RESSALVA

Atendendo solicitação do(a) autor(a), o texto completo desta tese será disponibilizado somente a partir de 21/10/2024.

UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
INSTITUTO DE BIOCIÊNCIAS DE BOTUCATU
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS BIOLÓGICAS (GENÉTICA)

André Luiz Molan

Construção de um Modelo Preditivo para o Prognóstico de Pacientes com Neuroblastoma baseado em Assinaturas Transcricionais

Botucatu, Setembro de 2022

UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
INSTITUTO DE BIOCIÊNCIAS DE BOTUCATU
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS BIOLÓGICAS (GENÉTICA)

André Luiz Molan

Construção de um Modelo Preditivo para o Prognóstico de Pacientes com Neuroblastoma baseado em Assinaturas Transcricionais

Tese apresentada ao Instituto de Biociências, Campus de Botucatu, UNESP, em preenchimento dos requisitos para a obtenção do título de Doutor no Programa de Pós-Graduação em Ciências Biológicas (Genética).

Área de Concentração: Genética
Orientador: Prof. Dr. José Luiz Rybarczyk Filho

Botucatu, Setembro de 2022.

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSANGELA APARECIDA LOBO-CRB 8/7500

Molan, André Luiz.
Construção de um modelo preditivo para o prognóstico de pacientes com neuroblastoma baseado em assinaturas transcricionais / André Luiz Molan. - Botucatu, 2022

Tese (doutorado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu
Orientador: José Luiz Rybarczyk Filho
Capes: 20205007

1. Aprendizado do computador. 2. Neuroblastoma. 3. Predição (Logica). 4. Controle preditivo. 5. Prognóstico.

Palavras-chave: aprendizado de máquina; assinaturas transcricionais; neuroblastoma; predição; prognóstico.

Agradecimentos

- Agradeço ao processo CAPES 88882.433297/2019-01 e aos processos CNPq 458810/2013-4 e 473789/2013-2 pelo suporte financeiro, fundamental para o desenvolvimento do trabalho;
- Agradeço ao meu orientador, Prof. Dr. José Luiz Rybarczyk Filho, pela orientação de elevada qualidade, apoio, paciência e incentivo;
- Agradeço à Profª. Drª. Agnes Alessandra Sekijima Takeda, fundamental para melhoria da qualidade do trabalho e apresentações orais;
- Agradeço aos amigos de laboratório, Giordano Bruno Sanches Seco, José Rafael Pilan e Giovanna Melato Bonança, os quais contribuíram enormemente com suas amizades e ideias para eventuais melhorias no trabalho;
- Agradeço aos meus pais, José Alberto Turini Molan e Maria Tereza de Oliveira Molan, pelo apoio e suporte incondicionais mesmo nos momentos mais difíceis.

Resumo

O neuroblastoma é o tumor sólido extracraniano mais comum em indivíduos com idade inferior a 15 anos. É uma doença extremamente heterogênea, podendo regredir ou evoluir de forma espontânea e, em alguns casos, se mostrar bastante agressiva. Alguns sistemas como o *International Neuroblastoma Staging System* (INSS) e o *International Neuroblastoma Pathology Classification* (INPC), permitem classificar o estágio da doença, relacionando-o a um prognóstico favorável ou desfavorável com base em características histológicas. Com o surgimento das tecnologias de sequenciamento NGS, associadas a técnicas como o RNA-seq, modelos preditivos robustos e versáteis puderam ser desenvolvidos e aplicados ao câncer. Várias abordagens para o estudo de neoplasias pediátricas, entretanto, foram adaptadas de pesquisas oncológicas relacionadas a adultos. Porém, cânceres infantis apresentam taxas de mutações recorrentes muito baixas em relação a adultos, demandando novas abordagens para melhor compreender vulnerabilidades e etiologias específicas. Nos últimos anos, um aumento de memória, poder de processamento e capacidade de armazenamento aumentaram a importância da computação para a biologia, tornando possível a implementação de abordagens capazes de resolver problemas antes considerados difíceis. Dentre elas, destaca-se o aprendizado de máquina, o qual permite modelar a relação entre um conjunto de valores observáveis (entradas) e um grupo de variáveis relacionadas a estes valores (saídas). Neste contexto, utilizamos regressão logística, redes neurais artificiais com *multilayer perceptron* e árvore de decisão, todas técnicas de aprendizado supervisionado, para a análise de assinaturas transcricionais em dados de RNA-seq de pacientes com neuroblastoma. Desenvolvemos diferentes modelos preditivos baseados nas variáveis sobrevivência, progressão do tumor, alto risco e classe tumoral, com o propósito de caracterizar prognósticos clínicos favoráveis e desfavoráveis. Aplicamos recursos de bibliotecas das linguagens *Python* e R na criação dos modelos, seleção de *features* e balanceamento dos dados. Dentre os principais resultados, conseguimos, por exemplo, predizer a sobrevivência (vivo ou morto) com 91,33% de acurácia. Além disso, para compreendermos biologicamente o melhor modelo encontrado, conduzimos uma análise de enriquecimento funcional em termos de processos biológicos do *Gene Ontology* e ontologias relacionadas à doenças - este, através da biblioteca DOSE da linguagem R. Isso nos permitiu associar as *features* (subconjunto de transcritos utilizados pelo modelo) a processos biológicos relacionados à funções de desenvolvimento e divisão celular, além de ontologias de doenças referentes ao neuroblastoma.

Abstract

Neuroblastoma is the most common extracranial solid tumor in children. It is an extremely heterogeneous disease, which may spontaneously regress or evolve and, in some cases, be quite aggressive. Some systems such as the *International Neuroblastoma Staging System* (INSS) and the *International Neuroblastoma Pathology Classification* (INPC), make it possible to classify the stage of the disease, relating it to a favorable or unfavorable prognosis based on histological characteristics. NGS sequencing technologies, associated to techniques such as RNA-seq, make it possible to develop robust and versatile predictive models applied to cancer studies. Several approaches to the study of pediatric neoplasms have been adapted from adult oncology research. However, childhood cancers have very low rates of recurrent mutations in relation to adults, requiring new approaches to better understand vulnerabilities and specific etiologies. In the last years, increased memory, processing power and storage capacity have increased the importance of computer science to biology, making it possible to implement approaches that can solve problems once considered too dificult. In this scenario, we highlight *machine learning* techniques, which allows us to model the relationship between a set of observable values (input) and a group of variables related to these values (outputs). In this context, we used logistic regression, artificial neural networks with multilayer perceptron and decision tree, all supervised learning techniques, for the analysis of transcriptional signatures in RNA-seq data from patients with neuroblastoma. We developed different predictive models based on the variables survival, tumor progression, high risk and tumor class in order to characterize favorable and unfavorable clinical outcomes. We applyed resources from Python and R programming languages libraries to create the models, select features and balance the datasets. Among the main results, we were able, for example, to predict survival (alive or dead) with 91.33% accuracy. Furthermore, in order to biologically understand the best model found, we conducted a functional enrichment analysis in terms of biological processes from Gene Ontology and ontologies related to diseases - this one, by DOSE library from R language. This made it possible to associate the features (subset of transcripts used by the model) to biological processes related to developmental functions and cell division, as well as disease ontologies related to neuroblastoma.

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

1.1 Câncer

De acordo com o *National Institute of Cancer* (NIH, 2020), câncer é o nome atribuído a um grupo de doenças em que algumas células multiplicam-se sem controle, conforme Figura 1.1, e se espalham nos tecidos adjacentes, podendo ocorrer em quase qualquer tipo de tecido. Normalmente as células crescem e se dividem com o intuito de formar novas células em função da necessidade do organismo. Quando envelhecem ou são danificadas, elas morrem e novas células tomam o seu lugar. Quando o câncer se desenvolve, porém, este processo é interrompido. As células se tornam anormais. As mais velhas ou danificadas sobrevivem e novas são formadas sem que haja necessidade. As células extra se dividem incessantemente e são capazes de originar crescimentos irregulares conhecidos como tumor. Muitos tipos de câncer formam massas de tecidos chamadas de tumores sólidos. Contudo, vale evidenciar que alguns tipos de câncer como leucemia, por exemplo, geralmente não são sólidos.

