PUCRS

ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
MESTRADO EM CIÊNCIA DA COMPUTAÇÃO

RODRIGO HENRICH

APRENDIZADO DE MÁQUINA APLICADO NA PREVISÃO DO TEMPO DE SOBREVIDA EM PACIENTES COM CÂNCER DE PULMÃO

Porto Alegre
2025

PÓS-GRADUAÇÃO - *STRICTO SENSU*

Pontifícia Universidade Católica
do Rio Grande do Sul

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

APRENDIZADO DE MÁQUINA APLICADO NA PREVISÃO DO TEMPO DE SOBREVIDA EM PACIENTES COM CÂNCER DE PULMÃO

RODRIGO HENRICH

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul.

Orientadora: Profª. Dra. Isabel Harb Manssour

Porto Alegre
2025

Ficha Catalográfica

H518a Henrich, Rodrigo

Aprendizado de máquina aplicado na previsão do tempo de sobrevida em pacientes com câncer de pulmão / Rodrigo Henrich. – 2025.

79 f.

Dissertação (Mestrado) – Programa de Pós-Graduação em Ciência da Computação, PUCRS.

Orientadora: Profa. Dra. Isabel Harb Manssour.

1. Tempo de Sobrevida. 2. Câncer de Pulmão. 3. Aprendizado de Máquina Supervisionado. 4. Classificação. I. Manssour, Isabel Harb. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS com os dados fornecidos pelo(a) autor(a).

Bibliotecária responsável: Clarissa Jesinska Selbach CRB-10/2051

RODRIGO HENRICH

APRENDIZADO DE MÁQUINA APLICADO NA PREVISÃO DO TEMPO DE SOBREVIDA EM PACIENTES COM CÂNCER DE PULMÃO

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Ciência da Computação do Programa de Pós-Graduação em Ciência da Computação, Escola Politécnica da Pontifícia Universidade Católica do Rio Grande do Sul.

Aprovado(a) em 28 de Março de 2025.

BANCA EXAMINADORA:

Prof. Dr. Sílvio César Cazella (PPGCS/UFCSPA)

Profª. Dra. Soraia Raupp Musse (PPGCC/PUCRS)

Profª. Dra. Isabel Harb Manssour (PPGCC/PUCRS - Orientadora)

APRENDIZADO DE MÁQUINA APLICADO NA PREVISÃO DO TEMPO DE SOBREVIDA EM PACIENTES COM CÂNCER DE PULMÃO

RESUMO

O câncer de pulmão é um dos tipos mais comuns e letais de câncer em todo o mundo. O diagnóstico precoce e o tratamento adequado desempenham um papel fundamental na redução da mortalidade associada a essa doença. A inteligência artificial tem se tornado uma ferramenta promissora em diversas áreas e também na medicina. Uma Revisão Sistemática da Literatura (RSL) sobre predição do tempo de sobrevida em pacientes com câncer revelou que, entre os 64 estudos analisados, 55 utilizaram alguma técnica de aprendizado de máquina. Neste contexto, este estudo propõe a aplicação de técnicas de aprendizado de máquina supervisionado para desenvolver e avaliar modelos capazes de classificar pacientes com câncer de pulmão de acordo com seu tempo de sobrevida (curta ou longa). Os resultados obtidos indicam que os algoritmos com melhor desempenho nesta tarefa foram o *Random Forest*, com acurácia de 87,12%, e o *Decision Tree*, com acurácia de 86,94%. Para compreender os resultados dos modelos de *Machine Learning* (ML), foi utilizado SHAP (SHapley Additive exPlanations), que permite interpretar a contribuição de cada variável para as previsões realizadas. A incorporação desses modelos na prática clínica pode apoiar a tomada de decisões e a personalização dos tratamentos dos pacientes.

Palavras-Chave: Tempo de Sobrevida, Câncer de Pulmão, Aprendizado de Máquina Supervisionado, Classificação.

