa o algoritmo de *Classification and Regression Trees* - CART (Árvores de classificação e regressão) que em tradução livre é árvores de classificação e regressão. Na classificação o resultado pertence a uma classe, enquanto a regressão apresenta resultados com base em números reais, isto é, pode ser usado para prever o valor de um imóvel (BREIMAN et al, 1984). Ainda existem as técnicas de *ensemble* que usam mais de uma árvore de decisão, conhecidas como *bagging* e *boosting*.

As árvores de decisão podem usar a entropia como medida para o ganho de informação, isto é, caracteriza o nível de pureza da árvore, quanto menor a entropia mais assertiva é a árvore. Para diminuir a entropia é usado o índice gini, quanto mais perto de zero mais puro é o nó (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Dentre as vantagens da árvore de decisão está a facilidade de entender e interpretar os resultados, versatilidade em lidar com dados numéricos e categóricos, robusto contra colinearidade, bom desempenho com grandes conjuntos de dados. Como desvantagem pode apresentar *overfitting* (JAMES et al., 2013).

2.2.6 *Random Forest*

Random Forest (floresta aleatória) é um *ensemble* de árvores de decisão, isto é, um conjunto de várias árvores de decisão oriundas do mesmo *dataset*. O número de árvores de decisão é definido pelo hiperparâmetro *max_samples* através dele podemos estipular se queremos que nosso algoritmo interprete 10 árvores ou 100 (BREIMAN, 2001).

Podemos usar esse modelo para resolver problemas de regressão e classificação. Para regressão é usado uma média de valores entre as árvores enquanto para resolver problemas de classificação é usado um sistema de votação entre as árvores, onde o rótulo mais votado é a predição do modelo (GARGE; BOBASHEV; EGGLESTON, 2013).

Esse tipo de algoritmo usa uma abordagem denominada *bagging* ou *bootstrap aggregation* (Figura 11), cria árvores de maneira aleatória usando os dados de um mesmo *dataset* mas uma é independente da outra, para obter uma predição cada árvore detém um resultado e ao final quando todas as árvores já foram analisadas e feito uma média para regressão ou a votação no caso de classificação (BORSTELMANN, 2020).

Dentre as vantagens de usar esse modelo é que ele consegue reduzir o *overfitting* o que deixa o modelo mais robusto. Também é possível identificar quais atributos foram mais importantes para a decisão através de ferramentas de explicabilidade. Porém como desvantagem é que apresenta um custo computacional mais alto do que usar apenas o modelo de *decision tree* visto que essa abordagem usa mais árvores (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

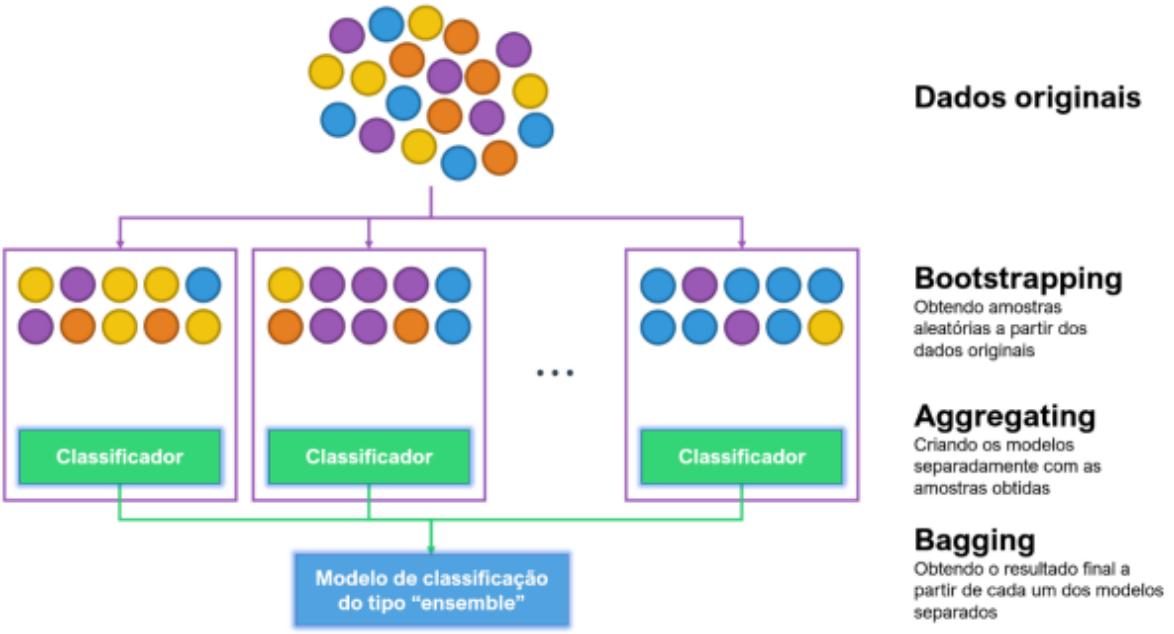


Figura 11. Esquema mostrando como funciona a criação de *ensemble* do tipo *bagging.* ADAPTADO DE: CARRICONDE, 2020.

2.2.7 *Gradient Boosting*

Gradient Boosting também faz parte de um *ensemble* de árvores de decisões (FRIEDMAN, 2001). Diferente do *random forest* que usa uma abordagem *bagging* o *gradient boosting* usa uma abordagem *boosting* como diz o nome.

O *boosting* (Figura 12) funciona de uma forma dependente de outras árvores, por exemplo, quando estimamos que no nosso modelo queremos ter 10 árvores, a primeira é criada uma com base no *dataset*, essa primeira árvore é treinada, e então é criada a segunda árvore com base nos erros da primeira e assim por diante até alcançarmos o número de árvores que definimos. Esse tipo de treino é chamado de *week learning* (FRIEDMAN, 1999). Além do *gradient boosting, outros* modelos usam a abordagem *boosting* como o *XGBoosting* (CHEN; GUESTRIN, 2016) e *Adaboost* (FREUND; SCHAPIRE, 1997).

Figura 12. Esquema de um modelo *boosting*. ADAPTADO DE: CARRICONDE, 2020.

2.2.8 *Logistic Regression*

Logistic regression (regressão logística) é um modelo de classificação apesar do nome ser regressão. É um modelo binário e por causa disso não se usa regressão linear junto com ele pois na regressão logística os rótulos de saída vão de 0 a 100%, indicando uma probabilidade (YU; HUANG; LIN, 2011). É um modelo fácil de implementar mas é passível de *overfitting* (EL SANHARAWI; NAUDET, 2013).

2.2.9 SVM

Support Vector Machine - SVM (Máquina de vetores de suporte) é um modelo de ML baseado em regressão linear, que podem ser usados para classificação, regressão e detecção de *outlier* (BISHOP, 2006). Esse modelo usa pontos de decisão que são chamados de vetores de suporte, para criar um espaço binário, ou seja o modelo mapeia exemplos de treinamento para esses vetores de suporte que podem maximizar a largura da lacuna entre as duas classes (PLATT, 2000).

O algoritmo também pode realizar uma classificação não linear usando kernels que são cálculos não lineares como cálculos sigmóides, polinomial, de função de base radial (*Radial basis function* - RBF). As funções de kernel retornam o produto interno entre dois pontos em um espaço de recurso adequado dando uma noção de similaridade (CHANG; LIN, 2011).

Dentre as vantagens de usar esse modelo é evitar o overfitting, que funciona bem em problemas de classificação pois consegue dividir bem o que pertence a qual classe. Porém tem um custo computacional alto (SMOLA; SCHÖLKOPF, 2004).

2.2.10 *Multi-layer perceptrons (redes neurais)*

O modelo de redes neurais imita o funcionamento de um neurônio, matematicamente para o neurônio funcionar é necessário que os dados de entrada que são encontrados na primeira camada (*input layer*) a qual envia os dados para as camadas escondidas (*hidden layer*) ganham um peso o qual é calculado por vários cálculos de regressão linear e não lineares, após isso os dados que passaram por todas as camadas são disponibilizados na camada de saída (*output layer)* (HINTON, 1989) (Figura 13). Uma rede neural simples tem até 3 camadas escondidas, mais que isso é considerado uma rede neural profunda e se enquadra no *deep learning* (GLOROT; BENGIO, 2010).