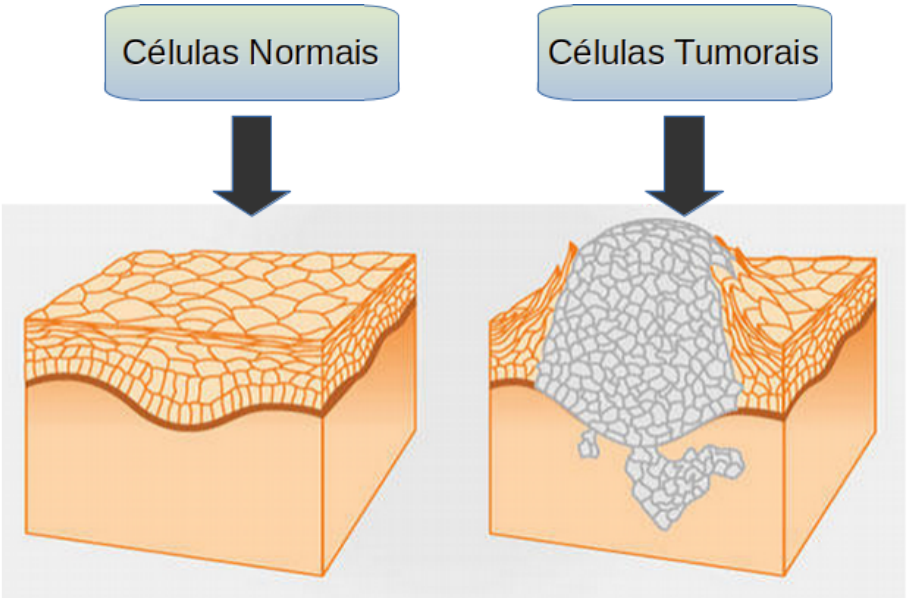

Figura 1.1: Tecido normal e tecido com proliferação descontrolada de células, formando um tumor. Fonte: Figura adaptada de (NIH, 2020).

Tumores cancerígenos são malignos, o que significa que eles podem se espalhar ao redor ou ainda invadir outros tecidos mais próximos (NIH, 2020). Além disso, uma vez que há o crescimento do tumor, as células locais podem viajar pelo sangue ou sistema linfático para locais distantes no corpo e formar novos tumores em outros órgãos. No entanto, diferente dos tumores malignos, tumores benignos não se espalham, mas podem crescer localmente e se tornarem grandes. Estes, quando removidos cirurgicamente, normalmente não retornam, ao contrário dos malignos.

Há evidências que mostram uma origem genética para a transformação de células de um estado benigno para maligno. Agentes mutagênicos como substâncias químicas e radiação ionizante são capazes de introduzir mutações nos genes e causarem câncer. Dentre as principais mutações, damos destaque às oncogênicas e às que atuam em genes supressores de tumor. As mutações oncogênicas presentes em uma célula cancerosa atuam, principalmente, no **ganho de função**, sugerindo duas características principais: na primeira, as proteínas codificadas por oncogenes são ativadas em células tumorais, ao passo que na segunda, a mutação precisa existir apenas em um alelo para que seja capaz de contribuir com a formação do tumor. Quanto às mutações nos genes supressores de tumor, tratam-se de mutações recessivas de **perda de função** e que devem estar presentes em ambos os alelos do gene. Ela faz com que os produtos gênicos codificados percam totalmente ou uma parte de sua atividade (GRIFFITHS et al., 2013).

No mundo, hoje, o câncer é uma das maiores causas de morte e surge como uma barreira para o aumento da expectativa de vida em todos os países do globo no século 21 (BRAY et al., 2018). De acordo com a organização mundial de saúde (WHO), o câncer é a primeira ou segunda maior responsável por óbitos de indivíduos com menos de 70 anos em 91 de 172 países, como mostra a Figura 1.2, e terceira ou quarta em um adicional de 22 países.

A incidência do câncer e a taxa de mortalidade por ele ocasionada crescem rapidamente mundo afora (BRAY et al., 2018). As razões são complexas, mas refletem, principalmente, o envelhecimento e crescimento populacional, além de fatores associados ao desenvolvimento socioeconômico.

Segundo estimativas do Ministério da Saúde (INCA, 2020), no Brasil, para o triênio 2020-2022, espera-se que ocorrerão em torno de 625 mil casos de câncer. Se forem excluídos os casos de câncer de pele não melanoma, o número cai para 450 mil. A predominância, contudo, será do câncer de pele não melanoma, seguido pelos cânceres de próstata e mama, cólon e reto, pulmão e estômago.

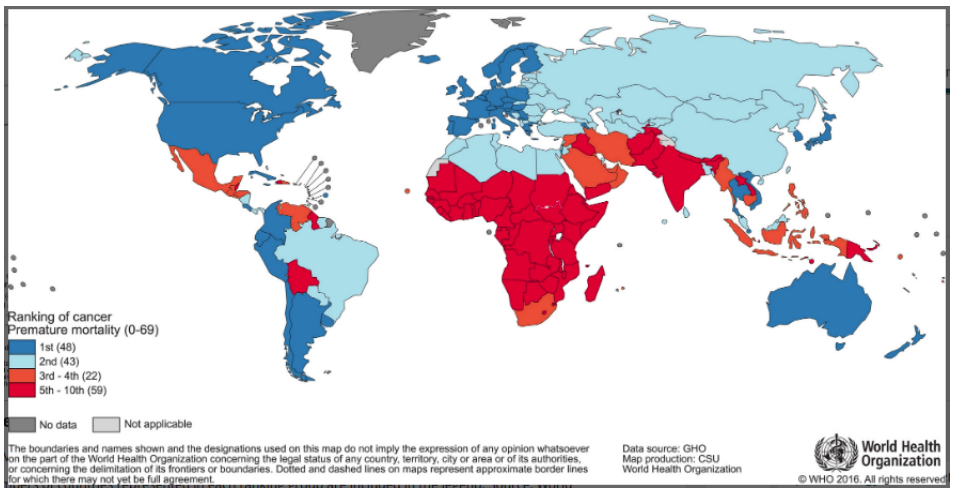

Figura 1.2: Mapa Global apresentando o câncer como principal causa de mortes de indivíduos com menos de 70 anos em 2015. O número de países representados em cada ranking descrito por diferentes cores é apresentado na legenda. Fonte: Organização Mundial de Saúde, 2015.

Podemos ainda destacar o câncer infanto-juvenil, que acomete crianças e adolescentes de 0 a 19 anos de idade e corresponde entre 1% e 4% de todos os tumores malignos na maioria das populações (nos países em desenvolvimento, a proporção é entre 3% e 4%, ao passo que em países desenvolvidos, ela não ultrapassa 1% (SPIRONELLO et al., 2020)). No Brasil, o câncer infanto-juvenil é considerado raro quando comparado ao câncer em adultos, consistindo entre 2% e 3% de todos os tumores registrados. De acordo com o INCA (Instituto Nacional do Câncer), a incidência foi de 12600 novos casos em 2017, sendo as regiões Sudeste e Nordeste aquelas com maior concentração (5300 e 2900 respectivamente). Ainda, de acordo com o INCA, para o triênio 2020, 2021 e 2022, estima-se que sejam diagnosticados 8640 novos casos de câncer infanto-juvenil, sendo 4130 em homens e 4150 em mulheres (137,87 novos casos por milhão em homens e 139,04 casos por milhão em mulheres). Na infância e adolescência, tumores como leucemias e linfomas, que atingem o sistema nervoso central, são os mais frequentes. Outros exemplos de tumores que atingem indivíduos nesta faixa etária são o tumor de *Wilms*, retinoblastoma, tumor germinativo, osteosarcomas, sarcomas e neuroblastoma.

1.2 Neuroblastoma

Tumores neuroblásticos são neoplasias derivadas de células da crista neural, as quais povoam, principalmente, gânglios simpáticos e medula adrenal (CASTLEBERRY, 1997). Por compartilharem a mesma origem, não são considerados como entidades tumorais distintas, sendo representados por diferentes níveis de diferenciação de células estromais e neuroblásticas (FENDLER et al., 2013). Um exemplo é o neuroblastoma, o tumor sólido extracraniano mais

comum em bebês e crianças. Sua incidência média é de 1 caso a cada 100000 crianças nos Estados Unidos e representa 8% de todas as malignidades diagnosticadas em pacientes com idade inferior a 15 anos, sendo responsável por aproximadamente 15% das mortes ligadas a cânceres pediátricos (YOUNG-JR et al., 1986). Trata-se de uma doença extremamente heterogênea, podendo, de forma espontânea, regredir ou evoluir - com ou sem terapia - ou ainda se mostrar bastante agressivo. Entretanto, foram identificados uma série de fatores responsáveis por tal heterogeneidade, ampliando o número de evidências biológicas e moleculares capazes de prever o comportamento clínico desse tipo de tumor (DAVIDOFF, 2012).