ABSTRACT

Lung cancer is one of the most common and lethal types of cancer worldwide. Early diagnosis and appropriate treatment play a fundamental role in reducing mortality associated with this disease. Artificial intelligence has become a promising tool in several areas, including medicine. A Systematic Literature Review (SLR) on survival time prediction in cancer patients revealed that, among the 64 studies analyzed, 55 used some machine learning technique. In this context, this study proposes applying supervised machine learning techniques to develop and evaluate models capable of classifying lung cancer patients according to their survival time (short or long). The results indicate that the algorithms with the best performance in this task were Random Forest, with an accuracy of 87.12%, and Decision Tree, with an accuracy of 86.94%. To understand the results of Machine Learning (ML) models, we used SHAP (SHapley Additive exPlanations), which allows interpreting the contribution of each variable to the predictions made. Incorporating these models into clinical practice can support decision-making and personalization of patient treatments.

Keywords: Survival Time, Lung Cancer, Supervised Machine Learning, Classification.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

3D CNN – *3D Convolutional Neural Network*

ACS – *American Cancer Society*

AI – Artificial Intelligence

AUC – *Area under the curve*

CART – *Classification and Regression Trees*

DP – *Deep learning*

CNN – *Convolutional Neural Network*

DT – *Decision Tree*

ECOG – *Eastern Cooperative Oncology Group*

GCO – Global Cancer Obervatory

GDC – Genomic Data Commons

NHGRI – National Human Genome Research Institute

KNN – *K-Nearest Neighbor*

ML – *Machine Learning*

MLP – *Multilayer Perceptron*

NCI – National Cancer Institute

NB – *Naive Bayes*

RF – *Random Forest*

RNA – Rede Neural Artificial

ROC – *Receiver Operating Characteristic*

RSL – Revisão Sistemática da Literatura

SBPT – Sociedade Brasileira de Pneumologia e Tisiologia

SVM – *Support Vector Machine*

TC – Tomografia Computadorizada

TCGA – *The Cancer Genome Atlas Program*

TNM – *Tumor Linfonodos Metástases*

SUMÁRIO

1. INTRODUÇÃO

Conforme a Organização Panamericana de Saúde (OPAS) [64], a incidência de câncer no mundo é um problema de saúde pública crescente e complexo, que tem um alto índice de mortalidade. A Figura 1.1 mostra dados mundiais sobre alguns tipos de câncer obtidos do *Global Cancer Observatory* (GCO) [22]. Neste gráfico estão relacionadas as incidências de casos e a taxa de mortalidade em milhões de pessoas. Os tipos de câncer mais comuns apontados pelo gráfico são os de pulmão e de mama. Sendo que o câncer de pulmão figura entre a maior parte dos casos e, além disso, o índice de mortalidade é alto devido ao fato de que os sintomas não estão presentes nos estágios iniciais da doença.