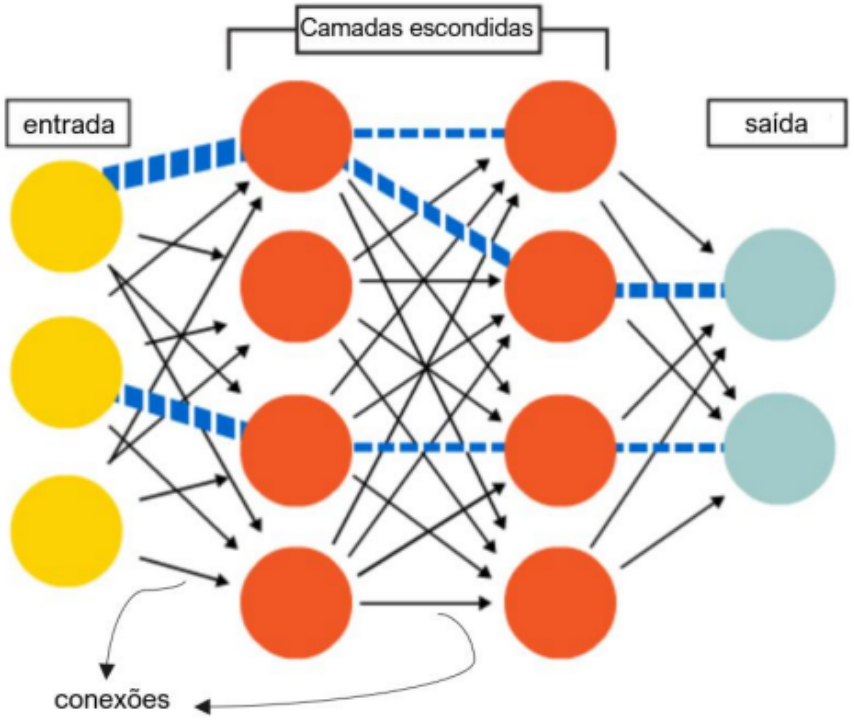


Figura 13. Esquema mostrando como funciona uma rede neural. ADAPTADO DE: CARRICONDE, 2020.

2.2.11 Métricas de classificação

Para avaliar o funcionamento dos modelos usados num experimento de AM, algumas métricas são usadas, as principais são *precision*, *recall*, f1-*score* e *accuracy* (CHURCHER et al., 2021). A *accuracy* (acurácia) refere-se ao quanto o modelo acertou nas previsões. É uma boa média geral de como o modelo funcionou, porém pode ser enganosa (GURUCHARAN, 2020). O *recall* indica a relação entre todos os positivos previstos e quais são realmente positivos e serve para observarmos os falsos negativos (YAO; ZHAO; FAN, 2006). F1-*score* é a pontuação média ponderada entre a precisão e o *recall*. F-Score baixo, é um indicativo de que ou a precisão ou o recall está baixo (CAHYO; HIDAYAT; ADHIPTA, 2016). *Precision* (precisão) é usada como uma medida para determinar o número de verdadeiros divididos por todas as previsões positivas, ou seja, nos diz entre todos os positivos qual a chance de haver um falso positivo (NIKHITA; JABBAR, 2019).

3. HIPÓTESE E OBJETIVOS

3.1 Hipótese

É possível se desenvolver um *framework* para construção de modelos preditivos de QSAR a partir de dados do PubChem BioAssays para a busca de moléculas com atividade antitumoral.

3.2 Objetivo Geral

Desenvolvimento de uma ferramenta de QSAR, baseada em *machine learning*, que permita a identificação de moléculas com possível atividade anti-tumoral

3.3 Objetivos Específicos

- Obter dados experimentais de estudos de HTS para os cânceres elencados como modelos de estudo;
- Avaliar diferentes algoritmos para o treinamento de modelos preditivos;
- Avaliar as características químicas que possuem maior contribuição no resultado;
- Disponibilizar os modelos preditivos identificados em uma interface web.

4. MATERIAL E MÉTODO

4.1 Obtenção dos dados

Os dados de compostos ativos e inativos foram obtidos através de um banco de dados denominado PubChem (WANG et al., 2012) (https://pubchem.ncbi.nlm.nih.gov/) que fica hospedado dentro da plataforma do NCBI (SAYERS et al., 2011). PubChem é um repositório público, gratuito, que contém dados de atividade biológica de pequenas moléculas e reagentes RNAi, além disso contém estruturas e propriedades químicas de moléculas de baixo peso molecular, que são amplamente usadas em pesquisas para novos fármacos.

Para validar a plataforma desenvolvida, foram selecionados 4 ensaios do PubChem *Bioassays,* sendo um de ensaios referentes e interações com proteínas específicas do câncer de mama (Receptor de Estrógeno Alfa), e 3 relacionados a inibidores de crescimento celular de linhagens de melanoma. Os estudos selecionados estão sumarizados na tabela 1.

Tabela 1. Conjuntos de dados de HTS obtidas no pubchem BioAssays para diferentes tipos de cânceres.

Nº	Tipo de câncer	BioAssays Code	Descrição	Compostos	
				Totais	Ativos
1	Câncer de mama	629	*Estrogen Receptor-alpha Coactivator Binding inhibitors*	86,095	1,151
4	Melanoma	33	*NCI human tumor cell line growth inhibition assay. Data for the UACC-257*	51,102	2,069

			Melanoma cell line		
5	Melanoma	35	*NCI human tumor cell line growth inhibition assay. Data for the SK-MEL-2 Melanoma cell line*	48,644	2,235
6	Melanoma	39	*NCI human tumor cell line growth inhibition assay. Data for the SK-MEL-28 Melanoma cell line*	50,951	1,837

4.2 Engenharia de Features

Para treinar os algoritmos capazes de inferir as características de interesse, foram dadas características estruturais e fisicoquímicas (“descritores”) os quais foram obtidos a partir dos dados de estrutura derivados do pubchem. Sendo para isso utilizada a ferramenta RDKit (https://www.rdkit.org/),uma biblioteca de quimioinformática. Foram usados 27 descritores numéricos dentre eles, descritores 3D, descritores topológicos (*graph*), e cada um representa algo específico da molécula. Os descritores usados podem ser vistos na tabela 2.

Tabela 2. Relação de descritores usados no trabalho.

Descriptor	Description
MimAbsPartialCharge	menor carga que um átomo pode ter
TPSA	é a soma da superfície de todos os átomos

	polares
ExactMolWt	peso molecular exato
MaxAbsPartialCharge	maior carga que um átomo pode ter
NumRadicalEletrons	número de elétrons radicais
MolLogP	estima a lipofilicidade
MolMR	refratividade molecular
HeavyAtomMolWt	peso da molécula ignorando os hidrogênios
NHOHCount	contagem de hidróxido de sódio
NunHAcceptors	receptores de hidrogênio
NunHDonors	doadores de hidrogênio
NunHeteroAtoms	número de heteroátomos
NumRotatableBonds	numero de atomos rotacionáveis
NumValenceElectrons	numero de eletrons na camada de valência
RingCount	contagem de anéis aromáticos
FpDensityMorgan1	*fingerprint* calculado pelo método de morgan considerando o raio igual a 1 (MORGAN, 1965)
FpDensityMorgan2	*fingerprint* calculado pelo método de morgan considerando o raio igual a 2 (MORGAN, 1965)
FpDensityMorgan3	*fingerprint* calculado pelo método de morgan considerando o raio igual a 3 (MORGAN, 1965)
BalabanJ	calcula o índice topológico J (BALABAN, 1982) da molécula reduzindo as chances de

	moléculas isômeras gerarem o mesmo valor
BertzCT	índice topológico da molécula que reflete a complexidade de ligações a heteroátomos (BERTZ, 1981)
Ipc	conteúdo de informação calculado através da matriz de adjacência (BONCHEV; TRINAJSTIĆ, 1977)
Chi	índice topológico calculados a partir das características dos elétrons de valência na molécula para interação de uma certa ordem (LIPKOWITZ, 1991)
Kappa1	índice topológico calculado a partir de sub-caminhos dentro da molécula (LIPKOWITZ, 1991)
HallkierAlpha	representa índices que consideram raios covalentes e o estado de hibridização para a forma da molécula (LIPKOWITZ, 1991)
NOCount	contagem de monóxido de nitrogênio
MolWt	peso molecular

4.3 Seleção de Features

Para selecionar as features e ver quais foram mais relevantes foi usado uma biblioteca python boruta SHAP (EOGHAN KEANY, 2020) que usa dois programas, um deles é o SHAP (https://shap.readthedocs.io/en/latest/index.html) e o outro o Boruta Package (KURSA; RUDNICKI, 2010). Após análise, as features mais relevantes foram usadas para testar os modelos.