O neuroblastoma é um tumor maligno embrionário com origem nas células da crista neural, a qual se forma entre a terceira e quarta semana do desenvolvimento embrionário humano, com algumas de suas células se diferenciando e migrando para criar o sistema nervoso simpático. Os tumores podem surgir em qualquer lugar ao longo do sistema nervoso simpático, mas são encontrados com maior frequência nas glândulas adrenais (em torno de 40% dos casos) ou em regiões do abdômen, tórax ou pélvis (HECK et al., 2009). Embora o neuroblastoma seja, algumas vezes, diagnosticado ainda na fase perinatal, não foram identificados de forma consistente influências do ambiente ou exposições parentais que impactam no surgimento da doença. Uma investigação da biologia molecular do neuroblastoma começa com a caracterização citogenética de linhagens celulares derivadas do tumor. A grande maioria dessas linhagens apresentam corpos de cromatina de minuto duplo (DMs - pequenos corpos que consistem na amplificação gênica em uma região extra cromossômica e são encontrados em uma variedade de células tumorais) ou regiões de coloração homogênea (HSRs), ambas representando amplificação do DNA e deleções no braço curto do cromossomo 1. Com isso, essas observações iniciais demonstraram que tanto o ganho quanto a perda de material genético são comuns durante a evolução do neuroblastoma, o que é consistente com os atuais conceitos sobre gênesis tumoral envolvendo ativação de oncogenes e inativação de genes supressores de tumores (RILEY et al., 2004).

Dentre as principais variáveis clínicas relacionadas ao neuroblastoma, podemos citar a idade e o estágio da doença no momento do diagnóstico como sendo as mais importantes. Esta última segue a *International Neuroblastoma Staging System* (INSS), um sistema cirúrgico-patológico que depende, especialmente, da ressecção completa do tumor primário e avaliação dos linfonodos ipsilaterais (DAVIDOFF, 2012). Além disso, *Shimada* e colaboradores (SHIMADA et al., 1984) criaram um sistema de classificação com base na idade do paciente e morfologia do tumor que, posteriormente, após revisão, passou a ser conhecido como *International Neuroblastoma Pathology Classification* (INPC) (PEUCHMAUR et al., 2003) e pode ser melhor descrito através da Figura 1.3. Tumores neuroblásticos são divididos em quatro subtipos histológicos com base no grau do estroma de *Schwann* adjacente: neuroblastoma (pobre

em estroma de *Schwann*), *ganglioneuroblastoma-intermixed* (rico em estroma de *Schwann*), ganglioneuroblastoma-nodular (composto - rico, pobre ou estroma de *Schwann* dominante) e ganglioneuroma (estroma de *Schwann* dominante). Tais tumores são relacionados a um de dois subgrupos de prognóstico - favorável e desfavorável - baseados nas características histológicas, que também incluem o grau de diferenciação do neuroblasto, morfologia de células neuroblásticas (índice mitose-*karyorrhexis*, MKI) e idade do paciente (CHATTEN et al., 1988). O sistema de classificação INPC se tornou amplamente aceito e provou ser muito útil quanto a predição de prognósticos, sendo sua importância confirmada em uma ampla análise também feita por *Shimada* e colaboradores (SHIMADA et al., 2001).

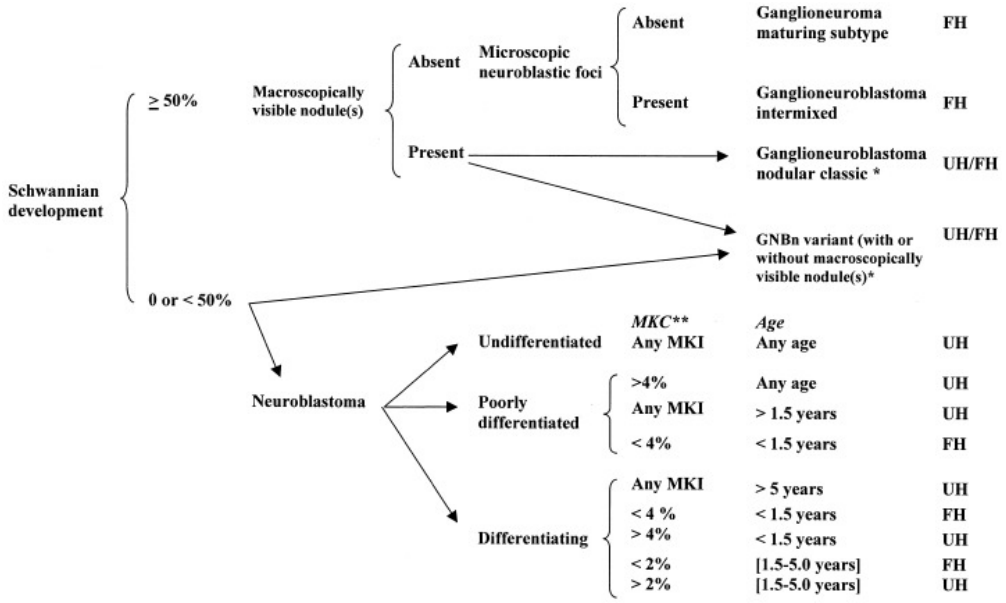

Figura 1.3: Revisão e classificação de *Shimada* (INPC) exemplificado a partir do desenvolvimento das células de *Schwannian*. FH: histologia favorável; UH: histologia desfavorável; GNBn: ganglioneuroblastoma nodular; MKC: *mitotic and karyorrhectic cells*; MKI: mitosis-karyorrhexis index; Asterisco único: avaliação de prognóstico de GNBn; Asterisco duplo: MKC = 2%, 100 de 5000 células, 4%, 200 de 5000 células. Fonte: (PEUCHMAUR et al., 2003).

Uma predição acurada sobre o curso da doença surge como pré-requisito fundamental para a estimativa de risco e planejamento de terapias individuais, combinando fatores clínicos e moleculares relacionados ao estágio do tumor, idade do paciente no momento do diagnóstico e amplificação genômica do status do proto-oncogene *MYCN* (ZHANG et al., 2015). Identificado em 1983, *MYCN*, membro da família *MYC*, é encontrado em aproximadamente 25% dos casos e está relacionado ao alto risco da doença e a um desfecho desfavorável para o paciente. Dessa forma, quando amplificado (superexpresso), este gene surge como o melhor marcador genético para o prognóstico do neuroblastoma (HUANG; WEISS, 2013).

Sendo de origem familiar ou esporádica, anormalidades cromossomais contribuem para o

desenvolvimento do neuroblastoma, ocasionando desregulação dos microRNAs (miRNAs). Os miRNAs são pequenas moléculas de RNA fita simples não-codificantes mas que agem diretamente na regulação da expressão gênica. Algumas atividades relacionadas ao silenciamento gênico controladas por miRNAs podem afetar diversas funções intracelulares, incluindo uma possível diferenciação de células troncos neurais. A participação dessas pequenas moléculas em diversos processos biológicos são, reconhecidamente, aspectos intimamente ligados à genética do câncer, permitindo, assim, que os miRNAs sejam empregados como biomarcadores associados ao prognóstico e caracterização de diferentes modelos tumorais (ZAMMIT; BARON; AYERS, 2018).

Moléculas de RNA longo não codificante (lncRNA) também podem ser utilizadas como biomarcadores. Ao sequenciarem transcriptomas de neuroblastoma de alto e baixo risco, *Pandey* e colaboradores (PANDEY et al., 2014) detectaram lncRNAs diferencialmente expressos. Eles identificaram o *NBAT-1* (*neuroblastoma associated transcript-1*) como um biomarcador capaz de predizer satisfatoriamente o comportamento clínico do neuroblastoma quanto à sua agressividade. Um nível reduzido de *NBAT-1* foi relacionado com o aumento da proliferação celular e invasão, além de afetar a diferenciação neuronal através da ativação do fatores de transcrição específicos.