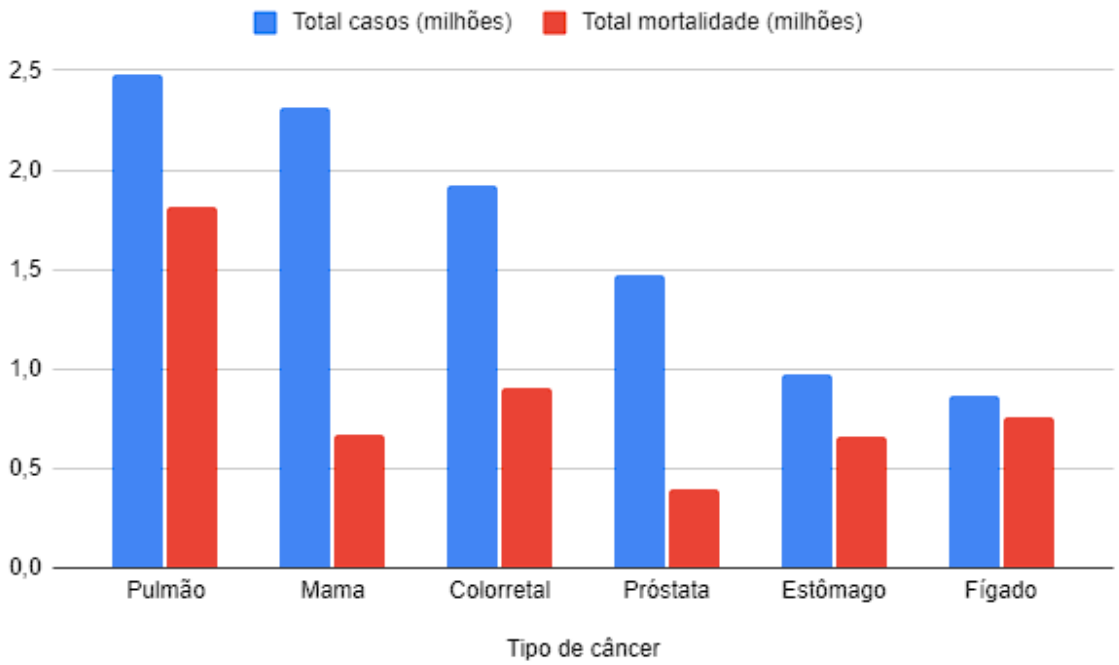


Figura 1.1 – Dados mundiais de incidência e mortalidade de câncer [22].

Estes elementos evidenciam a importância do desenvolvimento de tecnologia voltada para auxiliar os profissionais de saúde no diagnóstico e na tomada de decisões sobre as abordagens terapêuticas a serem adotadas para cada paciente. Nesse contexto, considera-se a possibilidade de incorporar técnicas de *Artificial Intelligence* (AI), ou Inteligência Artificial, uma área em constante evolução e cada vez mais integrada ao cotidiano, com potencial para beneficiar diversas áreas, inclusive a medicina preventiva e curativa.

Diversos estudos relacionam o aprendizado de máquina à análise do tempo de sobrevida de pacientes com câncer [23, 63, 41], utilizando diferentes tipos de informações, como dados clínicos, imagens de tomografia computadorizada, dados de expressão gênica ou combinações de diferentes tipos de informações [63, 71, 29]. A principal dificuldade relatada na literatura é a obtenção de um volume significativo de dados, essencial para análises

mais precisas [23, 5, 78]. Além disso, a presença de dados faltantes exige técnicas robustas de pré-processamento. Outro desafio apontado é a necessidade de testar diferentes modelos de aprendizado de máquina [62, 86, 46] e validar os modelos desenvolvidos em conjuntos de dados independentes [57, 53, 2]. Nesse contexto, o presente estudo explora o potencial do *Machine Learning* (ML), ou Aprendizado de Máquina, um ramo da AI que permite que computadores aprendam a partir de dados e experiências, para prever o tempo de sobrevida de pacientes com câncer de pulmão. A pesquisa busca responder à seguinte questão: qual modelo de aprendizado de máquina apresenta o melhor desempenho na tarefa de prever o tempo de sobrevida de pacientes com câncer de pulmão? Buscando responder a esta pergunta, o objetivo da pesquisa é desenvolver e avaliar modelos de aprendizado de máquina supervisionado capazes de classificar pacientes com câncer de pulmão de acordo com o tempo de sobrevida (curto ou longo), a partir de dados clínicos.

Para atingir o objetivo proposto, foram elaborados os seguintes objetivos específicos:

- Realizar uma Revisão Sistemática da Literatura para identificar as abordagens mais utilizadas na predição do tempo de sobrevida em pacientes com câncer.
- Realizar a busca, seleção e preparação de um conjunto de dados clínicos de pacientes com câncer de pulmão, contemplando as etapas de coleta, limpeza, pré-processamento e balanceamento de classes.
- Selecionar e aplicar algoritmos de aprendizado de máquina supervisionado, incluindo *Random Forest* (RF), *Logistic Regression* (LR), *K-Nearest Neighbors* (KNN), *Decision Tree* (DT) e *Support Vector Machine* (SVM), para a tarefa de classificação.
- Comparar o desempenho dos modelos utilizando métricas como acurácia, precisão, recall, F1-score, AUC e matriz de confusão.
- Explicar os resultados obtidos com os modelos de ML utilizando o método SHAP (*SHapley Additive exPlanations*) [56], identificando os atributos mais relevantes para a decisão dos modelos.

