UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Seleção de genes para a predição da sobrevida em pacientes com câncer de mama

Khennedy Bacule dos Santos

Dissertação de Mestrado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP
SÃO CARLOS

Khennedy Bacule dos Santos

Seleção de genes para a predição da sobrevida em pacientes com câncer de mama

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências – Ciências de Computação e Matemática Computacional. *EXEMPLAR DE DEFESA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientadora: Profa. Dra. Mariana Cúri
Coorientador: Dr. Israel Tojal da Silva

USP – São Carlos
Abril de 2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Bacule dos Santos, Khennedy
B116s Seleção de genes para a predição da sobrevida em pacientes com câncer de mama / Khennedy Bacule dos Santos; orientadora Mariana Cúri; coorientador Israel Tojal da Silva. -- São Carlos, 2024.
94 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. Analise da sobrevida. 2. modelo de cox. 3. redução de dimensionalidade. 4. expressão gênica. I. Cúri, Mariana, orient. II. Tojal da Silva, Israel, coorient. III. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Khennedy Bacule dos Santos

Gene selection for predicting survival in breast cancer patients

Master dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master Program in Computer Science and Computational Mathematics. *EXAMINATION BOARD PRESENTATION COPY*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Profa. Dra. Mariana Cúri
Co-advisor: Dr. Israel Tojal da Silva

USP – São Carlos
April 2024

AGRADECIMENTOS

Agradeço primeiramente a minha esposa, que vem me apoiando nos últimos dois anos a manter o foco e a superar os diversos obstáculos da vida.

Agradeço a minha orientadora, Profa. Dra. Mariana Cúri pela oportunidade, atenção, todo conhecimento e paciência. Ao meu coorientador Dr. Israel Tojal por toda atenção e conhecimentos na área do câncer.

Ao **Conselho Nacional de Desenvolvimento Cientifico e Tecnológico**, pelo apoio financeiro na bolsa de mestrado.

"Nós só podemos ver um pouco do futuro, mas o suficiente para perceber que há muito a fazer."
(Alan Turing)

RESUMO

SANTOS, K. B. **Seleção de genes para a predição da sobrevida em pacientes com câncer de mama**. 2024. 94 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

Milhares de pessoas sofrem com o câncer, tornando-a uma das doenças que mais matam pessoas ao redor do mundo. Esta doença se carateriza por modificações na estrutura do DNA, o que impacta na produção descontrolada das células. Neste estudo abordamos uma predição da sobrevida para pacientes com câncer de mama nos estágios I, II e III, levando em consideração informações clinicas e genéticas. Para isto, o método de Cox, uma regressão capaz de estimar a função de risco, é usada para predizer a sobrevida dos pacientes. Devido a alta dimensionalidade da informação genética e as limitações do modelo Cox, são abordados métodos para a redução dos dados. Abordamos três maneiras para a redução de dimensionalidade, consistindo na penalização lasso na regressão de Cox, seleção por similaridade na expressão genética, com o algoritmo de agrupamento K-means, e a redução da dimensionalidade por meio da rede neural AutoEncoder, baseado nos grupos de similaridade. A partir dos experimentos, constatamos que a informação genética colabora para a criação de melhores preditores, em que as três abordagens de redução da dimensionalidade, apresentaram um melhor C-index, quando comparado ao método abordando apenas informações clinicas. Ao decorrer desta pesquisa, também verificamos que o material genético, além de aumentar o risco da sobrevida em alguns casos, há ocorrência do efeito de proteção. Ao final, propomos baseado nos resultados obtidos, uma possível evolução para a criação de um método capaz de otimizar o erro na predição da sobrevida, interpretar suas decisões e lidar com a alta dimensionalidade dos dados.

Palavras-chave: analise da sobrevida, modelo de Cox, redução de dimensionalidade, expressão gênica.

ABSTRACT

SANTOS, K. B. **Gene selection for predicting survival in breast cancer patients**. 2024. 94 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

Thousands of people suffer from cancer, making it one of the most deadly diseases worldwide. This disease is characterized by changes in the DNA structure, which impacts the uncontrolled production of cells. In this study, we approach a survival prediction for patients with breast cancer in stages I, II, and III, considering clinical and genetic information. For this, the Cox method, a regression capable of estimating the risk function, is used to predict patient survival. Due to the high dimensionality of the genetic information and the limitations of the Cox model, methods for data reduction are used. We approach three ways to reduce dimensionality: lasso penalty in Cox regression, selection for similarity in gene expression with the K-means clustering algorithm, and dimensionality reduction through the AutoEncoder neural network. From the experiments, we found that genetic information contributes to the creation of better predictors. The three approaches to dimensionality reduction presented a better C-index when compared to the method that addressed only clinical information. In the course of this research, we verified that the genetic material increases the risk of survival in some cases, but we found it also has a protective effect. Finally, based on the results, we propose a possible evolution towards creating a method capable of optimizing the error in survival prediction, interpreting their decisions, and dealing with the high dimensionality of the data.

Keywords: survival analysis, Cox model, dimensionality reduction, gene expression.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

C-index	*Concordance index*
FPKM	*Fragments Per Kilobase per Million mapped fragments*
INCA	Instituto Nacional de Câncer
LASSO	*Least Absolute Shrinkage and Selection Operator*
mRMR	*Min-Redundancy and Max-Relevance Algorithm*
NCI	*National Cancer Institute*
NGS	*Next-Generation Sequencing*
PCA	*Principal Component Analysis*
RSF	*Random Survival Forests*
SVM	*Support Vector Machine*
TCGA	*The Cancer Genoma Atlas*

LISTA DE SÍMBOLOS

δ_i — Indica a ocorrência ou não do evento para o individuo i

$\lambda(t)$ — Função de risco no tempo

β — Variáveis explicativas

ε — Erro residual

$\hat{\beta}$ — Variáveis explicativas estimadas

h — Espaço latente

ϕ — Rede codificadora

ψ — Rede decodificadora

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

O câncer é uma doença multifatorial que atinge toda população mundial, acometendo cerca de 19,3 milhões de pessoas no ano de 2020 (SUNG *et al.*, 2021). De acordo com a Organização Mundial da Saúde (OMS, do inglês *World Health Organisation*), o câncer com maior taxa de incidência é o de mama, com cerca de 47/100 mil pessoas diagnosticadas pelo carcinoma na mama. Em segundo, está o câncer de próstata com uma incidência de 30/100 mil. De maneira geográfica, o câncer de mama prevalece com maior incidência em 107 países, conforme ilustrado na Figura 1,

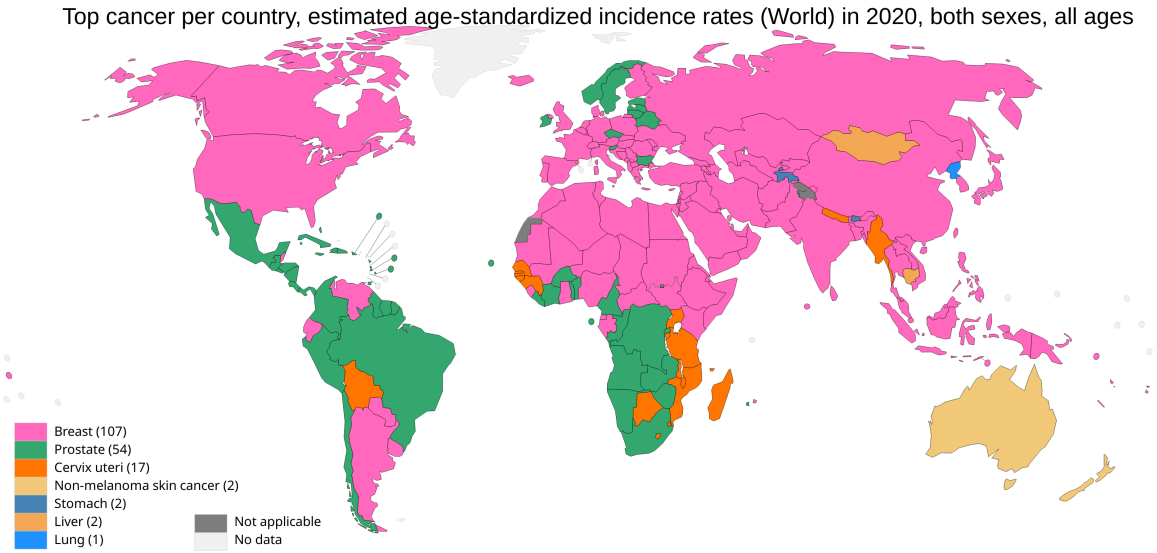


Figura 1 – Mapa geográfico mundial demonstrando os canceres com maior incidência por país.

Fonte: World Health Organisation (2023).

Em geral, a incidência e mortalidade por câncer estão crescendo rapidamente em todo o mundo (SUNG *et al.*, 2021). Isso reflete tanto o envelhecimento e o crescimento da população quanto às mudanças na prevalência e distribuição dos principais fatores de risco para o câncer,

muitos dos quais estão associados ao aumento da expectativa de vida, fatores ambientes, genéticos e comportamentais.

O diagnóstico de uma neoplasia de mama se dá por etapas. É usual o autodiagnóstico, no qual a própria pessoa tem alguma suspeita da doença, por meio de uma avaliação visual e sensitiva da região da mama e tecidos próximos. Em qualquer evidência de anomalia, se faz necessária a procura de um médico. Com um autodiagnóstico ou não, a procura de um médico é a primeira etapa, no qual é realizado a mesma avaliação anterior, visual e sensitiva. Com achados clínicos relevantes, é necessário realizar uma mamografia ou ultrassom, para obtenção de imagens do tecido. O uso da mamografia é mais comum e, em casos necessários, é feito o ultrassom. Resultados suspeitos julgados por um médico especializado são encaminhados para a realização de uma biópsia do tecido que apresenta a anomalia. Ao longo dessa sequência, é feito o diagnóstico. As etapas não precisam obrigatoriamente respeitar tal sequência, podendo seguir diferentes fluxos, de acordo com o critério do médico, sempre buscando o diagnóstico o mais precocemente possível.

Apesar da alta taxa de sobrevivência em pacientes com diagnóstico precoce, há a premissa da doença existir e já ter iniciado no indivíduo. Nos últimos anos, se tornou mais frequente e viável a análise do perfil genético, o que abre uma nova forma de prevenir o câncer, antes mesmo de sua manifestação. Dessa maneira, é possível estimar com certo grau de certeza se haverá a ocorrência da doença ou não. A aplicação deste tipo de método, por mais que ainda não amplamente disseminada, vem sendo cada vez mais viável, uma vez que a taxa de custos para o sequenciamento do material genético está reduzindo (National Human Genome Research Institute, 2023a).

Sendo o câncer uma doença genética, a compreensão dos aspectos moleculares tem um grande potencial para auxiliar na avaliação de biomarcadores de diagnóstico, prognóstico e preditivo. Por outro lado, a complexidade da informação genética e suas relações cria a necessidade de desenvolver e aplicar métodos analíticos inteligentes na pesquisa básica e translacional. Abordagens neste sentido são recorrentes, incluindo aquelas buscando encontrar biomarcadores (YU *et al.*, 2018; LIANG *et al.*, 2022; ZHANG *et al.*, 2022; XIAO *et al.*, 2022a).

Nesta pesquisa, é feita a análise da sobrevida de pacientes com câncer de mama não metastático, com o intuito de encontrar importantes traços genéticos capazes de auxiliar no prognóstico da doença. A inclusão do material genético do paciente permite criar um prognóstico mais assertivo. Neste contexto, considerando a alta dimensionalidade dos dados, o foco deste trabalho é criar abordagens de seleção dos genes relevantes para a doença e, posteriormente, analisar seu impacto na sobrevida. Além dos dados genéticos, são introduzidas informações básicas, como avaliação clínica e do tumor, usuais na prática disgnóstica.

Para a análise da sobrevida é utilizado o modelo de Riscos Proporcionais de Cox, um modelo de regressão para estimação do risco de óbito de acordo com o perfil clínico e genético do indivíduo. Devido à alta dimensão dos dados genéticos, aplicá-los diretamente ao modelo

é inviável. Desta maneira, para lidar com a informação genética, serão abordadas três formas para a redução de dimensionalidade: modelo com penalização lasso no algoritmo de estimação, seleção por agrupamento de genes com expressões similares e a redução por meio de redes neurais não supervisionadas. Esta última metodologia, em particular, é uma proposta inédita e combina os resultados da análise de agrupamento dos genes com o modelo de rede neural profunda, chamado *Autoencoder*.

Na literatura, diversos trabalhos têm objetivos similares a este. Jiang (2020), por exemplo, seleciona informações genéticas baseado em um modelo de regressão com penalização lasso. Abordagens com o uso de redes neurais estão cada vez mais comuns, devido à facilidade de lidar com a alta dimensão dos dados. A pesquisa de Tong *et al.* (2020) propõe o uso de *Auto Enconders*, modelo de rede neural profunda também adotado neste trabalho, para realizar a integração de diferentes camadas de informação molecular, contornando a alta dimensão dos dados para analisar a sobrevida.

Estes dois trabalhos e vários outros mencionados no Capítulo 3 evidenciam a necessidade de métodos para lidar com a alta dimensão de dados em estudos de câncer. Tais dados estão cada vez mais facilmente obtidos pela evolução dos métodos de sequenciamento genético, pelo aprimoramento do poder computacional e redução dos custos. Um importante desafio é que a taxa de crescimento da informação genética ultrapassa a lei de *Moore*, ou seja, temos uma taxa maior de geração de dados comparado ao de poder computacional. Uma previsão é que os dados genômicos tenham a escala de ZettaBytes em 2025 (STEPHENS *et al.*, 2015).

Modelos de aprendizagem de máquina são promissores nesse sentido, dada sua capacidade de identificar padrões complexos nos dados. No entanto, métodos de aprendizagem de máquina, em especial modelos baseado em rede neurais, possuem baixa interpretabilidade. Como uma das contribuições inéditas deste trabalho, propomos uma abordagem híbrida do método de agrupamento e do modelo de redes neurais, tornando-o parcialmente interpretável.

No decorrer deste capítulo, tem-se uma breve seção esclarecendo aspectos biológicos da doença estudada, seguida de uma seção com os objetivos e, finalmente, a organização do texto subsequente.

1.1 Biologia do câncer de mama

Todos os cânceres surgem como resultado de alterações no DNA das células. Essas alterações são geradas por eventos exógenos e endógenos. Por exemplo, a Figura 2 ilustra danos causados por mecanismos exógenos resultante da exposição a agentes radioativos.

No aspecto biológico, os genes são representados por uma sequência variável em tamanho, específica em bases nitrogenadas e encontram-se distribuídos ao longo do genoma que, por sua vez, contém aproximadamente 3 bilhões de sequências de nucleotídeos. No genoma humano

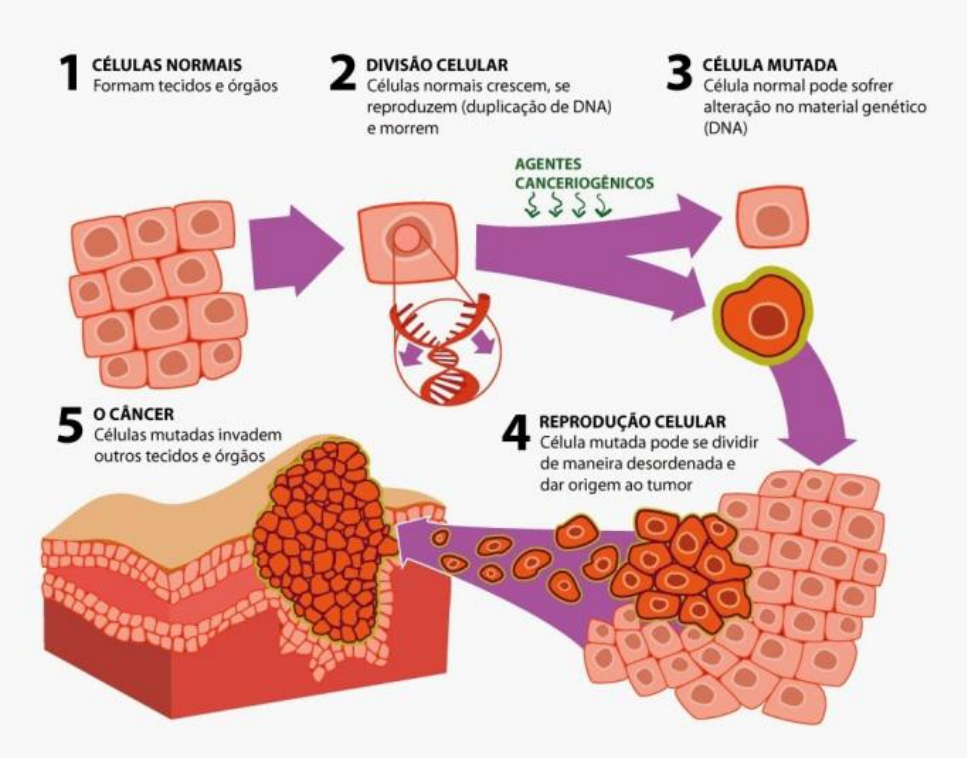


Figura 2 – Fluxo demostrando o processo de tumorigênese a partir do acúmulo de mutações no DNA das células.

Fonte: Hospital Hélio Angotti (2023).

(National Human Genome Research Institute, 2023b), há cerca de 130 mil genes. Destes, um total de 19207 são codificadores de proteínas, enquanto o resto são RNAs não codificadores, incluindo miRNA, lncRNA, snRNA, rRNA, dentre outros.

Particularmente, o câncer de mama é uma doença complexa e se apresenta de forma heterogênea nos vários subtipos histológicos e moleculares (The Cancer Genome Atlas Network, 2012). No aspecto histológico, a classificação do carcinoma de mama compreende nos subgrupos i) *in situ*, caracterizado pela proliferação de células neoplásicas dentro dos ductos mamários (DCIS) ou dentro dos lóbulos (LCIS) e ii) invasivo, onde as células tumorais já ultrapassaram a barreira celular da camada basal, conforme ilustra a Figura 3.

No contexto molecular, os tumores de mama são classificados em cinco sub tipos: Luminal A (tumores que apresentam receptores de estrogênio e progesterona positivos), Luminal B (possuem receptores estrogênio e/ou progesterona positivos), HER2 (não apresentam expressão dos receptores hormonais, mas têm a expressão da proteína HER2), Triplo Negativo (não possuem nem expressão hormonal, nem a proteína HER2, sendo negativo, portanto para estrogênio, progesterona e HER2 - **TNBC**, *Triple-Negative Breast Cancer*) e Normal (compartilha o mesmo perfil que o Luminal A, mas difere em níveis de expressão) (SUN *et al.*, 2021).

A compreensão das relações entre os aspectos moleculares do câncer e as características clínicas da doença tornou-se, portanto, uma das principais prioridades na pesquisa em

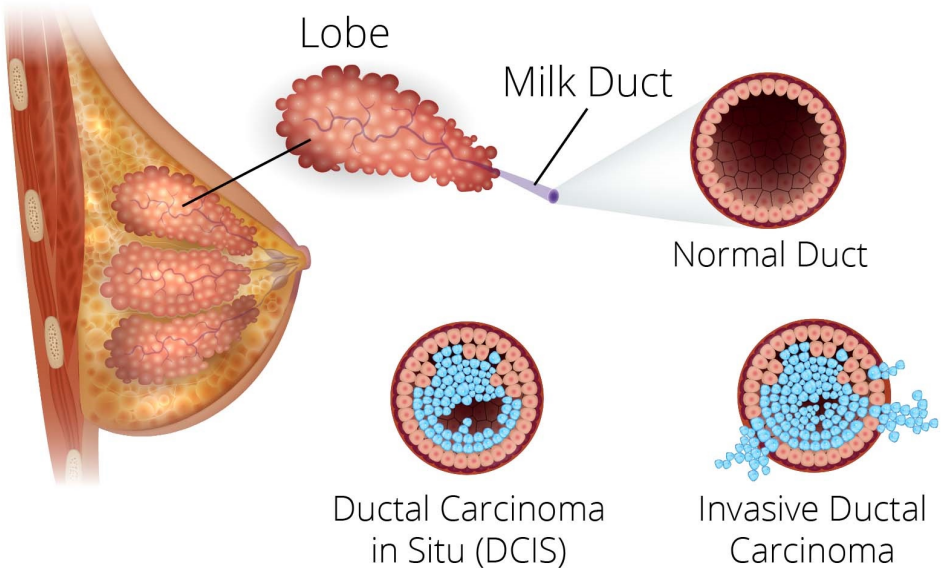


Figura 3 – Tipos histológicos do câncer de mama, DCIS e LCIS.

Fonte: Rocky Mountain Cancer Centers (2023).

câncer. Nos últimos anos, avanços nas tecnologias de sequenciamento de última geração (**NGS**, *Next Generation Sequencing*) permitiram o estudo aprofundado dos principais determinantes moleculares de um conjunto amplo de neoplasias (MARDIS, 2019; WEINSTEIN *et al.*, 2013).

Embora a classificação dos subtipos de câncer de mama tenham ampliado o conhecimento a respeito da biologia do tumor, visando um diagnóstico preciso e, consequentemente, um prognóstico mais assertivo, a complexidade dessa doença ainda traz desafios que requerem uma avaliação mais ampla considerando um conjunto de informações clínicas e patológicas.

Com o avanço das técnicas de sequenciamento visando compreender o processo de tumorigênese, dados genéticos estão cada vez mais disponíveis e podem ser representados em sistemas computacionais. Essa evolução trouxe a possibilidade de estruturar a informação genética de pacientes com câncer. A representação da atividade de um dado gene se dá por meio da leitura do total das moléculas (RNAs) que foram expressas em um dado contexto biológico no tecido de interesse, seja ao diagnóstico, durante o tratamento ou outra condição qualquer. De maneira resumida, a expressão gênica é o processo pelo qual as informações contidas no DNA são transformadas em instruções para a produção de proteínas ou outras moléculas, envolve a transcrição do DNA em RNA mensageiro (mRNA), seguida de uma tradução em proteínas.

Em geral, as etapas básicas no fluxo de análise envolvendo expressão gênica compreendem uma sequência estruturada de etapas (CONESA *et al.*, 2016; KUKURBA; MONTGOMERY, 2015), denominada de *pipeline* (Figura 4). A primeira etapa é selecionar/isolar células ou populações de tecidos das quais o RNA será extraído, seja a partir de uma biópsia, peça cirúrgica ou linhagens celular. Em seguida, todo o RNA (às vezes também chamado de RNA total) é extraído das células. Para mitigar os efeitos de lote (*Batch effect*) (ZHANG; PARMIGIANI;

JOHNSON, 2020), é importante que todos os procedimentos laboratoriais, tanto para o grupo controle quanto para o(s) grupo(s) experimental(is), sejam realizados no mesmo dia, no mesmo laboratório e pela mesma pessoa. O próximo grande passo seria isolar o conjunto específico de RNAs. Por exemplo, a maioria dos estudos direciona os esforços para investigar o RNA mensageiro e, assim, torna-se necessário remover o RNA ribossômico (rRNA) e novas classes de RNA não codificadores(ncRNA), incluindo os micro RNA (miRNA). No próximo passo, o RNA total ou mRNA é primeiro transcrito em um molécula de DNA complementar (cDNA - *complementary DNA*). O cDNA é então amplificado por PCR (*The Polymerase Chain Reaction*) e sequenciamento pelas plataformas de próxima geração (*Next Generation Sequencing - NGS*), que produzirá leituras de extremidade única ou emparelhada. Em seguida, as etapas básicas no fluxo de trabalho de bioinformática para análises de RNA é verificar a qualidade das leituras de sequência, remover leituras curtas, leituras de baixa qualidade e sequências residuais do adaptador. As leituras remanescentes são alinhadas/comparadas ao genoma de referência e, considerando as coordenadas do alinhamento, é realizada quantificação da abundância de leituras que foram anotadas em cada gene. Isso geralmente é chamado de geração de contagem, que nada mais é do que uma grande tabela contendo os nomes de cada gene e quantas leituras foram mapeadas/atribuídas aos genes. Finalmente, filtros adicionais, normalização e análises posteriores são realizadas nas amostras de interesse.

A avaliação de uma assinatura de expressão gênica contendo grupo de genes e sua associação com variáveis clínicas específicas tem o potencial para estabelecer um diagnóstico ou prognóstico mais assertivo (SUPPLITT *et al.*, 2021). Ademais, a detecção de uma assinatura de expressão gênica possibilita a classificação dos tumores em diferentes grupos permitindo, assim, um tratamento personalizado ao direcionar a terapia dos pacientes (VIJVER *et al.*, 2002; CARDOSO *et al.*, 2016).

Estudos anteriores reforçam a importância da utilização de assinaturas genéticas preditivas para auxiliar na decisão clínica. Métodos de inteligência artificial, especificamente *deep learning*, tem se mostrado como ferramenta poderosa com potencial para predizer fenótipos a partir de perfis de expressão gênica (LIU; YAO, 2022). Por outro lado, esses métodos são vistos como 'caixa preta', onde as predições são fornecidas sem qualquer explicação. As exigências para que esses modelos se tornem interpretáveis são cada vez maiores, principalmente na área médica (LONDON, 2019).

A abordagem predominante para estimar a expressão gênica inclui métodos de sequenciamento de próxima geração (NGS). O sequenciamento de RNA, também conhecido como RNA-Seq, é um método NGS que envolve a conversão de moléculas de RNA em DNA complementar (cDNA) e a determinação da sequência de nucleotídeos no cDNA para análise e quantificação da expressão gênica. Comparado com outras abordagens, o RNA-Seq oferece várias vantagens, incluindo maior especificidade, resolução e maior sensibilidade à expressão diferencial(SLATKO; GARDNER; AUSUBEL, 2018). Devido à natureza das medidas de quan-

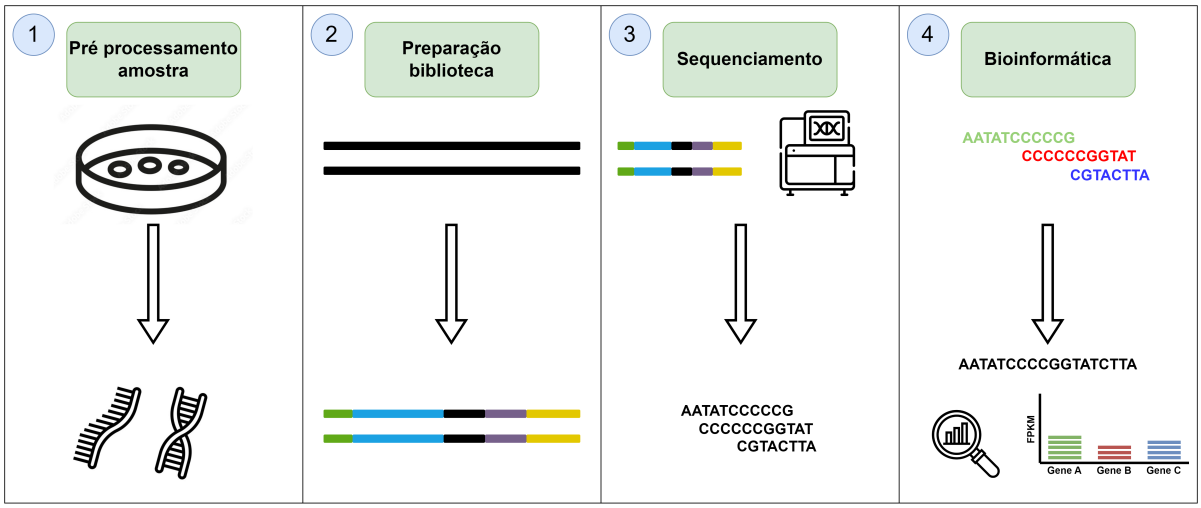


Figura 4 – Uma representação simplificada das principais etapas envolvidas no processamento de dados do sequenciamento total de RNA

Fonte: Elaborada pelo autor.

tificação, o nível de expressão de um gene poder ser representado considerado os seguintes métodos de normalização: TPM (*transcripts per million*) e RPKM/FPKM (*Reads/Fragments Per Kilobase per Million mapped fragments*) que considera, por exemplo, o tamanho do gene e o total de transcritos observados na amostra biológica (CONESA *et al.*, 2016) .

1.2 Objetivos

O principal objetivo desta pesquisa é fazer uso dos dados moleculares afim de aprimorar as análises preditivas da sobrevida em câncer de mama. Em combinação com dados clínicos, tradicionalmente adotados para diagnóstico e prognóstico da doença, os dados moleculares podem melhorar a predição e auxiliar no diagnóstico. Para tanto, utilizamos um modelo de sobrevida, que para possibilitar o uso dos dados genéticos necessitam de uma redução de dimensão dos dados. Nesse contexto, diferentes abordagens são adotadas nesta pesquisa.