4.4 Treinamento de Modelos

Os modelos usados para treinar a ferramenta foram o *Gradient boosting classifier*, *Random forest classifier*, *MPLClassifier*, *Logistic regression* e *Decision tree classifier.* Os parâmetros escolhidos foram selecionado com base no *grid search* que é uma função do scikit-learn que prevê quais os melhores parâmetros a serem usados para cada um dos modelos, que podem ser conferidos na tabela 3.

Tabela 3. Parâmetros usados para o treinamento dos modelos.

Model	Description	hyperparameters
Gradient Boosting Classifier	modelo *ensemble* de árvores de decisão que usa a técnica *boosting*	n_estimators (10 - 201 : 10) max_depth (1 - 21)
Random Forest Classifier	modelo *ensemble* de árvores de decisão que usa a técnica *bagging*	max_depth (1 - 21)
MLPClassifier	rede neural	learning_rate: (0.0001, 0.001, 0.01)
Decision Tree Classifier	árvores de decisão	,max_depth (1 - 21) criterion (“gini”, “entropy”)

4.6 Provisionamento

Os modelos preditivos que apresentarem melhor performance foram provisionados em uma interface *web* implementada com uso do *framework* Flask (https://flask.palletsprojects.com/), hospedada na plataforma Google Cloud Platform.

5. RESULTADOS

O nosso principal resultado é a criação da ferramenta *BioAssays Model Builder* - BAMBU, além dos resultados dos modelos usados para validar o objetivo da ferramenta, que é distinguir entre moléculas ativas e inativas.

5.1 Comandos do Bambu

BAMBU é uma ferramenta desenvolvida para uma dissertação de mestrado no OMIXLAB da Universidade Federal de Pelotas, dentro do programa de pós-graduação em biotecnologia, Foi desenvolvida para auxiliar pesquisadores no desenvolvimento de novos fármacos antitumorais, anti-alzheimer ou outras doenças.

5.2 bambu-preprocess

Bambu usa dados obtidos do PubChem em formato CSV (tabelas) e SDF (estruturas). Além de algumas abordagens como SMOTE e Tomek Links usadas para balancear o conjunto de dados. Os dados experimentais são armazenados em `--assays_csv` e os dados estruturais são armazenados em `--assays_sdf`. Os arquivos em CSV representam dados tabulares e podem ser abertos no excel enquanto os dados em SDF representam a estrutura da molécula. Os resultados desses dados após os modelos terem decorrido são armazenados em `--output_csv`. No comando do bambu-preprocess podemos escolher quais são os nossos dados de entradas, como balancear esses dados e onde eles serão salvos. Durante o pré-processamento é feito de forma aleatória a separação dos dados em conjuntos de treino e teste, 75% e 25% respectivamente.

```
usage: bambu-preprocess [-h] --assays_csv ASSAYS_CSV --assays_sdf ASSAYS_SDF
                        --output_csv OUTPUT_CSV
                        [--balancing_strategy {random_undersampling,random_oversampling,smote}]
                        [--remove_tomek_links]

optional arguments:
  -h, --help            show this help message and exit
  --assays_csv ASSAYS_CSV
                        path to PubChem BioAssays CSV file
  --assays_sdf ASSAYS_SDF
                        path to PubChem BioAssays SDF file
  --output_csv OUTPUT_CSV
                        path to output CSV
  --balancing_strategy {random_undersampling,random_oversampling,smote}
                        data balancing strategy
  --remove_tomek_links  remove tomek links
```

Figura 14. Mensagem de pré-processamento. Nesse código podemos escolher quais as opções de balanceamento e quais os dados que serão pré-processados.

Para o balanceamento dos *datasets* (*balancing_strategy)* temos 4 abordagens diferentes:

- *Undersampling*: Remove exemplos da classe majoritária para corresponder à classe minoritária e equilibrar o conjunto de dados (Figura 15a).
- *Oversampling*: adiciona mais exemplos à classe minoritária para corresponder à classe majoritária equilibrando o conjunto de dados (Figura 15b).

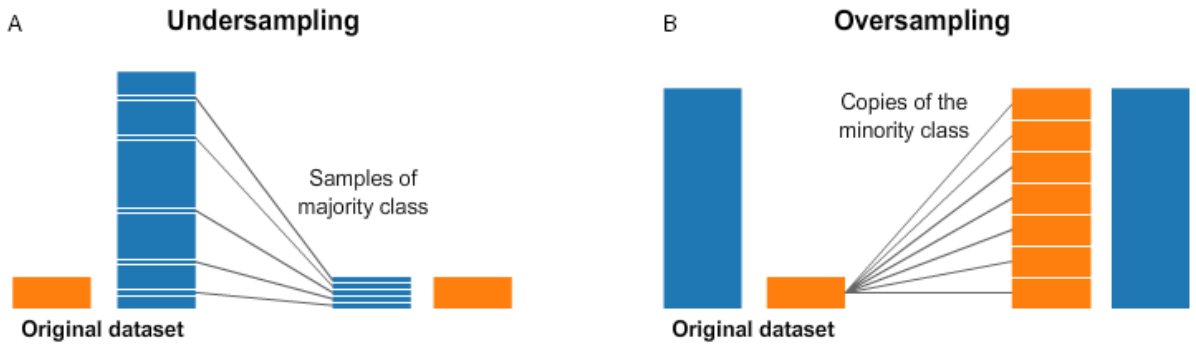


Figura 15. Esquema para exemplificar como funciona o *undersampling* e *oversampling*. A) *Undersampling*, onde ocorre a retirada de exemplos. B) *Oversampling* onde ocorre a adição de exemplos a classe minoritária (AL-SERW, 2021).

- Tomek links: pares de observações são selecionados e aqueles da classe majoritária são excluídos a fim de equilibrar o conjunto de dados (Figura 16). Essa abordagem é indicada para ser usada junto de algoritmos baseados em regressão melhor, visto que ela visa separar em classes, criando uma linha que melhor faz a separação entre as classes.
- SMOTE: *Synthetic Minority Oversampling Technique* (Técnica de sobreamostragem de minoria sintética), sintetiza elementos da classe minoritária, onde elementos já existem. É uma maneira simples de gerar amostras sintéticas aleatoriamente a partir dos atributos de instâncias da classe minoritária (Figura 17).

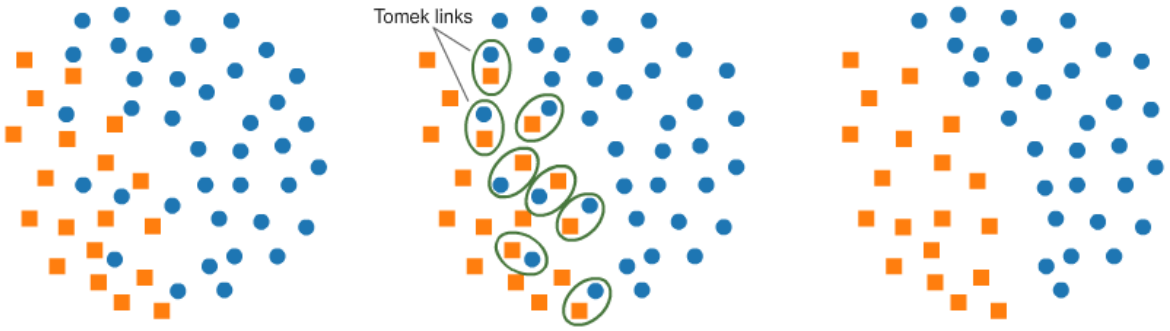


Figura 16. Esquema de como funciona o tomek links (ALENCAR, 2021).

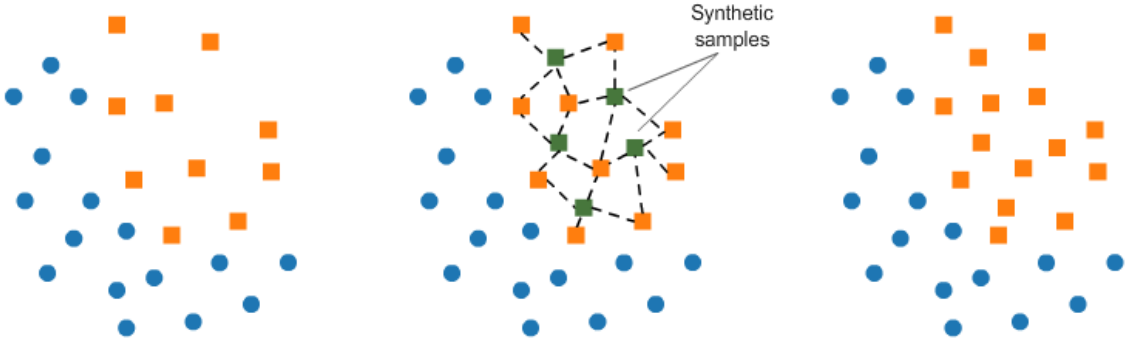


Figura 17. Esquema de como funciona o SMOTE (ALENCAR, 2021).