Os algoritmos foram implementados de duas formas, sendo a primeira com os parâmetros padrão definidos pelos desenvolvedores. Nesta configuração, os melhores desempenhos de classificação foram obtidos pelos algoritmos DT (acurácia de 86,9%) e RF (acurácia de 86,7%). Os mesmos modelos foram testados aplicando customização de parâmetros, usando a ferramenta *Grid Search Cross Validation*. Após a definição dos parâmetros, os melhores classificadores foram os algoritmos RF (acurácia de 87,12%) e KNN (acurácia de 92,79%). Os resultados dão um direcionamento para a escolha dos melhores algoritmos de ML para classificação de pacientes com câncer de pulmão. Outro

ponto importante obtido por meio da análise usando SHAP foi a identificação dos atributos que possuem um maior impacto na tomada de decisão dos modelos de ML.

O restante do documento está estruturado em cinco capítulos, cada um abordando um aspecto da pesquisa. O próximo capítulo aborda os principais assuntos relacionados ao aprendizado de máquina e técnicas de avaliação de desempenho. O terceiro capítulo apresenta os resultados da Revisão Sistemática da Literatura (RSL) realizada para mapear os estudos relacionados ao tempo de sobrevida em pacientes com câncer. O quarto capítulo detalha a metodologia adotada, incluindo os procedimentos de obtenção dos dados, etapas de pré-processamento aplicadas, os algoritmos de aprendizado de máquina empregados e as métricas utilizadas para avaliar o desempenho dos modelos. O capítulo seguinte apresenta e analisa os resultados obtidos, discutindo a eficácia das técnicas aplicadas e comparando os resultados com estudos correlatos. Por fim, o último capítulo apresenta as considerações finais do estudo.

2. FUNDAMENTAÇÃO TEÓRICA

Este capítulo aborda os principais tópicos necessários para a compreensão do contexto desta pesquisa. Nas próximas seções, serão explorados temas sobre aprendizado de máquina, incluindo alguns dos algoritmos mais citados nos trabalhos selecionados.

2.1 *Machine learning*

As tarefas de *Machine learning* (ML), podem ser classificadas de diferentes formas. Em Faceli et al. [17] os autores separam as técnicas de aprendizado de máquina em preditivas ou descritivas, conforme ilustra a Figura 2.1. Tarefas preditivas são aquelas empregadas em um conjunto de dados anotados, ou seja, que possuem as respostas para o atributo alvo. O termo supervisionado está relacionado ao fato de já conhecermos as respostas para um determinado conjunto de dados, o que permite saber se o algoritmo está chegando às respostas corretas, o que levaria ao fim do treinamento e validação, ou se estão incorretas, o que levaria a ajustes no algoritmo adotado até obter resultados satisfatórios, o que é chamado de convergir em aprendizado de máquina.