Utilizamos dados de pacientes com câncer de mama, em estágios não avançado (estágios I, II e III) e focamos em modelos para a redução dos dados genéticos, sendo um deles inédito, a principal contribuição deste trabalho.

Para alcançar o principal objetivo desta pesquisa, os seguintes passos foram seguidos:

- realização do pré processamento adequado aos dados, removendo informações com ruídos.
- aplicação de métodos para redução de dimensionalidade do material genético.
- criação de um modelo base utilizando dados clínicos, com o intuito de comparar os modelos criados adicionando-se a informação genética.

- avaliação dos resultados dos modelos de sobrevida.
- validação dos resultados genéticos afim de confirmar a assertividade dos resultados.

1.3 Organização

Este trabalho apresenta no Capítulo 2 a parte teórica que fundamenta esta pesquisa. A maneira que os métodos são construídos é uma parte essencial para o entendimento da sua aplicabilidade no contexto do câncer de mama.

No Capítulo 3, podemos ver quais são as atuais metodologias utilizadas na literatura para tratar dados clínicos e genéticos. Os dados utilizados são apresentados no Capítulo 4.

Algumas propostas foram escolhidas para serem adotadas no escopo deste trabalho e são discutidas no Capítulo 5. Processos como o tratamento dos dados e a forma de avaliação dos modelos são discutidos também. Os resultados são apresentados no Capítulo 6, em que se discute se a informação genética agrega na predição da sobrevida.

Por fim no Capítulo 7, teremos a conclusão, em que discutiremos os resultados, seus impactos e as limitações encontradas. Na conclusão ainda apresentaremos os trabalhos futuros, em que abordaremos melhorias para esta pesquisa, com o intuito de evoluir a análise realizada.

CAPÍTULO

2

FUNDAMENTAÇÃO TEÓRICA

Nesse capítulo, vamos discutir sobre a base teórica dos métodos para a análise da sobrevida, seleção e redução de dimensão de variáveis.

Em análise de sobrevivência, é necessário entender os conceitos que remetem a termos como falha, tempo de falha, censura e alguns outros que serão abordados no decorrer deste capítulo (Seção 2.1). Após a apresentação desses conceitos básicos, é introduzido um dos modelos mais usados na área para descrever as funções de sobrevida e de risco de óbito em função do tempo de vida dos pacientes: o modelo de Riscos Proporcionais de Cox (Seção 2.2). Na Seção 2.3, falaremos sobre a forma de avaliação de modelos de sobrevida.

As últimas três seções abordarão três temas para a seleção de variáveis e redução da dimensionalidade dos dados genéticos. A Seção 2.4, conceitua a penalização L1, conhecida como lasso, no modelo de Cox, para selecionar os fatores genéticos mais associados à sobrevida. Na Seção 2.5, seguimos com o método de agrupamento chamado K-means, para a seleção dos genes com similaridade na expressão. Por fim, na Seção 2.6, fundamentaremos uma estrutura de rede neural não supervisionada, denominada *Autoencoder*, que será utilizada para a redução de dimensionalidade dos genes em conjunto do método de agrupamento.

2.1 Análise de sobrevivência

Em 1662, a primeira tabela de sobrevivência foi criada por John Graunt, sendo um dos principais marcos para a grande área da analise de sobrevivência (CAMILLERI, 2019). Desde então, esse tipo de análise está relacionado à mortalidade. Entretanto, nos últimos anos, diversas áreas vem aplicando modelos para a análise de sobrevivência, como por exemplo na engenharia, para determinar a vida útil de equipamentos, e na economia, para análises do aspecto financeiro.

A análise de sobrevivência tem como principal característica o estudo do tempo até a ocorrência de um evento de interesse. Óbito é geralmente o evento de interesse, dando nome

ao método. Mas outros eventos tais como quebra de um equipamento, ocorrência de um fato (remissão de doença, infarto, etc) também são outras tipos de eventos de interesse. A ideia é caracterizar o comportamento da probabilidade de sobrevivência (i.e., probabilidade de não ocorrência do evento) ao longo do tempo e estudar fatores que interferem nessa ocorrência.

Usualmente, esse evento de interesse, também denominado falha, ocorre uma única vez para cada unidade da amostra. Porém, há situações em que ocorre de forma recorrente, mas que não serão tratadas no escopo deste trabalho.

O tempo de falha é o tempo entre o início do estudo de cada unidade da amostra e a ocorrência do evento. Neste trabalho, o tempo de falha é definido como o período entre o diagnóstico da doença e a ocorrência do óbito do paciente por câncer de mama. Geralmente, este tempo está relacionado a um tempo real, como horas, dias, meses, etc. Há casos, como a aplicação na engenharia, que tal definição pode mudar, em que a unidade pode ser quilômetros, por exemplo.

Porém, há situações em que nem todas as unidades da amostra são observadas até a ocorrência da falha (do óbito). Nestes casos, aparece a definição de censura. A censura na análise de sobrevivência é um termo que foi introduzido para tratar dados em que não houve a ocorrência do evento de interesse até o limite de tempo observado. Essa informação é relevante na análise, visto que por algum tempo essas unidades foram observadas e a falha não ocorreu. Esse dado pode e deve ser aproveitado no modelo estatístico de alguma maneira. Uma censura, por exemplo, pode ocorrer quando é iniciado um estudo com 10 pacientes e um deles morre por um outro motivo, que não o de interesse. Ou mesmo pacientes que sobreviveram até o último momento da pesquisa, também caracterizam censuras.

Podemos definir diferentes tipos de censura: o Tipo I, Tipo II e o tipo aleatório. A Figura 5 mostra as diferentes configurações que podem aparecer no cenário de análise de sobrevivência, sendo elas: (a) quando não existe censura, todos atingiram o evento de falha, (b) censura tipo I, quando o estudo é finalizado em um tempo pré-determinado e os indivíduos sem a ocorrência de falha até aquele tempo são tratados como censuras, (c) censura tipo II, quando o estudo atinge um número pré-determinado de eventos de falha e (d) censura aleatória, ou seja, quando por algum motivo o individuo teve outro desfecho e a falha não foi observada até um tempo t_i. Na maioria dos casos, principalmente em casos de pesquisa na área médica, a censura mais comum é a do tipo aleatória. Nela, os dados censurados geralmente costumam estar associados ao término do estudo antes da ocorrência da falha, ou quando ocorre outro evento não esperado, por exemplo morte por acidente de carro e o estudo é sobre óbitos de paciente com câncer.

A partir deste momento, podemos então definir a estrutura dos dados de sobrevivência. Define-se o par (t_i, δ_i), onde o individuo i $(i = 1, ..., n)$ possui a representação do tempo até a ocorrência da falha ou censura em t_i e a indicação se o evento de interesse foi observado ou

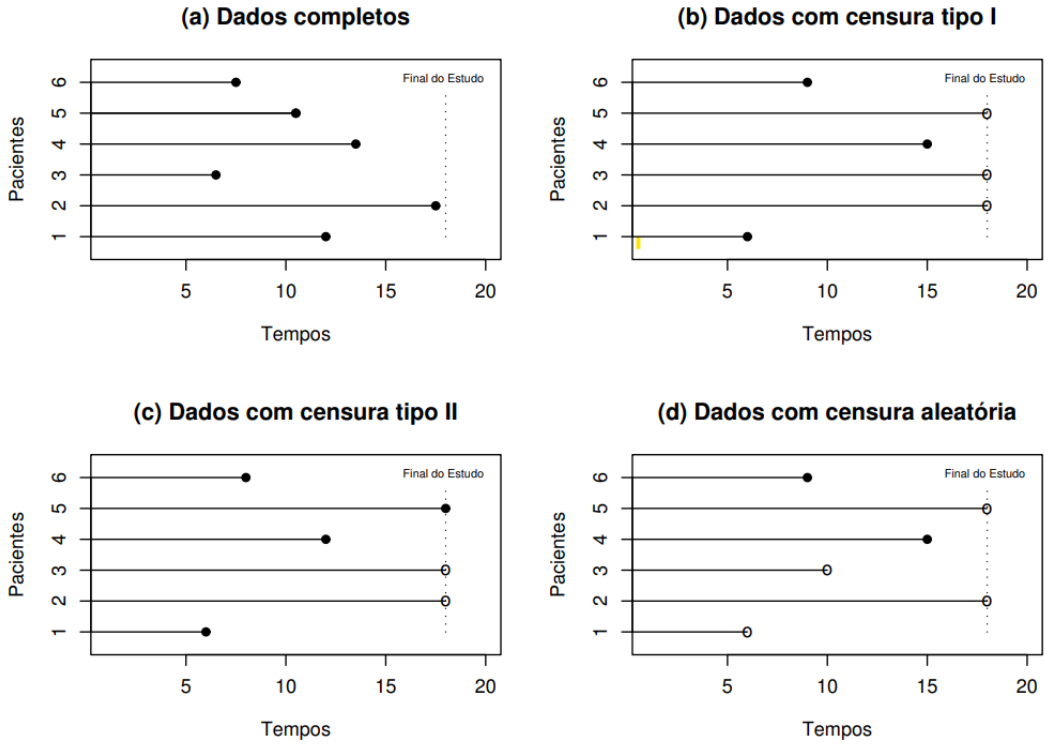


Figura 5 – Tipos de censura na analise de sobrevivência.

Fonte: E. (2006, pag. 7).

ocorreu uma censura em δ_i. Dessa maneira:

$$\delta_i = \begin{cases} 1 \text{ se } t_i \text{ é um tempo de falha} \\ 0 \text{ se } t_i \text{ é um tempo censurado.} \end{cases} \tag{2.1}$$

Para o estudo dos fatores associados ao tempo de falha, devemos introduzir mais um termo, referente a tais características. Obtemos, então, a tripla $(t_i, \delta_i, \boldsymbol{x}_i)$. Neste caso, $\boldsymbol{x}_i$ se refere a um vetor de características do individuo, tais como idade, estadiamento do câncer e tumor.

Algumas funções do tempo de falha são importantes de ser definidas. Um dos principais componentes da análise de sobrevida é a função que descreve a probabilidade de sobrevivência do indivíduo, ao longo do tempo t. Essa função é conhecida como função de sobrevivência, definida da seguinte forma:

$$S(t) = P(T \geq t). \tag{2.2}$$

Podemos então atribuir a função $F(t) = 1 - S(t)$ como a probabilidade de um indivíduo (ou objeto) não sobreviver no tempo t.

Uma outra importante definição é a taxa de falha, que mostra a forma em que a taxa instantânea de falha muda com o tempo. Em estudos clínicos podemos ver quanto um grupo de pacientes com uma determinada doença possui de variação, descrevendo a evolução da doença.

Essa taxa é definida por:

$$\lambda(t) = \lim_{\Delta_t \to 0} \frac{P(t \leq T < t + \Delta_t | T \geq t)}{\Delta_t}. \tag{2.3}$$

A partir de tais funções relativas ao tempo de falha, podemos associar covariáveis que as influenciam através de um modelo de regressão apropriado para lidar com censuras. Na seção seguinte o modelo de Cox, um dos mais usados para tal fim, é apresentado.

2.2 Modelo de Riscos Proporcionais de Cox

Na análise de sobrevivência, diversos modelos foram propostos para identificar fatores que influenciam na sobrevida. Alguns deles supõem distribuições de probabilidade conhecidas para o tempo de falha. Num contexto mais geral, sem assumir nenhuma distribuição específica para o tempo de falha, há o modelo de Cox, utilizado neste trabalho para predizer o risco de óbito dos pacientes com câncer de mama.

O modelo de Cox é uma regressão que relaciona a taxa de falha entre dois grupos de indivíduos diferentes como constante. Suponha que temos dois grupos de estudo: um com o câncer em estágio 1 e outro no estágio 2. A função de taxa de falha para o grupo 1 é $\lambda_0(t)$ e para o grupo 2, $\lambda_1(t)$. Podemos definir que

$$\frac{\lambda_0(t)}{\lambda_1(t)} = K, \tag{2.4}$$

sendo K a razão entre as taxas de falha dos grupos 1 e 2, suposta constante no tempo. Generalizando para mais de dois grupos, definidos a partir de n covariáveis, o modelo de Cox define o risco da seguinte forma:

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t) \exp(\beta_1 x_1 + ... + \beta_n x_n), \tag{2.5}$$

para cada $\boldsymbol{x}_j$, $j = 1, ..., n$, que representa alguma característica dada pelo estudo, e um parâmetro β_j atrelado. O modelo de Cox é chamado de semi-paramétrico pelo fato de possuir dois componentes: um não paramétrico e outro paramétrico. A parcela dependente das covariáveis, características dos indivíduos, é a parte paramétrica, enquanto o fator $\lambda_0(t)$ é a parcela não paramétrica.

A denominação “riscos proporcionais"dada para este modelo vem do fato de que a taxa de falha entre dois indivíduos é considerada constante. Anteriormente, vimos que a razão dos riscos de óbito entre dois grupos (em estágios 1 e 2, respectivamente) é K, independente do tempo t. Neste caso mais geral de n covariáveis, temos que os indivíduos i e i^* possuem a seguinte relação de taxa de falha:

$$\frac{\lambda_i(t)}{\lambda_{i^*}(t)} = \frac{\lambda_0(t) \exp\{\boldsymbol{x}_i' \beta\}}{\lambda_0(t) \exp\{\boldsymbol{x}_{i^*}' \beta\}} = \exp\{\boldsymbol{x}_i' \beta - \boldsymbol{x}_{i^*}' \beta\}. \tag{2.6}$$

Se essa razão é maior (ou menor) do que 1, pode-se dizer o indivíduo i tem um risco de óbito maior (ou menor) do que o indivíduo i^*, de maneira constante no tempo.

Com o modelo de Cox podemos tirar diversas informações acerca da influência das covariáveis na sobrevida (i.e., na taxa ou risco de falha)[1]. A significância de cada parâmetro β no modelo diz sobre a importância da covariável correspondente na sobrevida. Nesse sentido, a estimação dos β's deve ser feita e poder-se-ia pensar em utilizar o método de máxima verossimilhança, classicamente. No entanto, para contornar a existência da componente não paramétrica $\lambda_0(t)$, o método de máxima verossimilhança parcial é adotado.

[1] Define-se a verossimilhança parcial como sendo a probabilidade de o indivíduo i vir a falhar no tempo t_i dividida pela soma dos riscos de falha de todos os indivíduos em risco até esse momento. Dessa maneira teremos:

$$\frac{\lambda_i(t)}{\sum_{j\in R(t_i)}\lambda_j(t)} = \frac{\lambda_0(t)\exp\{\boldsymbol{x}_i'\beta\}}{\sum_{j\in R(t_i)}\lambda_0(t)\exp\{\boldsymbol{x}_j'\beta\}} = \frac{\exp\{\boldsymbol{x}_i'\beta\}}{\sum_{j\in R(t_i)}\exp\{\boldsymbol{x}_j'\beta\}}, \tag{2.7}$$

em que $R(t_i)$ representa o conjunto das observações sob risco no tempo t_i. Repare que o componente não paramétrico simplifica, possibilitando a estimação.

Supondo independência entre os indivíduos, a verossimilhança parcial fica dada por:

$$L(\beta) = \prod_{i=1}^{k}\frac{\exp\{\boldsymbol{x}_i'\beta\}}{\sum_{j\in R(t_i)}\exp\{\boldsymbol{x}_j'\beta\}}, \tag{2.8}$$

em que o objetivo é maximar essa função em relação aos parâmetros para obtermos seus estimadores, seguindo a relação suposta no modelo.

Existem diversas maneiras de solucionar problemas de maximização. Na maioria dos casos, o problema é resolvido buscando solucionar um sistema de equações. Neste caso, minimiza-se a função $-\ell(\beta)$, que consiste no $-log(L(\beta))$, que facilita o processo matemático. A aplicação do log se dá por alguns motivos, sendo o principal o produtório definido pela função de probabilidade, já que na aplicação da função log o produto é transformado em soma. A representação matemática e computacional simplifica, tem mesma solução e facilita a aplicação dos métodos numéricos.

Com esse entendimento do método Cox e a forma usada para estimação de seus parâmetros, podemos partir para a seção seguinte, que apresenta uma métrica para avaliar o modelo apresentado.

2.3 Método de avaliação na sobrevida

Nas análises de sobrevida, o índice de concordância ou C-Index (do inglês, *Concordance Index*) é uma métrica muito utilizada e será a principal adotada neste trabalho, juntamente com

[1] Na estatística os termos taxa ou função de risco e taxa ou função de falha são sinônimos. A menção de sobrevida no modelo de Cox diz respeito ao risco de falha da observação.

o AIC (*Akaike Information Criterion*), como forma comparativa entre as soluções executadas (HARRELL FRANK E. *et al.*, 1982). Tal escolha se deve ao levantamento bibliográfico realizado, onde trabalhos relacionados fizeram uso desta mesma métrica.

O C-Index é definido baseado nas informações da sobrevida, analisando a qualidade de predição do modelo e a lógica de representação dos dados. Para sua construção, suponha η como o risco de ocorrência do evento. No caso do modelo de Cox, seria a função λ, ou seja, o risco de falha para o individuo i é η_i e a informação do tempo do individuo i é denotado por T_i.

A métrica é construída comparando-se os pares de indivíduos i e j e fazendo-se uma contagem, da seguinte forma:

- se i e j são censuras, são desconsiderados na computação da contagem,
- se i e j são falhas, é considerado um par válido para computar a contagem,
- se i e j são um falha e outro censura, temos que analisar os tempos de falha/censura para considerá-lo como um par válido para computar a contagem.

A contagem é referente ao número de pares corretos encontrados no conjunto de dados, considerando o modelo em questão. Um par de indivíduos i e j é denotado como correto se para o caso do tempo de falha $T_j > T_i$ for acompanhado por uma maior probabilidade do paciente i morrer do que j. Esse cálculo deve levar em consideração se os tempos T_i e T_j são tempos de falha ou de censura.

Por essa razão, não se pode computar a contagem quando os dois indivíduos são censurados. Nesse caso, os tempos T_i e T_j seriam tempos de censura e nada se pode dizer a respeito dos tempos de falha de ambos os indivíduos, relativamente um ao outro.

Quando temos duas falhas, por outro lado, é possível comparar seus dois tempos e taxa de falha, pois temos a informação de dois indivíduos com $\delta = 1$. Com os seus tempos de falha e a probabilidade de falha dada pelo modelo, o C-index considera a predição correta se o individuo com maior T tiver o menor η. Em caso contrário, a métrica será computa como um par incorreto.

Essa regra se torna lógica se pensarmos em pacientes: dado dois pacientes i e j, que foram a óbito em T_i e T_j, respectivamente, tal que $T_j > T_i$, ou seja o paciente i morreu mais cedo do que o j. Se o modelo resultar numa maior probabilidade de óbito do paciente i do que j, o par é classificado como correto.

Seguindo essa lógica podemos inferir para casos em que um dos indivíduos é uma censura: tenhamos i como a censura e j como falha. Assim temos que no caso de $T_i < T_j$, não é considerado um par válido, isso porque não podemos inferir uma regra para o dado censurado. Quando $T_i > T_j$ é possível entender que até o tempo T_i não foi detectada a falha, mas para o j tivemos uma falha no tempo T_j, assim esperamos que o preditor entregue uma maior

probabilidade de falha para o individuo j, computando assim um par correto. E caso contrário, é considerado um par incorreto.

Dessa maneira, podemos construir uma relação lógica para a métrica C-index, tal que:

$$C_{index} = \frac{\sum_{i,j} T_j < T_i \cdot \eta_j > \eta_i \cdot \delta_j}{\sum_{i,j} T_j < T_i \cdot \delta_j}. \tag{2.9}$$

Quando as operações lógicas são verdades, é retornado o valor 1, e quando são falsas, o valor 0. Dessa maneira, vemos que para contemplar a análise de censura, esta deve ter o maior tempo e o δ_j é inserido, obrigando o individuo j a ser sempre uma falha.

O C-index é computado para cada modelo ajustado. Caso ele assuma o valor 0.5, podemos interpretar que o modelo acertou 50% dos casos e errou a outra metade, ou seja, é um modelo aleatório. Quanto mais a métrica se aproxima do valor 1, melhor é o modelo. Repare que um C-index igual a 1, implica que o modelo acertou todos os pares analisados.

A métrica AIC também avalia o ajuste do modelo, mas diferente do C-Index a avaliação é feita pela quantidade de parâmetros do modelo e a função de verossimilhança. O AIC é definido da seguinte maneira:

$$AIC = 2k - 2\ln(\hat{L}). \tag{2.10}$$

Onde k é a quantidade de variáveis do modelo e $\hat{L}$ é o valor da função de verossimilhança. Quanto menor o valor do AIC melhor é seu ajuste, conforme visto na sua fórmula, essa métrica penaliza modelos com grande número de variáveis. Um maior valor da verossimilhança reduz o valor do AIC. Logo essa métrica pondera a simplicidade do modelo com o seu poder de predição.

Com a apresentação da análise da sobrevida, do modelo de Cox e a sua avaliação, resta apresentar a fundamentação dos métodos que serão usados para selecionar ou reduzir a dimensionalidade dos dados, importantes para contemplar os dados genéticos.

2.4 LASSO - *Least Absolute Shrinkage and Selection Operator*

A penalização lasso foi utilizada neste trabalho com o intuito de redução de dimensionalidade dos dados. Mais especificamente dos dados genéticos, visto que possuem uma alta dimensão, cerca de 19 mil, quando filtrados em genes que codificam proteínas.

O lasso é uma penalização muito utilizada em modelos de regressão, particularmente na regressão linear, mas também em modelos de regressão generalizados e também modelo de Cox, de forma análoga. Para melhor entender essa penalização, suponha a seguinte relação na

regressão linear:

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{pi} + \varepsilon_i, \tag{2.11}$$

em que $i = 1, 2, ...n$, Y_i representa a variável resposta para a i-ésima observação, β_0 é o intercepto, X_{ji}, para $j = 1, ..., p$, são as variáveis explicativas (covariáveis), em que j representa a característica em questão e i o indivíduo, ε_i representa o erro associado ao i-ésimo indivíduo. De maneira mais usual, podemos reescrever a expressão dada para a regressão linear em notação matricial como:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{2.12}$$

Na regressão lasso, o mais usual é utilizar o método de mínimos quadrados para estimar o vetor $\boldsymbol{\beta}$, obtendo o termo da regressão linear e a penalização lasso

$$\hat{\beta} = \arg\min_{\beta} \left(\sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right), \lambda \geq 0. \tag{2.13}$$

O termo $\lambda \sum_{j=1}^{p} |\beta_j|$ penaliza os β's estimados diferentemente de zero, visto que o modelo tende a minimizar a primeira parcela que são os mínimos quadrados. Quando introduzimos a segunda parcela, pressionamos para que o conjunto solução tenha valores baixos para β, na medida do possível e viabilizado pelos dados. No caso de alguns dos valores estimados serem exatamente iguais a zero implica que as covariáveis correspondentes não são importantes para explicar a variação da resposta. O método possui uma parâmetro λ, que estipula uma flexibilidade quanto à penalização. Quando temos o valor de λ igual a 0, teremos a regressão linear tradicional, sem penalização. Já quando temos $\lambda \to \infty$, os β's terão suas estimativas zeradas. Assim sendo, é um parâmetro que exigirá um estudo para sua escolha, por isso denominado hiperparâmetro.

A penalização lasso também é chamada de penalização L1 e, na descrição acima, explicamos sua aplicação em um modelo de regressão linear. A aplicação para a redução de genes via regressão linear é uma possibilidade viável, tendo a literatura um trabalho relacionado com esta aplicação (JIANG, 2020). Entretendo, neste trabalho não usaremos a penalização lasso com a regressão linear, mas sim com o modelo de Cox.

Para a adição da penalização lasso ao modelo de Cox, usaremos a mesma lógica vista acima. Desta maneira, teremos

$$arg\min_{\beta} -\ell(\beta) + \lambda|\beta|. \tag{2.14}$$

2.5 K-means

Métodos de agrupamento usualmente são usados para a redução de dimensionalidade, visto que a sua proposta é agrupar dados que possuem alguma similaridade. Nesta seção, vamos mostrar a fundamentação do método não supervisionado K-means com o intuito de

obter um agrupamento de variáveis e posterior seleção de uma representante de cada grupo formado (TANG *et al.*, 2017). Neste trabalho, em particular, usaremos o método para selecionar caraterísticas (variáveis) da expressão genica.

O método K-means necessita da pré-definição da quantidade de grupos que serão criados. Dessa forma, a tarefa de identificar o número K ideal de grupos necessita de uma análise. Veremos o método de Elbow, que apoia esta decisão. Quando definido o número de grupos, o algoritmo é iniciado criando K pontos aleatórios, chamados de centroides. Em alguns casos são utilizados métodos heurísticos para a seleção de bons pontos iniciais.

Ao definir os centroides, o algoritmo busca realizar o agrupamento dos indivíduos (ou de variáveis). Isto é feito de acordo com o centroide mais próximo do individuo (ou característica). Neste caso, a maneira mais usual de calcular essa relação é usando a distância Euclidiana.

Desta maneira podemos obter a seguinte função de otimização, que o K-means busca minimizar:

$$\arg\min_{S} \sum_{i=1}^{K} \sum_{x \in S_i} ||x - \mu_i||^2, \tag{2.15}$$

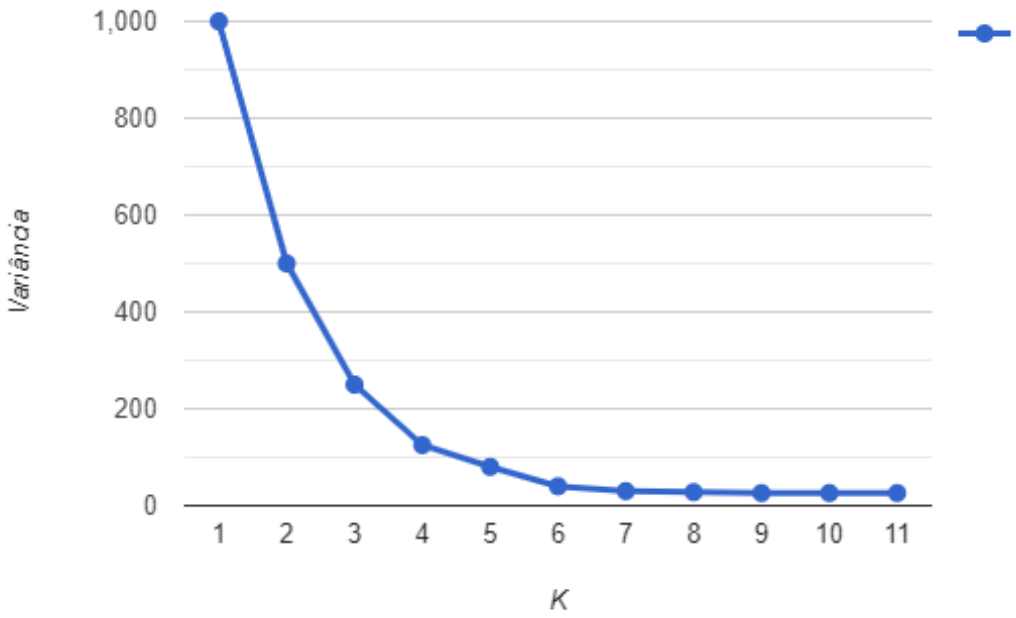
em que μ_i e S_i são o i-ésimo centroide e o conjunto de indivíduos mais próximos do centroide, respectivamente. Ao minimizar essa função, podemos considerar que temos a etapa de atualização dos centroides. A sua atualização é baseada no ponto médio dentro do conjunto S. Após isso, são atualizados os conjuntos que especificam o grupo dos indivíduos no conjunto S. O critério de parada é definido quando não há mudanças nos centroides. Neste caso, interpretamos que o método convergiu para alguma solução.

Geralmente, em aplicações de análise de agrupamento, não se sabe o número de grupos que se pretende identificar. Desta maneira, existem algumas alternativas para buscar um valor de K ideal para os dados. O método que utilizamos neste trabalho é o método de Elbow. Sua aplicação consiste em criar diversos K-means variando o número K de grupos. Ao ajustar o modelo com os dados, obtemos a variância dos grupos e, assim, teremos uma relação do K com a variância do método.

Quando o método de Elbow mostrar uma alta variância podemos inferir que o número de grupos é pequeno. Quando a variância se estabiliza, não tendo mudanças ao aumentar o número de grupos, podemos dizer que ocorreu um *overffing*. Desta maneira, no gráfico, esperamos selecionar o ponto médio entre essas duas ocorrências. Por essa razão que o método de Elbow é também chamado de método do cotovelo, devido à sua sugestão de melhor número de grupos se encontrar na curva decrescente da variância.