Perfis de expressão gênica baseados em microarranjo também têm sido amplamente utilizados em pesquisas sobre câncer a fim de identificar biomarcadores que permitam prever um desfecho clínico, como diagnóstico, prognóstico e previsão de resposta a um dado tratamento (BEER et al., 2002; GLAS et al., 2005; GLINSKY et al., 2004). Com o surgimento das tecnologias de sequenciamento NGS (*Next Generation Sequencing*), a análise de transcriptomas eucarióticos sofreu uma revolução. Entre as principais técnicas aplicadas por meio de tais tecnologias está o RNA-seq, capaz de buscar por padrões de expressão gênica globais além de possibilitar a descoberta de novos genes, variantes alternativas, transcritos quiméricos e expressão de alelos específicos (SU et al., 2014; FERREIRA et al., 2013; OZSOLAK; MILOS, 2011). Vale reforçar, ainda, que sua versatilidade e robustez permitem que seja usado para o desenvolvimento de modelos preditivos baseados em expressão gênica aplicado ao estudo de câncer (VOLINIA; CROCE, 2013). Em um trabalho de *Zhang* e colaboradores (ZHANG et al., 2015), foram comparados modelos de predição clínica baseados em RNA-seq e microarranjo aplicados a neuroblastoma. A partir dos perfis de expressão gênica de 498 pacientes com a doença, demostrou-se que o RNA-seq supera o microarranjo quanto a determinação de características transcriptômicas do câncer. No entanto, referente à predição do seu desfecho clínico, ambas as técnicas apresentaram desempenho similares.

Várias abordagens para o estudo de neoplasias pediátricas foram adaptadas de pesquisas oncológicas relacionadas a adultos. Entretanto, existem diferenças entre as que afetam crianças e aquelas que atingem adultos. Um modelo de análise comum envolve a identificação de alterações genéticas, investigando seus papéis quanto ao surgimento do tumor, progressão e resistência a ação de determinadas drogas (SALAZAR et al., 2016). Porém, grande parte dos cânceres infantis apresentam uma taxa de mutações recorrentes muito baixa em relação aos carcinomas em adultos, necessitando de novas perspectivas para melhor compreender vulnerabilidades e etiologias específicas (ADAMSON et al., 2014; LAWRENCE et al., 2013). É preciso, também, destacar o fato de que, ao contrário das doenças que atingem adultos, a maior parte dos tumores que aflige crianças são raros, limitando significativamente o financiamento de pesquisas relacionadas a eles (ADAMSON et al., 2014).

A complexidade no desenvolvimento do neuroblastoma é muito grande. Consórcios de pesquisadores, ao longo dos anos, se dedicaram em determinar os diferentes estágios e características da doença. Para tanto, o uso das novas tecnologias de sequenciamento teve papel fundamental. Isso, contudo, fez com que surgissem gigantescos conjuntos de dados biológicos. Dessa forma, cada vez mais, a exigência por métodos estatísticos e computacionais, em especial aqueles abrangidos pelo conceito de *big data*, se tornou essencial para a condução de uma análise de dados eficaz (SALAZAR et al., 2016).

1.3 Sequenciamento NGS e RNA-seq

Sequenciamento paralelo massivo, também conhecido como *Next-generation sequencing* (NGS), é uma maneira eficaz para a obtenção de informações genômicas sobre o câncer. Grande parte das tecnologias ligadas ao NGS concentram-se no sequenciamento por síntese. Os fragmentos de DNA são ligados a uma *array* e através de uma enzima chamada DNA polimerase é feita a adição sequencial de nucleotídeos marcados. Em seguida, uma câmera de alta resolução faz a captura dos sinais emitidos por cada nucleotídeo associando-o à coordenadas de tempo e espaço. Cada sequência, então, é inferida por um *software*, gerando uma sequência de DNA contínua chamada de *read* (GAGAN; ALLEN, 2015).

Existe uma gama de diferentes plataformas de sequenciamento NGS. O que há de comum entre elas é o poder de sequenciar milhões de fragmentos de DNA paralelamente (BEHJATI; TARPEY, 2013). O sequenciamento NGS, consiste no agrupamento de vários métodos como preparação do modelo, sequenciamento, obtenção de imagem e análise dos dados. A combinação de protocolos específicos é que diferencia uma tecnologia da outra e determina o

formato dos dados gerados por cada plataforma (METZKER, 2010).

Na última década, nos deparamos com um estado dinâmico dos genomas. Mudanças quantitativas e qualitativas de RNA passaram a ser caracterizadas em espécies que vão de organismos modelos a humanos (OZSOLAK; MILOS, 2011). Neste contexto, surge a técnica de sequenciamento de RNA (RNA-seq), cujo protocolo pode ser visto na Figura 1.4.

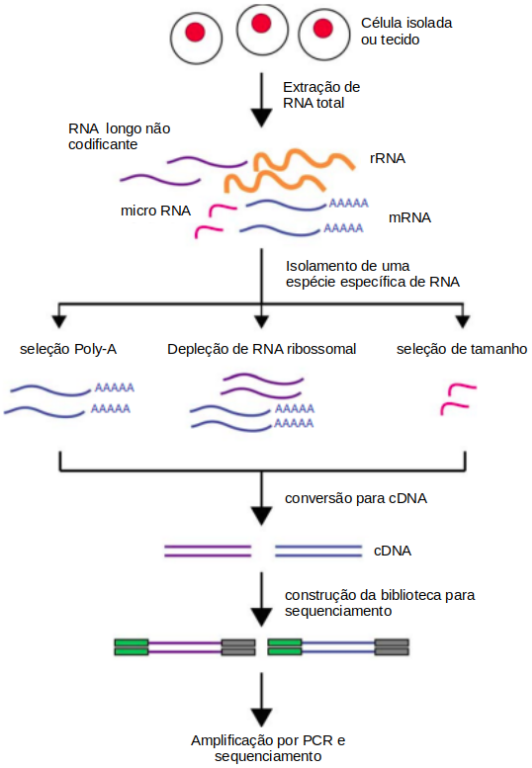

Figura 1.4: Visão global do protocolo de RNA-seq. Inicialmente, ocorre a extração de RNA do material biológico desejado (células únicas ou amostras de tecidos). Em seguida, moléculas de RNA são isoladas por meio de protocolos específicos, como o protocolo de seleção *poly-A*, para enriquecer para transcritos poliadenilados, ou o protocolo para remoção de RNA ribossomal. Logo após, o RNA é convertido em DNA complementar (cDNA) por meio da enzima transcriptase reversa e adaptadores são ligados às extremidades dos fragmentos de cDNA. Por fim, ocorre a amplificação via PCR e, então, a biblioteca está pronta para o sequenciamento. Fonte: Figura adaptada de (KUKURBA; MONTGOMERY, 2015).

O RNA-seq é uma técnica que permite a avaliar a expressão gênica, bem como descobrir novos transcritos, através do sequenciamento de RNA. Primeiramente, seleciona-se a amostra desejada, podendo ser célula única ou parte de um dado tecido. Em seguida, por meio de protocolos específicos, isola-se o RNA e, com o auxilio da enzima transcriptase reversa, converte-se o RNA em cDNA, pois a molécula de RNA é instável e não consegue ser sequenciada em seu estado original. Os fragmentos de cDNA são marcados por meio de sequências específicas, chamadas de adaptadores e, com isso, a biblioteca está pronta para ser sequenciada (KUKURBA; MONTGOMERY, 2015).

Comparado à técnicas anteriores, como *Sanger* e microarranjo, o RNA-seq proporciona maiores cobertura e resolução do transcriptoma. Além de quantificar a expressão gênica, os dados obtidos pela técnica facilitam a descoberta de novos transcritos, identificação de genes com *splicing* alternativo e detecção da expressão de alelos específicos (KUKURBA; MONTGOMERY, 2015). Além disso, a quantidade de dados gerada é muito grande, exigindo técnicas computacionais e estatísticas para sua manipulação e compreensão.

1.4 Aprendizado de Máquina

Com avanços significativos quanto à memória, poder de processamento e capacidade de armazenamento, a importância da computação para a biologia subiu de patamar. Diferentes técnicas passaram a ser implementadas, trazendo novas perspectivas à análise de dados e resolução de problemas antes consideradas distantes. Entre os principais métodos desenvolvidos, destaca-se o aprendizado de máquina (DEO, 2015).