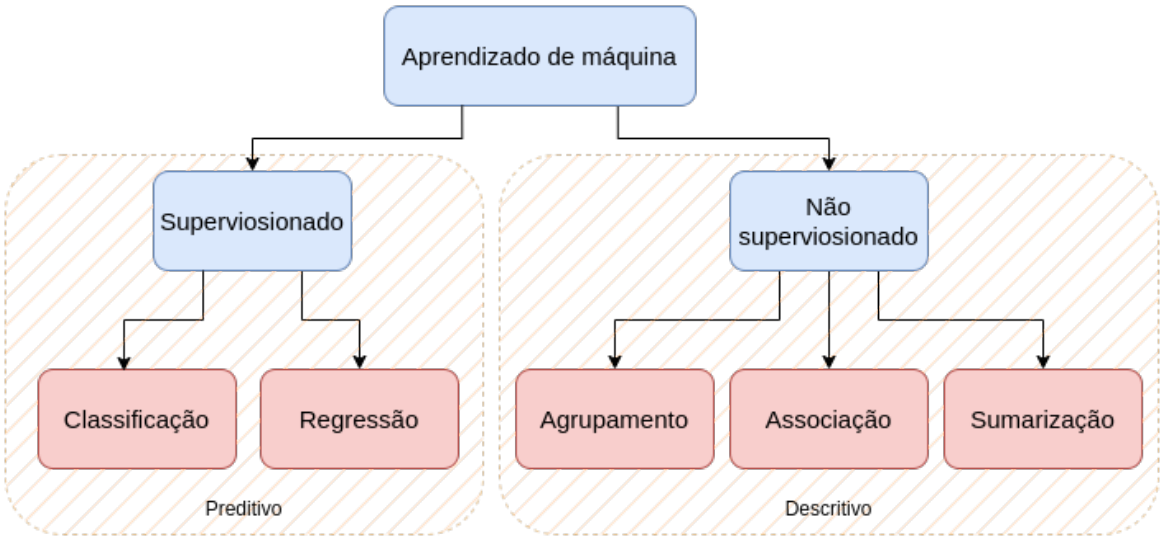


Figura 2.1 – Apresentação clássica de aprendizado de máquina [17].

O aprendizado supervisionado trabalha com tarefas de regressão e classificação, estas se diferem pelo tipo de variável resultante. A classificação vai entregar um resultado discreto, como, por exemplo, classificar e-mail entre spam e não spam ou presença ou ausência de um objeto em uma imagem. Já a regressão vai entregar um resultado contínuo, como a previsão do preço de um imóvel ou previsão da temperatura, por exemplo. O aprendizado não supervisionado se aplica a tarefas descritivas, que podem ser separadas em

três subcategorias: agrupamento, associação e sumarização. Tarefas de agrupamento têm por objetivo agrupar os dados por alguma similaridade; as tarefas de associação têm por objetivo buscar associações entre os atributos das instâncias de um banco de dados; e sumarização tem por objetivo gerar uma descrição simples do conjunto de dados.

No aprendizado supervisionado os dados devem ser processados pelo algoritmo e o resultado comparado com o valor anotado, esse processo ocorre até que o algoritmo chegue a resultados próximos aos anotados. Para realizar o processo, o conjunto de dados anotados é dividido em treinamento e teste. Finalizado o treinamento e teste do algoritmo, ele deve estar pronto para obter resultados coerentes para um conjunto de dados novo, que ele não conheça. O processo é ilustrado pela Figura 2.2. O termo supervisionado está relacionado ao fato de já conhecermos as respostas para um determinado conjunto de dados, o que permite saber se o algoritmo está chegando às respostas corretas, o que levaria ao fim do treinamento, ou se estão incorretas, o que levaria a ajustes no algoritmo adotado até obter resultados satisfatórios.