Conforme mostra a Figura 6, podemos selecionar o K com o valor 4, visto que à sua direita a variância mostra que a adição de novos grupos não agrega no modelo e à sua esquerda a variância é alta, mostrando o baixo número de grupos para os dados utilizados.

Figura 6 – Exemplificação método de Elbow



Fonte: Elaborada pelo autor.

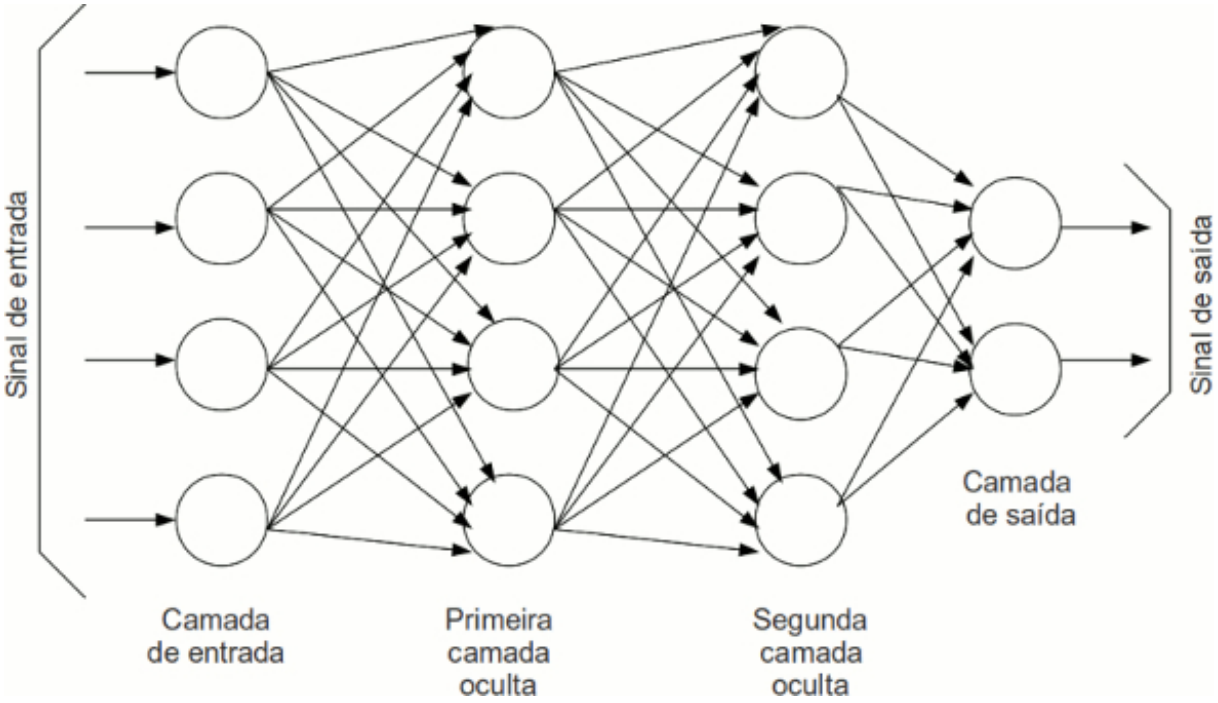
2.6 Autoencoder

As redes neurais são modelos matemáticos baseados no sistema nervoso de um animal. O sistema nervoso contém um conjunto de neurônios interligados, que se comunicam através de pulsos elétricos, chamados de sinapses. A estrutura de um neurônio biológico é formada por três partes fundamentais: (i) os dendritos, responsáveis por receber informações, (ii) o corpo celular, que realiza o processamento das informações recebidas, e por último, (iii) o axônio, com a função de distribuir a informação processada. Um ser humano possui bilhões de neurônios com milhares de ligações entre eles, criando uma rede enorme que possibilita o raciocínio.

No âmbito computacional, a ideia é criar uma estrutura semelhante. Na Figura 7, podemos visualizar uma estrutura de uma rede neural artificial. Os sinais de entrada para a rede são os dados. O processamento na rede é feito por multiplicações vetoriais, de uma matriz de pesos com os valores de entrada e, ao passar pelo que é chamado de função de ativação, é processada a saída. Neste caso, o aprendizado se dá pelo ajuste desta matriz de pesos ao decorrer do envio de dados: esta é a fase de treinamento da rede neural. A definição de uma rede neural profunda se dá pela quantidade de camadas internas que a rede possui, no qual redes com duas ou mais camadas internas são classificadas como profundas (ROSENBLATT, 1958).

Como mencionado, o aprendizado de uma rede neural artificial ocorre pela atualização

Figura 7 – Arquitetura de uma rede neural



Fonte: REDES... (2017).

da matriz de pesos. Na maioria dos casos, a atualização de pesos é feita pelo algoritmo *back-propagation* e o uso do gradiente descendente. O algoritmo busca a solução que minimiza o erro da rede, caracterizado pelo contraste dos valores observados da variável resposta e seus valores preditos na saída da rede. O uso do gradiente descendente é para determinar a direção de correção da rede no espaço (RUMELHART; HINTON; WILLIAMS, 1986) em cada iteração realizada.

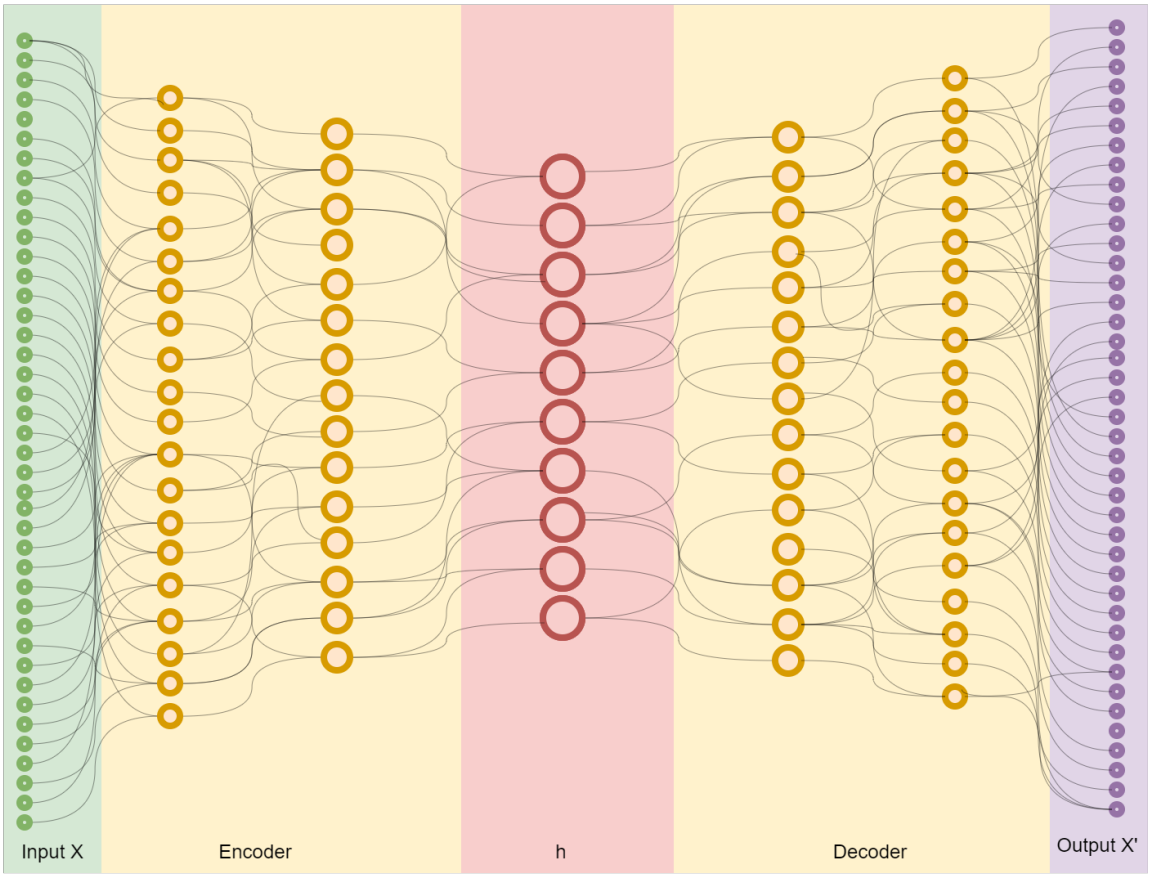
O *Autoencoder* é uma arquitetura de rede neural que utiliza o principio apresentado acima, com duas peculiaridades. Primeiramente, é um método aplicado a situações em que as predições são referentes aos próprios valores de entrada da rede, caracterizando uma situação de aprendizado “não supervisionado", como usualmente chamado na linguagem de aprendizado de máquina. Adicionalmente, o número de neurônios nas camadas de entrada e de saída é o mesmo, além das camadas intermediárias usualmente terem um número menor de neurônios. Nesta arquitetura, a distribuição dos neurônios ao longo das camadas possibilita uma redução da dimensionalidade dos dados de entrada, seguida de uma ampliação para a recuperação dos dados originais. Assim sendo, a usaremos para a redução da quantidade dos dados genéticos (BANK; KOENIGSTEIN; GIRYES, 2020).

São definida duas redes neurais convencionais no *Autoencoder*. Elas são denominadas de rede codificadora (*encoder*) e decodificadora (*decoder*), conforme Figura 8. A rede codificadora tem a função de realizar a compreensão dos dados de entrada, ela deve "aprender"a representar os

dados em uma menor dimensão (menos neurônios do que a camada de entrada). Enquanto a rede decodificadora tem o papel de reconstruir as informações que a rede codificadora comprimiu. Com esta rede o aprendizado da rede codificadora se torna possível, visto que quanto menor o erro da decodificadora melhor está a compreensão de dados. Trabalhando juntas, temos um sistema que "aprende"a representar os dados em uma menor dimensão.

Repare que além destas redes, temos o espaço h, a camada escondida central, muito importante porque é de onde obtemos as informações reduzidas.

Figura 8 – Arquitetura de um *Autoencoder*



Fonte: Elaborada pelo autor.

A única restrição existente para esta arquitetura é quanto a entrada e saída das redes, que devem ser compatíveis. Isto significa que a entrada do codificador terá o mesmo tamanho de saída do decodificar, e a saída do codificador tem o mesmo tamanho da entrada do decodificador.

A rede codificadora busca reduzir a entrada X, que possui n variáveis (ou nós na camada de entrada), para um tamanho m, onde $m < n$. Dessa forma, vamos reduzir a dimensão dos dados. O decodificador tem o papel de restaurar a informação que foi dada de entrada na sua saída, da segunda rede. A entrada da segunda rede é o resultado da compreensão da rede codificadora.

Podemos definir matematicamente o *Autoencoder* em duas partes, da seguinte forma:

$$\begin{aligned} &\phi : X \rightarrow F \\ &\psi : F \rightarrow X \\ &\phi, \psi = \arg\min_{\phi,\psi} ||X - (\psi \circ \phi)X||^2, \end{aligned} \tag{2.16}$$

sendo ϕ a rede codificadora e ψ a rede decodificadora. A busca de ambas as funções significa minimizar o erro quadrático entre a entrada e a saída. Neste sentido, a rede é atualizada de acordo com seu erro e taxas selecionadas para o treinamento.

CAPÍTULO

3

TRABALHOS RELACIONADOS

Neste capítulo discutiremos sobre os atuais trabalhos publicados que possuem relação com a pesquisa em questão. Na busca dos trabalhos relacionados, foram procuradas pesquisas que tratam qualquer tipo de câncer, não especificamente o câncer de mama, o abordado nesta pesquisa. Isto porque o objetivo é analisar os métodos para tratar problemas relacionados ao câncer, no âmbito da sobrevida e complexidade de dados.

Não diferente da maioria das áreas, abordagens com algoritmos de aprendizado de máquina vem crescendo. Na PubMed em 2022 tiveram 2561 trabalhos, relacionados aos termos *machine learning AND cancer*, em títulos e resumos, o que representa um grande crescimento, visto que em 2015 tiveram apenas 149 trabalhos com os mesmos termos.

A analise de sobrevivência possui um amplo campo de aplicabilidade, não apenas na saúde, e isso fez com que diversas abordagens fossem propostas nas últimas décadas. Por mais que diversos métodos foram propostos para casos específicos, a maioria dos métodos são aplicáveis de maneira geral, visto que todos seguem a mesma fundamentação, a da analise de sobrevivência.

O trabalho de Wang, Li e Reddy (2017) realizou um levantamento da literatura, demonstrando o amplo campo de aplicação e fundamentação dos métodos de analise de sobrevivência. Adicionalmente, definiu uma taxonomia para os métodos de analise de sobrevivência, ilustrada na Figura 9.

De acordo com Wang, Li e Reddy (2017), os métodos estatísticos são segmentados em três modalidades. Os não paramétricos, que foram os primeiros propostos e mais simplistas, visto que em suas análises não são supostas caraterísticas sobre a distribuição dos dados. O método semi paramétrico, utilizado nesta pesquisa, que supõem que os dados satisfazem certas condições não tão estritas. Por fim, os métodos paramétricos, capazes de obter bons resultados, sob a condição dos dados satisfazerem algumas suposições, incluindo sua distribuição de probabilidade, o que limita sua aplicabilidade em problemas reais.

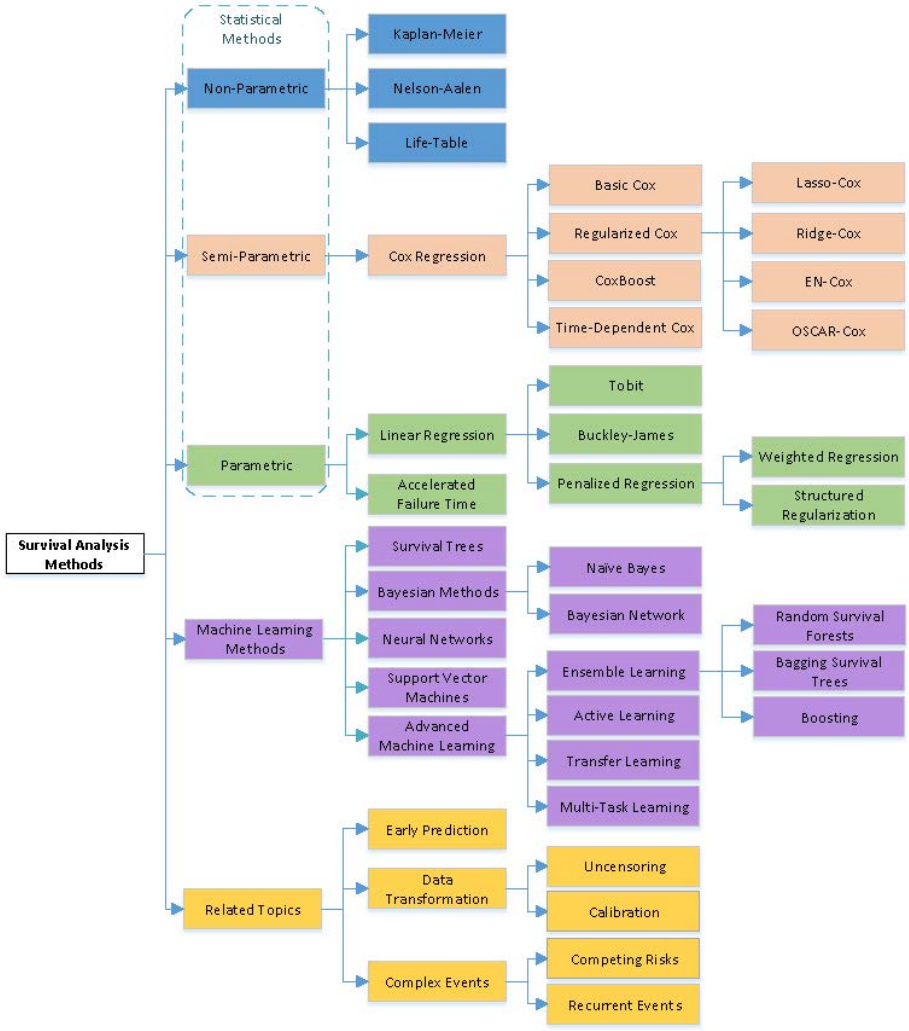


Figura 9 – Taxonomia dos métodos aplicados a análise de sobrevivência

Fonte: Wang, Li e Reddy (2017).

Os métodos de aprendizagem de máquina são diversos: desde modelos Bayesianos até modelos baseados em redes neurais. Os trabalhos relacionados nesta seção são abrangentes, considerando por exemplo múltiplos eventos e recorrência, para ampliar a compreensão do contexto da ciência de dados moderna na análise de sobrevivência de forma geral.

Neste trabalho, são aplicados métodos de aprendizado de máquina para redução da dimensionalidade dos dados, seguida do uso dos resultados obtidos num modelo de Cox. Mais especificamente, adotaremos o método K-means e de redes neurais, numa abordagem inédita, como um pré processamento para a análise da sobrevida no modelo de Cox (COX, 1972).

De acordo com o escopo desta pesquisa e com a classificação dos métodos mencionados, realizamos um levantamento da literatura, buscando publicações científicas que estudaram a sobrevida de pacientes com câncer e lidaram com alta dimensão de dados. A Tabela 1 relaciona as publicações encontradas classificadas em duas diferentes abordagens de análise: os trabalhos

que fazem o uso do modelo semi paramétrico de Cox, como este trabalho, e os que estudam a sobrevida através de modelos em aprendizagem de máquina. Adicionalmente, destacam-se os métodos adotados para lidar com a alta dimensão de dados e o(s) tipo(s) de câncer a que se referem. Note que na maioria das publicações, apenas um conjunto de dados é utilizado.

Tabela 1 – Trabalhos relacionados que realizam análise da sobrevida em pacientes com câncer que tratam alta dimensão de dados.

Método de análise da sobrevida	Pesquisa	Método(s) de seleção ou/e redução utilizado	Dados
Semi Paramétrico Modelo de Cox	Chai *et al.* (2021)	*Denoising Autoencoder Network*	Mama, pulmão, colorretal e outros 10
	Li *et al.* (2021)	LASSO WCGNA	Cervical
	Wang e Liu (2020)	*R-LASSO*	Glioblastoma, esôfago e Colorretal
	Hira *et al.* (2021)	*Variational Autoencoder Network*	Ovário
	Chaudhary *et al.* (2018)	*Autoencoder Network*	Fígado
	Este trabalho	LASSO, K-Means e Autoencoder Network baseado em grupos	Mama
Aprendizagem de máquina	Ramirez *et al.* (2021)	*Graph Neural Network*	Mama, pulmão, colorretal e outros 10
	Katzman *et al.* (2018)	*Deep Neural Network*	Mama
	Jiang *et al.* (2020)	*Variational Autoencoder and Multi-view Factorization Autoencoder*	Pulmão
	Ching, Zhu e Garmire (2018)	*Deep Neural Network*	Mama, pulmão, colorretal e outros 7
	Tong *et al.* (2020)	*Autoencoder Network*	Mama

A grande vantagem do modelo de Cox está relacionada à fácil interpretabilidade, no qual os métodos de análise da sobrevida baseado em modelos de aprendizagem de máquina possuem maior limitação. Por outro lado, os modelos para a analise da sobrevida com aprendizagem de máquina possuem uma melhor performance, na sua maioria avaliado pelo C-Index.

Antes de aprofundarmos nos achados das pesquisas levantadas, é importante destacar o poder dos algoritmos baseados em redes neurais, em especial o *Autoencoder*. Devido à alta dimensão dos dados, algoritmos que buscam relações lineares, como é o caso do PCA (do inglês *Principal Component Analysis*), não entregam bons resultados, quando comparados aos modelos neurais. Um exemplo disto, é a pesquisa feita por Khalili *et al.* (2016) com o intuito de tratar da alta dimensão dos dados de expressão genética. Eles concluem que o método *Autoencoder* possui uma capacidade consideravelmente maior do que o PCA. Na mesma linha, Hu *et al.* (2023), Zhao *et al.* (2021), Wang e Wang (2020) reforçam os resultados favoráveis obtidos com *Autoencoder* para a redução de dimensionalidade da informação genética.

As duas subseções subsequentes, detalham cada pesquisa levantada da Tabela 1, divididas pelo método adotado para a analise da sobrevida: modelo semi paramétrico e modelos de aprendizagem de máquina.

3.1 Pesquisas com uso de modelos semi paramétrico

3.1.1 Chai et al. (2021) Denoising Autoencoder Network

O modelo neural *Denoising Autoencoder Network* é uma rede *Autoencoder* com a inserção de ruídos. A informação de entrada da rede sofre alterações (ruídos) e o modelo deve ser capaz de reconstruir os dados de entrada removendo os ruídos. Além do ruído inserido, o estudo aplica a penalização lasso para evitar *overfitting*, efeito que faz a rede "decorar"o conjunto de dados.

Essa estratégia é utilizada para redução de dimensionalidade de quatro dados genômicos diferentes, sendo mRNA, miRNA, metilação, CNV (do inglês *Copy Number Variation*). Após a redução, o modelo de Cox é aplicado para predição dos riscos. Desta forma, a metodologia aplicada consegue estimar o risco de óbito dos indivíduos, mas não consegue determinar quais os fatores genéticos associados. Devido a essa falta de interpretabilidade, é aplicado o método XGboost baseado nos riscos estimados pelo Cox e nos dados mRNA. Com a seleção das características dada pela regressão da árvore, é possível encontrar biomarcadores com a alta associação da rede.

A Figura 10 ilustra a metodologia aplicada pela pesquisa, no qual a etapa A consiste em realizar a redução e aplicação do modelo de Cox, a etapa B representa a seleção de características relevantes na determinação da sobrevida e, por fim, em C tem-se a associação dos dados para determinar biomarcadores.

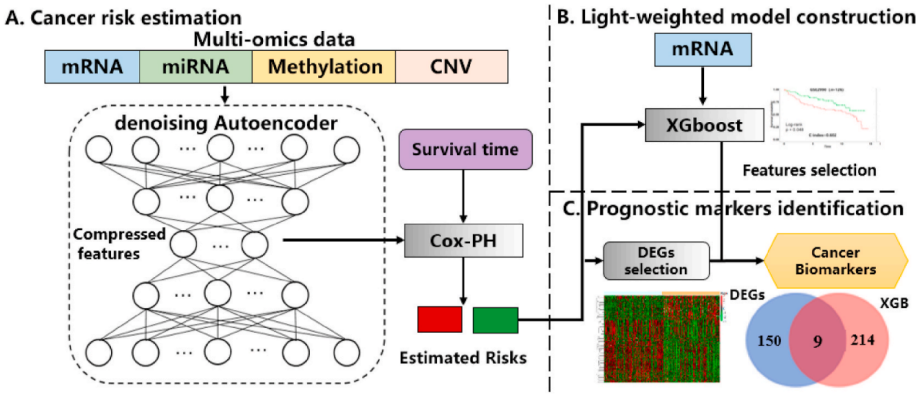


Figura 10 – Metodologia proposta por Chai *et al.* (2021).

Fonte: Chai *et al.* (2021).

O resultado da pesquisa mostra que o método proposto, quando comparado com outros tais como PCA e modelo de Cox penalizado, possui um *C-Index* superior em todos os casos, exceto para o câncer colorretal, dentro dos 15 conjuntos de dados considerados. É sugerida, como trabalhos futuros, a inserção de dados clínicos e uso de imagens patológicas nas análises.

3.1.2 Li et al. (2021) ASSO

O método de penalização lasso ainda é muito utilizado com bons resultados para este tipo de aplicação. Na pesquisa em questão, utiliza-se o modelo de Cox, com penalização lasso e WGCNA (do inglês *Weighted Correlation Network Analysis*), no qual é feita uma analise de uma rede ponderada que identifica correlação entre as expressões genéticas.

Através do agrupamento por WGCNA, são identificados 6 módulos de genes altamente correlacionados. A importância dos genes dentro dos módulos na sobrevida é feita com o Cox penalizado com o lasso. Em seguida, o modelo de Cox multivariado é adotado, identificando um

biomarcador de 6 genes. Diferente da maioria dos trabalhos, este não utiliza o C-Index, mas sim a curva ROC para a análise dos resultados em 1, 5, 10 e 15 anos de sobrevivência.

3.1.3 Wang e Liu (2020) R-LASSO

Esta pesquisa, assim como a anterior, utiliza o lasso e a expressão genética enfatizando o fato do modelo de Cox penalizado não tratar relações de processo biológico. O lasso no modelo de Cox trata apenas as variáveis para melhor descrever a sobrevida, mas não leva em consideração a relação que os genes têm entre si em um processo biológico.

A partir deste argumento, é proposta uma alteração na penalização lasso na qual é inserido um componente que considera a relação entre os genes. Esta é obtida através dos conjuntos de dados KEGG (Kyoto Encyclopedia of Genes and Genomes) e HuRI (Human Reference Interactome) representados em um grafo.

A partir deste grafo, é inserida uma parcela na penalização lasso multiplicando a covariavel pela sua representação no grafo de tal forma que quanto maior sua relação topológica, menor é a penalização. Dessa maneira, a tendencia é que o lasso penalize genes com baixa interação no grafo, ou seja, que tenham baixa relevância em processos biológicos.

Para validar a modelagem proposta, três conjuntos de dados foram utilizados. A nova proposta obteve valores de C-Index maiores ou próximos aos do método apenas com o lasso.

3.1.4 Hira et al. (2021) Variational Autoencoder Network

Variational Autoencoder Network é um modelo semelhante ao *Autoencoder Network*. A diferença é a normalização dos dados da camada latente, que permite a geração de informação, característica que não é aplicada na pesquisa. A proposta é realizar a redução de dimensionalidade de três diferentes fontes, CNVs (do inglês *Copy Number Variation*), mRNA e metilação.

A pesquisa possui diferentes abordagens após a redução de dimensionalidade, uso de algoritmos de agrupamento e classificação para determinar relações da camada latente, analise com o modelo de Cox, classificador SVM para grupos definidos do agrupamento. Por fim, como ilustrado na Figura 11, que contempla toda metodologia, é feita uma analise de correlação para obter biomarcadores (iii). A correlação é feita de maneira linear de acordo com cada valor de entrada, que possui significância clinica na camada latente.

Tal abordagem difere da proposta de nossa pesquisa, no qual temos a interpretação da redução de dimensionalidade diretamente no modelo *Autoencoder* e a determinação dos biomarcadores ("reduzidos"no *Autoencoder*) mais relacionados à sobrevida dos pacientes no modelo de Cox.

Por outro lado, a pesquisa discutida (HIRA *et al.*, 2021) traz uma abordagem distinta, na qual é possível trazer no agrupamento informação de grupos classificados, que podem tornar a

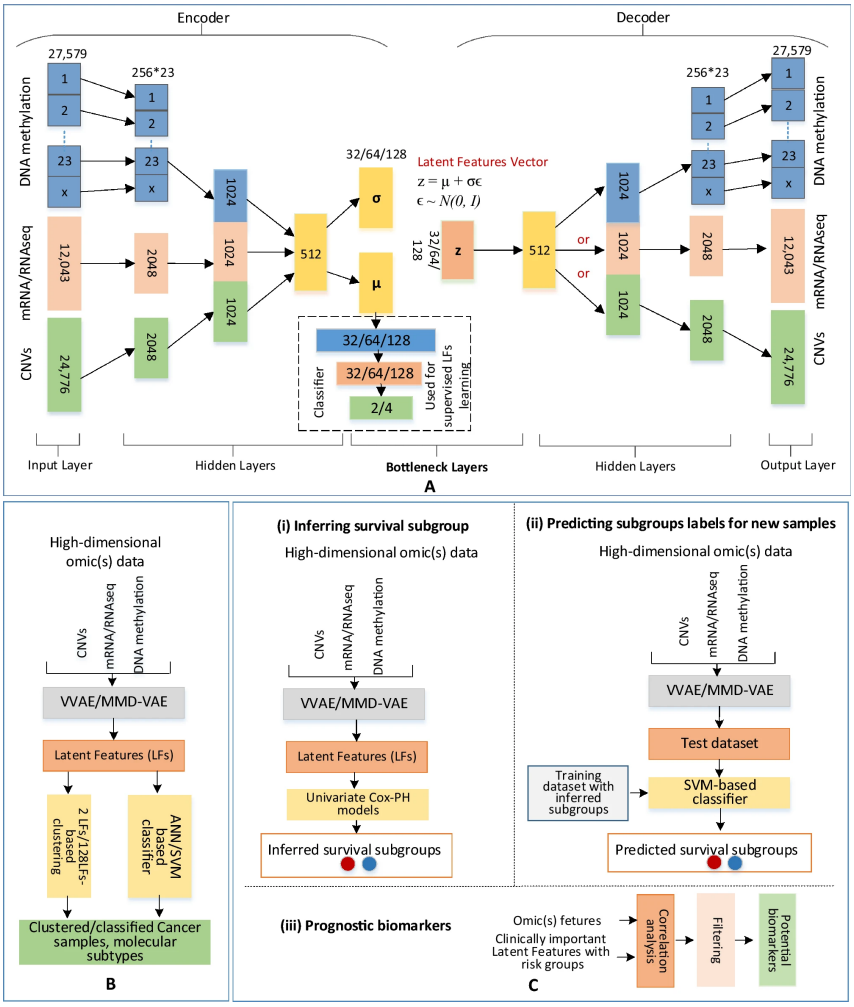


Figura 11 – Metodologia proposta por Hira *et al.* (2021).