Para iniciarmos nossa discussão sobre aprendizado de máquina, é preciso salientar que ele não é a mesma coisa que inteligência artificial (IA). A IA é um conceito muito amplo e inclui o aprendizado de máquina como um de seus recursos (TECNOBLOG, 2020), o que pode ser observado a partir da Figura 1.5.

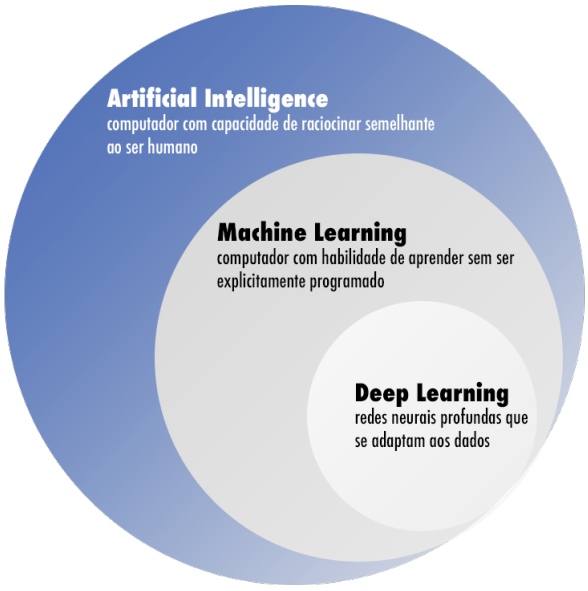

Figura 1.5: Diferença entre aprendizado de máquina e inteligência artificial. Fonte: (CAUDURO, 2020).

Inteligência artificial consiste em mecanismos computacionais que, para resolver um problema, se baseiam no comportamento humano. Eles são capazes de buscar por padrões ou

tendências em um dado conjunto de dados, fazer análises mais apuradas e, então, tirar conclusões para a tomada de decisões. *Aprendizado de máquina* implica em um sistema capaz de modificar o seu comportamento de forma autônoma baseado na sua própria experiência e com o mínimo de interação humana. Ele possui um conjunto de regras geradas a partir do reconhecimento de padrões encontrados dentro dos dados analisados.

A implementação de um sistema de aprendizado de máquina pode ser conduzido de diferentes formas. Qualquer projeto de implantação de software passa por fases como estudo de viabilidade, desenvolvimento, teste e validação/homologação. Somente após todo este processo ele poderá ser disponibilizado para a resolução dos problemas que demandaram, inicialmente, a sua construção. Levando-se em conta que o problema já tenha sido definido, é preciso coletar os dados das mais variadas fontes, explorá-los de forma que possam ser limpos e transformados para, só então, aplicarmos um determinado algoritmo de aprendizado, treinando-o e testando-o em linguagens de programação como *R*, *Python* ou *Java*, por exemplo (SCIENCE, 2020). Entretanto, existem diferentes maneiras de implementar um modelo e que pode ser categorizado ao longo de duas dimensões, como mostra a Figura 1.6.

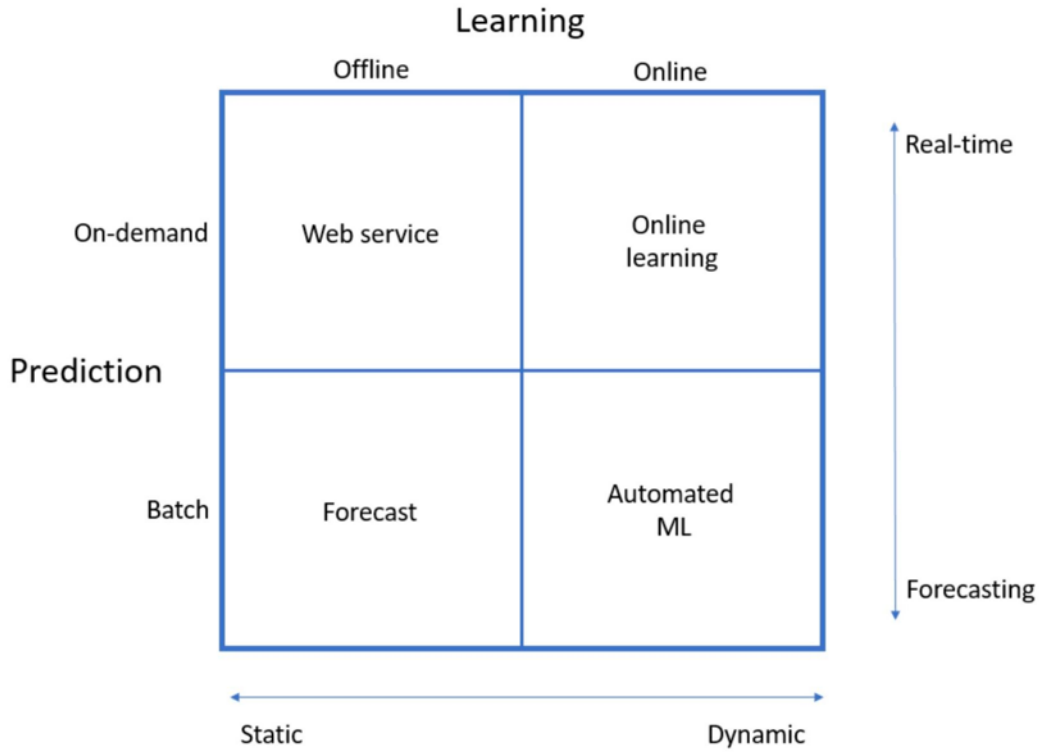

Figura 1.6: Matriz bidimensional exemplificando a implementação de um sistema de aprendizado de máquina. Fonte: (SCIENCE, 2020).

Temos quatro formas de implementar um modelo de aprendizado de máquina: *Web Service*, *Forecast*, *Online Learning* e *Automated Machine Learning*, no entanto, existem apenas

duas formas de treinamento: *offline* e *online*. No primeiro, o modelo, inicialmente, utiliza dados históricos como fonte de treinamento e, depois de implementado, permanece constante. Já no segundo, o modelo é constantemente atualizado ao passo que novos dados vão sendo inseridos. Isto é útil, em especial, para séries temporais que utilizam dados da área de finanças ou mesmo de sensores, por exemplo. Além disso, vale frisar as diferentes formas que os algoritmos fazem previsões: *batch prediction* (previsões em lote) e *on-demand prediction* (previsões sob demanda). Nas previsões em lote, o algoritmo cria uma tabela de previsões com base nos seus respectivos dados de entrada, sendo indicado para dados que não dependem do tempo ou cujas saídas não precisem ser a todo momento atualizadas. Quanto às previsões sob demanda, as previsões são feitas em tempo real e utilizam apenas os dados de entrada que estão disponíveis no momento da solicitação (SCIENCE, 2020).

Podemos, ainda, dividir o aprendizado de máquina em duas grandes áreas: *supervisionado* e *não-supervisionado*. No primeiro, temos variáveis de *input* (X) e *output* (Y) e é utilizado um algoritmo que aprende uma função de mapeamento $Y = f(X)$ a partir do *input* e *output*. O objetivo é aproximar esta função de forma que seja possível predizer de maneira acurada as variáveis de *output Y* a partir de novos dados de *input* (x). Didaticamente, chama-se aprendizado supervisionado porque o processo de aprendizado de um algoritmo partindo de um conjunto de dados de treinamento pode ser entendido como um professor supervisionando o aprendizado de seus alunos. Nós sabemos quais são as respostas corretas, o algoritmo faz as predições utilizando os dados de treinamento e então o professor as corrige. O aprendizado só termina quando é atingido um nível de performance de predições satisfatório. *Classificação* e *Regressão* são exemplos de técnicas deste tipo. No modelo não-supervisionado, temos apenas os dados de *input* (X) sem a presença de variáveis de *output* correspondentes. Seu objetivo é modelar uma estrutura subjacente ou determinada distribuição presente nos dados com o propósito de aprender mais sobre estes. Também de um jeito mais didático, ao contrário dos modelos supervisionados, não há uma resposta correta e não há um professor. Os algoritmos são deixados por conta própria para que consigam encontrar estruturas interessantes presentes nos dados. Exemplos de algoritmos desse tipo são a *Clusterização* e a *Associação* (BROWNLEE, 2020).