5.3 bambu-train-model

O modelo de treino é o resultado do teste do modelo, neste caso, estamos executando os dados de pré-processamento em um modelo e realizando um teste para ver a precisão do modelo (Figura 18).

Argumentos:

- `--preprocess_csv`: refere-se a arquivos que já foram pré-processados e agora podem ser testados por modelos.
- `--output_model`: refere-se ao arquivo que será produzido contendo o modelo treinado.

```
usage: bambu-train-model [-h] --preprocess_csv PREPROCESS_CSV --output_model
                         OUTPUT_MODEL

optional arguments:
 -h, --help            show this help message and exit
 --preprocess_csv PREPROCESS_CSV
                       file generated by bambu-preprocess
 --output_model OUTPUT_MODEL
                       path to output model
```

Figura 18. Mensagem do bambu-train-model. Aqui podemos escolher os dados e qual o modelo a ser treinado.

5.4 bambu-server

Bambu-server é onde foi desenvolvido o link para acesso a plataforma (http://104.154.74.231/), a pagina é aberta diretamente na interface onde é colocada a molécula que será analisada como na figura 19. Quando a molécula no formato SMILES é analisada, o resultado (Figura 20) indica se o potencial atividade biológica da molécula analisada, se for acima de 0.5 é considerado uma molécula ativa, caso contrário é uma molécula inativa. Além disso, podemos observar o *shapley value* (Figura 21) que indica a contribuição de cada descritor para a predição do modelo. A

molécula analisada para gerar essas *screenshots* foi o hidrocloreto de 1-N-[3-[2-cloroetil(etil)amino]fenil]-4-N-[3-(dimetilaminametil)fenil]benzeno-1,4-dicarboxamida, de SMILES (CCN(CCCl)C1=CC=CC(=C1)NC(=O)C2=CC=C(C=C2)C(=O)NC3=CC=CC(=C3)CN(C)C.Cl).

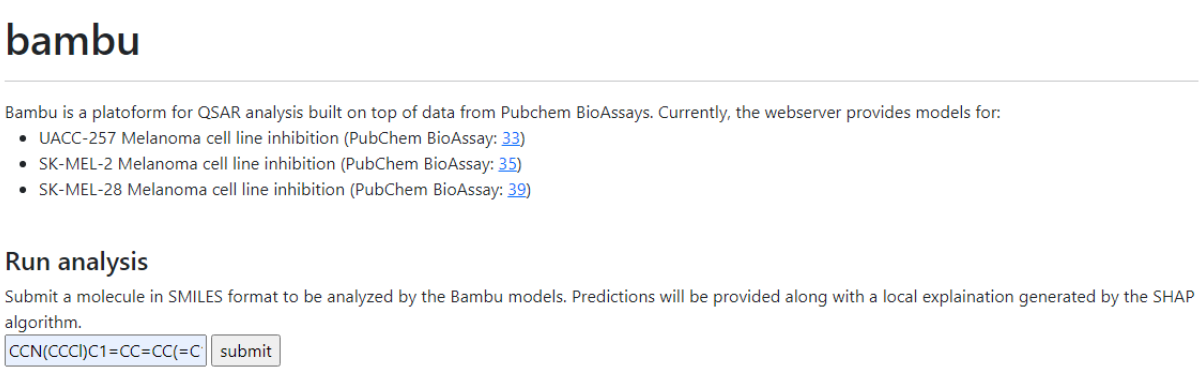
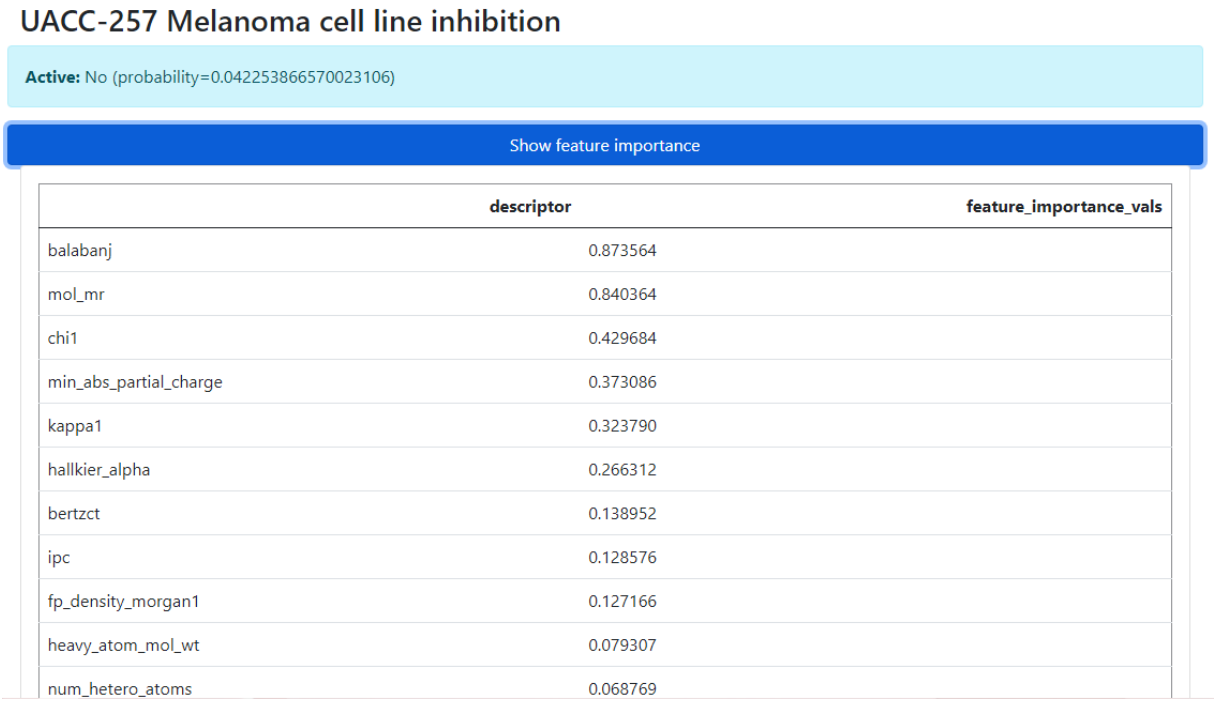


Figura 19. Interface da plataforma web bambu. Nessa barra é colocada a molécula a ser analisada.

Figura 20. Resultados Bambu para a molécula 1-N-[3-[2-cloroetil(ethyl)amino]fenil]-4-N-[3-(dimetilaminametil)fenil]benzeno-1,4-dicarboxamida, de SMILES (CCN(CCCl)C1=CC=CC(=C1)NC(=O)C2=CC=C(C=C2)C(=O)NC3=CC=CC(=C3)CN(C)C.Cl)



descriptor	feature_importance_vals
balabanj	0.873564
mol_mr	0.840364
chi1	0.429684
min_abs_partial_charge	0.373086
kappa1	0.323790
hallkier_alpha	0.266312
bertzct	0.138952
ipc	0.128576
fp_density_morgan1	0.127166
heavy_atom_mol_wt	0.079307
num_hetero_atoms	0.068769

Figura 21. Resultados SHAP para as predições geradas para a molécula 1-N-[3-[2-cloroetil(etil)amino]fenil]-4-N-[3-(dimetilaminametil)fenil]benzeno-1,4-dicarboxamida, de SMILES (CCN(CCCl)C1=CC=CC(=C1)NC(=O)C2=CC=C(C=C2)C(=O)NC3=CC=CC(=C3)CN(C)C.Cl, para o modelo preditivo para inibidores do crescimento da linhagem celular UACC-257.