Fonte: Hira *et al.* (2021).

análise mais fiel ao risco ou a outros determinantes de interesse.

3.1.5 Chaudhary et al. (2018) Autoencoder Network

Igualmente à nossa pesquisa, a pesquisa de Chaudhary *et al.* (2018) faz o uso do modelo *Autoencoder Network* para reduzir a dimensionalidade da informação genética. De início, uma das suas diferenças está em empregar o uso de três camadas de informação biológica, metilação, MiRNA-seq e RNA-seq, esta última utilizada em nossa pesquisa.

A Figura 12 ilustra a metodologia proposta pelos autores. Como primeiro passo, é criada uma rede *Autoencoder* para a redução das milhares de variáveis biológicas. Posteriormente, as informações reduzidas são analisadas em um modelo de Cox univariado, que com o risco de óbito estimado é agrupado no modelo de K-Means. Depois de definidos os grupos de sobrevivência, é feita uma seleção das características explicativas utilizando a ANOVA, adotando o valor F para ranquear as características. As características mais significativas são entradas para o modelo

SVM, que realiza a classificação dos grupos de risco, definidos pela análise entre Cox e K-Means.

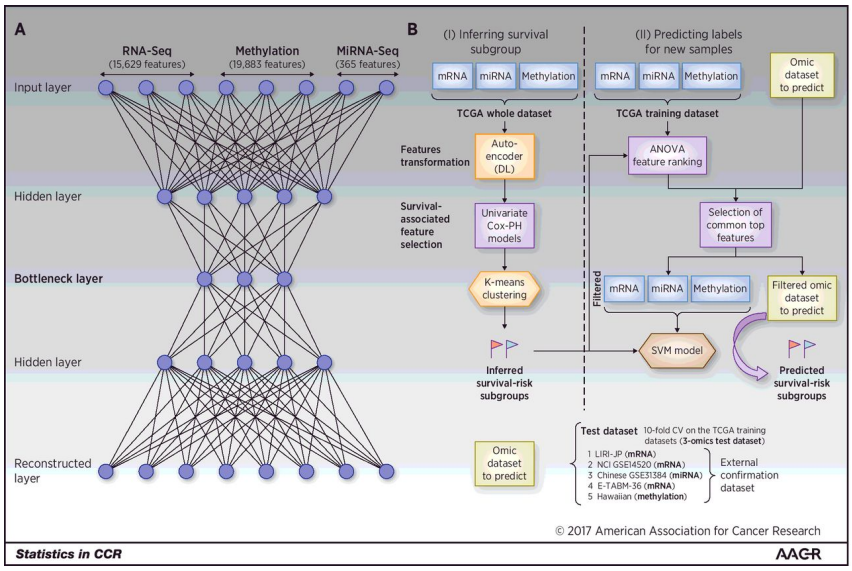


Figura 12 – Metodologia proposta por Chaudhary *et al.* (2018).

Fonte: Chaudhary *et al.* (2018).

Em comparação com nossa proposta, temos apenas em comum o uso do *Autoencoder*. O modelo de Cox foi adotado para um diferente propósito, baseando-se em um modelo mais robusto (SVM) para classificar pacientes em risco. Entretanto, a metologia aplicada serve para validar que o uso do *Autoencoder*, para a redução de dimensionalidade obtém bons resultados. Outro aprendizado obtido desta pesquisa é o uso do K-Means para o agrupamento de risco, que pode se tornar um projeto futuro para evolução de outra metodologia.

3.2 Pesquisas com uso de modelos em aprendizagem de máquina

3.2.1 Ramirez et al. (2021) Graph Neural Network

Cada vez mais aplicam-se algoritmos de aprendizagem de máquina em diferentes contextos. A maior motivação é a capacidade destes modelos lidarem com a alta dimensão de dados. Por este motivo, a adaptação destes modelos no âmbito da analise da sobrevida, na sua maioria modelos neurais, estão ficando populares. A pesquisa em discussão (RAMIREZ *et al.*, 2021) propõe uma abordagem para lidar com os dados de expressão genica e clinica.

A metodologia difere na maneira de representar os dados de expressão. Os dados são transformados em um grafo, partindo de uma matriz de correlação, como ilustrado na Figura 13 (a entrada é um grafo). Com o grafo construído, é inserida mais uma camada de informações: um grafo de interações da GeneMania, trazendo relações entre os genes. As informações da camada genética são unidas em uma camada oculta.

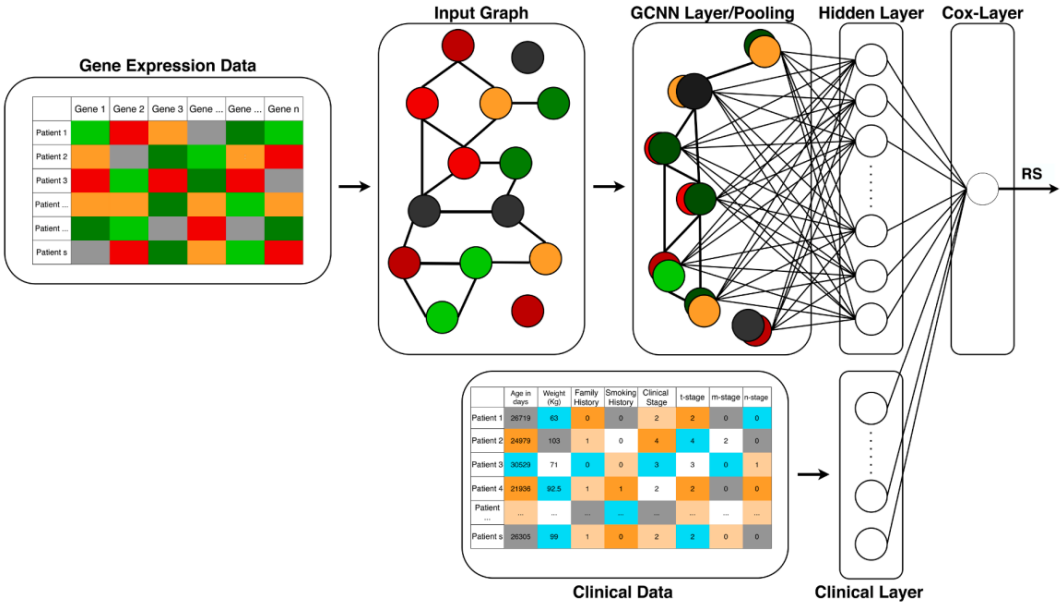


Figura 13 – Metodologia proposta por Ramirez *et al.* (2021).

Fonte: Ramirez *et al.* (2021).

A partir da camada oculta dos dados genéticos e a camada de dados clínicos, é criada uma camada denominada de Camada Cox com apenas uma saída, predizendo o valor do risco de óbito. Esta camada é baseada no conceito do modelo de Cox, no qual trata-se de uma regressão, mas que neste caso os parâmetros de estimação são definidos por uma rede neural. Para que a estimação faça sentido, no âmbito da sobrevida, é necessário realizar ajuste na função de perda da rede neural. No estudo, foi utilizado um pacote em Python chamado de Tensorcox, que busca otimizar a verossimilhança parcial.

A pesquisa foi aplicada em 13 diferentes tipos de câncer, incluindo o de mama, no qual foi concluído que a inserção de informação clinica melhora a análise de sobrevivência.

3.2.2 Katzman et al. (2018) Deep Neural Network

Na maioria das pesquisas, a metodologia está totalmente relacionada a um interesse biológico ou clinico, como a busca de biomarcadores. No entanto, o trabalho de (KATZMAN *et al.*, 2018) tem apenas o interesse de propor uma modelagem, baseada no modelo Cox. A estrutura do trabalho, diferentemente dos outros, traz um viés de modelagem mais matemático. Para melhor validar a modelagem proposta, chamada de DeepSurv (ilustrado na Figura 14), são feitas analises com dados simulados e reais. Além disso, é discutida a linearidade do modelo de Cox e demonstrada a superioridade da modelagem proposta possivelmente devido a essa limitação do modelo tradicional.

Além do fator de implementar um modelo não linear para sobrevida, o estudo busca entregar recomendações de tratamento, através de grupos com diferenças na taxa de sobrevivência.

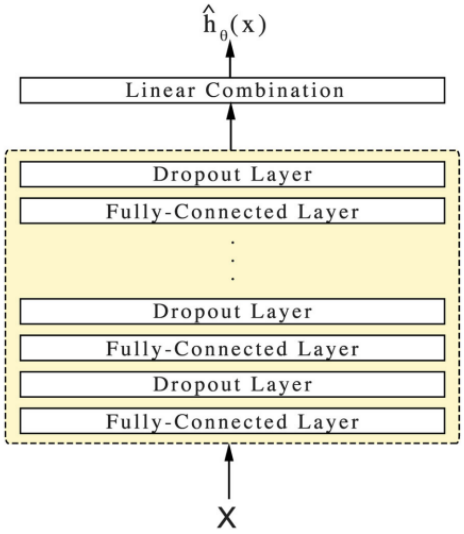


Figura 14 – Modelo proposto por Katzman *et al.* (2018).

Fonte: Katzman *et al.* (2018).

O trabalho proposto (KATZMAN *et al.*, 2018) conta com mais de 500 citações, evidenciando sua importância. Na avaliação da abordagem, a principal métrica de comparação é o C-Index, no qual comparado ao modelo *Random Forest Survival*, obteve um valor superior com os dados de câncer de mama.

A proposta de um modelo neural absorver a predição do risco de sobrevida está na definição da função objetivo. A função proposta segue a lógica da função de verossimilhança parcial. Diferente do modelo clássico, no qual utiliza os estimadores do modelo de Cox na função de verossimilhança, o modelo neural utiliza os pesos. Além dessa equivalência, é utilizada a regularização L2.

A maneira de predizer o risco de sobrevida no modelo neural, é um objeto de estudo futuro para nosso trabalho. Além de obtermos uma redução de dimensionalidade dos dados de expressão genética, teremos a relação da redução com a sobrevida.

3.2.3 Jiang et al. (2020) Variational Autoencoder e Multi-view Factorization Autoencoder

Variational Autoencoder e *Multi-view Factorization Autoencoder* são modelos neurais que codificam as entradas em uma menor dimensão, similares ao utilizado neste trabalho, o *Autoencoder*. Tratar a alta dimensão de dados é uma tarefa complexa. No trabalho de Jiang *et al.* (2020), são tratados quatro conjuntos de dados de alta dimensão: CNV, metilação, MiRNA e mRNA. Propuseram uma arquitetura mais sofisticada para lidar com múltiplas entradas de alta dimensão. A Figura 15 ilustra as três abordagens aplicadas, sendo a primeira (A) o modelo proposto por Ching, Zhu e Garmire (2018), trabalho que será detalhada em uma subseção. Os modelos ilustrados em B e C são as abordagens propostas pelo trabalho, sendo o *Multi-view*

Factorization Autoencoder e *Variational Autoencoder*, respectivamente.

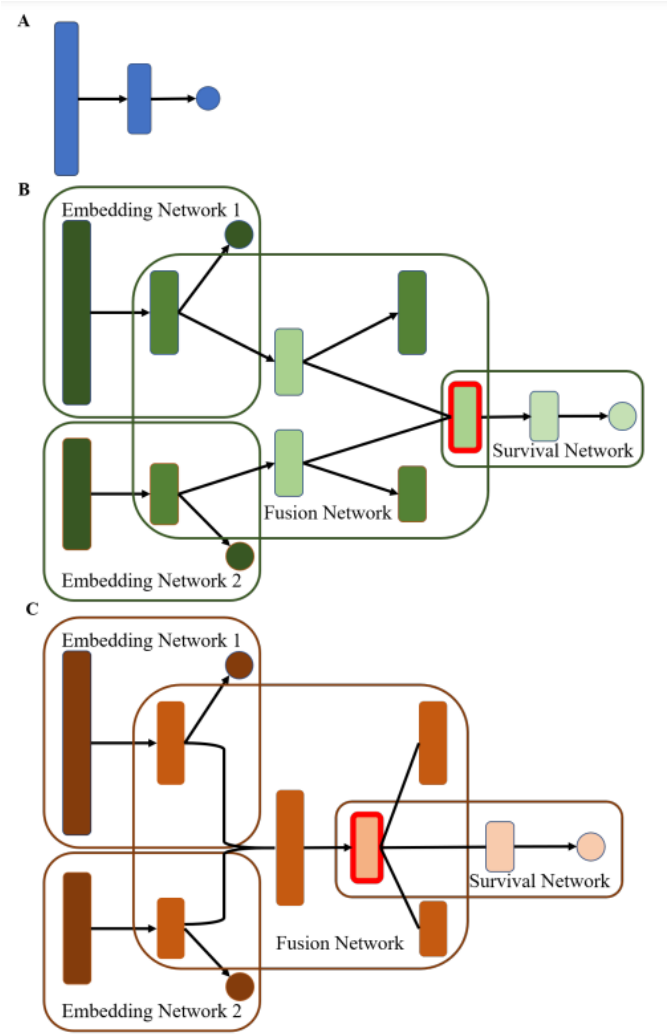


Figura 15 – Modelos propostos por Jiang *et al.* (2020).

Fonte: Jiang *et al.* (2020).

Ambas abordagens propostas (B, C), da esquerda para a direita, seguem a entrada de dados, etapa em que é realizada a redução de dimensionalidade dados. Nesta etapa, é aplicada como função de perda a verossimilhança parcial. Para o treinamento da primeira etapa, das redes *embedding*, todas as outras são congeladas. Da mesma forma, na segunda etapa da rede de *fusion*, o restante é congelado. Nela, é feita a união dos dados, seja por meio de uma *view* (B) ou pela concatenação (C). Por fim, na rede *survival*, é usado o modelo de (A).

Para validar o modelo, os dados utilizados são de câncer de pulmão e a métrica adotada é o C-Index. O modelo proposto com uso da *Multi-view Factorization Autoencoder* obteve um melhor resultado. A partir da conclusão, podemos levar em consideração o uso de diferentes redes, com etapas bem definidas para suavizar o grande volume de dados. A arquitetura da rede é de fato um instrumento de estudo futuro, podendo trazer ganhos significativos, como a pesquisa discutida mostrou.

3.2.4 Ching, Zhu e Garmire (2018) Deep Neural Network

A pesquisa de Ching, Zhu e Garmire (2018) tem mais de 200 citações, sendo um dos primeiros trabalhos a propor o uso de redes neurais com a incorporação da função de verossimilhança parcial.

A arquitetura da rede é convencional, como ilustrado na Figura 16. A proposta é tratar a alta dimensão de dados com as camadas ocultas, similar a rede *Encoder*, mas que ao final será reduzida a uma combinação linear, que tem como resposta o risco do paciente. A combinação linear é baseada na última camada oculta, no qual para cada nó é representado como um estimador, similarmente ao modelo de Cox. A otimização, então, é feita a partir dos resultados de saída da rede.

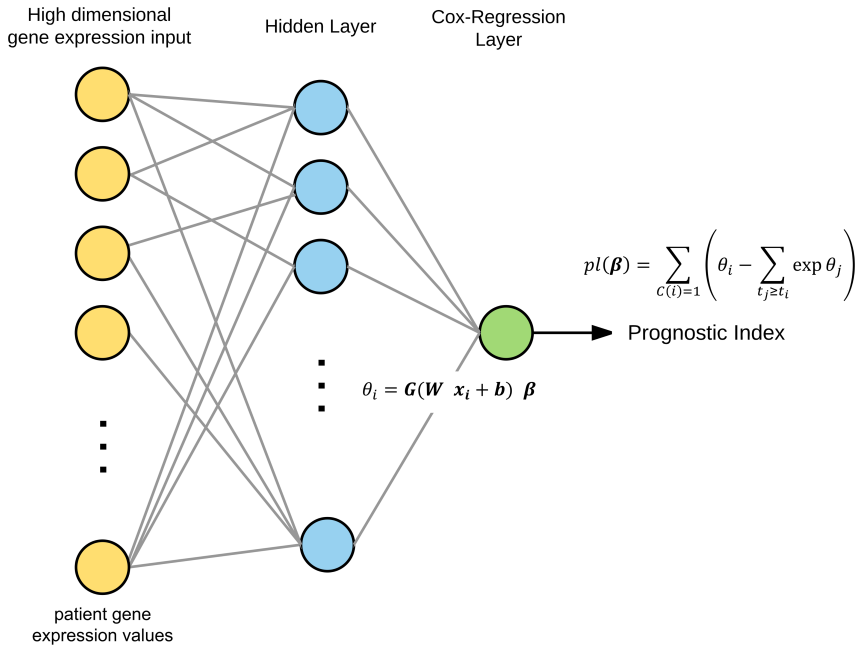


Figura 16 – Modelo proposto por Ching, Zhu e Garmire (2018).

Fonte: Ching, Zhu e Garmire (2018).

3.2.5 Tong et al. (2020) Autoencoder Network

O trabalho de Tong *et al.* (2020) emprega também o modelo *Autoencoder*, além do mesmo conjunto de dados usado nesta pesquisa. Embora as metodologias estejam relacionadas, existem diferenças que serão o principal ponto de discussão nesta subseção.

A Figura 17 ilustra a proposta. Como dito acima, a base de dados é a mesma, porém outras três fontes são utilizadas: metilação, miRNA e CNV, e não foram considerados os dados clínicos (inclusos em nossa proposta). É feito um pré processamento de dados, normalização e a separação em quatro *folds*.

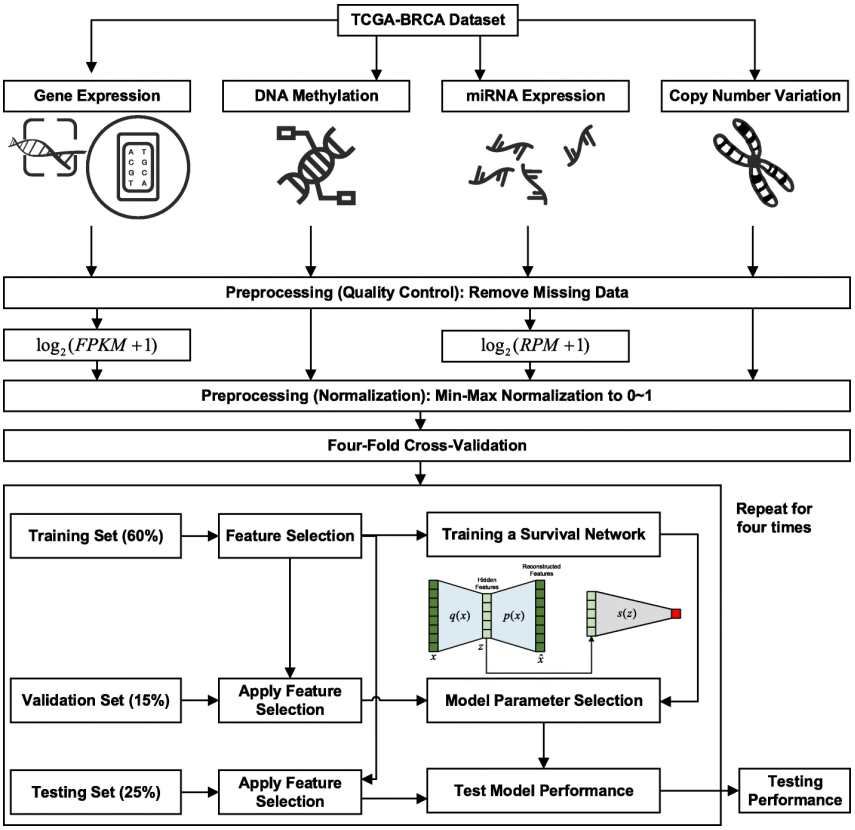


Figura 17 – Modelo proposto por Tong *et al.* (2020).

Fonte: Tong *et al.* (2020).

A partir dos dados refinados, é feita uma seleção de características explicativas através do método PCA, selecionando os primeiros 100 componentes principais. O que leva ao uso da rede *Autoencoder*, abordada de duas formas distintas. Na primeira, é feita para cada fonte de dados (*-omic*) uma rede e ao final é concatenada a informação (*ConcatAE*). Na segunda abordagem, as redes são unidas e a informação de entrada para cada rede deve ser reconstruída se baseando também nas outras redes (*CrossAE*).

Para validar a arquitetura proposta, o estudo realiza a aplicação na base MNIST, um conjunto de dados conhecido na área de aprendizagem de máquina, composto por imagens de números. O modelo obteve bons resultados, com acurácia e precisão acima de 0.95. Mas devido o contexto da sobrevida, no qual foi utilizada uma rede neural (CHING; ZHU; GARMIRE, 2018), o estudo faz uso da métrica C-Index.

Os resultados mostraram que a abordagem de concatenar as informações (*ConcatAE*) foi melhor. A pesquisa discutida entrega uma forma de analisar a sobrevida com diversas fontes de dados, mas antes de utilizar o modelo neural aplicou método linear (PCA). Um objeto de estudo futuro para nossa pesquisa é abordar o uso de mais fontes de dados, com o uso de múltiplas redes que são concatenadas, como também o uso do modelo de Cox baseado no modelo neural.

3.3 Considerações finais

A quantidade de pesquisas relacionadas a este trabalho são consideráveis. Na sua grande maioria, trabalhos recentes, dos últimos 3 ou 4 anos. Cada vez mais, novas abordagens estão sendo propostas. As que foram discutidas neste capítulo são aquelas mais aderentes a essa pesquisa e com maior citação na comunidade científica. Existem outros trabalhos relacionados, mas devido ao grande volume e a particularidades que fogem do contexto desta pesquisa, não foram abordados neste levantamento bibliográfico.

Para mais detalhes, o leitor pode encontrar o tratamento de valores censurados (ZHANG *et al.*, 2022; WANG *et al.*, 2021), no qual é dada importância ao tratamento dos dados e a alta dimensão. Abordagens para construção de grafos de interação genética (XING *et al.*, 2022; JUEXIN *et al.*, 2021; YUAN; BAR-JOSEPH, 2020), que buscam encontrar padrões em processos biológicos.

Pesquisas nestes contextos trazem também maneiras de evoluir este trabalho, tais como na forma de relacionar grupos e interações genéticas, na evolução de tratamento de variáveis faltantes e outros fatores. Tais tópicos ainda serão instrumentos de novas pesquisas futuras, fora do escopo desta dissertação.

A partir do vasto levantamento bibliográfico realizado, algumas decisões foram tomadas para este trabalho. O uso da métrica C-Index para avaliação e comparação dos modelos, valor que independe do formato do modelo, foi uma das decisões principais. A aplicação do algoritmo K-Means com uso da distancia Euclidiana (OH; PARK; ZHANG, 2021; LU *et al.*, 2004) é outro exemplo, assim como parte do pré processamento de dados (TONG *et al.*, 2020).

No capítulo seguinte, será apresentado o conjunto de dados, tais como a fonte dos dados, a estrutura e algumas estatísticas. Posteriormente, a metodologia é apresentada, na qual grande parte foi discutida aqui na apresentação dos trabalhos relacionados.

CAPÍTULO

4

CONJUNTO DE DADOS

Os dados utilizados neste trabalho são disponibilizados pela *National Cancer Institute* (NCI), disponíveis no site <https://portal.gdc.cancer.gov>. Os dados provém do projeto *The Cancer Genoma Atlas* (TCGA), iniciado em 2005 com o intuito de catalogar diferentes dados para diversos tipos de câncer, como dados clínicos, genéticos, mutações e outros. Neste projeto utilizamos dois conjuntos de dados. A primeira base contém informações clínicas e a segunda, as informações genéticas de diversos pacientes. A base utilizada é denominada TCGA-BRCA, que são dados de pacientes com câncer de mama.

4.1 Dados clínicos

Os dados clínicos possuem uma grande variedade de informações, com 154 caraterísticas sendo uma delas para a identificação do paciente. O conjunto todo possuí 1219 registros. Não abordaremos todas as características clinicas devido a grande quantidade de dados faltantes e informações com mesmo significado.

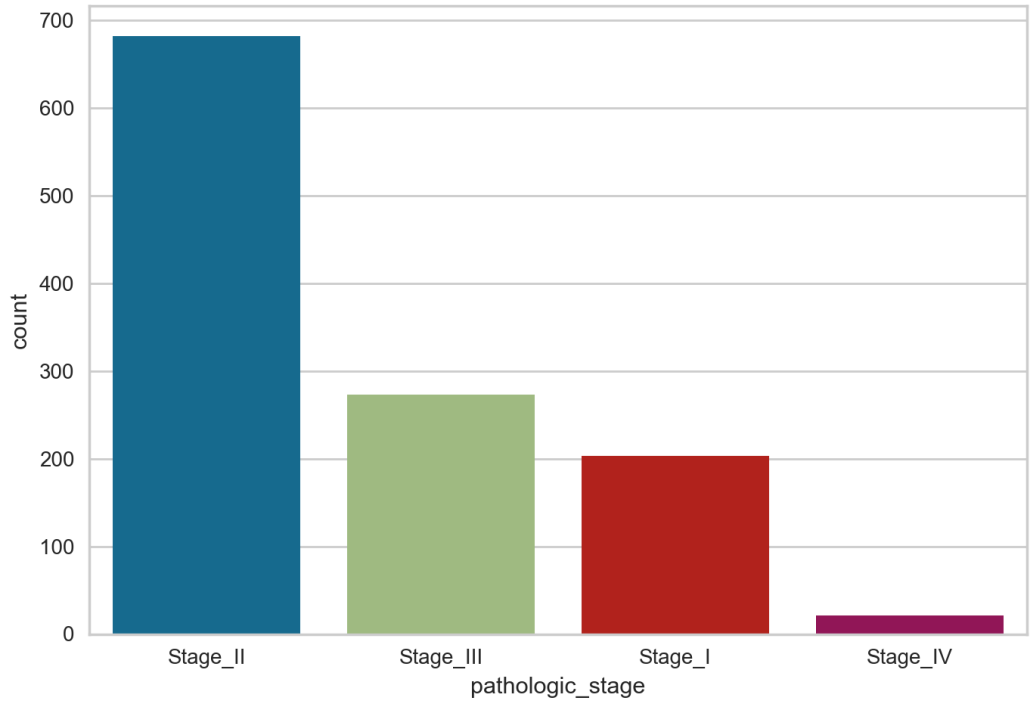
A grande maioria dos pacientes é do sexo feminino, compondo 98%. Os outros 2% são do sexo masculino, sendo apenas 13 homens com ocorrência de câncer de mama. Posteriormente o sexo masculino foi totalmente removido pelo pré processamento de dados, pois não se enquadraram em nossos critérios de inclusão no estudo.

Uma das variáveis mais importantes nesta base é a **pathologic stage**. Nela, temos a informação do estadiamento do câncer e interpretar a evolução do carcinoma: se é inicial ou um quadro mais avançado. A maneira como é determinado o estadiamento é um processo de várias etapas, que envolve analisar os tipos de células do câncer, sua localização, tamanho, entre outras variáveis.

De acordo com a Figura 18, os dados possuem uma maior concentração de casos em estadiamento II. Pacientes no estágio II possuem o câncer em desenvolvimento, que está se

espalhando para tecidos próximos do tumor. A distinção entre os estágios I, II e III são mais sutis, diferentemente do estágio IV, que mostra uma alta agressividade do câncer que se espalhou por outras partes do corpo, tornando-o um câncer metastático.