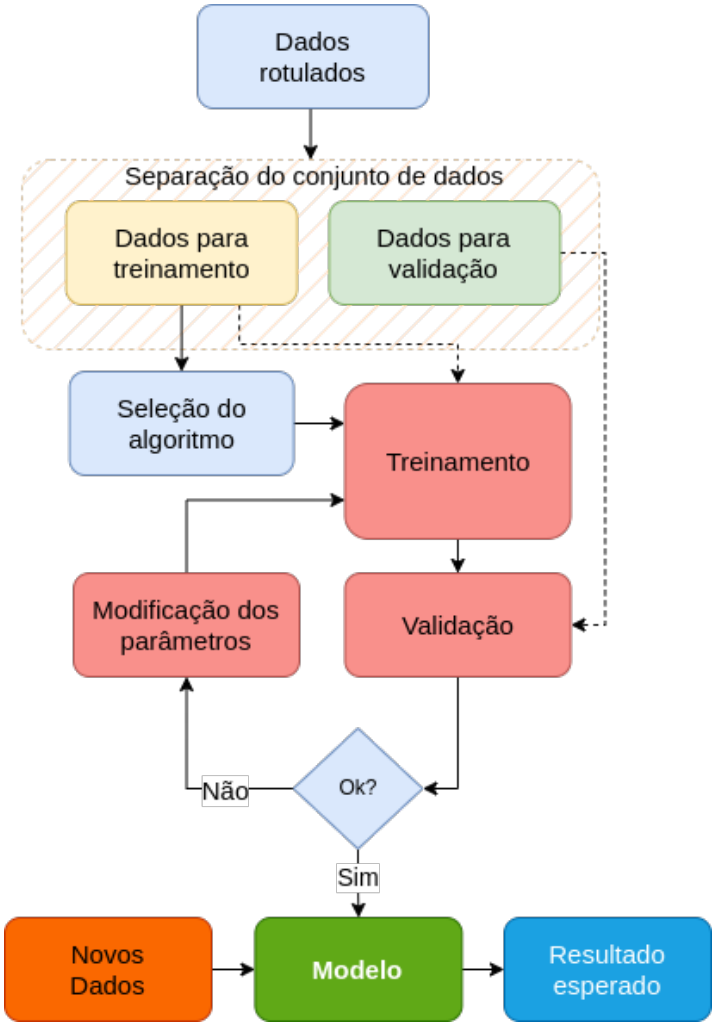


Figura 2.2 – Fluxo do processo de aprendizado supervisionado [17].

2.2 Algoritmos de aprendizado de máquina

Nesta seção são apresentados alguns dos algoritmos de aprendizado de máquina que são mais citados nos estudos pesquisados.

2.2.1 *Support Vector Machine*

Segundo Faceli et al. [17], o algoritmo *Support Vector Machine* (SVM), Máquina de Vetores de Suporte, é diretamente baseado na Teoria do Aprendizado Estatístico proposta por Vapnik [84]. Essa teoria estabelece princípios a serem seguidos para obter um classificador com uma boa capacidade de generalização. Em sua implementação mais simples, o SVM consegue lidar com problemas linearmente separáveis. Desta forma, o algoritmo consegue selecionar as instâncias pertencentes a uma das duas classes, traçando uma linha de separação entre elas. Essa linha busca garantir uma margem o mais larga possível entre as duas classes. A Figura 2.3 ilustra esse processo de separação. Neste exemplo, temos instâncias pertencentes à classe -1 ou classe +1 e as linhas H1 e H2 representam os hiperplanos, que são linhas que marcam a fronteira entre as classes.