5.5 Resultados dos treinamentos dos modelos

Os resultados obtidos através dos modelos podem ser observados nas tabelas 4, 5, 6 e 7. As tabelas apresentam qual foi o algoritmo usado e qual a estratégia de balanceamento usada, como dados obtidos temos a precisão, *recall*, f1-*score* e a acurácia.

A precisão nos diz entre todos os positivos qual a chance de haver um falso positivo, já o *recall* indica a relação entre todos os positivos previstos e quais são realmente positivos e serve para observarmos os falsos negativos. O f1-*score* é a média harmônica entre a precisão e o *recall* e a acurácia indica o quão bem o modelo funcionou.

Tabela 4. Resultados do treinamento e seleção de modelos para inibição do crescimento da cultura celular UACC-257 de melanoma (PubChem BioAssays: 33).

algoritmo (sklearn)	balanceamento	precisão	recall	f1-score	acurácia
DecisionTreeClassifier	*no balancing strategy*	98.15	13.73	24.09	56.74
GradientBoostingClassifier	*no balancing strategy*	99.03	**26.42**	**41.72**	**63.08**
RandomForestClassifier	*no balancing strategy*	**100.00**	16.58	28.44	58.29
MLPClassifier	*no balancing_strategy*	0.00	0.00	0.00	50.00
DecisionTreeClassifier	*random undersampling*	100.00	17.79	30.21	58.90
GradientBoostingClassifier	*random undersampling*	95.77	17.04	28.94	58.15
RandomForestClassifier	*random undersampling*	**100.00**	14.79	25.76	57.39
MLPClassifier	*random undersampling*	64.93	**56.14**	60.22	62.91
DecisionTreeClassifier	*random oversampling*	98.39	14.73	25.63	57.25
GradientBoostingClassifier	*random oversampling*	99.06	25.36	40.38	62.56
RandomForestClassifier	*random oversampling*	98.46	15.46	26.72	57.61
MLPClassifier	*random oversampling*	35.90	3.38	6.18	48.67
DecisionTreeClassifier	*smote*	94.74	8.80	16.11	54.16
GradientBoostingClassifier	*smote*	98.90	22.00	36.00	60.88
RandomForestClassifier	*smote*	98.18	13.20	23.28	56.48
MLPClassifier	*smote*	0.00	0.00	0.00	50.00

Tabela 5. Resultados do treinamento e seleção de modelos para inibição do crescimento da cultura celular SK-MEL-2 de melanoma (PubChem BioAssays: 35).

algoritmo (sklearn)	balanceamento	precisão	recall	f1-score	acurácia
DecisionTreeClassifier	*no balancing strategy*	95.29	22.01	35.76	60.46
GradientBoostingClassifier	*no balancing strategy*	99.07	28.80	**44.63**	**64.27**
RandomForestClassifier	*no balancing strategy*	**100.00**	17.12	29.23	58.56
MLPClassifier	*no balancing strategy*	0.00	0.00	0.00	50.00
DecisionTreeClassifier	*random undersampling*	**100.00**	3.99	7.67	51.99
GradientBoostingClassifier	*random undersampling*	98.88	23.40	37.85	61.57
RandomForestClassifier	*random undersampling*	98.55	18.09	30.56	58.91
MLPClassifier	*random undersampling*	0.00	0.00	0.00	50.00

DecisionTreeClassifier	*random oversampling*	94.12	15.80	27.06	57.41
GradientBoostingClassifier	*random oversampling*	96.33	25.93	40.86	62.47
RandomForestClassifier	*random oversampling*	98.18	13.33	23.48	56.54
MLPClassifier	*random oversampling*	63.57	67.65	65.55	64.44
DecisionTreeClassifier	*smote*	95.08	14.61	25.33	56.93
GradientBoostingClassifier	*smote*	**100.00**	28.21	44.01	64.11
RandomForestClassifier	*smote*	**100.00**	19.14	32.14	59.57
MLPClassifier	*smote*	74.22	**42.07**	53.70	63.73

Tabela 6. Resultados do treinamento e seleção de modelos para inibição do crescimento da cultura celular SK-MEL-28 de melanoma (PubChem BioAssays: 39).

algoritmo (sklearn)	**balanceamento**	**precisão**	**recall**	**f1-score**	**acurácia**
DecisionTreeClassifier	*no balancing strategy*	92.86	7.49	13.87	53.46
GradientBoostingClassifier	*no balancing strategy*	**100.00**	**24.50**	**39.35**	**62.25**
RandomForestClassifier	*no balancing strategy*	**100.00**	14.12	24.75	57.06
MLPClassifier	*no balancing strategy*	33.33	3.75	6.74	48.13
DecisionTreeClassifier	*random undersampling*	94.59	10.51	18.92	54.95
GradientBoostingClassifier	*random undersampling*	95.95	21.32	34.89	60.21
RandomForestClassifier	*random undersampling*	98.18	16.22	27.84	57.96
MLPClassifier	*random undersampling*	0.00	0.00	0.00	50.00
DecisionTreeClassifier	*random oversampling*	**100.00**	7.74	14.36	53.87
GradientBoostingClassifier	*random oversampling*	**100.00**	20.92	34.60	60.46
RandomForestClassifier	*random oversampling*	**100.00**	13.75	24.18	56.88
MLPClassifier	*random oversampling*	34.88	4.30	7.65	48.14
DecisionTreeClassifier	*smote*	**100.00**	0.83	1.65	50.42
GradientBoostingClassifier	*smote*	**100.00**	20.83	34.48	60.42
RandomForestClassifier	*smote*	**100.00**	15.56	26.92	57.78
MLPClassifier	*smote*	0.00	0.00	0.00	50.00

Tabela 7. Resultados do treinamento e seleção de modelos para inibidoras do Receptor de Estrógeno identificada a partir de Time-resolved fluorescence energy transfer (TR-FET) (PubChem BioAssays: 639).

algoritmo (sklearn)	balanceamento	precisão	recall	f1-score	acurácia
DecisionTreeClassifier	*no balancing strategy*	78.57	3.27	6.29	51.19
GradientBoostingClassifier	*no balancing strategy*	82.35	4.17	7.93	51.64
RandomForestClassifier	*no balancing strategy*	**100.00**	1.19	2.35	50.60
MLPClassifier	*no balancing strategy*	0.00	0.00	0.00	50.00
DecisionTreeClassifier	*random undersampling*	0.00	0.00	0.00	50.00
GradientBoostingClassifier	*random undersampling*	84.21	4.62	8.77	51.88
RandomForestClassifier	*random undersampling*	**100.00**	0.29	0.58	50.14
MLPClassifier	*random undersampling*	**10.00**	0.29	0.56	48.84
DecisionTreeClassifier	*random oversampling*	50.00	0.26	0.52	50.00
GradientBoostingClassifier	*random oversampling*	90.32	**7.31**	**13.53**	**53.26**
RandomForestClassifier	*random oversampling*	**100.00**	0.26	0.52	50.13
MLPClassifier	*random oversampling*	0.00	0.00	0.00	50.00
DecisionTreeClassifier	*smote*	0.00	0.00	0.00	50.00
GradientBoostingClassifier	*smote*	**100.00**	3.45	6.67	51.72
RandomForestClassifier	*smote*	**100.00**	0.57	1.14	50.29
MLPClassifier	*smote*	50.00	0.29	0.57	50.00

6. DISCUSSÃO

QSAR é uma abordagem computacional para analisar compostos químicos, usando cálculos matemáticos que conseguem prever o potencial farmacológico desses compostos, sendo amplamente usado para a descoberta de novos fármacos (EUGENE N. MURATOV, 2020). É um dos primeiros campos de pesquisa que prioriza a curadoria dos dados e a validação dos modelos desenvolvidos usando essa abordagem. As etapas de curadoria química incluem a identificação e a correção de erros estruturais para um conjunto de compostos, fazendo a retirada de dados incompletos, para isso algumas ferramentas computacionais como o RDKit podem ser usadas (FOURCHES; MURATOV; TROPSHA, 2016).

A construção de um modelo QSAR inclui a coleta de dados que pode ser realizada através de bancos de dados de HTS públicos como o PubChem e o chEMBL, ou de dados privados que o próprio pesquisador tenha, após é feita a curadoria dos dados com o RDKit, por exemplo. Para a construção do modelo é usado um conjunto de dados a partir dos dados obtidos e curados nas etapas anteriores, então são usados algoritmos de ML, que dependendo do objetivo pode ser usado para classificação ou regressão e por fim é feita a validação do modelo que usa conjuntos externos e avalia as métricas de confusão que são a precisão, *recall*, f1-*score* e acurácia (SPIEGEL; SENDEROWITZ, 2020).