Figura 18 – Quantidade de pacientes por estadiamento

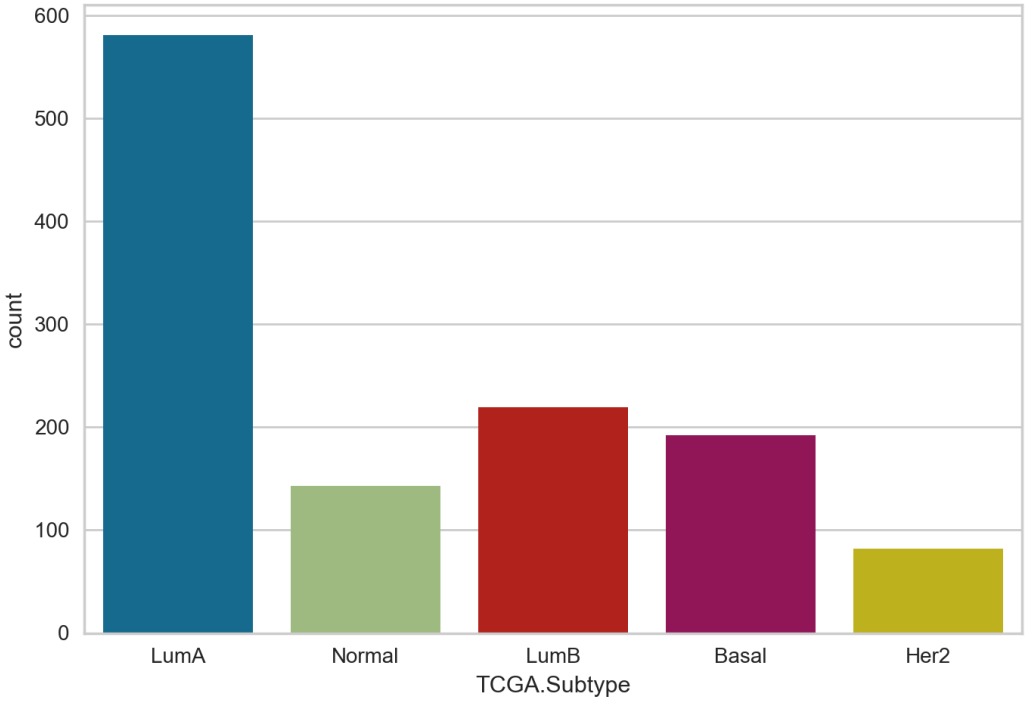


Fonte: Elaborada pelo autor.

Um outro fator ligado ao câncer é a variável **subtype**, que caracteriza os subtipos do câncer de mama. Essa classificação é feita em nível genético, de acordo com as células cancerígenas. A tipagem do câncer é importante para determinar tratamentos, podendo estar ligados a agressividade do câncer. Por exemplo, o subtipo Basal demonstra ser mais agressivo comparado aos subtipos LumA e LumB (YERSAL, 2014). Na base de dados utilizada nesta pesquisa, a maioria dos casos se concentram no subtipo LumA. Em segundo lugar, aparece o subtipo LumB. A Figura 19 mostra o histograma dos subtipos, em que pode-se notar que Her2 é o subtipo com a menor ocorrência na base.

Quando mencionamos pessoas com câncer, a idade é um fator que possui alta relevância, pois indivíduos com maior idade possuem uma menor resistência a mutações, o que implica na maior probabilidade de evolução de células cancerígenas (BENZ, 2008). Porém, nem sempre a ocorrência da doença em pessoas mais novas determina um melhor prognóstico (CHEN *et al.*, 2016). Na Figura 20, vemos que a grande maioria dos pacientes possuem idade entre 60 e 80 anos.

Figura 19 – Quantidade de pacientes por subtipos



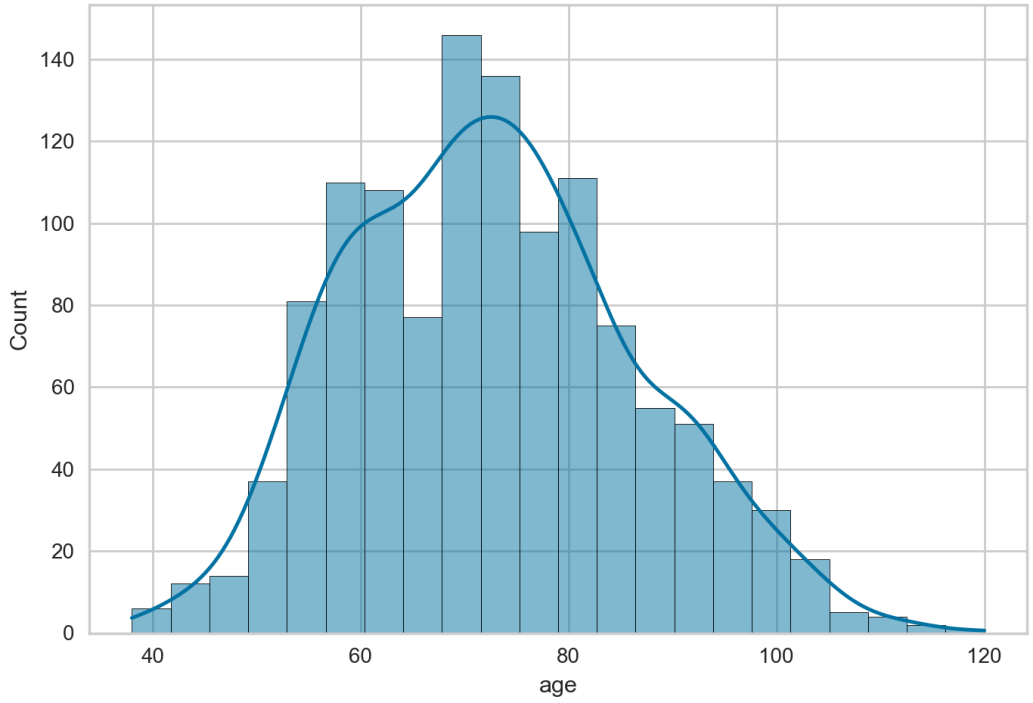
Fonte: Elaborada pelo autor.

As idades mostram como os pacientes estão distribuídos, mas não revela uma importante informação: quais são as idades com estágios iniciais e avançados. De maneira intuitiva, podemos sugerir que casos mais avançados acontecem com maior frequência em pacientes com idades maiores, mas a Figura 21 mostra que a idade não possui uma alta relação com o estadiamento, visto que em idades avançados houve maior frequência de pacientes com estágio I do que com estágio IV.

Uma importante característica no estudo da sobrevida é a informação quanto ao tempo de monitoramento do paciente. Neste estudo, possuímos registros de até 8605 dias, ou seja, temos pacientes sendo acompanhados por mais de 23 anos. Por mais que o conjunto de dados possua a ocorrência de longos registros, a maioria se concentra nos primeiros 2000 dias, correspondendo a um acompanhamento de quase 6 anos (vide Figura 22).

Observa-se a ocorrência de 199 eventos no conjunto de dados, neste caso a ocorrência de óbitos, o que representa 16% do total. O restante são casos de censura, representando a maioria das informações. Devido ao nosso estudo não utilizar dados do estadiamento IV, tal número é reduzido, que após a remoção de informações ausentes do acompanhamento, resulta em 44 eventos restantes.

Figura 20 – Histograma da idade dos pacientes



Fonte: Elaborada pelo autor.

As características descritas nesta seção, sobre os dados clínicos, serão incluídas na predição da sobrevida. A ilustração da base de dados apresentada na Tabela 2, mostra a estrutura discutida, sendo uma parcela do conjunto de treinamento do modelo.

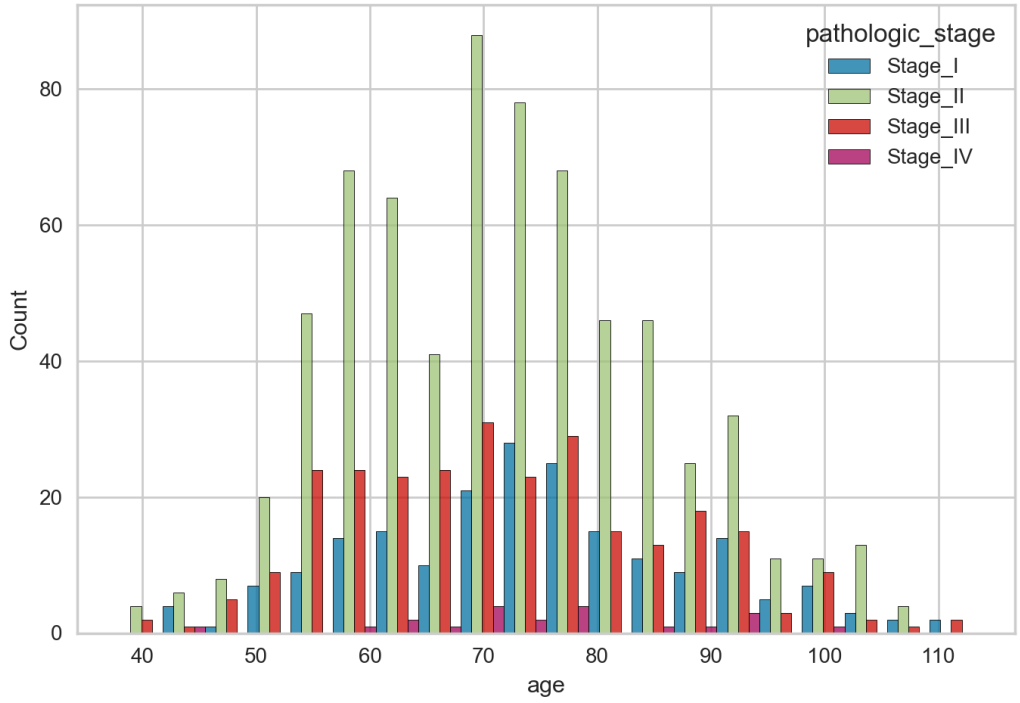
Tabela 2 – Parte da amostra do conjunto dos dados clínicos

bcr_patient_barcode	days_last_follow_up	TCGA.Subtype	pathologic_stage	age
TCGA-E9-A295	375	LumA	Stage_II	82
TCGA-BH-A0HF	727	LumA	Stage_I	91
TCGA-A8-A08H	0	Normal	Stage_II	81
TCGA-A7-A6VW	285	Basal	Stage_II	57
TCGA-EW-A1OY	908	LumB	Stage_II	76

4.2 Dados genéticos

Os dados genéticos dos pacientes referem-se aos níveis de expressão dos seus genes. Esses dados são provenientes de um processo chamado de sequenciamento de nova geração ou NGS (do inglês, *Next-Generation Sequencing*). Com o NGS, tornou-se fácil e barato o

Figura 21 – Histograma da idade dos pacientes por estagiamento



Fonte: Elaborada pelo autor.

sequenciamento genético. Desta maneira, podemos obter dados de expressão gênica de todos os genes presentes na região do tumor.

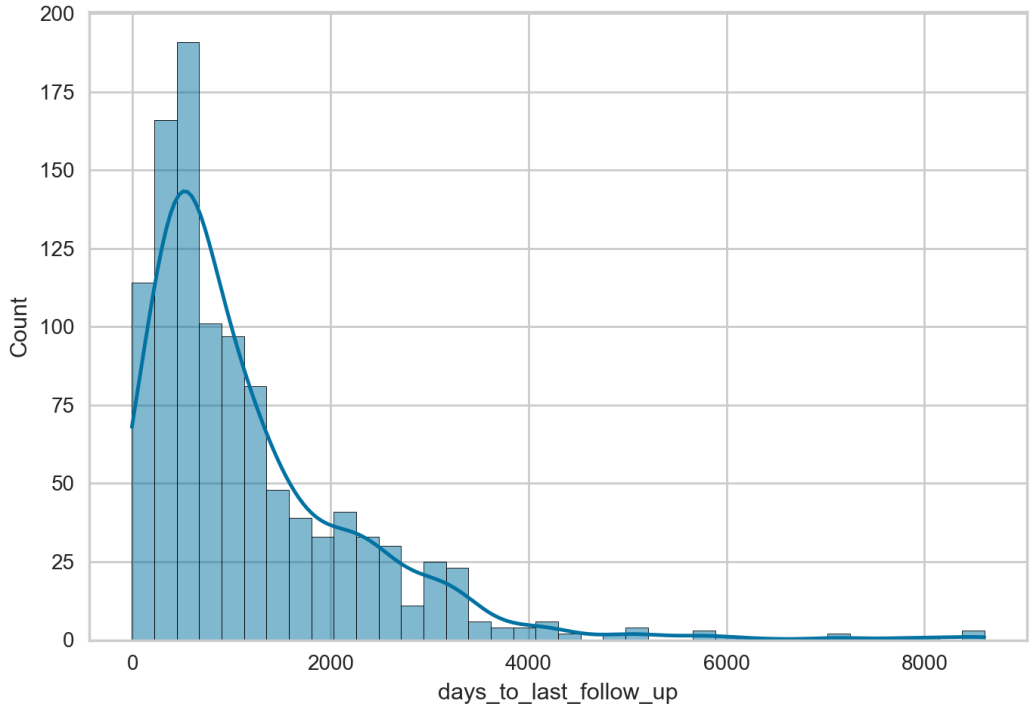
Na base de dados, são representados 60483 genes separados em dois grandes grupos, por suas funções. Cerca de 19595 genes estão associados em processos que codificam proteínas. Esses genes possuem papel fundamental, visto que as proteínas estão presentes em diversos processos biológicos do corpo, como a criação de anticorpos. Os genes restantes, aqueles que não são codificadores de proteínas, estão relacionados a processos internos da célula.

Os dados estão estruturados em formato de matriz matriz, em que o paciente representa uma dimensão e os genes, a outra. Os elementos da matriz contêm os valores de expressão do gene. A Tabela 3 é uma amostra dos dados de expressão genética. Nela, podemos ver os pacientes representado nas colunas e os genes nas linhas.

Tabela 3 – Amostra do conjunto de dados de expressão genica

gene_id	TCGA-3C-AAAU	TCGA-3C-AALI	TCGA-3C-AALJ	TCGA-3C-AALK	TCGA-4H-AAAK
ENSG00000000003.13	2.476427	2.420345	12.549974	12.210248	11.642803
ENSG00000000005.5	0.021430	0.032542	1.403222	0.046197	0.190965
ENSG00000000419.11	26.827419	39.609199	46.289504	21.397945	24.470661
ENSG00000000457.12	2.789583	11.113364	2.941420	4.333150	3.647937
ENSG00000000460.15	1.118863	2.401440	2.024126	1.648454	1.428980

Figura 22 – Histograma do tempo em dias de acompanhamento dos pacientes



Fonte: Elaborada pelo autor.

O processo de digitalização destes dados envolve diversos processos químicos, em equipamentos específicos. O sequenciamento é feito na contagem de genes que são posteriormente normalizados por meio do método FPKM (*Fragments Per Kilobase per Million mapped fragments*). Os genes, conforme amostra, são baseados na representação disposta pela anotação Ensembl. Entretanto, essa representação não é muito adotada no meio cientifico e será realizado o processamento para obter o *Symbol ID* equivalente, mais comumente utilizado.

CAPÍTULO

5

METODOLOGIA

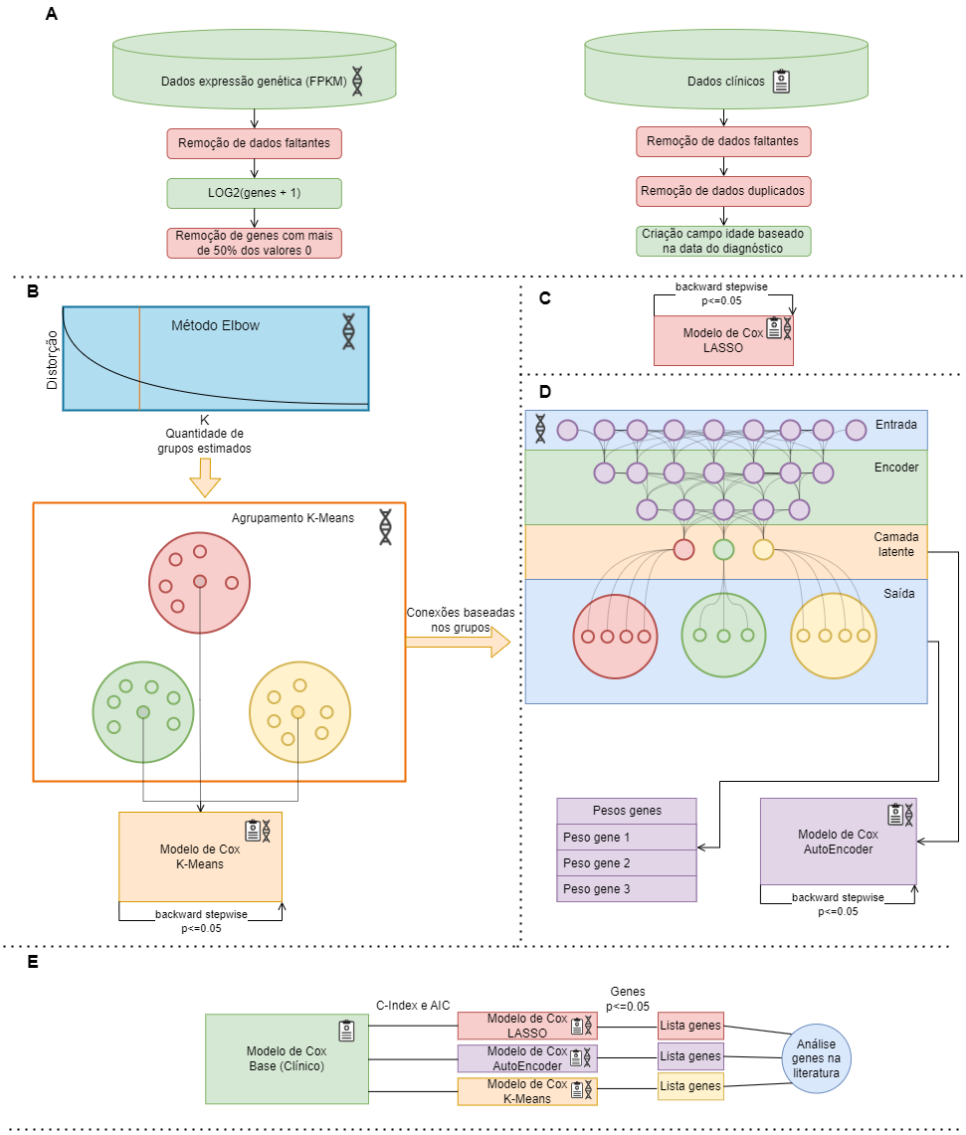
No capítulo de metodologia vamos discutir as etapas envolvidas para atingir o objetivo do trabalho. De maneira macro, será abordado a construção das abordagens para redução de dimensionalidade, as métricas e a analise de sobrevivência. De forma geral, a metodologia desta pesquisa consiste em aplicar o método de Cox utilizando duas camadas de informação: clínica e genética.

Este trabalho se concentra em métodos para tratar a informação genética, para possibilitar a analise de sobrevida. A metodologia é ilustrada na Figura 23, em que diversas etapas estão envolvidas.

Estas etapas são: (A) o pré processamento dos dados, em que as descrições em vermelho são processamentos que reduzem a quantidade de registros da base e as verdes são transformações feitas, (B) abordagem com o método de agrupamento K-Means, onde um gene de cada grupo será selecionado como seu representante para ser inserido no modelo de Cox, (C) abordagem do modelo de Cox com a penalização lasso, (D) a abordagem que utiliza a metodologia de rede neural *Autoencoder*, definindo as ligações da camada latente a partir do agrupamento do K-Means da etapa B, inserindo as informações da camada latente no modelo de Cox e, por fim, (E) a comparação dos resultados destes métodos com os obtidos com o modelo de Cox com os dados clínicos e com penalização lasso quando se consideram os dados genéticos.

Todas as etapas serão discutidas nas próximas seções, com exceção da etapa E, que é discutida em cada uma das etapas do Capítulo 6, de maneira dissertativa, apresentando os resultados dos modelos e os comparando, com base nas métricas C-Index e AIC. E buscando referências para avaliar os resultados com a literatura, baseado nas informações genéticas obtidas.

Figura 23 – Metodologia para a redução de dimensionalidade dos genes para aplicação no modelo de Cox. (A) pré processamento dos dados, (B) seleção de genes baseado nos grupos do K-Means, (C) modelo de Cox com penalização lasso, (D) modelo *Autoencoder* baseado no grupo ótimo do K-Means e (E) avaliação dos modelos de Cox gerados com modelo base (clínico).



Fonte: Elaborada pelo autor.

5.1 Pré processamento

Como visto no Capítulo 4, os dados clínicos possuem uma grande quantidade de características: cada paciente contêm 154 informações clínicas. Entretanto cerca de 70% dos dados são observações faltantes.

O número de registros é de 1219, porém há casos de repetições de pacientes. Assim, foram selecionados apenas pacientes com registros únicos e os últimos registros daqueles pacientes que tinham repetição. Após essa seleção, o número de registros é de 1098, garantindo a unicidade de registro por paciente, e consequente independência entre as unidades amostrais.

Na base de dados são encontradas informações de pacientes com estágio metastático (IV). Estes serão filtrados, visto que o objetivo é analisar pacientes em estágios não terminais.

Na Tabela 4, são mostrados os dados clínicos que serão considerados na predição da sobrevida dos pacientes. Ao todo, foram consideradas 5 caraterísticas como variáveis explicativas e as informações **days_to_last_follow_up** e **vital_status.clinical**, que são necessárias para as definições do evento e tempo de seguimento no modelo de Cox.

Tabela 4 – Caraterísticas clinicas usadas para a predição da sobrevida

Atributos	Tipo	Descrição
days_to_last_follow_up	discreta	dias de acompanhamento do paciente
vital_status.clinical	nominal	estado do paciente, vivo ou óbito
TCGA.Subtype	nominal	subtipos do câncer de mama
pathologic_stage	ordinal	estágio do câncer
age	discreta	idade do paciente

As características **TCGA.Subtype** e **pathologic_stage** são variáveis que não estão em um domínio numérico e são tratadas como categorizadas para que seja possível sua inserção no modelo de Cox. A forma utilizada para este tratamento foi a técnica *one-hot-enconding*. Neste algoritmo, criam-se n novas características binárias, cada uma representando uma categoria de resposta possível. Assim, para as novas variáveis criadas, a respectiva categoria de resposta é representada como 1, e 0 para as demais respostas.

No modelo de Cox, apenas $n-1$ dessas novas caraterísticas são inseridas como variáveis explicativas, sendo uma delas (a que não foi inserida no modelo) considerada a categoria de referência para a comparação dos riscos de óbito. Quando um paciente possuir o valor referencia, todas as outras $n-1$ características criadas deverão estar com o valor 0.

Nos dados genéticos, foram aplicados os seguintes processamentos: filtragem para genes que codificam proteínas e remoção de genes com mais de 50% de valores faltantes.

A base de dados foi separada em 4 *folds*, no qual cada *fold* possui o mesmo número de eventos (óbitos), devido a alta taxa de censura na base. Para a escolha do modelo foi obtido 4 resultados, realizando 4 iterações, em que 3 *folds* são para treinamento e o *fold* restante para teste, dos 4 resultados obtidos foi realizado a média da métrica C-Index para seleção.

5.2 Abordagem por agrupamento

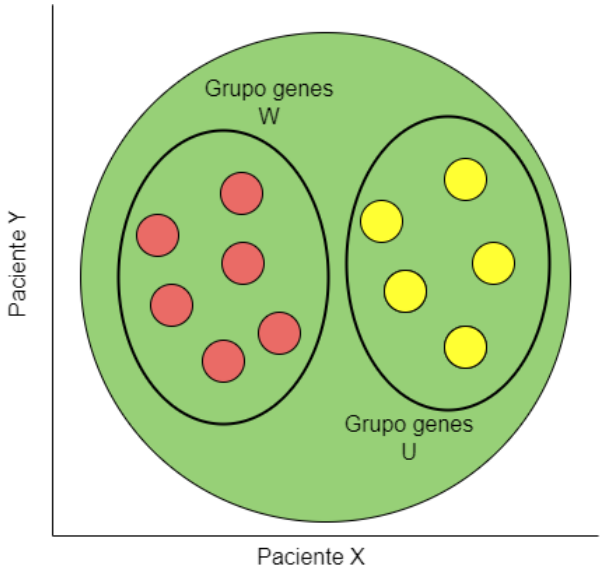
Nesta seção discutiremos sobre a seleção de variáveis utilizando o método de agrupamento, com o algoritmo K-means. Este método representa seus dados em um plano $\mathbb{R}^p$, onde p é a dimensão dos dados. Após a redução do número de genes, temos $p = 17433$.

Os pacientes contribuem com sua expressão gênica para o possível agrupamento de genes. Dessa maneira possuímos $i = (1,2,...,1004)$ pacientes, onde o paciente p_i conta com

18656 genes de expressão proteica. Podemos considerar que a entrada para o modelo K-means consiste na matriz X_{gp}, onde g são os genes e p os pacientes.

Um gene é representado por um ponto u_g no plano $\mathbb{R}^p$, formado pelo vetor $u_g = (p_{11}, p_{21}, ..., p_{i1})$. Desta maneira, o agrupamento consiste em identificar genes que possuem similaridade na sua expressão, baseado nos pacientes com câncer de mama. A Figura 24 exemplifica a abordagem. Vemos dois grupos formados por genes, e a identificação dos grupos de genes é baseada nos níveis genéticos expressos pelos pacientes.

Figura 24 – Abordagem utilizando K-means



Fonte: Elaborada pelo autor.

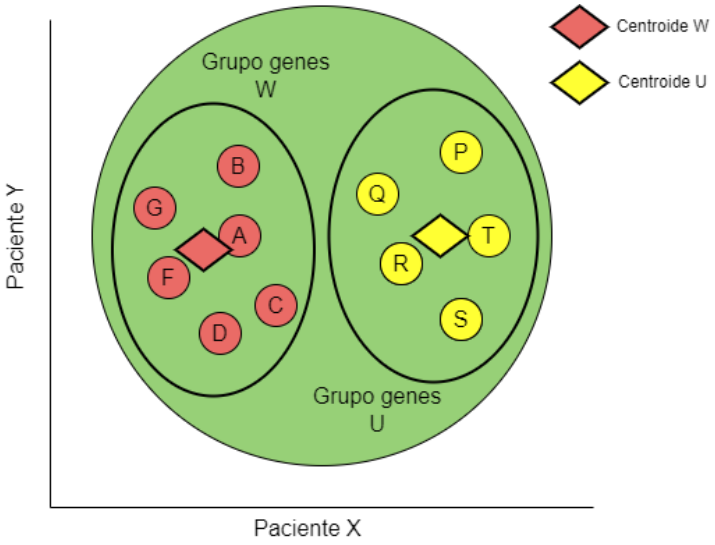
O método de agrupamento adotado utiliza a distância Euclidiana para estimar os grupos que serão formados. Uma das tarefas dentro do algoritmo K-means é definir o valor de k, que representa a quantidade de grupos. O método utilizado para estimar esse valor é o Elbow. Nesse método, escolhe-se o k como um equilíbrio entre o número de grupos e a variância criada pelos grupos definidos. Ou seja, se com um determinado número de grupos tem-se alta variância, podemos considerar que há a necessidade de inserir um novo grupo, assim o valor de k deve ser maior. O contrário é quando a variância não muda significativamente quando há a inserção de novos grupos, pois está ocorrendo uma especificidade, e isto ocasiona *overfitting*.

A escolha do k, então, é dada no momento em que a inserção de novos grupos ainda sejam relevantes, ao mesmo tempo que não especifique os dados. Ao estimar a quantidade de grupos, obteremos um agrupamento de genes e dessa maneira iremos realizar o processo de seleção dos genes.

O K-means ao iniciar sua execução cria k pontos μ, que são chamados centroides. Ao finalizar a sua execução, estes pontos estarão ao centro do seu respectivo grupo, de maneira que o ponto μ seja o ponto médio do grupo. Iremos considerar que o gene mais próximo do centroide

será o selecionado para a o modelo de Cox. Podemos ver na Figura 25 que o gene selecionado no grupo W consiste no gene A e no grupo U o gene T, pois eles são os genes mais próximos do centroide.

Figura 25 – Seleção dos genes com o K-means



Fonte: Elaborada pelo autor.

Com essa abordagem, ao escolher o gene mais próximo do centroide, conseguimos remover alguns possíveis *outliers* na solução, visto que os genes escolhidos são aqueles que mais se aproximam do centro do seu grupo, já que *outliers* são pontos que fogem do centro do grupo.

5.3 Abordagem com penalização

A abordagem por penalização utiliza o Lasso. Esta penalização, como visto no Capítulo 2, realiza alterações nos coeficientes, afim de minimizar o erro da regressão. O lasso força o modelo a selecionar apenas importantes características, ou seja, se a aquela característica não agrega na regressão, e tem um valor de penalização maior que o seu retorno na função objetiva, sua estimativa de β tende a ser zerada.