Ressaltamos, entretanto, a existência de uma área mista: o aprendizado de máquina semi-supervisionado. Pertencem a esta classe problemas em que se dispõe de uma quantidade grande de dados de *entrada* (X) e apenas alguns destes dados apresentam suas respectivas *saídas*. Muitos dos problemas de aprendizado reais são mistos, uma vez que pode ser caro ou consumir muito tempo para atribuir uma saída. Entradas sem saídas são mais fáceis e baratas de serem coletadas e armazenadas. Para resolver problemas desse tipo, é possível utilizar técnicas presentes nos aprendizados supervisionados e não-supervisionados. Podem ser utilizados tanto

algoritmos que aprendam a partir das saídas presentes e então façam as respectivas predições ou, então, utilizar algoritmos que irão fazer descobertas a partir da estrutura dos dados existentes (BROWNLEE, 2020).

Diversas áreas científicas possuem, em comum, a necessidade de modelar a relação entre um conjunto de valores observáveis (entradas) e um grupo de variáveis relacionadas a tais valores (saídas). Trata-se de um modelo matemático que permite predizer o valor de variáveis desejadas com base na observação das entradas. O aprendizado de máquina disponibiliza um conjunto de técnicas que tornam possível a construção de um modelo computacional dessa relação *entrada*/*saída*. O processo de construção deste modelo passa por uma técnica de validação cruzada, segue para uma etapa de treinamento e por fim é testado em um conjunto de dados que ainda não foi utilizado, de forma que resultados mais precisos e confiáveis sejam alcançados. Um modelo treinado pode gerar novas perspectivas sobre como os valores de entrada são transformados em saídas, podendo ser utilizado para realizar predições a partir de um novo conjunto de dados distintos dos dados de treinamento. Os algoritmos de aprendizado requerem grandes quantidades de dados de treinamento. Contudo, tais dados precisam ser exemplos representativos, de forma que a saída identificada seja mais consistente (BASTANLAR; OZUYSAL, 2014).

5 *Conclusão*

Em lingugem R, partindo de uma metanálise com base no índice de sobrevivência dos pacientes (vivo ou óbito) por meio do método de *Stouffer*, fomos capazes de selecionar 6 grupos gênicos (*features*) para que pudessem ser testados como possíveis assinaturas transcricionais em um modelo de *machine learning* supervisionado embasado na técnica de *regressão logística*. Cada grupo foi exaustivamente testado. Utilizamos dados normalizados, 2 algoritmos de balanceamento amostral (*MWMOTE* e *RWO*), dois modelos de regressão (*GLM* e *LogitBoost*) e duas formas de distribuição dos dados para treinamento/teste (60%/40% e 70%/30%), resultando na construção de 48 modelos.

Os dados apresentavam *labels* com quantidades de pacientes desiguais - o número de vivos era superior ao de mortos. Para tanto, fizemos um balanceamento utilizando os algoritmos *MWMOTE* e *RWO*, o que gerou dados sintéticos a partir do conjunto de expressão pertencente ao *label* minoritário e igualou o número de vivos e mortos. Após a realização de um procedimento de validação cruzada e aplicação dos modelos de regressão *GLM* e *LogitBoost*, encontramos percentuais de acurácia, sensibilidade e especificidade elevados em grande parte dos grupos, além de p-valores extremamente baixos para a validação da acurácia.

Dentre os melhores resultados encontrados, podemos citar o Grupo I, contendo 963 *features*. Verificamos 92,31% de acurácia, 87,18% de sensibilidade, 97,44% de especificidade e p-valor de acurácia de $1{,}4 \times 10^{-44}$ para uma distribuição 70% treinamento e 30% teste, algoritmo de balanceamento *RWO* e modelo de regressão *LogitBoost*.

Realizamos, ainda, um enriquecimento funcional com o *Gene Ontology* em termos de *processos biológicos* de cada um dos grupos de *features*, entretanto, os grupos V e VI não apresentaram nenhuma ontologia significativa quando consideramos um valor de *cutoff* de 10^{-3}. Dando maior atenção ao Grupo I, encontramos diversas ontologias relacionadas à divisão celular sendo a GO:0010720 (regulação positiva do desenvolvimento celular), GO:0000280 (divisão nuclear) e GO:0022008 (neurogênesis).

Quanto aos modelos construídos em *Python*, implementamos os métodos de regressão

logística, redes neurais artificiais com *multilayer perceptron* e árvore de decisão. Utilizamos, em complemento à metanálise, *feature selection*, *univariate selection* e expressão diferencial para a seleção das *features*. Os conjuntos de treinamento foram balanceados com os métodos *oversampling* (SWMOTE) e *undersampling* (amostragem com redução do *label* majoritário). Referente às variáveis testadas, além de sobrevivência, consideramos progressão do tumor, alto risco do tumor e classe tumoral. No total, criamos 1180 modelos.

Aplicamos, ainda, um método de aprendizado não supervisionado com *k-means*, entretanto, não conseguimos identificar diferentes *labels* além dos que já existiam. O melhor modelo encontrado foi aquele com 2040 *features* selecionadas a partir do método univariado e com balanceamento do tipo *oversampling*, apresentando acurácia de 91,33% e área sob a curva ROC de 94%. Conduzimos, ainda, uma análise de enriquecimento funcional com base em processos biológicos do *Gene Ontology* e bancos de dados de doenças - este, através do pacote DOSE da linguagem R. Podemos destacar a presença dos grupos *response to xenobiotic stimulus* (GO:0009410), *regulation of glial cell differentiation* (GO:0045685) e *mononuclear cell proliferation* (GO:0032943), processos normalmente associados a um tumor com prognóstico ruim. Quanto às ontologias de doenças, destacamos a ontologia C4725671, *High-Risk Neuroblastoma*, com p-valor 0.016 e 8 transcritos anotados, encontrada por meio do banco de dados *Disease Gene Network*.

Dessa forma, a partir dos resultados obtidos com a criação de 48 modelos em R e 1180 em *Python* (ao todo, 1228), vemos que o aprendizado de *máquina supervisionado* se mostrou uma ferramenta eficaz no processo de predição de *labels* de variáveis associadas ao neuroblastoma. Embora tenhamos encontrado diversos modelos com bom desempenho preditivo, os melhores resultados foram obtidos com regressão logística aplicada à sobrevivência do paciente, tanto em R como em *Python*, alcançando percentuais de acurácia acima de 90%. Entretanto, para elaborarmos um modelo mais confiável e que, eventualmente, possa ser empregado na rotina clínica como ferramenta para auxiliar no prognóstico de pacientes, precisamos ampliar o espaço amostral, testar novas formas de balanceamento e seleção de *features* e conduzir uma análise biológica mais aprofundada acerca dos conjuntos de *features* utilizados.

As técnicas de aprendizado de máquina podem ser aplicadas sobre conjuntos de dados de diferentes áreas do conhecimento humano. Na biologia, em especial quando tratamos de assuntos relacionados à saúde, modelos de aprendizado podem ser implementados para auxiliar a rotina clínica no que diz respeito ao diagnóstico e prognóstico de doenças. O neuroblastoma, por se tratar de um tumor altamente preditivo, permite o uso do aprendizado de máquina na busca por assinaturas transcricionais capazes de contribuir para o planejamento de tratamentos

personalizados e com maior eficácia.

Referências Bibliográficas

ADAMSON, P. C. et al. Drug discovery in paediatric oncology: roadblocks to progress. *Nature Reviews Clinical Oncology*, Nature Publishing Group, v. 11, n. 12, p. 732, 2014.

ALEXA, A.; RAHNENFÜHRER, J. Gene set enrichment analysis with topgo. *Bioconductor Improv*, v. 27, 2009.

AMBROS, I.; AMBROS, P. Schwann cells in neuroblastoma. *European Journal of Cancer*, Elsevier, v. 31, n. 4, p. 429–434, 1995.

AMBROS, I. M. et al. Role of ploidy, chromosome 1p, and schwann cells in the maturation of neuroblastoma. *New England Journal of Medicine*, Mass Medical Soc, v. 334, n. 23, p. 1505–1511, 1996.

BADR, W. *Having an imbalanced dataset? Here is how you can fix it.* 2019. `https://towardsdatascience.com/having-an-imbalanced-dataset-here-is-how-you-can-solve-it-1640568947eb`. Acessado em: 20-08-2022.

BAELDUNG. *Multiclass Classification Using Support Vector Machines.* 2021. `https://www.baeldung.com/cs/svm-multiclass-classification#:~:text=In%20its%20most%20simple%20type,into%20multiple%20binary%20classification%20problems.` Acessado em: 19-09-2022.