Bambu é uma ferramenta web que usa dados de HTS obtidos do Pubchem, desenvolvida através da linguagem de programação python, e algoritmos de ML, similar ao bambu, existe a ferramenta DPubChem (SOUFAN et al., 2018) que também é baseada em QSAR e utiliza os dados do pubchem, o intuito dessa ferramenta é o mesmo que a do bambu, auxiliar pesquisadores a acelerar o processo de conhecimento sobre moléculas de interesse. Porém a principal diferença entre as duas ferramentas é o fato que de o bambu apresenta uma funcionalidade adicional que é a parte de explicabilidade, que permite ver quais as *features* que mais contribuíram para o resultado, a partir da análise de SHAP (https://shap.readthedocs.io/en/latest/index.html), o que permite que o pesquisador entenda melhor o resultado predito pelo modelo. Além disso, a ferramenta pode ser

executada localmente a partir do link do repositório disponível no github (https://github.com/omixlab/cancer-drugs-ml), as dependências para execução do projeto podem ser instaladas com a conda (https://docs.conda.io/en/latest/).

Quatro estudos foram usados para validar o funcionamento da ferramenta, 33, 35, 39 e 629. Com os dados obtidos vimos que os melhores modelos para predição foram os modelos baseados em *ensamble* como o *random forest* e o *gradient boosting* que chegaram em uma precisão alta, acima de 95 com exceção do estudo 629 que usou o modelo de *gradient boosting* sem estratégia de balanceamento, usando a estratégia de *oversampling e undersampling*.

O estudo 33 de inibição do crescimento da cultura celular UACC-257 de melanoma teve os melhores resultados quando usou o algoritmo de *random forest* usando as estratégias de balanceamento *undersampling* e sem estratégia, porém podemos observar que o *recall* não foi tão bom quanto esperado pois em nenhum a predição superou 50% com exceção do algoritmo que usou o algoritmo de MPLClassifier usando a estratégia de balanceamento *undersampling* porém não mostrou uma precisão tão interessante quanto outros algoritmos, e isso pode ter sido causado por ter sido usado uma rede neural menos complexa, apenas com 3 camadas, ou por não ter tido exemplos suficientes para aumentar a complexidade da rede neural, de qualquer forma, tanto para aumentar o *recall* podemos aumentar o número de exemplos, ou adicionar mais *fingerprints* e outros descritores (NEVES et al., 2021).

O estudo 35 de inibição do crescimento da cultura celular SK-MEL-2 de melanoma mostrou 100% de precisão em 4 situações diferentes e apresentou os mesmos critérios que o estudo anterior, mostrando uma precisão alta e um *recall* mais baixo. O estudo 39 de inibição do crescimento da cultura celular SK-MEL-28 de melanoma foi o com maior número de situações onde a precisão alcançou 100%, mas como nos demais o *recall* não foi tão alto. Já o estudo 629 de inibidoras do RE identificada a partir de Transferência de energia de fluorescência resolvida no tempo (*Time-resolved fluorescence energy transfer* - TR-FET) 3 situações onde a precisão alcançou 100% mas foi o estudo que apresentou os piores índices de *recall*.

Dados os resultados, observamos que é interessante adicionar mais exemplos, mais descritores e novos *fingerprints* para melhorar tanto as predições que usam as redes neurais quanto o *recall* de todos os estudos (XIE, 2010). O *recall* pode melhorar ainda mais a ferramenta, dessa forma algumas abordagens para isso devem ser vistas como adicionar novos descritores e avaliar a correlação com os alvos.

7. CONCLUSÃO GERAL

A ferramenta bambu é uma ferramenta promissora e que pode auxiliar pesquisadores na área de *drug discovery* e facilitar a busca por novas moléculas. Alguns ajustes para melhorar as métricas de predição são essenciais, bem como melhorias na interface, mas se mostrou eficiente e cumpre o papel que foi proposto que é conseguir discriminar entre moléculas bioativas com atividade antitumoral.

8. REFERÊNCIAS

ADES, F. et al. Luminal B Breast Cancer: Molecular Characterization, Clinical Management, and Future Perspectives. **Journal of Clinical Oncology**, v. 32, n. 25, p. 2794–2803, 2014.

AHMAD, A. (ED.). Breast Cancer Metastasis and Drug Resistance: Challenges and Progress. **Springer International**. v. 1152, 2019.

AKBANI, R. et al. Genomic Classification of Cutaneous Melanoma. **Cell**, v. 161, n. 7, p. 1681–1696, 18 jun. 2015.

ALENCAR, R. Resampling strategies for imbalanced datasets. **Kaggle**. 2021. Disponível em: <https://kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets>. Acesso em: 20 out. 2021.

AL-SERW, N. A.-R. Undersampling and oversampling: An old and a new approach. **Analytics Vidhya**. 2021. Disponível em: <https://medium.com/analytics-vidhya/undersampling-and-oversampling-an-old-and-a-new-approach-4f984a0e8392>. Acesso em: 20 out. 2021.

AMBURE, P. et al. Development of Generalized QSAR Models for Predicting Cytotoxicity and Genotoxicity of Metal Oxides Nanoparticles. **International Journal of Quantitative Structure-Property Relationships**, v. 5, p. 15–32, 2020.

BADVE, S. et al. Basal-like and triple-negative breast cancers: a critical review with an emphasis on the implications for pathologists and oncologists. **Modern Pathology**, v. 24, n. 2, p. 157–167, 2011.

BALABAN, A. T. Highly discriminating distance-based topological index. **Chemical Physics Letters**, v. 89, n. 5, p. 399–404, 1982.

BAÑULS, J. Estrategias para disminuir el retraso en el diagnóstico del melanoma lentiginoso acral. **Actas Dermo-Sifiliográficas**, v. 109, n. 9, p. 764, 2018.

BEDARD, P. L. et al. Small molecules, big impact: 20 years of targeted therapy in oncology. **The Lancet**, v. 395, n. 10229, p. 1078–1088, 2020.

BENITEZ, M. L. R. et al. Mycobacterium bovis BCG in metastatic melanoma therapy. **Applied Microbiology and Biotechnology**, v. 103, n. 19, p. 7903–7916, 2019.

BERTZ, S. H. The first general index of molecular complexity. **Journal of the American Chemical Society**, v. 103, n. 12, p. 3599–3601, 1981.

BERWICK, M. et al. Melanoma Epidemiology and Prevention. In: KAUFMAN, H. L.; MEHNERT, J. M. (Eds.). **Springer International**. v. 167p. 17–49, 2016.

BI, Q. et al. What is Machine Learning? A Primer for the Epidemiologist. **American Journal of Epidemiology**, v. 188, n. 12, p. 2222–2239, 2019.

BISHOP, C. M. Pattern recognition and machine learning. **Springer.** 2006.

BONCHEV, D.; TRINAJSTIĆ, N. Information theory, distance matrix, and molecular branching. **The Journal of Chemical Physics**, v. 67, n. 10, p. 4517–4533, 1977.

BORSTELMANN, S. M. Machine Learning Principles for Radiology Investigators. **Academic Radiology**, v. 27, n. 1, p. 13–25, 2020.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

BRISTOW; BOWLING, I. JONATHAN. Dermoscopy as a technique for the early identification of foot melanoma. **Journal of Foot and Ankle Research**. 2009.

BÜHLMANN, S.; REYMOND, J.-L. ChEMBL-Likeness Score and Database GDBChEMBL. **Frontiers in Chemistry**, v. 8, p. 46, 2020.

CAHYO, A. N.; HIDAYAT, R.; ADHIPTA, D. Performance comparison of intrusion detection system based anomaly detection using artificial neural network and support vector machine. **AIP Conference Proceedings**, v. 1755, n. 1, p. 070011, 2016.

CHANG, C.-C.; LIN, C.-J. LIBSVM: A library for support vector machines. **ACM Transactions on Intelligent Systems and Technology**, v. 2, n. 3, p. 1–27, 2011.

CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, p. 785–794, 13 ago. 2016.

CHERKASOV, A. et al. QSAR Modeling: Where Have You Been? Where Are You Going To? **Journal of Medicinal Chemistry**, v. 57, n. 12, p. 4977–5010, 2014.

CHURCHER, A. et al. An Experimental Analysis of Attack Classification Using Machine Learning in IoT Networks. **Sensors (Basel, Switzerland)**, v. 21, n. 2, p. 446, 2021.