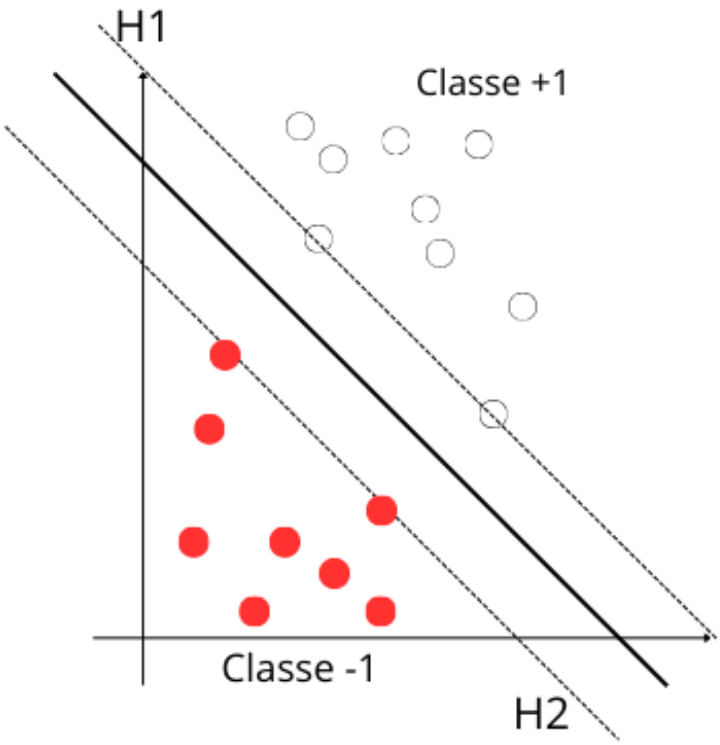


Figura 2.3 – Modelo de classificação simples com SVM.

2.2.2 *Decision Tree*

Decision Tree (DT), ou Árvores de decisão, é um algoritmo relativamente intuitivo. Harrison [26] faz um comparativo entre uma consulta médica e uma árvore de decisão, pois um médico decide entre um diagnóstico ou outro baseado em uma série de perguntas realizadas ao paciente. A principal vantagem deste algoritmo é a capacidade de lidar com dados não numéricos, o que exige pouco pré-processamento dos dados na maioria dos casos, além de permitir a interpretação dos resultados, pois é possível percorrer cada nodo da árvore para entender o porquê de determinada resposta.

Para apresentar um exemplo simples, considere os dados apresentados na Tabela 2.1. Nela temos uma relação de dias e, com base nas informações climáticas, temos a missão de decidir ir ou não à praia.

Tabela 2.1 – Dados para construção da *Decision Tree*.

Dia	Sol	Vento	Ir para praia
1	Sim	Sim	Não
2	Sim	Sim	Não
3	Sim	Não	Sim
4	Não	Não	Não
5	Não	Sim	Não
6	Não	Sim	Não

A Figura 2.4 mostra uma possível DT criada a partir dos dados da Tabela 2.1. Na árvore, fica fácil visualizar a tomada de decisão sobre ir ou não para a praia baseada nos dados climáticos históricos.

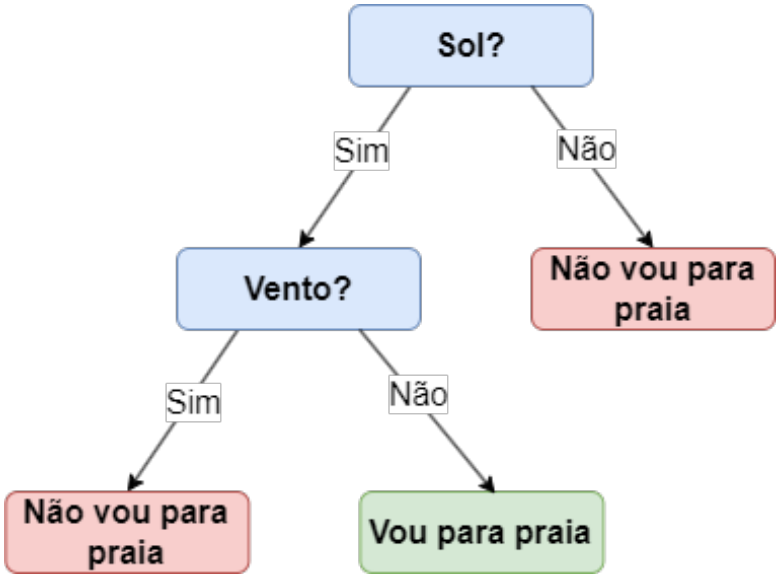


Figura 2.4 – Exemplo de árvore de decisão.

2.3 *Random Forest*

Hartshorn [27] apresenta o algoritmo Random Forest (RF), Floresta aleatória, como uma técnica de ML supervisionado. O termo *Forest* está relacionado com o fato do algoritmo criar um conjunto de DTs, sempre diferentes entre si, já que a escolha dos nós da árvore ocorre de forma aleatória a cada nova árvore criada.