BALI, R. et al. *R: Unleash Machine Learning Techniques*. [S.l.]: Packt Publishing Ltd, 2016.

BAO, L. et al. Boosted near-miss under-sampling on svm ensembles for concept detection in large-scale imbalanced datasets. *Neurocomputing*, Elsevier, v. 172, p. 198–206, 2016.

BARUA, S. et al. Mwmote–majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 26, n. 2, p. 405–425, 2012.

BASTANLAR, Y.; OZUYSAL, M. mirnomics: Microrna biology and computational analysis. Springer, v. 1107, p. 105–128, 2014.

BATISTA, G. E.; PRATI, R. C.; MONARD, M. C. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, ACM New York, NY, USA, v. 6, n. 1, p. 20–29, 2004.

BEER, D. G. et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature medicine*, Nature Publishing Group, v. 8, n. 8, p. 816, 2002.

BEHJATI, S.; TARPEY, P. S. What is next generation sequencing? *Archives of Disease in Childhood-Education and Practice*, Royal College of Paediatrics and Child Health, v. 98, n. 6, p. 236–238, 2013.

BERWANGER, B. et al. Loss of a fyn-regulated differentiation and growth arrest pathway in advanced stage neuroblastoma. *Cancer cell*, Elsevier, v. 2, n. 5, p. 377–386, 2002.

BISONG, E. Introduction to scikit-learn. In: *Building machine learning and deep learning models on Google cloud platform*. [S.l.]: Springer, 2019. p. 215–229.

BLOOMFIELD, P.; STEIGER, W. L. *Least absolute deviations: theory, applications, and algorithms*. [S.l.]: Springer, 1983.

BOTTENSTEIN, J. E.; SATO, G. H. Growth of a rat neuroblastoma cell line in serum-free supplemented medium. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 76, n. 1, p. 514–517, 1979.

BRAY, F. et al. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, Wiley Online Library, v. 68, n. 6, p. 394–424, 2018.

BROWNLEE, J. *Supervised and Unsupervised Machine Learning Algorithms.* 2020. `https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/`. Acessado em: 16-08-2020.

CANGELOSI, D. et al. Logic learning machine creates explicit and stable rules stratifying neuroblastoma patients. *BMC bioinformatics*, Springer, v. 14, n. 7, p. 1–20, 2013.

CARLSON, M. et al. org. hs. eg. db: Genome wide annotation for human. *R package version*, v. 3, n. 2, p. 3, 2019.

CASTLEBERRY, R. P. European journal of cancer. Elsevier Science, v. 33, n. 9, p. 1430–1438, 1997.

CAUDURO, M. *Deep Learning: o motor dos negócios na era da inteligência artificial.* 2020. `https://medium.com/huia/inteligencia-artificial-uma-corrida-desleal-80bfa53075ed`. Acessado em: 17-08-2020.

CHATTEN, J. et al. Prognostic value of histopathology in advanced neuroblastoma: a report from the childrens cancer study group. *Human pathology*, Elsevier, v. 19, n. 10, p. 1187–1197, 1988.

CHAWLA, N. V. et al. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, v. 16, p. 321–357, 2002.

CONSORTIUM, G. O. The gene ontology (go) database and informatics resource. *Nucleic acids research*, Oxford University Press, v. 32, n. suppl_1, p. D258–D261, 2004.

COOPER, H. M. Statistically combining independent studies: A meta-analysis of sex differences in conformity research. *Journal of personality and social psychology*, American Psychological Association, v. 37, n. 1, p. 131, 1979.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine learning*, Springer, v. 20, n. 3, p. 273–297, 1995.

D'ANTONIO, M. et al. Network of cancer genes (ncg 3.0): integration and analysis of genetic and network properties of cancer genes. *Nucleic acids research*, Oxford University Press, v. 40, n. D1, p. D978–D983, 2012.

DAVIDOFF, A. M. Seminars in pediatric surgery. Elsevi, v. 21, p. 2–14, Fevereiro 2012.

DEO, R. C. Machine learning in medicine. *Circulation*, Am Heart Assoc, v. 132, n. 20, p. 1920–1930, 2015.

FENDLER, W. P. et al. High 123 i-mibg uptake in neuroblastic tumours indicates unfavourable histopathology. *European journal of nuclear medicine and molecular imaging*, Springer, v. 40, n. 11, p. 1701–1710, 2013.

FERREIRA, P. G. et al. Transcriptome characterization by rna sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic leukemia. *Genome research*, Cold Spring Harbor Lab, p. gr–152132, 2013.

FRIEDMAN, J. et al. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, Institute of Mathematical Statistics, v. 28, n. 2, p. 337–407, 2000.

GAGAN, J.; ALLEN, E. M. V. Next-generation sequencing to guide cancer therapy. *Genome medicine*, BioMed Central, v. 7, n. 1, p. 1–10, 2015.

GLAS, A. M. et al. Gene expression profiling in follicular lymphoma to assess clinical aggressiveness and to guide the choice of treatment. *Blood*, Am Soc Hematology, v. 105, n. 1, p. 301–307, 2005.

GLINSKY, G. V. et al. Gene expression profiling predicts clinical outcome of prostate cancer. *The Journal of clinical investigation*, Am Soc Clin Investig, v. 113, n. 6, p. 913–923, 2004.

GOH, K.-I. et al. The human disease network. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 104, n. 21, p. 8685–8690, 2007.

GRIFFITHS, A. et al. Introdução à genética. 10a edição. *Edditora Guanabara, editor*, p. 503–506, 2013.

HAHSLER, M.; PIEKENBROCK, M.; DORAN, D. dbscan: Fast density-based clustering with r. *Journal of Statistical Software*, v. 91, p. 1–30, 2019.

HALL, P.; GILL, N. *An introduction to machine learning interpretability*. [S.l.]: O'Reilly Media, Incorporated, 2019.

HART, P. The condensed nearest neighbor rule (corresp.). *IEEE transactions on information theory*, Citeseer, v. 14, n. 3, p. 515–516, 1968.

HECK, J. E. et al. The epidemiology of neuroblastoma: a review. *Paediatric and perinatal epidemiology*, Wiley Online Library, v. 23, n. 2, p. 125–143, 2009.

HUANG, M.; WEISS, W. A. Neuroblastoma and mycn. *Cold Spring Harbor perspectives in medicine*, Cold Spring Harbor Laboratory Press, v. 3, n. 10, p. a014415, 2013.

IBM. *What is a decision tree?* 2022. `https://www.ibm.com/topics/decision-trees`. Acessado em: 19-08-2022.

INCA. *Incidencia de cancer no Brasil - Estimativa 2020.* 2020. https://www.inca.gov.br/sites/ufu.sti.inca.local/files/media/document/estimativa-2020-incidencia-de-cancer-no-brasil.pdf. Acessado em: 15-08-2020.

KOROBOV, M.; LOPUHIN, K. *Eli5*. 2020. https://eli5.readthedocs.io/en/latest/. Acessado em: 20-08-2022.

KUHN, M. The caret package. *R Foundation for Statistical Computing, Vienna, Austria. URL https://cran. r-project. org/package= caret*, Citeseer, 2012.

KUKURBA, K. R.; MONTGOMERY, S. B. Rna sequencing and analysis. *Cold Spring Harbor Protocols*, Cold Spring Harbor Laboratory Press, v. 2015, n. 11, p. pdb–top084970, 2015.

LANGFELDER, P.; HORVATH, S. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, Springer, v. 9, n. 1, p. 559, 2008.

LAWRENCE, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, Nature Publishing Group, v. 499, n. 7457, p. 214, 2013.

METZKER, M. L. Sequencing technologies. *Nature reviews genetics*, Nature Publishing Group, v. 11, n. 1, p. 31–46, 2010.

MEYER, D.; WIEN, F. Support vector machines. *The Interface to libsvm in package e1071*, v. 28, p. 20, 2015.

MICHALOS, A. C. *Encyclopedia of quality of life and well-being research*. [S.l.]: Springer Netherlands Dordrecht, 2014. 302-308 p.

MOHAMMED, R.; RAWASHDEH, J.; ABDULLAH, M. Machine learning with oversampling and undersampling techniques: overview study and experimental results. In: IEEE. *2020 11th international conference on information and communication systems (ICICS)*. [S.l.], 2020. p. 243–248.