CLARK, W. H.; ELDER, D. E.; VAN HORN, M. The biologic forms of malignant melanoma. **Human Pathology**, v. 17, n. 5, p. 443–450, 1986.

DEO, R. C. Machine Learning in Medicine. **Circulation**, v. 132, n. 20, p. 1920–1930, 2015.

DEWANE, M. E. et al. Melanoma on chronically sun-damaged skin: Lentigo maligna and desmoplastic melanoma. **Journal of the American Academy of Dermatology**, v. 81, n. 3, p. 823–833, 2019.

DRUKER, B. J. et al. Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. **The New England Journal of Medicine**, v. 344, n. 14, p. 1031–1037, 2001.

EL GHISSASSI, F. et al. A review of human carcinogens—Part D: radiation. **The Lancet Oncology**, v. 10, n. 8, p. 751–752, 2009.

EL SANHARAWI, M.; NAUDET, F. Comprendre la régression logistique. **Journal Français d'Ophtalmologie**, v. 36, n. 8, p. 710–715, 2013.

ELDER, D. E. Thin Melanoma. **Archives of Pathology & Laboratory Medicine**, v. 135, n. 3, p. 342–346, 2011.

EOGHAN KEANY. BorutaShap : A wrapper feature selection method which combines the Boruta feature selection algorithm with Shapley values. **Zenodo**. 2020.

ESCRIG SOS, J.; GÓMEZ QUILES, L.; MAIOCCHI, K. The 8th Edition of the AJCC-TNM Classification: New Contributions to the Staging of Esophagogastric Junction Cancer. **Cirugía Española**. v. 97, n. 8, p. 432–437, 2019.

Estatísticas de câncer. Disponível em: <https://www.inca.gov.br/numeros-de-cancer>. Acesso em: 22 maio. 2021.

EUGENE N. MURATOV, J. B. QSAR without borders. **Chemical Society reviews**, v. 49, n. 11, p. 3525, 2020.

FALK DELGADO, A.; ZOMMORODI, S.; FALK DELGADO, A. Sentinel Lymph Node Biopsy and Complete Lymph Node Dissection for Melanoma. **Current Oncology Reports**, v. 21, n. 6, p. 54, 2019.

FORMAN, S. B. et al. Is superficial spreading melanoma still the most common form of malignant melanoma? **Journal of the American Academy of Dermatology**, v. 58, n. 6, p. 1013–1020, 2008.

FOURCHES, D.; MURATOV, E.; TROPSHA, A. Trust, But Verify II: A Practical Guide to Chemogenomics Data Curation. **Journal of chemical information and modeling**, v. 56, n. 7, p. 1243–1252, 2016.

FREUND, Y.; SCHAPIRE, R. E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. **Journal of Computer and System Sciences**, v. 55, n. 1, p. 119–139, 1997.

FRIEDMAN, J. Stochastic Gradient Boosting. **Computational Statistics & Data Analysis**, v. 38, p. 367–378, 1999.

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. **The Annals of Statistics**, v. 29, n. 5, p. 1189–1232, 2001.

FRIGAULT, M. J. et al. Tisagenlecleucel CAR T-cell therapy in secondary CNS lymphoma. **Blood**, v. 134, n. 11, p. 860–866, 2019.

GARGE, N. R.; BOBASHEV, G.; EGGLESTON, B. Random forest methodology for model-based recursive partitioning: the mobForest package for R. **BMC bioinformatics**, v. 14, p. 125, 2013.

GASTEIGER, J. Chemoinformatics: Achievements and Challenges, a Personal View. **Molecules**, v. 21, n. 2, p. 151, 2016.

GINI, G. QSAR: What Else? In: NICOLOTTI, O. Computational Toxicology. Methods in Molecular Biology. **Springer New York**. v. 1800p. 79–105, 2018.

GLOROT, X.; BENGIO, Y. Understanding the difficulty of training deep feedforward neural networks. **Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics**. **In: PROCEEDINGS OF THE THIRTEENTH INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND STATISTICS. JMLR Workshop and Conference Proceedings**. 2010. Disponível em: <https://proceedings.mlr.press/v9/glorot10a.html>. Acesso em: 30 set. 2021.

GOLDSTEIN, A. M. et al. High-risk Melanoma Susceptibility Genes and Pancreatic Cancer, Neural System Tumors, and Uveal Melanoma across GenoMEL. **Cancer Research**, v. 66, n. 20, p. 9818–9828, 2006.

GUAN, L. et al. ADMET-score – a comprehensive scoring function for evaluation of chemical drug-likeness Electronic supplementary information (ESI) available. **MedChemComm**, v. 10, n. 1, p. 148–157, 30 nov. 2018.

GURUCHARAN, M. K. **Machine Learning Basics: K-Nearest Neighbors Classification**. 2020. Disponível em: <https://towardsdatascience.com/machine-learning-basics-k-nearest-neighbors-classification-6c1e0b209542>. Acesso em: 22 nov. 2021.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The Elements of Statistical Learning. **Springer New York**. 2009.

HEAKAL, Y.; KESTER, M.; SAVAGE, S. Vemurafenib (PLX4032): An Orally Available Inhibitor of Mutated BRAF for the Treatment of Metastatic Melanoma. **Annals of Pharmacotherapy**, v. 45, n. 11, p. 1399–1405, 2011.

HERRERO E, FERNÁNDEZ-MEDARDE A. Advanced targeted therapies in cancer: Drug nanocarriers, the future of chemotherapy. **Eur J Pharm Biopharm**. v. 93, p. 52-79, 2015.

HINTON, G. E. Connectionist learning procedures. **Artificial Intelligence**, v. 40, n. 1–3, p. 185–234, 1989.

INMAN, B. A. et al. Atezolizumab: A PD-L1-Blocking Antibody for Bladder Cancer. **Clinical Cancer Research: An Official Journal of the American Association for Cancer Research**, v. 23, n. 8, p. 1886–1890, 2017.

JAMES, G. et al. An Introduction to Statistical Learning. **Springer New York**. v. 103, 2013.

KURSA, M. B.; RUDNICKI, W. R. Feature Selection with the **Boruta** Package. **Journal of Statistical Software**, v. 36, n. 11, 2010.

KWON, S. et al. Comprehensive ensemble in QSAR prediction for drug discovery. **BMC Bioinformatics**, v. 20, n. 1, p. 521, 2019.

LARSON, R. S.; OPREA, T. I. (Eds.). . Bioinformatics and Drug Discovery. Methods in Molecular Biology. **Springer New York**. v. 1939p. 11–35, 2019.

Lewandowska AM, Rudzki M, Rudzki S, Lewandowski T, Laskowska B. Environmental risk factors for cancer – review paper. **Ann Agric Environ Med**. 26(1):1-7, 2019.

LIM, W. A.; JUNE, C. H. The Principles of Engineering Immune Cells to Treat Cancer. **Cell**, v. 168, n. 4, p. 724–740, 2017.

LIN, X.; LI, X.; LIN, X. A Review on Applications of Computational Methods in Drug Screening and Design. **Molecules**, v. 25, n. 6, p. 1375, 2020.

LIPKOWITZ, K. B. (ED.). Reviews in computational chemistry. **VCH**. 1991.

LO, Y.-C. et al. Machine learning in chemoinformatics and drug discovery. **Drug discovery today**, v. 23, n. 8, p. 1538–1546, 2018.

LONGO, C.; PELLACANI, G. Melanomas. **Dermatologic Clinics**, v. 34, n. 4, p. 411–419, 2016.

LUKE, J. J. et al. Targeted agents and immunotherapies: optimizing outcomes in melanoma. **Nature Reviews Clinical Oncology**, v. 14, n. 8, p. 463–482, 2017.

MENZIES, S. W. et al. Dermoscopic evaluation of nodular melanoma. **JAMA dermatology**, v. 149, n. 6, p. 699–709, 2013.

MISIR, A. F. et al. Primary malignant melanoma. **Saudi Medical Journal**, v. 37, n. 4, p. 446–449, 2016.

MORGAN, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. **Journal of Chemical Documentation**, v. 5, n. 2, p. 107–113, 1965.

NAGAI, H.; KIM, Y. H. Cancer prevention from the perspective of global cancer

burden patterns. **Journal of Thoracic Disease**, v. 9, n. 3, p. 448–451, 2017.

NAMIKAWA, K.; YAMAZAKI, N. Targeted Therapy and Immunotherapy for Melanoma in Japan. **Current Treatment Options in Oncology**, v. 20, n. 1, p. 7, 2019.