Não muito diferente do método de agrupamento, o lasso também possui um hiperparâmetro, que é o fator de penalização. A determinação deste hiperparâmetro é complexa e na maioria dos estudos a sua busca é feita realizando testes empíricos. Em aprendizagem de máquina, a maioria dos métodos possuem hiperparâmetros e é comum realizar um processo chamado de *tuning* ou *hyperparameter optimization* para a seleção desses valores. Neste caso, o lasso possui apenas um hiperparâmetro, o que torna mais fácil de encontrar valores adequados.

A maneira de seleção do hiperparâmetro, λ, está totalmente ligada à predição da sobrevida, de maneira que com um valor de λ_k teremos um valor de erro E_k. Então, de maneira linear, podemos variar n valores de λ e aquele que apresentar o menor erro será o selecionado.

Neste caso, é necessário definir qual será o valor de n. Afim de atingir todo o espaço de busca, iremos definir um n suficientemente grande tal que faça todos as estimativas dos parâmetros de regressão, β's, serem zeradaos e a medida que o valor de $n \rightarrow 0$ as características serão selecionadas.

O passo de tamanho entre os valores de λ é definido como 0.1, uma variação razoável, visto que pequenas variações podem não alterar a função objetivo e altos valores perdem soluções.

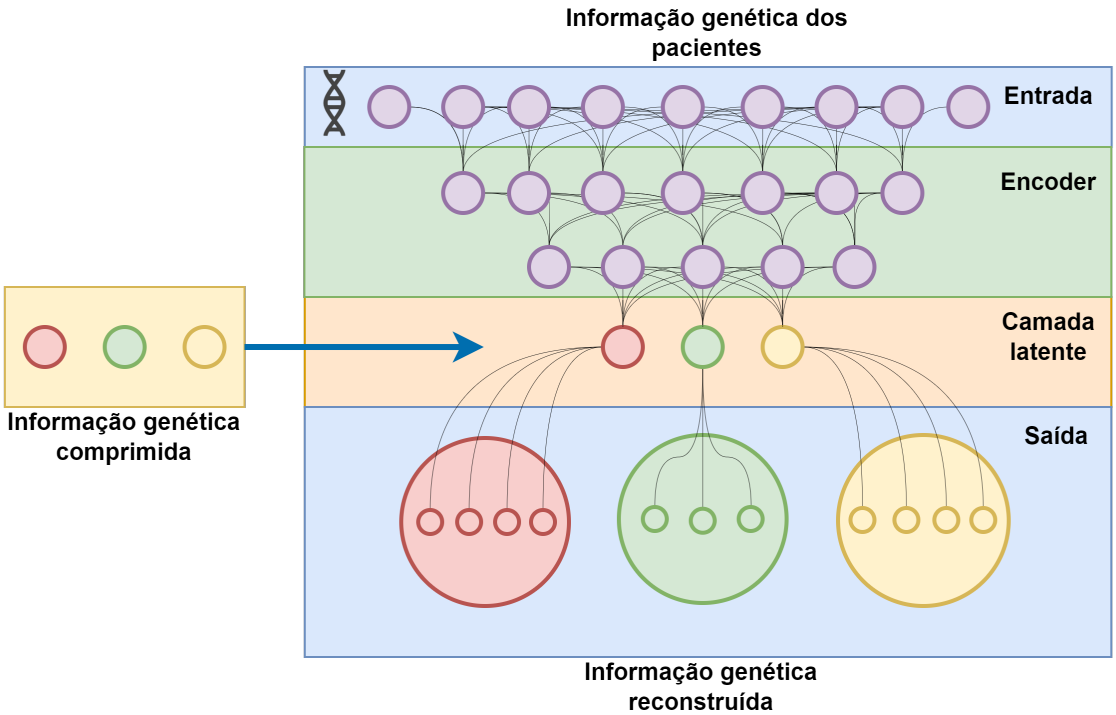
Para determinar quais serão as características selecionadas para o modelo de Cox, serão consideradas apenas características com estimativas de β diferentes de zero. Assim, todas aquelas em que a penalização zerou o valor de β serão desconsideradas na seleção.

5.4 Abordagem por rede neural baseada em grupos

O *Autoencoder* é uma arquitetura de rede neural que consegue realizar a compressão de dados, visto que a arquitetura da rede neural pode ser configurada para comprimir os dados e depois restaurá-los na camada de saída. Desta forma, a execução do algoritmo *backpropagation* está totalmente atrelada ao erro da reconstrução dos dados.

Diante disso, a missão é comprimir os dados sobre as informações genéticas dos pacientes com câncer de mama. A Figura 26 mostra de maneira simplificada a aplicação da rede.

Figura 26 – Abordagem utilizando *Autoencoder* para redução de dimensionalidade da expressão gênica



Fonte: Elaborada pelo autor.

A rede neural busca a reconstrução dos valores observados de expressão dos genes na sua camada de saída. O erro para que a rede convirja está totalmente ligado à distância da expressão gênica que o modelo produz com os valores reais dos pacientes, os valores da entrada.

Diferente dos métodos anteriores, não teremos a seleção dos genes para a análise da sobrevida. Serão obtidos grupos de variáveis (latentes) que representam um determinado conjunto de genes predominantemente. As variáveis latentes são calculadas nos nós da camada latente da rede, como ilustrado, com a informação genética comprimida.

Ao representarmos na camada de entrada os genes, seu número de nós deve ser definido como a a quantidade de genes, ou seja, N é igual a 18656, o número de genes que estão ligados a processos de produção proteica. Na camada de saída, é feita a reconstrução dos valores de entrada. Assim, o tamanho da saída é o mesmo da entrada, então $N' = N$.

A rede Encoder, responsável por realizar a compressão dos dados, é definida como uma rede neural profunda, com duas ou mais camadas ocultas. Enquanto a rede Decoder, possui algumas particularidades, proposta inovadora deste trabalho, essenciais para obter a interpretabilidade do modelo.

A rede de decodificação, possui apenas uma camada, a de saída. Como ilustrado, após a camada latente, para cada variável comprimida existe uma rede neural. Cada rede neural é criada a partir do conjunto de agrupamento do K-Means. Ou seja, a abordagem *Autoencoder* busca comprimir as informações genéticas de maneira agrupada, providas da abordagem anterior.

A interpretabilidade é possível devido à ligação de cada informação comprimida a uma rede baseada nos grupos. Dessa forma, devido à unicidade de representação de cada grupo, teremos uma matriz de pesos de dimensão 1. Os pesos estão totalmente associadas à saída que representa o valor de um gene. A interpretação de cada peso é a influencia que o respectivo gene tem para a reconstrução da informação.

Devido ao agrupamento dos genes serem baseados nos perfis de expressão, é esperado que cada grupo tenha um desvio padrão baixo. O que implica em uma maior chance de *overfitting* na rede. Afim de evitar isto, é utilizada uma técnica chamada *dropout*, uma maneira de, em tempo de treinamento, zerar alguns neurônios para a rede se adaptar.

Essa abordagem está totalmente relacionada à anterior (K-Means) e seguiremos a recomendação *Elbow* para sua execução. Dessa maneira, para um determinado valor K estimado, teremos essa abordagem com tamanho de K variáveis latentes.

Ainda alguns pontos precisam ser definidos, tais como a quantidade de camadas na rede Decoder, número de neurônios, otimizador, funções de ativação, taxa de aprendizado, entre outros hiperparâmetros. Os hiperparâmetros são definidos com base em testes empíricos, apresentados no capítulo seguinte.

CAPÍTULO

6

RESULTADOS E DISCUSSÃO

Neste capítulo iremos apresentar quais foram os resultados obtidos em cada método mencionado no Capítulo 5. Ao longo do texto, verificaremos se a informação genética agrega na análise da sobrevida, identificando seu nível de importância no desfecho dos pacientes. Também poderemos identificar biomarcadores relacionados à sobrevida, verificando se os genes com maior importância também estão citados como tal na literatura científica.

Dividimos o estudo em quatro abordagens. A primeira consiste na implementação do modelo de Cox utilizando apenas os dados clínicos como variáveis explicativas. Na segunda abordagem, veremos os resultados do penalizador lasso na regressão de Cox, incluindo os dados clínicos e genéticos. Em seguida, apresentamos os resultados da seleção por agrupamento dos genes. Por fim, veremos a abordagem utilizando a rede neural Autoencoder, com a arquitetura baseada no agrupamento obtido.

A informação do subtipo do câncer foi considerada nas analises mas não está presente em nenhum dos resultados, devido à baixa relevância estatística (valor p), de acordo com o nível descritivo obtido no modelo de Cox. Desta forma, as únicas informações relevantes clínicas foram o estadiamento do câncer e idade do paciente.

6.1 Clínicos

O primeiro método, utilizou apenas características clinicas. A implementação utilizou o pré processamento dos dados clínicos e o modelo de regressão de Cox.

Na análise da sobrevida é muito utilizada a visualização por meio do *forest plot*. Na Figura 27, temos esta visualização ilustrando os resultados da aplicação do modelo de Cox com os dados clínicos. A interpretação das estimativas dos parâmetros de regressão β trazem informações relevantes acerca da sobrevida. Nota-se que o estágio **II** possui um risco de óbito maior do que o **I**, e no estágio **III** um risco maior ainda, o que faz sentido, visto que é um câncer

em um estágio mais avançado.

Ainda na Figura 27, notamos o valor P, que representa o nível descritivo do teste de hipóteses que verifica a significância da variável explicativa em questão no modelo de Cox, ou seja, se a característica em questão altera ou não o risco de óbito de forma significativa. Conforme foi mencionado na metodologia, é utilizado o método de *stepwise* para seleção das variáveis no modelo. Neste contexto, todas as informações no modelo têm significância estatística.

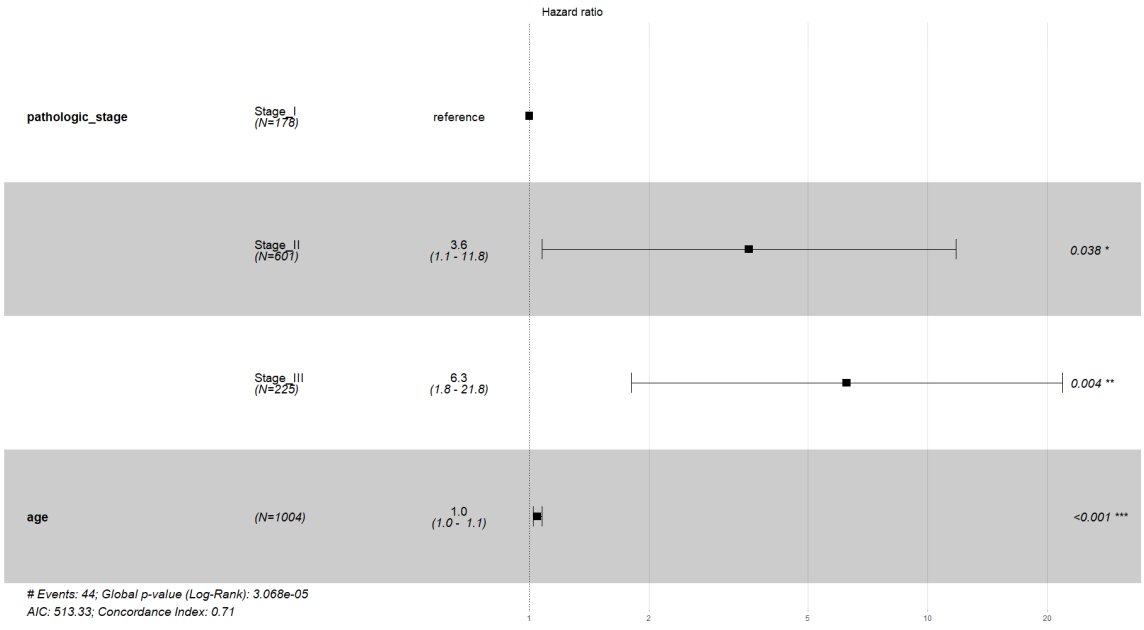


Figura 27 – *Forest plot* do modelo de Cox clínico

Fonte: Elaborada pelo autor.

O modelo de Cox utilizando as informações clinicas, obteve um C-Index de 0.71. Neste caso, o valor estimado mostra que o modelo com apenas informações clínicas conseguiu determinar em alguns casos a predição corretamente, mas na maioria das vezes não obteve um bom resultado, visto que um modelo com 0.5 de C-Index é considerado uma abordagem aleatória. A outra métrica adotada é o critério de informação de Akaike (AIC), que resultou em 513.33, uma métrica útil para fins de comparação com outros modelos.

6.2 Penalização LASSO

O ajuste do modelo de Cox com a penalização lasso mostrou que ao inserir as variáveis da expressão gênica obtemos melhores resultados. Lembrando da estratégia adotada em que variamos o valor da penalização λ, podemos analisar os resultados das diversas análises. A Figura 28 traz a quantidade de genes selecionados pelo lasso no modelo de Cox, baseado na métrica C-Index, já com a execução do *stepwise*.

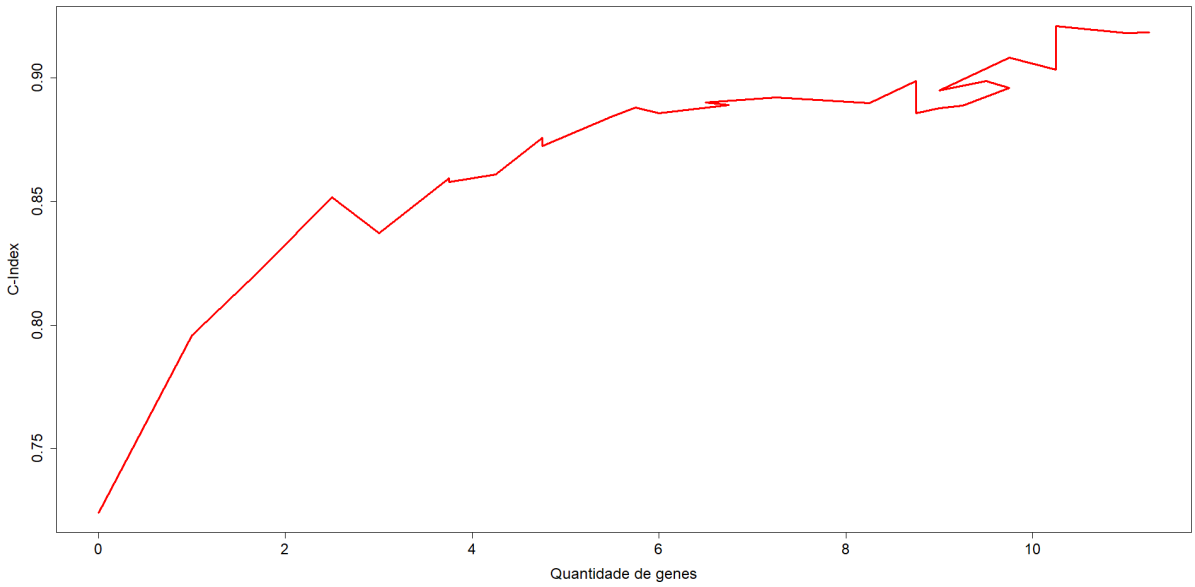


Figura 28 – Resultado da quantidade final de genes pela métrica C-Index do modelo de Cox com penalização lasso.

Fonte: Elaborada pelo autor.

De acordo com os resultados, o melhor C-index resultou em um modelo com 13 genes, com o método de *stepwise* aplicado. Para fins de comparação, o modelo sem a realização do *stepwise* resultou em 30 genes, ou seja, 17 tiveram um valor de P acima de 5%. Percebe-se na Figura 28, que os melhores resultados forma encontrados quando considerou-se um maior número de genes. Isto demonstra dois importantes fatos. O primeiro é quanto a importância do material genético na sobrevida, visto que a inclusão das características do material genético melhora os resultados do modelo. O segundo fato é que podemos considerar diminuir a penalização para a inserção de mais genes. Entretanto isto foi realizado e resultou em colinearidade entre as expressões gênicas. Por essa razão decidimos pela limitação do parâmetro λ, com o maior valor sendo 0.033 e o menor 0.019.

A Figura 29apresenta os resultados do modelo de Cox com a penalização lasso. O C-Index obtido foi de 0.94, superior ao modelo apenas com as variáveis clínicas. Podemos, então, dizer que o modelo com os genes selecionados por lasso, traz uma melhor predição da sobrevida, adicionando informação significativa aos dados clínicos a respeito do risco de óbito. O AIC (373.66) também demonstra ser melhor comparado ao modelo apenas com dados clínicos, já que quanto menor o valor, melhor o modelo.

A característica clinica **pathologic_stage** resultou em um valor de P alto, principalmente para o estágio **II**. Entretanto, mantivemos tal característica no modelo para efeito de comparação com as outras abordagens. Diferente da característica **subtype** que foi removida, já que em todos os métodos resultou em valores altos de P, maiores que 0.5.

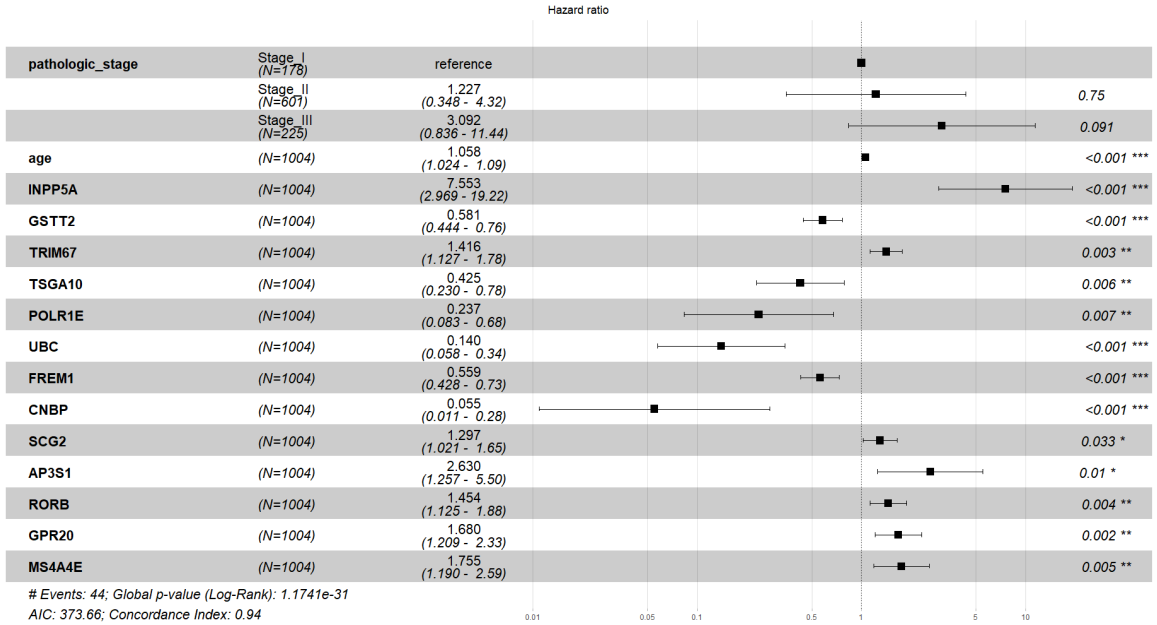


Figura 29 – *Forest plot* do modelo de Cox com penalização lasso.

Fonte: Elaborada pelo autor.

Identificamos, a partir da Figura 29, que o gene **INPP5A** representa alto risco quando há a sua expressão (ZHOU *et al.*, 2022a). O aumento da expressão desse gene, em média, eleva o risco de morte em 7 vezes. Além deste, existem outros 6 genes que estão relacionados ao risco de morte. O segundo gene com maior risco é o **AP3S1**, que é encontrado na literatura com alta relação ao pior prognóstico, em diversos canceres (WU *et al.*, 2022).

Para melhor visualizar o risco, no qual a população utilizada possui com o gene **INPP5A**, criamos uma curva de sobrevivência, visualizada na Figura 30. Separamos toda a população em dois grupos, com alta e baixa expressão. A divisão foi feita a partir da mediana, sendo que indivíduos acima da mediana são de alto risco. A visualização corrobora os resultados do modelo de Cox e os achados anteriores da literatura.

Além dos genes que trazem uma piora no quadro dos pacientes, existem 6 genes que estão relacionados ao efeito de proteção. A proteção é o efeito oposto ao aumento de risco de óbito, de tal forma que quanto maior a expressão dos genes com efeito protetivo, menor é o risco de morte. O gene com o maior efeito protetivo é o **CNBP**, reduzindo o risco em média de 18 vezes, já que 1/0.055 resulta em aproximadamente 18.8.

Um outro gene encontrado na literatura cuja expressão traz o efeito de proteção é o **FREM1**. O trabalho de Li *et al.* (2020) demonstra que a alta expressão do gene resulta em um prognóstico favorável para pacientes com câncer de mama.

Os genes restantes não possuem muitas referências na literatura, mas ainda é possível avaliar por meio de análises feitas pelo próprio projeto que catalogou os dados, o TCGA. Dos genes selecionados pelo lasso, a grande maioria está aderente ao que o TCGA apresenta, com

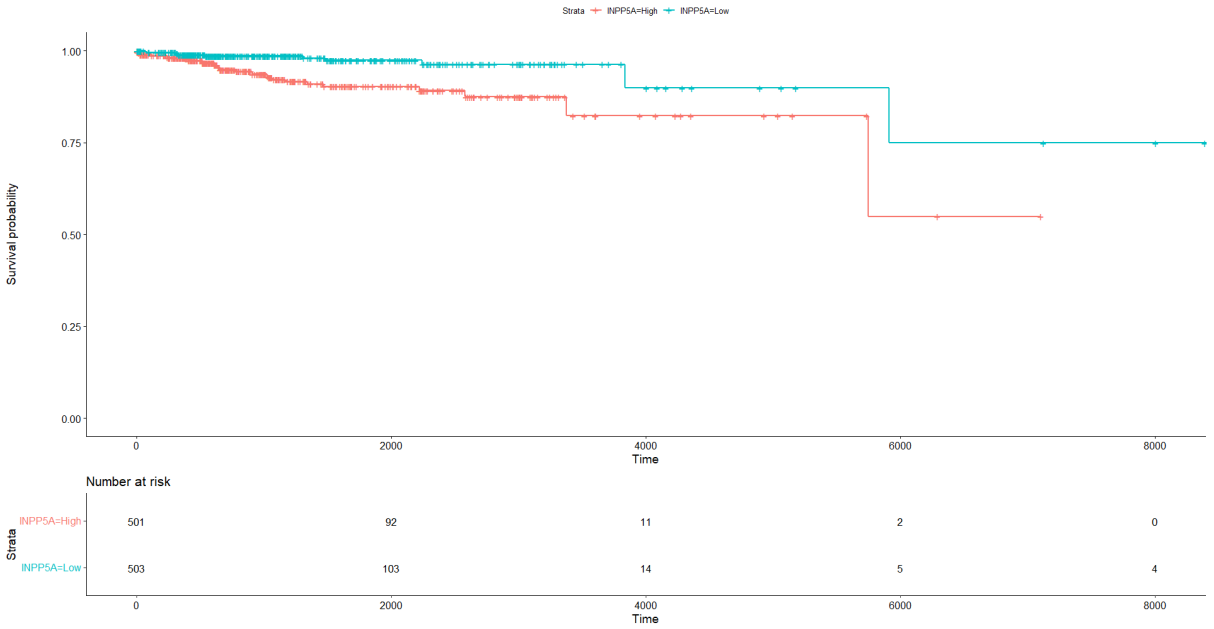


Figura 30 – Curva da sobrevida para o gene **INPP5A**, *High* são indivíduos com expressão superior a mediana e *Low* abaixo.

Fonte: Elaborada pelo autor.

apenas 4 não refletindo as analises.

6.3 K-Means

A proposta utilizando a abordagem do K-Means obteve o segundo melhor resultado. A primeira etapa para esta abordagem consiste em analisar a execução do método *Elbow*, assim teremos uma estimativa da quantidade de grupos necessários. Posteriormente, seguimos com a execução do K-means, selecionando um gene de cada grupo definido para incluir no modelo de Cox.

O método *Elbow* retornou uma estimativa aproximada de 24 grupos, ou seja, uma diminuição de 17433 para 24 genes, representando uma redução de mais de 99% do conjunto. Podemos ver na Figura 6 o resultado dado pelo método *Elbow*, construído definindo de 2 até 150 grupos. De qualquer forma, para o ajuste do modelo, iremos analisar outros valores de k, para que seja possível determinar se o método *Elbow* neste tipo de aplicação é realmente eficaz.

Assim executamos o K-means para a seleção de genes com diferentes valores de k. De acordo com o resultado do *Elbow*, estimamos um intervalo entre os valores 15 e 100. Dessa maneira teremos de 15 genes selecionados até 100 genes, com um incremento de 1 grupo (gene) a cada execução.

Após o agrupamento, com a seleção dos genes que estão mais próximos dos centroides, e a aplicação do método *stepwise* para a seleção das variáveis explicativas no modelo de Cox,

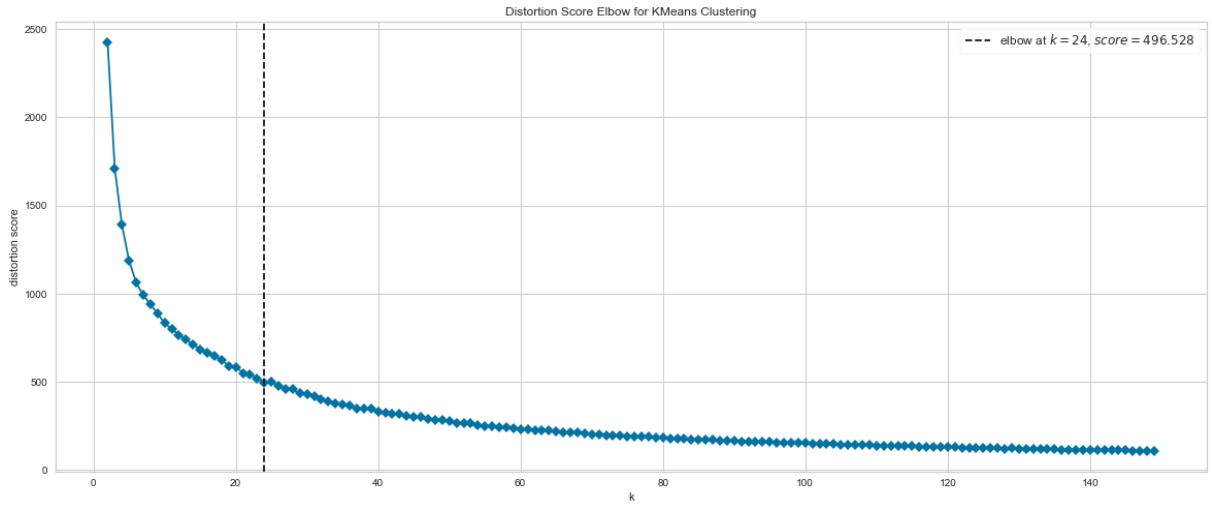


Figura 31 – Resultado método de *Elbow*

Fonte: Elaborada pelo autor.

podemos ver na Figura 32 os valores C-Index resultantes. Para cada execução, tivemos um número de genes selecionados, considerados estatisticamente relevantes, por meio do *stepwise*. Além disto, o resultado ilustrado contempla a incorporação das características clínicas no modelo. Notamos que considerar uma maior quantidade de genes no modelo de Cox tende a aumentar o valor do C-Index.

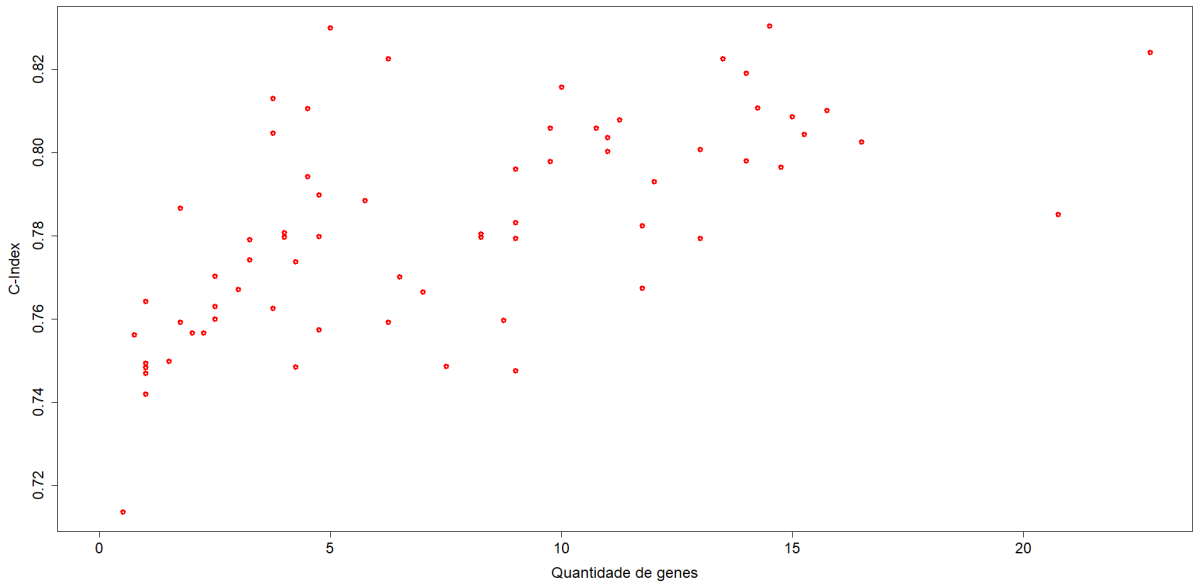


Figura 32 – Resultado Cox com seleção de variáveis com K-Means

Fonte: Elaborada pelo autor.