MUSA, J. et al. Mybl2 (b-myb): a central regulator of cell proliferation, cell survival and differentiation involved in tumorigenesis. *Cell death & disease*, Nature Publishing Group, v. 8, n. 6, p. e2895–e2895, 2017.

NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, Wiley Online Library, v. 135, n. 3, p. 370–384, 1972.

NICK, T. G.; CAMPBELL, K. M. Logistic regression. *Topics in biostatistics*, Springer, p. 273–301, 2007.

NIH. *What is Cancer?*. 2020. https://www.cancer.gov/about-cancer/understanding/what-is-cancer. Acessado em: 15-08-2020.

NORIEGA, L. Multilayer perceptron tutorial. *School of Computing. Staffordshire University*, Citeseer, 2005.

OZSOLAK, F.; MILOS, P. M. Rna sequencing: advances, challenges and opportunities. *Nature reviews genetics*, Nature Publishing Group, v. 12, n. 2, p. 87–98, 2011.

PANDEY, G. K. et al. The risk-associated long noncoding rna nbat-1 controls neuroblastoma progression by regulating cell proliferation and neuronal differentiation. *Cancer cell*, Elsevier, v. 26, n. 5, p. 722–737, 2014.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

PEIXOTO, F. *A simple overview of Multilayer Perceptron (MLP).* [S.l.]: Analitics Vidhya, 2020. https://www.analyticsvidhya.com/blog/2020/12/mlp-multilayer-perceptron-simple-overview/#. Acessado em: 19-08-2022.

PEUCHMAUR, M. et al. Revision of the international neuroblastoma pathology classification: confirmation of favorable and unfavorable prognostic subsets in ganglioneuroblastoma, nodular. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, Wiley Online Library, v. 98, n. 10, p. 2274–2281, 2003.

RILEY, R. D. et al. A systematic review of molecular and biological tumor markers in neuroblastoma. *Clinical Cancer Research*, AACR, v. 10, n. 1, p. 4–12, 2004.

ROBINSON, M. D.; MCCARTHY, D. J.; SMYTH, G. K. edger: a bioconductor package for differential expression analysis of digital gene expression data. *bioinformatics*, Oxford University Press, v. 26, n. 1, p. 139–140, 2010.

SALAZAR, B. M. et al. Neuroblastoma, a paradigm for big data science in pediatric oncology. *International journal of molecular sciences*, Multidisciplinary Digital Publishing Institute, v. 18, n. 1, p. 37, 2016.

SCHRIML, L. M. et al. Disease ontology: a backbone for disease semantic integration. *Nucleic acids research*, Oxford University Press, v. 40, n. D1, p. D940–D946, 2012.

SCHULTE, J. H. et al. Deep sequencing reveals differential expression of microrna s in favorable versus unfavorable neuroblastoma. *Nucleic acids research*, Oxford University Press, v. 38, n. 17, p. 5919–5928, 2010.

SCIENCE. *Data Science Academy: Como publicar um modelo de machine learning em producao?*. 2020. http://datascienceacademy.com.br/blog/como-publicar-um-modelo-de-machine-learning-em-producao/. Acessado em: 17-08-2020.

SCIKIT-LEARN. *Feature selection using SelectFromModel.* 2022. https://scikit-learn.org/stable/modules/feature_selection.html#select-from-model. Acessado em: 19-09-2022.

SHIMADA, H. et al. Histopathologic prognostic factors in neuroblastic tumors: definition of subtypes of ganglioneuroblastoma and an age-linked classification of neuroblastomas. *JNCI: Journal of the National Cancer Institute*, Oxford University Press, v. 73, n. 2, p. 405–416, 1984.

SHIMADA, H. et al. International neuroblastoma pathology classification for prognostic evaluation of patients with peripheral neuroblastic tumors: a report from the children's cancer group. *Cancer*, Wiley Online Library, v. 92, n. 9, p. 2451–2461, 2001.

SINAGA, K. P.; YANG, M.-S. Unsupervised k-means clustering algorithm. *IEEE access*, IEEE, v. 8, p. 80716–80727, 2020.

SMYTH, G. K. Limma: linear models for microarray data. In: *Bioinformatics and computational biology solutions using R and Bioconductor*. [S.l.]: Springer, 2005. p. 397–420.

SPICEWORKS. *What is Logistic Regression? Equation, Assumptions, Types and Best Practices.* 2022. `https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/`. Acessado em: 19-09-2022.

SPIRONELLO, R. A. et al. Mortalidade infantil por câncer no brasil. *Saúde e Pesquisa*, v. 13, n. 1, 2020.

SPRENT, P. Fisher exact test. In: *International encyclopedia of statistical science*. [S.l.]: Springer, 2011. p. 524–525.

STEVENS, E. A.; MEZRICH, J. D.; BRADFIELD, C. A. The aryl hydrocarbon receptor: a perspective on potential roles in the immune system. *Immunology*, Wiley Online Library, v. 127, n. 3, p. 299–311, 2009.

STOUFFER, S. A. et al. The american soldier: Adjustment during army life.(studies in social psychology in world war ii), vol. 1. Princeton Univ. Press, 1949.

SU, Z. et al. A comprehensive assessment of rna-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nature biotechnology*, Nature Publishing Group, v. 32, n. 9, p. 903, 2014.

TECNOBLOG. *Machine learning: o que é e por que é tão importante.* 2020. `https://tecnoblog.net/247820/machine-learning-ia-o-que-e/`. Acessado em: 17-08-2020.

TUSZYNSKI, J.; TUSZYNSKI, M. J. The catools package. *R package version*, p. 1–8, 2007.

VOLINIA, S.; CROCE, C. M. Prognostic microrna/mrna signature from the integrated analysis of patients with invasive breast cancer. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 110, n. 18, p. 7413–7417, 2013.

WATTERS, J. J. et al. Rapid membrane effects of steroids in neuroblastoma cells: effects of estrogen on mitogen activated protein kinase signalling cascade and c-fos immediate early gene transcription. *Endocrinology*, Oxford University Press, v. 138, n. 9, p. 4030–4033, 1997.

WILLIAMS, A. *A beginners practical guide: apply k-means clustering algorithm*. 2020. `https://levelup.gitconnected.com/a-beginners-practical-guide-apply-k-means-clustering-algorithm-d57295699e61`. Acessado em: 20-08-2022.

WU, P.-Y. et al. Novel endogenous ligands of aryl hydrocarbon receptor mediate neural development and differentiation of neuroblastoma. *ACS Chemical Neuroscience*, ACS Publications, v. 10, n. 9, p. 4031–4042, 2019.

WU, S. et al. $l1$-norm batch normalization for efficient training of deep neural networks. *IEEE transactions on neural networks and learning systems*, IEEE, v. 30, n. 7, p. 2043–2051, 2018.

WYTHOFF, B. J. Backpropagation neural networks: a tutorial. *Chemometrics and Intelligent Laboratory Systems*, Elsevier, v. 18, n. 2, p. 115–155, 1993.

YOON, K. J.; DANKS, M. K. Cell adhesion molecules as targets for therapy of neuroblastoma. *Cancer biology & therapy*, Taylor & Francis, v. 8, n. 4, p. 306–311, 2009.

YOUNG-JR, J. L. et al. Cancer incidence, survival, and mortality for children younger than age 15 years. *Cancer*, Wiley Online Library, v. 58, n. S2 S2, p. 598–602, 1986.

YU, G. et al. Dose: an r/bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*, Oxford University Press, v. 31, n. 4, p. 608–609, 2015.

ZAMMIT, V.; BARON, B.; AYERS, D. Mirna influences in neuroblast modulation: An introspective analysis. *Genes*, Multidisciplinary Digital Publishing Institute, v. 9, n. 1, p. 26, 2018.

ZHANG, H.; LI, M. Rwo-sampling: A random walk over-sampling approach to imbalanced data classification. *Information Fusion*, Elsevier, v. 20, p. 99–116, 2014.

ZHANG, T.; RAMAKRISHNAN, R.; LIVNY, M. Birch: A new data clustering algorithm and its applications. *Data mining and knowledge discovery*, Springer, v. 1, n. 2, p. 141–182, 1997.

ZHANG, W. et al. Comparison of rna-seq and microarray-based models for clinical endpoint prediction. *Genome biology*, BioMed Central, v. 16, n. 1, p. 133, 2015.