NEVES, B. J. et al. Automated Framework for Developing Predictive Machine Learning Models for Data-Driven Drug Discovery. **Journal of the Brazilian Chemical Society**, v. 32, p. 110–122, 2021.

NIKHITA; JABBAR, M. N. K Nearest Neighbor Based Model for Intrusion Detection System. **International Journal of Recent Technology and Engineering**, v. 8, n. 2, p. 2258–2262, 2019.

NOBLE, S.; GOA, K. L. Aldesleukin (recombinant interleukin-2). **BioDrugs: Clinical Immunotherapeutics, Biopharmaceuticals and Gene Therapy**, v. 7, n. 5, p. 394–422, 1997.

O'NEILL, C. H.; SCOGGINS, C. R. Melanoma. **Journal of Surgical Oncology**, v. 120, n. 5, p. 873–881, 2019.

ONITILO, A. A.; WITTIG, J. A. Principles of Immunotherapy in Melanoma. **Surgical Clinics of North America**, v. 100, n. 1, p. 161–173, 2020.

PARDOLL, D. M. The blockade of immune checkpoints in cancer immunotherapy. **Nature reviews. Cancer**, v. 12, n. 4, p. 252–264, 2012.

PARK, S. J. et al. Comparative efficacy of sunitinib versus sorafenib as first-line treatment for patients with metastatic renal cell carcinoma. **Chemotherapy**, v. 58, n. 6, p. 468–474, 2012.

PATEL, L. et al. Machine Learning Methods in Drug Discovery. **Molecules**, v. 25, p. 5277, 2020.

PLATT, J. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. **Adv. Large Margin Classif.**, v. 10, 2000.

PYZER-KNAPP, E. O. et al. What Is High-Throughput Virtual Screening? A Perspective from Organic Materials Discovery. **Annual Review of Materials Research**, v. 45, n. 1, p. 195–216, 2015.

READ, J.; WADT, K. A. W.; HAYWARD, N. K. Melanoma genetics. **Journal of Medical Genetics**, v. 53, n. 1, p. 1–14, 2016.

REDELMAN-SIDI, G.; GLICKMAN, M. S.; BOCHNER, B. H. The mechanism of action of BCG therapy for bladder cancer--a current perspective. **Nature Reviews. Urology**, v. 11, n. 3, p. 153–162, 2014.

RILEY, R. S. et al. Delivery technologies for cancer immunotherapy. **Nature reviews. Drug discovery**, v. 18, n. 3, p. 175–196, 2019.

ROJAS, K.; STUCKEY, A. Breast Cancer Epidemiology and Risk Factors. **Clinical Obstetrics & Gynecology**, v. 59, n. 4, p. 651–672, 2016.

ROSENBERG, S. A. IL-2: The First Effective Immunotherapy for Human Cancer. **Journal of immunology (Baltimore, Md. : 1950)**, v. 192, n. 12, p. 5451–5458, 2014.

RUSSNES, H. G. et al. Breast Cancer Molecular Stratification: From Intrinsic Subtypes to Integrative Clusters. **The American Journal of Pathology**, v. 187, n. 10, p. 2152–2162, 2017.

SAIYED, F. K.; HAMILTON, E. C.; AUSTIN, M. T. Pediatric melanoma: incidence, treatment, and prognosis. **Pediatric Health, Medicine and Therapeutics**, v. 8, p. 39–45, 2017.

SAYERS, E. W. et al. Database resources of the National Center for Biotechnology Information. **Nucleic Acids Research**, v. 39, n. Database issue, p. D38–D51, 2011.

SCHALKA; STEINER, S. D. Consenso Brasileiro de Fotoproteção - Recomendações da SBD - Guia para Dermatologistas. **SBD**. 2013. Disponível em: <https://www.sbd.org.br/dermatologistas/publicacoes/visualizar.aspx?id=4>. Acesso em: 27 ago. 2021.

SCOTT, L. J. Nivolumab: A Review in Advanced Melanoma. **Drugs**, v. 75, n. 12, p. 1413–1424, 2015.

SHALEV-SHWARTZ, S.; BEN-DAVID, S. Understanding Machine Learning: From Theory to Algorithms. **Cambridge University Press**, 2014.

SMOLA, A. J.; SCHÖLKOPF, B. A tutorial on support vector regression. **Statistics and Computing**, v. 14, n. 3, p. 199–222, 2004.

SONDAK, V. K. et al. Ipilimumab. **Nature Reviews Drug Discovery**, v. 10, n. 6, p. 411–412, 2011.

SOUFAN, O. et al. DPubChem: a web tool for QSAR modeling and high-throughput virtual screening. **Scientific Reports**, v. 8, n. 1, p. 9110, 2018.

SPIEGEL, J.; SENDEROWITZ, H. Evaluation of QSAR Equations for Virtual Screening. **International Journal of Molecular Sciences**, v. 21, n. 21, p. 7828, 2020.

TAKAHASHI, J.; NAGASAWA, S. Immunostimulatory Effects of Radiotherapy for Local and Systemic Control of Melanoma: A Review. **International Journal of**

Molecular Sciences, v. 21, n. 23, p. 9324, 2020.

TSANG, J. Y. S.; TSE, G. M. Molecular Classification of Breast Cancer. **Advances in Anatomic Pathology**, v. 27, n. 1, p. 27–35, 2020.

TSIMBERIDOU, A.-M. Targeted therapy in cancer. **Cancer Chemotherapy and Pharmacology**, v. 76, n. 6, p. 1113–1132, 2015.

VAN DE WATERBEEMD, H.; GIFFORD, E. ADMET in silico modelling: towards prediction paradise? **Nature Reviews Drug Discovery**, v. 2, n. 3, p. 192–204, 2003.

VANDIST, K.; STORMS, G.; VAN DEN BUSSCHE, E. Semisupervised category learning facilitates the development of automaticity. **Attention, Perception, & Psychophysics**, v. 81, n. 1, p. 137–157, 2019.

VAREKI; GARRIGÓS, S. M. ; C. Biomarkers of response to PD-1/PD-L1 inhibition. **Elsevier Enhanced Reader**. v. 116, p. 116-124, 2017.

VERMA, J.; KHEDKAR, V.; COUTINHO, E. 3D-QSAR in Drug Design - A Review. **Current Topics in Medicinal Chemistry**, v. 10, n. 1, p. 95–115, 2010.

VODUC, K. D. et al. Breast Cancer Subtypes and the Risk of Local and Regional Relapse. **Journal of Clinical Oncology**, v. 28, n. 10, p. 1684–1691, 2010.

WANG, Y. et al. PubChem's BioAssay Database. **Nucleic Acids Research**, v. 40, p. D400–D412, 2012.

WARR, W. A. A CADD-alog of strategies in pharma. **Journal of Computer-Aided Molecular Design**, v. 31, n. 3, p. 245–247, 2017.

WEBER, P. et al. Dermatoscopy of Neoplastic Skin Lesions: Recent Advances, Updates, and Revisions. **Current Treatment Options in Oncology**, v. 19, n. 11, p. 56, 2018.

WEILER, L.; THOMAS, L.; DALLE, S. Mélanome nodulaire pigmenté. **Annales de Dermatologie et de Vénéréologie**, v. 140, n. 12, p. 827–828, 2013.

XIE, X.-Q. Exploiting PubChem for Virtual Screening. **Expert opinion on drug discovery**, v. 5, n. 12, p. 1205, 2010.

YANG, A. S.; CHAPMAN, P. B. The History and Future of Chemotherapy for Melanoma. **Hematology/oncology clinics of North America**, v. 23, n. 3, p. 583, 2009.

YANG, J. Y. Promoting synergistic research and education in genomics and bioinformatics. **BMC Genomics.** 2008.

YAO, J.; ZHAO, S.; FAN, L. An Enhanced Support Vector Machine Model for

Intrusion Detection. **Rough Sets and Knowledge Technology**. **Lecture Notes in Computer Science**. v. 4062p. 538–543, 2006.

YEO, S. K.; GUAN, J.-L. Breast Cancer: Multiple Subtypes within a Tumor? **Trends in cancer**, v. 3, n. 11, p. 753–760, 2017.

YU, H.-F.; HUANG, F.-L.; LIN, C.-J. Dual coordinate descent methods for logistic regression and maximum entropy models. **Machine Learning**, v. 85, n. 1–2, p. 41–75, out. 2011.