Chegamos à mesma conclusão que a penalização lasso: um maior número de genes resulta em um melhor C-Index. E da mesma forma, ao especificar um número alto de grupos ao

método K-Means, os genes começam a apresentar colinearidade. Foi observado que a partir de 75 grupos há a existência de colinearidade, o que resulta em um modelo de Cox não adequado.

O melhor modelo ajustado com a abordagem do K-Means foi encontrado com o valor de k igual a 65, ou seja, 65 genes selecionados, que ao final resultou em apenas 14 relevantes, com o valor P menor ou igual a 5%. Os genes e outras informações do modelo são visualizados na Figura 33. Podemos identificar o grupo dos genes, a quantidade de genes pertencentes ao grupo e o gene selecionado, aquele mais próximo do centroide.

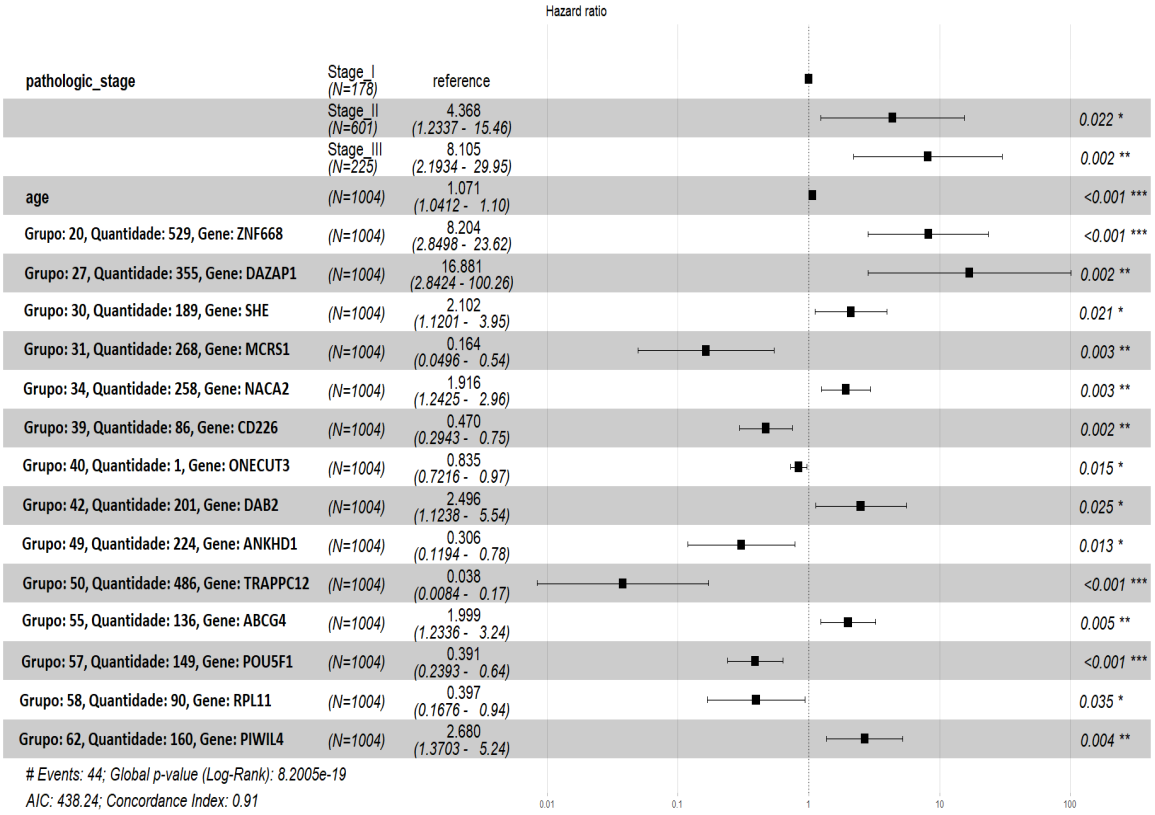


Figura 33 – *Forest plot* do modelo de Cox com seleção de genes a partir do método K-Means

Fonte: Elaborada pelo autor.

Diferentemente da penalização lasso, a variável clinica **pathologic_stage** resultou em valores P significativos. Além disto, foram encontrados genes com efeito de proteção. Ao todo, 7 genes estão relacionados a melhor sobrevida, enquanto que os outros 7 representam um perfil de piora.

O gene com maior valor ligado ao risco é o **DAZAP1**, que na literatura está relacionado ao pior prognostico, como o trabalho de Zhou *et al.* (2022b) indica, no qual realizou um estudo sobre o câncer das células plasmáticas. O estudo aborda experimentos que relaciona a expressão do gene a evolução da doença, resultando no rápido avanço do câncer de acordo com o nível de expressão, e por consequência na ocorrência do evento.

A metodologia com o uso do K-Means, resultou em um C-Index e AIC de 0.86 e 481.07,

respectivamente. Valores inferiores ao método com penalização lasso, porém com uma melhora considerável quando comparado ao modelo apenas com vari[aveis clínicas.

6.4 Autoencoder

A abordagem com uso do *Autoencoder*, uma arquitetura de rede neural utilizada para reduzir a dimensão dos dados, necessita de diversas definições, os hiperparâmetros, diferente da penalização lasso e K-Means, que foi necessário ajustar apenas um hiperparâmetro. É ilustrado na Figura 34 a arquitetura da rede e os hiperparâmetros avaliados.

A rede codificadora (encoder) possuindo 2 camadas ocultas, com 70% e 20% de neurônios, respectivamente, com o uso da função de ativação Tanh. Outros tamanhos de rede foram considerados, de maneira empírica, no qual a ilustrada obteve a melhor adequação, de acordo com o erro médio quadrático (MSE, do inglês *Mean Squred Error*).

A camada latente, camada com o menor número de neurônios, é considerado o valor da quantidade de grupos dado pelo K-Means, no qual K foi definido como 65, logo a quantidade de neurônios definidos na camada latente é de 65. A função de ativação, na camada latente e de saída é utilizado a linear. Entre a camada latente e de saída, há a aplicação do *Dropout*, com o valor de 20%, técnica utilizada para evitar *overfitting*, no qual na fase de treinamento são desativados alguns neurônios, com a probabilidade de 20% deste evento ocorrer.

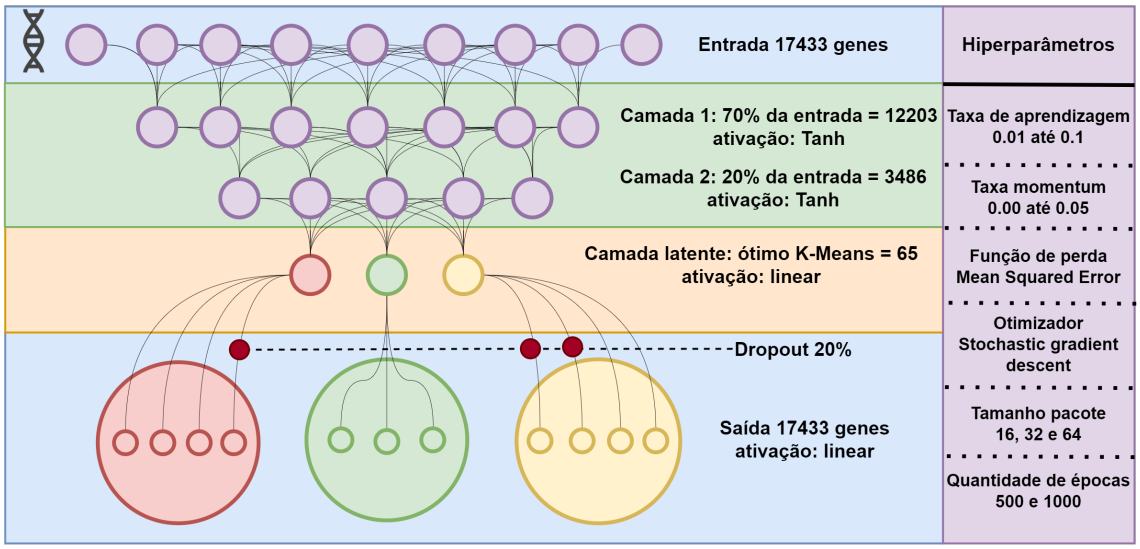


Figura 34 – Arquitetura e hiperparâmetros do *Autoencoder*.

Fonte: Elaborada pelo autor.

Além da estrutura da rede, existem algumas outras importantes definições, como a taxa de aprendizagem, em que foi considerado valores entre 0.01 e 0.1, e constatou a melhor adequação com o valor 0.1, e momentum 0.05. Devido a natureza da rede, no qual o objetivo é reconstruir

os dados de entrada é comum o uso da função de perda MSE, no qual a rede foi otimizada com o uso do SGD (do inglês Stochastic Gradient Descent). O tamanho dos pacotes (*batch size*) e a quantidade de épocas da rede, tiveram melhor retorno com tamanho 32 e 1000, respectivamente.

O menor erro quadrático médio, obtido pela rede definida com os dados de validação foi de 1.69, devido a natureza da métrica ser refletida com os dados, podemos ter a noção a partir do quantis (0%, 25%, 50%, 75%, 100%), com os seguintes valores (0, 0.33, 1.43, 2.4, 11.52). De acordo com os valores, em que também a distribuição dos dados são uniformes, é considerado que a rede teve uma performance razoável. Devido a natureza da rede decodificadora possuir apenas uma camada, e não densa, é esperado que a adequação da rede decodificadora trará uma redução do erro, o que foi validado, porém a proposta necessita da rede possuir a arquitetura desenhada para a sua interpretação.

Após o treinamento da rede neural, cada paciente teve sua atribuição de valor para cada grupo codificado, que unificados aos dados clínicos, executamos o modelo de Cox. A Figura 35 entrega os grupos codificados relevantes, com valor de P menor ou igual a 5%. Dos 65 grupos codificados, 12 tiverem relevância, destes 4 apresentam um perfil com o valor β maior que 1, o que em geral resulta em maior risco, mas devido a natureza da abordagem, a análise da sobrevida necessita de outra informação, a valoração dos pesos de cada grupo.

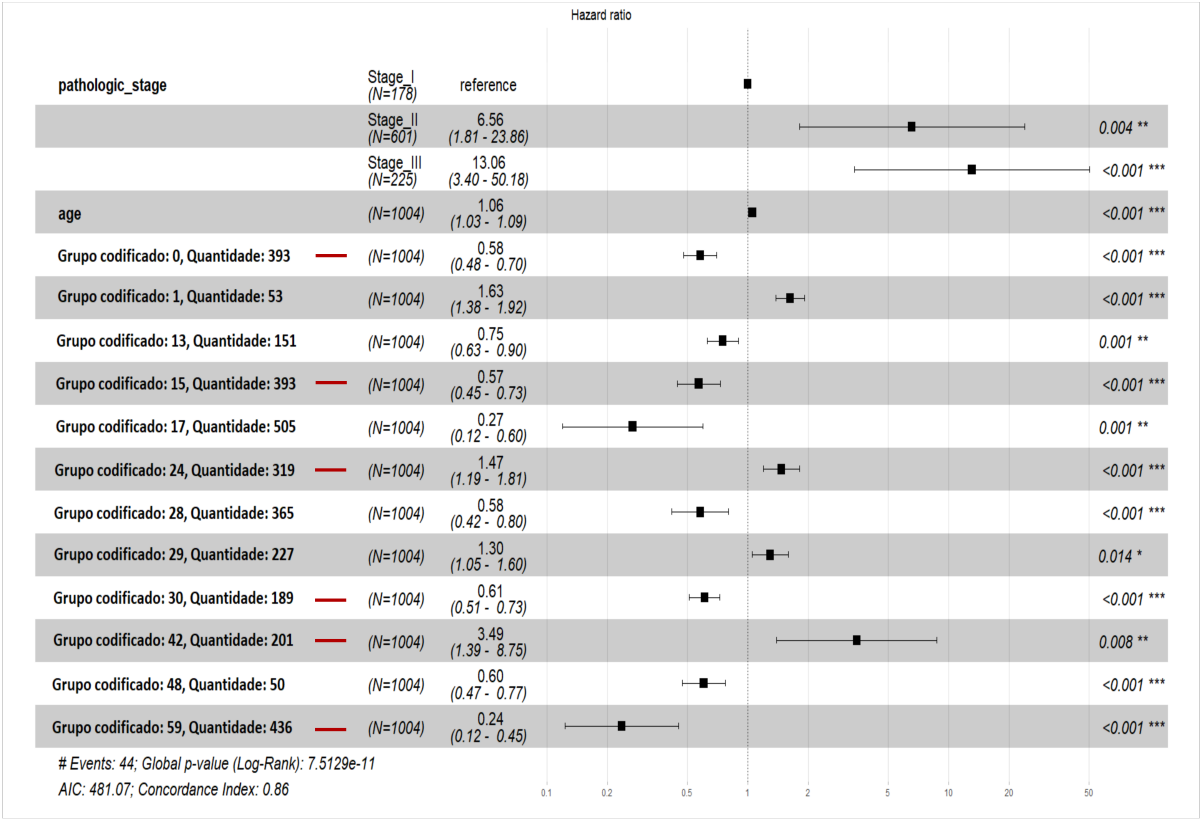


Figura 35 – *Forest plot* do modelo de Cox com redução de genes com *Autoencoder*, baseado nos grupos do método K-Means.

Fonte: Elaborada pelo autor.

Antes de aprofundarmos no entendimento dos grupos codificados, as variáveis clinicas resultaram em valores significativos, em que o **Stage_III**, o estágio mais avançado do estudo, apresentou um risco de em média 13 vezes maior que o estágio inicial. A idade, de maneira semelhante aos outros métodos tem baixa influencia no risco de morte.

Na Figura 35, é possível ver o grupo codificado e a quantidade de genes do grupo em questão, informação semelhante a abordagem K-Means. Devido a introdução de uma rede neural, método usado para codificar e decodificar os genes, para a interpretação do *forest plot* se torna necessário entender como é feito a decodificação do gene.

Nesse sentido, podemos definir a seguinte lógica: Se o gene possuí um peso negativo associado, quanto maior sua expressão menor é a saída da camada latente, logo teremos um efeito contrário do esperado, no qual se o valor de β indica menor que 1, teremos o aumento de risco e caso maior a redução. O peso sendo positivo não há alteração. Na Tabela 5 é apresentado os sinais dos grupos, neste caso todos os pesos de um mesmo grupo tiveram o mesmo sinal, o que facilita a analise.

Tabela 5 – Perfil dos sinais dos pesos

Grupo	**Sinal do pesos**
0	Negativos
1	Positivos
13	Positivos
15	Negativos
17	Positivos
24	Negativos
28	Positivos
29	Positivos
30	Negativos
42	Negativos
48	Positivos
59	Negativos

De acordo com a analise, os grupos 0, 15, 24, 30, 42 e 59 possuem o efeito contrário do apresentado no *forest plot*. Para melhor visualização, estes grupos estão marcados com o sinal $-$ em vermelho na Figura 35. Dessa forma, interpretamos que o grupo 0, por exemplo, que possuí um efeito de proteção ilustrado, mas na realidade possui um efeito de risco.

Dessa maneira, os genes com os maiores pesos de cada grupo, em modulo, são aqueles que apresentam maior risco ou efeito protetivo. Elencamos os 3 pesos mais relevantes para cada grupo, com seus respectivos genes na Tabela 6, a partir desta é possível analisar na literatura a existência de influência destes genes no câncer, afim de validar a consistência dos resultados e consequentemente o método proposto.

No grupo 0, qual representa o perfil de pesos negativos, e por este motivo estão associados ao maior risco de morte. Os genes **LONRF1**, **ZNF275**, **PIERCE2**, representam o maior risco.

Tabela 6 – Top 3 dos valores de pesos do *Autoencoder*, considerando apenas os grupos codificados com relevância.

Gene	Peso associado	Grupo
LONRF1	-0.14	0
ZNF275	-0.14	0
PIERCE2	-0.13	0
CFAP276	0.41	1
ODAD1	0.37	1
EFCAB12	0.35	1
MS4A1	0.24	13
PARP15	0.24	13
TRAT1	0.23	13
CBX8	-0.18	15
WDR25	-0.17	15
GSTZ1	-0.16	15
SLC9A7	0.12	17
NGLY1	0.12	17
TM2D1	0.12	17
MAB21L4	-0.30	24
ADGRG6	-0.26	24
HOXC11	-0.24	24
BAIAP3	0.28	28
CST2	0.27	28
RGL3	0.26	28
FCMR	0.19	29
STEAP1	0.19	29
PTK2B	0.18	29
ARHGAP40	-0.31	30
TCN1	-0.28	30
TP63	-0.27	30
COL14A1	-0.19	42
NFIL3	-0.18	42
LY96	-0.18	42
AK7	0.35	48
CYP2A6	0.34	48
PYY	0.33	48
LAMTOR2	-0.13	59
EIF4EBP1	-0.12	59
GPAA1	-0.12	59

Isto porque são os genes com menores pesos, logo quanto menor é a saída da camada latente, maior será a reconstrução. Os genes **ZNF**, são proteínas que possuem a propriedade de se ligarem ao DNA, e estão relacionadas ao pior prognóstico do câncer de mama, quando superexpressos (LEI *et al.*, 2022).

Em contrapartida o grupo 13, no qual possui pesos positivos e está relacionado ao efeito protetivo, o gene **MS4A1** é o que representa maior efeito, o que reflete com recentes estudos

feitos com este gene para o câncer de mama (LI; FANG, 2021). De maneira similar, o gene **TRAT1**, ainda do grupo 13, tem relação ao atraso do avanço do câncer de pulmão, quando há a sua alta expressão (XIAO *et al.*, 2022b).

Os resultados dos demais grupos, pelo menos um gene por grupo foi encontrado uma referencia, quais se enquadraram com os resultados, nem todos relacionados ao carcinoma de mama. O grupo 1, no qual representa maior risco (LEI *et al.*, 2022). Grupo 15, com maior risco (CHUNG *et al.*, 2016; CAI *et al.*, 2022). Grupo 17, com menor risco (WANG *et al.*, 2021; ZOLEKAR *et al.*, 2018). Grupo 24, com menor risco (AN *et al.*, 2022). Grupo 28, com menor risco (GONG *et al.*, 2020). Grupo 29, com maior risco (AL-JUBOORI *et al.*, 2019). Grupo 30, com maior risco (LIU *et al.*, 2020). Grupo 42, com menor risco (MALVIA *et al.*, 2023). Grupo 48, com menor risco (ZHANG *et al.*, 2021; GRISÉ *et al.*, 1999). Grupo 59, com maior risco (RUTKOVSKY *et al.*, 2019; GE; ZHANG; YANG, 2022)

Entretanto, o gene **BRCA2**, um gene conhecido na literatura por estar relacionado ao pior prognóstico (WANG *et al.*, 2018), está inserido no grupo 24, que apresentou efeito protetivo. Por mais que há um gene de efeito contrário no grupo, ainda existe a necessidade de análise do peso, no qual o gene **BRCA2** apresentou um valor de -0.04, sendo um dos genes com os valores mais próximos de zero. Genes com pesos próximos de zero possuem baixa importância para a analise da sobrevida, isto não necessariamente significa ter efeito contrário, como no exemplo apresentado, e sim que não tiverem importância para a restauração da informação.

Devido a abordagem por meio do *Autoencoder* entregar uma variedade de pesos, de diversos genes, há a intersecção entre o método lasso e K-Means. O gene **SCG2**, por exemplo, encontrado no método lasso como um fator de risco, está presente no grupo 30, com um peso de -0.1, que ainda representa um grau de importância, e em ambas soluções o gene está relacionado ao risco de óbito.

Entre o método K-Means e o *Autoencoder*, os grupos 30 e 42 aparecem, no qual o grupo 30 apresenta maior risco em ambos, e o grupo 42 com divergência, o método *Autoencoder* apresenta como um grupo protetivo, enquanto no K-Means relacionado ao risco. Mas para o *Autoencoder*, os genes selecionados pelo K-Means possuem baixa influencia, de acordo com os pesos, neste sentido a comparação não é fiel.

Por mais que as analises, providas dos pesos da rede neural estejam aderentes com a literatura, o resultado do C-Index (0.86) e AIC (481.07) não apresentou uma melhora comparada as abordagens lasso e K-Means, mas ainda foi superior ao método com apenas os dados clínicos. Além disto com a rede neural conseguimos ter uma analise mais completa, comparado ao K-Means, que seleciona apenas um gene por grupo. No capítulo seguinte iremos tirar as conclusões dos resultados, e discutiremos sobre as maneiras de evoluir a abordagem proposta, afim de ajustar o método proposto para uma melhor adequação a sobrevida.

CAPÍTULO

7

CONCLUSÃO

Com os resultados obtidos podemos afirmar que a inserção da informação da expressão gênica melhora a predição da sobrevida, visto que todos os resultados foram melhores comparados ao do modelo de Cox com apenas informações clinicas. Nesta pesquisa, obtivemos uma melhoria de 0.23 na métrica C-index e uma diferença de 139.67 na métrica AIC em relação ao modelo utilizando apenas variáveis clínicas. Assim, a informação genética entrega uma melhor análise da sobrevida dos pacientes com câncer de mama.

A penalização lasso é uma abordagem que insere uma restrição na função objetiva, que foi aplicada na regressão de Cox. Isto mostra que para essa metodologia ainda trataremos o problema de maneira linear, mas que ainda está apenas relacionada a função de verossimilhança, o que faz a otimização ser aplicada na sobrevida e possivelmente justifica obtermos o melhor resultado para esta abordagem.

O método K-means é um método robusto para diversas aplicações, e que neste caso obteve um melhor resultado comparado ao modelo clinico. Mas existem algumas limitações no K-means, como agrupamento de *outliers* e a especificação do valor de k. Tentamos tratar tais limitações com algumas abordagens, utilizando o método Elbow e selecionando as características mais próximas do centroide.

A abordagem utilizando Autoencoder, com as ligações da camada latente baseadas no agrupamento do K-Means, demonstrou ser melhor que o método com apenas informações clinicas, entregando analises fieis a literatura com os pesos associados à camada de decodificação. Entretanto, essa metodologia não obteve um melhor C-Index e AIC que o lasso e K-Means, o que deve estar associado ao erro propagado do método K-Means e a não capacidade de otimizar a rede para a sobrevida, como a penalização lasso.

Neste trabalho, mostramos que as informações genéticas agregam a predição da sobrevida, e que ao interpretar os valores estimados pela regressão de Cox, vimos que em sua maioria a informação genética possui uma maior relevância na sobrevida, do que os dados clínicos

se unicamente avaliados. Revelando além, de informações que apresentam um aumento na probabilidade do paciente vir a óbito, informações que possuem o efeito contrário, ou seja pacientes com estas caraterísticas devem reduzir a sua chance de vir a óbito.

Podemos então concluir que a informação genética possui uma alta importância para a analise da sobrevida, neste caso estudado em pacientes com câncer de mama. Os dados de expressão genica, demonstram de maneira geral até uma maior importância nas informações clinicas. E que também revelamos a existência de fatores para a redução do risco de óbito. Com isto, podemos motivar a evolução do estudo, buscando melhores formas de determinar e mostrar as relações complexas que o material genético possui no câncer, doença que impacta todo o mundo.

7.1 Trabalhos futuros

Como concluímos, a evolução deste trabalho se torna relevante, isto porque vimos que a informação genética de fato possui importância na predição da sobrevida.

Na literatura, soluções utilizando redes neurais para abordagens na sobrevida são diversas, como visto no Capítulo 3. Entretendo na maioria das propostas, não é considerado a interpretação da rede, o que neste trabalho foi resolvido com o uso de apenas uma camada no decoder. Mas ainda os trabalhos relacionados com uso de redes neurais trazem uma diferença, o uso da função de verossimilhança na otimização da rede. Sendo um dos principais pontos para evolução deste trabalho, que por sua vez deve entregar melhores resultados e analises para a sobrevida. Para isso, podemos se embasar em soluções como a de Katzman *et al.* (2018), para a criação de um modelo neural levando em consideração a sobrevida.

Além disto, o método K-Means é amplamente aplicado na área, no qual diferentes métricas são aplicadas. Neste estudo aplicamos a distância Euclidiana, que como visto no Capítulo 3 é aplicado, mas atualmente diversos trabalhos propõem o uso de outras funções de distância, que refletem na correlação e na interação biológica entre os genes (HANDHAYANI; HIRYANTO, 2015; BOTIA *et al.*, 2017), que por sua vez entregam um melhor resultado comparado a Euclidiana.

Outro ponto a ser considerado é o uso de outras bases de dados, devido a metodologia ser agnóstica ao tipo de câncer é possível realizar a aplicação de várias outras bases de dados para avaliar os modelos propostos. Também é considerado realizar o uso de outras informações, que em alguns trabalhos relacionados foram utilizados, como metilação e mutação (TONG *et al.*, 2020), que podem agregar na analise da sobrevida.

REFERÊNCIAS

AL-JUBOORI, S. I.; VADAKEKOLATHU, J.; IDRI, S.; WAGNER, S.; ZAFEIRIS, D.; PEARSON, J. R.; ALMSHAYAKHCHI, R.; CARAGLIA, M.; DESIDERIO, V.; MILES, A. K.; BOOCOCK, D. J.; BALL, G. R.; REGAD, T. PYK2 promotes HER2-positive breast cancer invasion. **Journal of Experimental & Clinical Cancer Research**, Springer Science and Business Media LLC, v. 38, n. 1, maio 2019. Disponível em: <https://doi.org/10.1186/s13046-019-1221-0>. Citado na página 84.

AN, W.; LIN, H.; MA, L.; ZHANG, C.; ZHENG, Y.; CHENG, Q.; MA, C.; WU, X.; ZHANG, Z.; ZHONG, Y.; WANG, M.; HE, D.; YANG, Z.; DU, L.; FENG, S.; WANG, C.; YANG, F.; XIAO, P.; ZHANG, P.; YU, X.; SUN, J.-P. Progesterone activates GPR126 to promote breast cancer development via the gi pathway. **Proceedings of the National Academy of Sciences**, Proceedings of the National Academy of Sciences, v. 119, n. 15, abr. 2022. Disponível em: <https://doi.org/10.1073/pnas.2117004119>. Citado na página 84.

BANK, D.; KOENIGSTEIN, N.; GIRYES, R. **Autoencoders**. arXiv, 2020. Disponível em: <https://arxiv.org/abs/2003.05991>. Citado na página 41.

BENZ, C. C. Impact of aging on the biology of breast cancer. **Critical Reviews in Oncology/Hematology**, v. 66, n. 1, p. 65–74, 2008. ISSN 1040-8428. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1040842807001795>. Citado na página 60.

BOTIA, J. A.; VANDROVCOVA, J.; FORABOSCO, P.; GUELFI, S.; D'SA, K.; HARDY, J.; LEWIS, C. M.; RYTEN, M.; WEALE, M. E. An additional k-means clustering step improves the biological features of wgcna gene co-expression networks. **BMC Systems Biology**, Springer Science and Business Media LLC, v. 11, n. 1, abr. 2017. Disponível em: <https://doi.org/10.1186/s12918-017-0420-6>. Citado na página 86.

CAI, W. L.; CHEN, J. F.-Y.; CHEN, H.; WINGROVE, E.; KURLEY, S. J.; CHAN, L. H.; ZHANG, M.; ARNAL-ESTAPE, A.; ZHAO, M.; BALABAKI, A.; LI, W.; YU, X.; KROP, E. D.; DOU, Y.; LIU, Y.; JIN, J.; WESTBROOK, T. F.; NGUYEN, D. X.; YAN, Q. Human WDR5 promotes breast cancer growth and metastasis via KMT2-independent translation regulation. **eLife**, eLife Sciences Publications, Ltd, v. 11, ago. 2022. Disponível em: <https://doi.org/10.7554/elife.78163>. Citado na página 84.

CAMILLERI, L. **History of survival analysis**. Allied Newspapers Ltd., 2019. Disponível em: <https://www.um.edu.mt/library/oar/handle/123456789/55748>. Citado na página 31.

CARDOSO, F.; VEER, L. J. van't; BOGAERTS, J.; SLAETS, L.; VIALE, G.; DELALOGE, S.; PIERGA, J.-Y.; BRAIN, E.; CAUSERET, S.; DELORENZI, M.; GLAS, A. M.; GOLFINOPOULOS, V.; GOULIOTI, T.; KNOX, S.; MATOS, E.; MEULEMANS, B.; NEIJENHUIS, P. A.; NITZ, U.; PASSALACQUA, R.; RAVDIN, P.; RUBIO, I. T.; SAGHATCHIAN, M.; SMILDE, T. J.; SOTIRIOU, C.; STORK, L.; STRAEHLE, C.; THOMAS, G.; THOMPSON, A. M.; HOEVEN, J. M. van der; VUYLSTEKE, P.; BERNARDS, R.; TRYFONIDIS, K.; RUTGERS, E.; PICCART, M. 70-gene signature as an aid to treatment decisions in early-stage breast cancer.

New England Journal of Medicine, Massachusetts Medical Society, v. 375, n. 8, p. 717–729, ago. 2016. Disponível em: <https://doi.org/10.1056/nejmoa1602253>. Citado na página 28.

CHAI, H.; ZHOU, X.; ZHANG, Z.; RAO, J.; ZHAO, H.; YANG, Y. Integrating multi-omics data through deep learning for accurate cancer prognosis prediction. **Computers in Biology and Medicine**, Elsevier BV, v. 134, p. 104481, jul. 2021. Disponível em: <https://doi.org/10.1016/j.compbiomed.2021.104481>. Citado nas páginas 13, 21, 47 e 48.

CHAUDHARY, K.; POIRION, O. B.; LU, L.; GARMIRE, L. X. Deep learning–based multi-omics integration robustly predicts survival in liver cancer. **Clinical Cancer Research**, American Association for Cancer Research (AACR), v. 24, n. 6, p. 1248–1259, mar. 2018. Disponível em: <https://doi.org/10.1158/1078-0432.ccr-17-0853>. Citado nas páginas 13, 21, 47, 50 e 51.

CHEN, H. long; ZHOU, M. qi; TIAN, W.; MENG, K. xin; HE, H. fei. Effect of age on breast cancer patient prognoses: A population-based study using the SEER 18 database. **PLOS ONE**, Public Library of Science (PLoS), v. 11, n. 10, p. e0165409, out. 2016. Disponível em: <https://doi.org/10.1371/journal.pone.0165409>. Citado na página 60.

CHING, T.; ZHU, X.; GARMIRE, L. X. Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. **PLOS Computational Biology**, Public Library of Science (PLoS), v. 14, n. 4, p. e1006076, abr. 2018. Disponível em: <https://doi.org/10.1371/journal.pcbi.1006076>. Citado nas páginas 13, 21, 47, 53, 55 e 56.

CHUNG, C.-Y.; SUN, Z.; MULLOKANDOV, G.; BOSCH, A.; QADEER, Z. A.; CIHAN, E.; RAPP, Z.; PARSONS, R.; AGUIRRE-GHISO, J. A.; FARIAS, E. F.; BROWN, B. D.; GASPAR-MAIA, A.; BERNSTEIN, E. Cbx8 acts non-canonically with wdr5 to promote mammary tumorigenesis. **Cell Reports**, Elsevier BV, v. 16, n. 2, p. 472–486, jul. 2016. Disponível em: <https://doi.org/10.1016/j.celrep.2016.06.002>. Citado na página 84.

CONESA, A.; MADRIGAL, P.; TARAZONA, S.; GOMEZ-CABRERO, D.; CERVERA, A.; MCPHERSON, A.; NIAK, M. W.; GAFFNEY, D. J.; ELO, L. L.; ZHANG, X.; MORTAZAVI, A. A survey of best practices for RNA-seq data analysis. **Genome Biol**, v. 17, p. 13, Jan 2016. Citado nas páginas 27 e 29.

COX, D. R. Regression models and life-tables. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley, v. 34, n. 2, p. 187–202, jan. 1972. Disponível em: <https://doi.org/10.1111/j.2517-6161.1972.tb00899.x>. Citado na página 46.

E., G. S. R. C. **Análise de sobrevivência aplicada**. [S.l.: s.n.], 2006. 392 p. Citado na página 33.

GE, S.; ZHANG, Q.; YANG, X. GPAA1 promotes the proliferation, invasion and migration of hepatocellular carcinoma cells by binding to RNA-binding protein SF3b4. **Oncology Letters**, Spandidos Publications, v. 23, n. 5, mar. 2022. Disponível em: <https://doi.org/10.3892/ol.2022.13280>. Citado na página 84.

GONG, P.-J.; SHAO, Y.-C.; HUANG, S.-R.; ZENG, Y.-F.; YUAN, X.-N.; XU, J.-J.; YIN, W.-N.; WEI, L.; ZHANG, J.-W. Hypoxia-associated prognostic markers and competing endogenous RNA co-expression networks in breast cancer. **Frontiers in Oncology**, Frontiers Media SA, v. 10, dez. 2020. Disponível em: <https://doi.org/10.3389/fonc.2020.579868>. Citado na página 84.

GRISÉ, K. R.; RONGIONE, A. J.; LAIRD, E. C.; MCFADDEN, D. W. Peptide YY inhibits growth of human breast cancerin vitroandin vivo. **Journal of Surgical Research**, Elsevier BV, v. 82, n. 2, p. 151–155, abr. 1999. Disponível em: <https://doi.org/10.1006/jsre.1998.5528>. Citado na página 84.

HANDHAYANI, T.; HIRYANTO, L. Intelligent kernel k-means for clustering gene expression. **Procedia Computer Science**, Elsevier BV, v. 59, p. 171–177, 2015. Disponível em: <https://doi.org/10.1016/j.procs.2015.07.544>. Citado na página 86.

HARRELL FRANK E., J.; CALIFF, R. M.; PRYOR, D. B.; LEE, K. L.; ROSATI, R. A. Evaluating the Yield of Medical Tests. **JAMA**, v. 247, n. 18, p. 2543–2546, 05 1982. ISSN 0098-7484. Disponível em: <https://doi.org/10.1001/jama.1982.03320430047030>. Citado na página 36.

HIRA, M. T.; RAZZAQUE, M. A.; ANGIONE, C.; SCRIVENS, J.; SAWAN, S.; SARKER, M. Integrated multi-omics analysis of ovarian cancer using variational autoencoders. **Scientific Reports**, Springer Science and Business Media LLC, v. 11, n. 1, mar. 2021. Disponível em: <https://doi.org/10.1038/s41598-021-85285-4>. Citado nas páginas 13, 21, 47, 49 e 50.

Hospital Hélio Angotti. **Sobre o cancer**. 2023. <https://www.helioangotti.com.br/paciente/sobre-o-cancer/>, Last accessed on 05/07/2023. Citado na página 26.

HU, Y.; ZHAO, Y.; SCHUNK, C. T.; MA, Y.; DERR, T.; ZHOU, X. M. ADEPT: Autoencoder with differentially expressed genes and imputation for robust spatial transcriptomics clustering. **iScience**, Elsevier BV, v. 26, n. 6, p. 106792, jun. 2023. Disponível em: <https://doi.org/10.1016/j.isci.2023.106792>. Citado na página 47.

JIANG, Q. Cancer classification and gene selection with machine learning method. In: ____. **Proceedings of the 2020 International Symposium on Artificial Intelligence in Medical Sciences**. New York, NY, USA: Association for Computing Machinery, 2020. p. 122–127. ISBN 9781450388603. Disponível em: <https://doi.org/10.1145/3429889.3429913>. Citado nas páginas 25 e 38.

JIANG, Y.; ALFORD, K.; KETCHUM, F.; TONG, L.; WANG, M. D. TLSurv. In: **Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics**. ACM, 2020. Disponível em: <https://doi.org/10.1145/3388440.3412422>. Citado nas páginas 13, 21, 47, 53 e 54.

JUEXIN; MA, A.; CHANG, Y.; GONG, J.; JIANG, Y.; QI, R.; WANG, C.; FU, H.; MA, Q.; XU, D. scGNN is a novel graph neural network framework for single-cell RNA-seq analyses. **Nature Communications**, Springer Science and Business Media LLC, v. 12, n. 1, mar. 2021. Disponível em: <https://doi.org/10.1038/s41467-021-22197-x>. Citado na página 57.

KATZMAN, J. L.; SHAHAM, U.; CLONINGER, A.; BATES, J.; JIANG, T.; KLUGER, Y. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. **BMC Medical Research Methodology**, Springer Science and Business Media LLC, v. 18, n. 1, Feb 2018. ISSN 1471-2288. Disponível em: <http://dx.doi.org/10.1186/s12874-018-0482-1>. Citado nas páginas 13, 21, 47, 52, 53 e 86.

KHALILI, M.; MAJD, H. A.; KHODAKARIM, S.; AHADI, B.; HAMIDPOUR, M. Prediction of the thromboembolic syndrome: an application of artificial neural networks in gene expression data analysis. **Archives of Advances in Biosciences**, Publisher: School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, v. 7, n. 2, p. 15–22, mar. 2016. ISSN 27831264. Disponível em: <https://doi.org/10.22037/jps.v7i2.11696>. Citado na página 47.

KUKURBA, K. R.; MONTGOMERY, S. B. RNA Sequencing and Analysis. **Cold Spring Harb Protoc**, v. 2015, n. 11, p. 951–969, Apr 2015. Citado na página 27.

LEI, T.; ZHANG, W.; HE, Y.; WEI, S.; SONG, X.; ZHU, Y.; LUO, G.; KUANG, Z.; LI, G.; ZHOU, Q.; SUN, Z.; XIAO, B.; LI, L. ZNF276 promotes the malignant phenotype of breast carcinoma by activating the CYP1b1-mediated wnt/β-catenin pathway. **Cell Death & Disease**, Springer Science and Business Media LLC, v. 13, n. 9, set. 2022. Disponível em: <https://doi.org/10.1038/s41419-022-05223-8>. Citado nas páginas 83 e 84.

LI, H. ning; LI, X. rui; LV, Z. tao; CAI, M. miao; WANG, G.; YANG, Z. fang. Elevated expression of FREM1 in breast cancer indicates favorable prognosis and high-level immune infiltration status. **Cancer Medicine**, Wiley, v. 9, n. 24, p. 9554–9570, out. 2020. Disponível em: <https://doi.org/10.1002/cam4.3543>. Citado na página 76.

LI, S.; FANG, Y. MS4a1 as a potential independent prognostic factor of breast cancer related to lipid metabolism and immune microenvironment based on TCGA database analysis. **Medical Science Monitor**, International Scientific Information, Inc., v. 27, out. 2021. Disponível em: <https://doi.org/10.12659/msm.934597>. Citado na página 84.

LI, S.; HAN, F.; QI, N.; WEN, L.; LI, J.; FENG, C.; WANG, Q. Determination of a six-gene prognostic model for cervical cancer based on WGCNA combined with LASSO and cox-PH analysis. **World Journal of Surgical Oncology**, Springer Science and Business Media LLC, v. 19, n. 1, set. 2021. Disponível em: <https://doi.org/10.1186/s12957-021-02384-2>. Citado nas páginas 21, 47 e 48.

LIANG, L.; LI, J.; YU, J.; LIU, J.; XIU, L.; ZENG, J.; WANG, T.; LI, N.; WU, L. Establishment and validation of a novel invasion-related gene signature for predicting the prognosis of ovarian cancer. **Cancer Cell International**, Springer Science and Business Media LLC, v. 22, n. 1, mar. 2022. Disponível em: <https://doi.org/10.1186/s12935-022-02502-4>. Citado na página 24.

LIU, G. jie; WANG, Y. jie; YUE, M.; ZHAO, L. mei; GUO, Y.-D.; LIU, Y. ping; YANG, H. chai; LIU, F.; ZHANG, X.; ZHI, L. hui; ZHAO, J.; SUN, Y.-H.; WANG, G. ying. High expression of TCN1 is a negative prognostic biomarker and can predict neoadjuvant chemosensitivity of colon cancer. **Scientific Reports**, Springer Science and Business Media LLC, v. 10, n. 1, jul. 2020. Disponível em: <https://doi.org/10.1038/s41598-020-68150-8>. Citado na página 84.

LIU, S.; YAO, W. Prediction of lung cancer using gene expression and deep learning with KL divergence gene selection. **BMC Bioinformatics**, v. 23, n. 1, p. 175, May 2022. Citado na página 28.

LONDON, A. J. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. **Hastings Cent Rep**, v. 49, n. 1, p. 15–21, Jan 2019. Citado na página 28.

LU, Y.; LU, S.; FOTOUHI, F.; DENG, Y.; BROWN, S. J. Incremental genetic k-means algorithm and its application in gene expression data analysis. **BMC Bioinformatics**, Springer Science and Business Media LLC, v. 5, n. 1, p. 172, 2004. Disponível em: <https://doi.org/10.1186/1471-2105-5-172>. Citado na página 57.

MALVIA, S.; CHINTAMANI, C.; SARIN, R.; DUBEY, U. S.; SAXENA, S.; BAGADI, S. A. R. Aberrant expression of col14a1, celrs3, and cthrc1 in breast cancer cyrsells. **Experimental Oncology**, National Academy of Sciences of Ukraine (Co. LTD Ukrinformnauka) (Publications), v. 45, n. 1, p. 28–43, 2023. Disponível em: <https://doi.org/10.15407/exp-oncology.2023.01.028>. Citado na página 84.

MARDIS, E. R. The Impact of Next-Generation Sequencing on Cancer Genomics: From Discovery to Clinic. **Cold Spring Harb Perspect Med**, v. 9, n. 9, 09 2019. Citado na página 27.

National Human Genome Research Institute. **DNA Sequencing Costs: Data**. 2023. <https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data>, Last accessed on 02/08/2023. Citado na página 24.

____. **The Human Genome Project**. 2023. <https://www.genome.gov/human-genome-project>, Last accessed on 06/07/2023. Citado na página 26.

OH, S.; PARK, H.; ZHANG, X. Hybrid clustering of single-cell gene expression and spatial information via integrated NMF and k-means. **Frontiers in Genetics**, Frontiers Media SA, v. 12, nov. 2021. Disponível em: <https://doi.org/10.3389/fgene.2021.763263>. Citado na página 57.

RAMIREZ, R.; CHIU, Y.-C.; ZHANG, S.; RAMIREZ, J.; CHEN, Y.; HUANG, Y.; JIN, Y.-F. Prediction and interpretation of cancer survival using graph convolution neural networks. **Methods**, Elsevier BV, v. 192, p. 120–130, ago. 2021. Disponível em: <https://doi.org/10.1016/j.ymeth.2021.01.004>. Citado nas páginas 13, 21, 47, 51 e 52.

REDES NEURAIS ARTIFICIAIS. 2017. Disponível em: <https://www.monolitonimbus.com.br/redes-neurais-artificiais/>. Acesso em: 02/03/2022. Citado na página 41.

Rocky Mountain Cancer Centers. **Breast Cancer Types & Hormone Receptors**. 2023. <https://www.rockymountaincancercenters.com/breast-cancer/types-hormone-receptors>, Last accessed on 05/07/2023. Citado na página 27.

ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. **Psychological Review**, p. 65–386, 1958. Citado na página 40.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. **Nature**, Springer Science and Business Media LLC, v. 323, n. 6088, p. 533–536, out. 1986. Disponível em: <https://doi.org/10.1038/323533a0>. Citado na página 41.

RUTKOVSKY, A. C.; YEH, E. S.; GUEST, S. T.; FINDLAY, V. J.; MUISE-HELMERICKS, R. C.; ARMESON, K.; ETHIER, S. P. Eukaryotic initiation factor 4e-binding protein as an oncogene in breast cancer. **BMC Cancer**, Springer Science and Business Media LLC, v. 19, n. 1, maio 2019. Disponível em: <https://doi.org/10.1186/s12885-019-5667-4>. Citado na página 84.

SLATKO, B. E.; GARDNER, A. F.; AUSUBEL, F. M. Overview of Next-Generation Sequencing Technologies. **Curr Protoc Mol Biol**, v. 122, n. 1, p. e59, Apr 2018. Citado na página 28.

STEPHENS, Z. D.; LEE, S. Y.; FAGHRI, F.; CAMPBELL, R. H.; ZHAI, C.; EFRON, M. J.; IYER, R.; SCHATZ, M. C.; SINHA, S.; ROBINSON, G. E. Big data: Astronomical or genomical? **PLOS Biology**, Public Library of Science, v. 13, n. 7, p. 1–11, 07 2015. Disponível em: <https://doi.org/10.1371/journal.pbio.1002195>. Citado na página 25.

SUN, N.; GAO, P.; LI, Y.; YAN, Z.; PENG, Z.; ZHANG, Y.; HAN, F.; QI, X. Screening and identification of key common and specific genes and their prognostic roles in different molecular subtypes of breast cancer. **Frontiers in Molecular Biosciences**, Frontiers Media SA, v. 8, fev. 2021. Disponível em: <https://doi.org/10.3389/fmolb.2021.619110>. Citado na página 26.

SUNG, H.; FERLAY, J.; SIEGEL, R. L.; LAVERSANNE, M.; SOERJOMATARAM, I.; JEMAL, A.; BRAY, F. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. **CA: a cancer journal for clinicians**, v. 71, n. 3, p. 209–249, 2021. ISSN 0007-9235. Disponível em: <http://dx.doi.org/10.3322/caac.21660>. Citado na página 23.

SUPPLITT, S.; KARPINSKI, P.; SASIADEK, M.; LACZMANSKA, I. Current Achievements and Applications of Transcriptomics in Personalized Cancer Medicine. **Int J Mol Sci**, v. 22, n. 3, Jan 2021. Citado na página 28.

TANG, X.; DONG, M.; BI, S.; PEI, M.; CAO, D.; XIE, C.; CHI, S. Feature selection algorithm based on k-means clustering. In: **2017 IEEE 7th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER)**. [S.l.: s.n.], 2017. p. 1522–1527. Citado na página 39.

The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. **Nature**, Springer Science and Business Media LLC, v. 490, n. 7418, p. 61–70, set. 2012. Disponível em: <https://doi.org/10.1038/nature11412>. Citado na página 26.

TONG, L.; MITCHEL, J.; CHATLIN, K.; WANG, M. D. Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis. **BMC Medical Informatics and Decision Making**, Springer Science and Business Media LLC, v. 20, n. 1, set. 2020. Disponível em: <https://doi.org/10.1186/s12911-020-01225-8>. Citado nas páginas 13, 21, 25, 47, 55, 56, 57 e 86.

VIJVER, M. J. van de; HE, Y. D.; VEER, L. J. van't; DAI, H.; HART, A. A.; VOSKUIL, D. W.; SCHREIBER, G. J.; PETERSE, J. L.; ROBERTS, C.; MARTON, M. J.; PARRISH, M.; ATSMA, D.; WITTEVEEN, A.; GLAS, A.; DELAHAYE, L.; VELDE, T. van der; BARTELINK, H.; RODENHUIS, S.; RUTGERS, E. T.; FRIEND, S. H.; BERNARDS, R. A gene-expression signature as a predictor of survival in breast cancer. **N Engl J Med**, v. 347, n. 25, p. 1999–2009, Dec 2002. Citado na página 28.

WANG, J.; CHEN, N.; GUO, J.; XU, X.; LIU, L.; YI, Z. Survnet: A novel deep neural network for lung cancer survival analysis with missing values. **Frontiers in Oncology**, v. 10, 2021. ISSN 2234-943X. Disponível em: <https://www.frontiersin.org/article/10.3389/fonc.2020.588990>. Citado na página 57.

WANG, J.; WANG, L. Prediction and prioritization of autism-associated long non-coding RNAs using gene expression and sequence features. **BMC Bioinformatics**, Springer Science and Business Media LLC, v. 21, n. 1, nov. 2020. Disponível em: <https://doi.org/10.1186/s12859-020-03843-5>. Citado na página 47.

WANG, P.; LI, Y.; REDDY, C. K. Machine learning for survival analysis: A survey. **CoRR**, abs/1708.04649, 2017. Disponível em: <http://arxiv.org/abs/1708.04649>. Citado nas páginas 45 e 46.

WANG, W.; LIU, W. Integration of gene interaction information into a reweighted Lasso-Cox model for accurate survival prediction. **Bioinformatics**, v. 36, n. 22-23, p. 5405–5414, 12 2020. ISSN 1367-4803. Disponível em: <https://doi.org/10.1093/bioinformatics/btaa1046>. Citado nas páginas 21, 47 e 49.

WANG, Z.; YANG, X.; SHEN, J.; XU, J.; PAN, M.; LIU, J.; HAN, S. Gene expression patterns associated with tumor-infiltrating CD4+ and CD8+ t cells in invasive breast carcinomas. **Human Immunology**, Elsevier BV, v. 82, n. 4, p. 279–287, abr. 2021. Disponível em: <https://doi.org/10.1016/j.humimm.2021.02.001>. Citado na página 84.

WANG, Z.; ZHANG, J.; ZHANG, Y.; DENG, Q.; LIANG, H. Expression and mutations of BRCA in breast cancer and ovarian cancer: Evidence from bioinformatics analyses. **International Journal of Molecular Medicine**, Spandidos Publications, set. 2018. Disponível em: <https://doi.org/10.3892/ijmm.2018.3870>. Citado na página 84.

WEINSTEIN, J. N.; COLLISSON, E. A.; MILLS, G. B.; SHAW, K. R. M.; OZENBERGER, B. A.; ELLROTT, K.; SHMULEVICH, I.; SANDER, C.; STUART, J. M. The cancer genome atlas pan-cancer analysis project. **Nature Genetics**, Springer Science and Business Media LLC, v. 45, n. 10, p. 1113–1120, set. 2013. Disponível em: <https://doi.org/10.1038/ng.2764>. Citado na página 27.

World Health Organisation. **Vision impairment and blindness, Fact Sheet Nº282**. 2023. <http://www.who.int/mediacentre/factsheets/fs282/fr/>, Last accessed on 05/07/2023. Citado na página 23.

WU, G.; CHEN, M.; REN, H.; SHA, X.; HE, M.; REN, K.; QI, J.; LIN, F. AP3s1 is a novel prognostic biomarker and correlated with an immunosuppressive tumor microenvironment in pan-cancer. **Frontiers in Cell and Developmental Biology**, Frontiers Media SA, v. 10, jul. 2022. Disponível em: <https://doi.org/10.3389/fcell.2022.930933>. Citado na página 76.

XIAO, M.; LIANG, X.; YAN, Z.; CHEN, J.; ZHU, Y.; XIE, Y.; LI, Y.; LI, X.; GAO, Q.; FENG, F.; FU, G.; GAO, Y. A DNA-methylation-driven genes based prognostic signature reveals immune microenvironment in pancreatic cancer. **Frontiers in Immunology**, Frontiers Media SA, v. 13, fev. 2022. Disponível em: <https://doi.org/10.3389/fimmu.2022.803962>. Citado na página 24.

XIAO, X.-Y.; GUO, Q.; TONG, S.; WU, C.-Y.; CHEN, J.-L.; DING, Y.; WAN, J.-H.; CHEN, S.-S.; WANG, S.-H. TRAT1 overexpression delays cancer progression and is associated with immune infiltration in lung adenocarcinoma. **Frontiers in Oncology**, Frontiers Media SA, v. 12, out. 2022. Disponível em: <https://doi.org/10.3389/fonc.2022.960866>. Citado na página 84.

XING, X.; YANG, F.; LI, H.; ZHANG, J.; ZHAO, Y.; GAO, M.; HUANG, J.; YAO, J. Multi-level attention graph neural network based on co-expression gene modules for disease diagnosis and prognosis. **Bioinformatics**, Oxford University Press (OUP), v. 38, n. 8, p. 2178–2186, fev. 2022. Disponível em: <https://doi.org/10.1093/bioinformatics/btac088>. Citado na página 57.

YERSAL, O. Biological subtypes of breast cancer: Prognostic and therapeutic implications. **World Journal of Clinical Oncology**, v. 5, n. 3, p. 412, 2014. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4127612/>. Citado na página 60.

YU, F.; QUAN, F.; XU, J.; ZHANG, Y.; XIE, Y.; ZHANG, J.; LAN, Y.; YUAN, H.; ZHANG, H.; CHENG, S.; XIAO, Y.; LI, X. Breast cancer prognosis signature: linking risk stratification to disease subtypes. **Briefings in Bioinformatics**, v. 20, n. 6, p. 2130–2140, 09 2018. ISSN 1477-4054. Disponível em: <https://doi.org/10.1093/bib/bby073>. Citado na página 24.

YUAN, Y.; BAR-JOSEPH, Z. GCNG: graph convolutional networks for inferring gene interaction from spatial transcriptomics data. **Genome Biology**, Springer Science and Business Media LLC, v. 21, n. 1, dez. 2020. Disponível em: <https://doi.org/10.1186/s13059-020-02214-w>. Citado na página 57.

ZHANG, X. ying; ZHOU, L. li; JIAO, Y.; LI, Y. qing; GUAN, Y. nuo; ZHAO, Y. chen; ZHENG, L. wen. Adenylate kinase 7 is a prognostic indicator of overall survival in ovarian cancer. **Medicine**, Ovid Technologies (Wolters Kluwer Health), v. 100, n. 1, p. e24134, jan. 2021. Disponível em: <https://doi.org/10.1097/md.0000000000024134>. Citado na página 84.

ZHANG, Y.; GUO, L.; DAI, Q.; SHANG, B.; XIAO, T.; DI, X.; ZHANG, K.; FENG, L.; SHOU, J.; WANG, Y. A signature for pan-cancer prognosis based on neutrophil extracellular traps. **Journal for ImmunoTherapy of Cancer**, BMJ, v. 10, n. 6, p. e004210, jun. 2022. Disponível em: <https://doi.org/10.1136/jitc-2021-004210>. Citado na página 24.

ZHANG, Y.; PARMIGIANI, G.; JOHNSON, W. E. : batch effect adjustment for RNA-seq count data. **NAR Genom Bioinform**, v. 2, n. 3, p. lqaa078, Sep 2020. Citado na página 28.

ZHANG, Z.; CHAI, H.; WANG, Y.; PAN, Z.; YANG, Y. Cancer survival prognosis with deep bayesian perturbation cox network. **Computers in Biology and Medicine**, Elsevier BV, v. 141, p. 105012, 2022. Disponível em: <https://doi.org/10.1016/j.compbiomed.2021.105012>. Citado na página 57.

ZHAO, Y.; DONG, Y.; SUN, Y.; CHENG, C. AutoEncoder-based computational framework for tumor microenvironment decomposition and biomarker identification in metastatic melanoma. **Frontiers in Genetics**, Frontiers Media SA, v. 12, maio 2021. Disponível em: <https://doi.org/10.3389/fgene.2021.665065>. Citado na página 47.

ZHOU, Q.; LIN, J.; YAN, Y.; MENG, S.; LIAO, H.; CHEN, R.; HE, G.; ZHU, Y.; HE, C.; MAO, K.; WANG, J.; ZHANG, J.; ZHOU, Z.; XIAO, Z. INPP5f translocates into cytoplasm and interacts with ASPH to promote tumor growth in hepatocellular carcinoma. **Journal of Experimental & Clinical Cancer Research**, Springer Science and Business Media LLC, v. 41, n. 1, jan. 2022. Disponível em: <https://doi.org/10.1186/s13046-021-02216-x>. Citado na página 76.

ZHOU, Y.; HUANGFU, S.; LI, M.; TANG, C.; QIAN, J.; GUO, M.; ZHOU, Z.; YANG, Y.; GU, C. DAZAP1 facilitates the alternative splicing of KITLG to promote multiple myeloma cell proliferation via ERK signaling pathway. **Aging**, Impact Journals, LLC, v. 14, n. 19, p. 7972–7985, out. 2022. Disponível em: <https://doi.org/10.18632/aging.204326>. Citado na página 79.

ZOLEKAR, A.; LIN, V. J. T.; MISHRA, N. M.; HO, Y. Y.; HAYATSHAHI, H. S.; PARAB, A.; SAMPAT, R.; LIAO, X.; HOFFMANN, P.; LIU, J.; EMMITTE, K. A.; WANG, Y.-C. Stress and interferon signalling-mediated apoptosis contributes to pleiotropic anticancer responses induced by targeting NGLY1. **British Journal of Cancer**, Springer Science and Business Media LLC, v. 119, n. 12, p. 1538–1551, nov. 2018. Disponível em: <https://doi.org/10.1038/s41416-018-0265-9>. Citado na página 84.


ICMC
USP
SÃO CARLOS