A criação de várias DTs para o funcionamento do algoritmo RF é importante para evitar *overfitting*, que ocorre quando o modelo perde a capacidade de generalização. O *overfitting* ocorre quando o modelo aprende padrões ou ruídos específicos do conjunto de treinamento que não se repetem nos dados novos, comprometendo a capacidade de generalização.

Ao receber uma instância a ser classificada, cada árvore fornece um resultado para essa instância. O resultado final será a classe mais votada pelas árvores que formam a floresta. O processo é ilustrado na Figura 2.5.

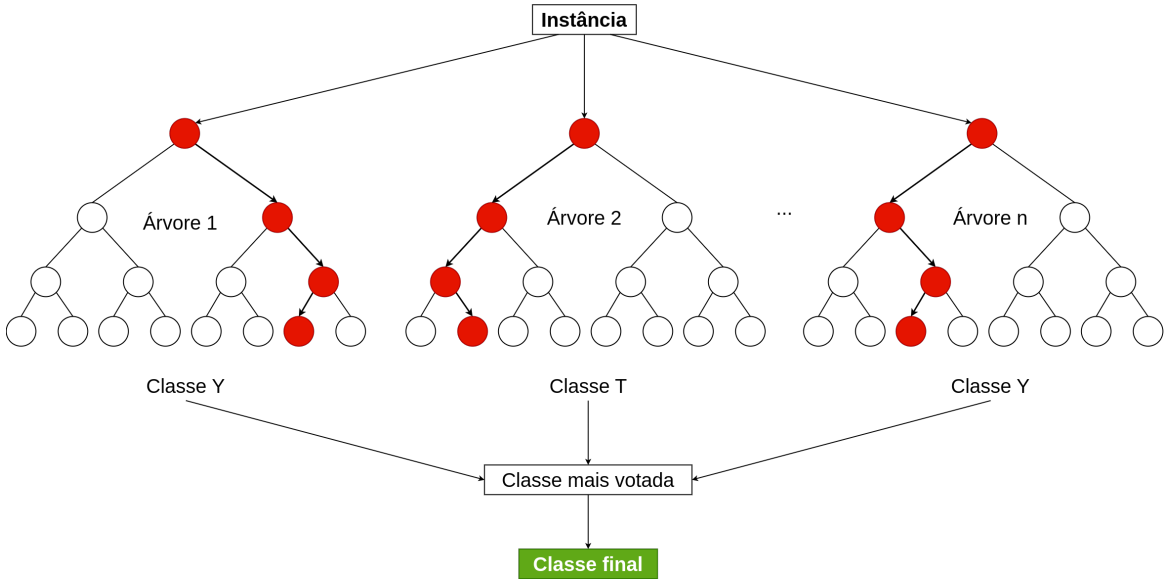


Figura 2.5 – Esquema de funcionamento do algoritmo *Random Forest*.

2.3.1 *K-Nearest Neighbors*

Izbicki e Santos [40] descrevem o algoritmo *K-Nearest Neighbor* (KNN), K-Vizinhos Mais Próximos, que é muito popular em aprendizado de máquina e possui um objetivo simples, da seguinte maneira: o KNN tenta classificar uma determinada instância de um conjunto de dados analisando a distância dele para os vizinhos mais próximos; a classe deste objeto será definida pela quantidade majoritária de votos, ou seja, ele terá a classe da maioria dos vizinhos mais próximos. Na Figura 2.6, temos a representação do funcionamento do

algoritmo KNN. Na Figura 2.6(a), é possível observar um conjunto de dados separado em duas classes. A Figura 2.6(b) ilustra a inclusão de um novo elemento para ser classificado. Aplicando o KNN, considerando o k=3, ou seja, considerando os 3 vizinhos mais próximos, é possível observar na Figura 2.6(c) que a quantidade de votos, ou vizinhos vermelhos, é maior que o número de vizinhos azuis. Desta forma, nosso novo objeto será considerado da classe vermelha, como mostrado na Figura 2.6(d).

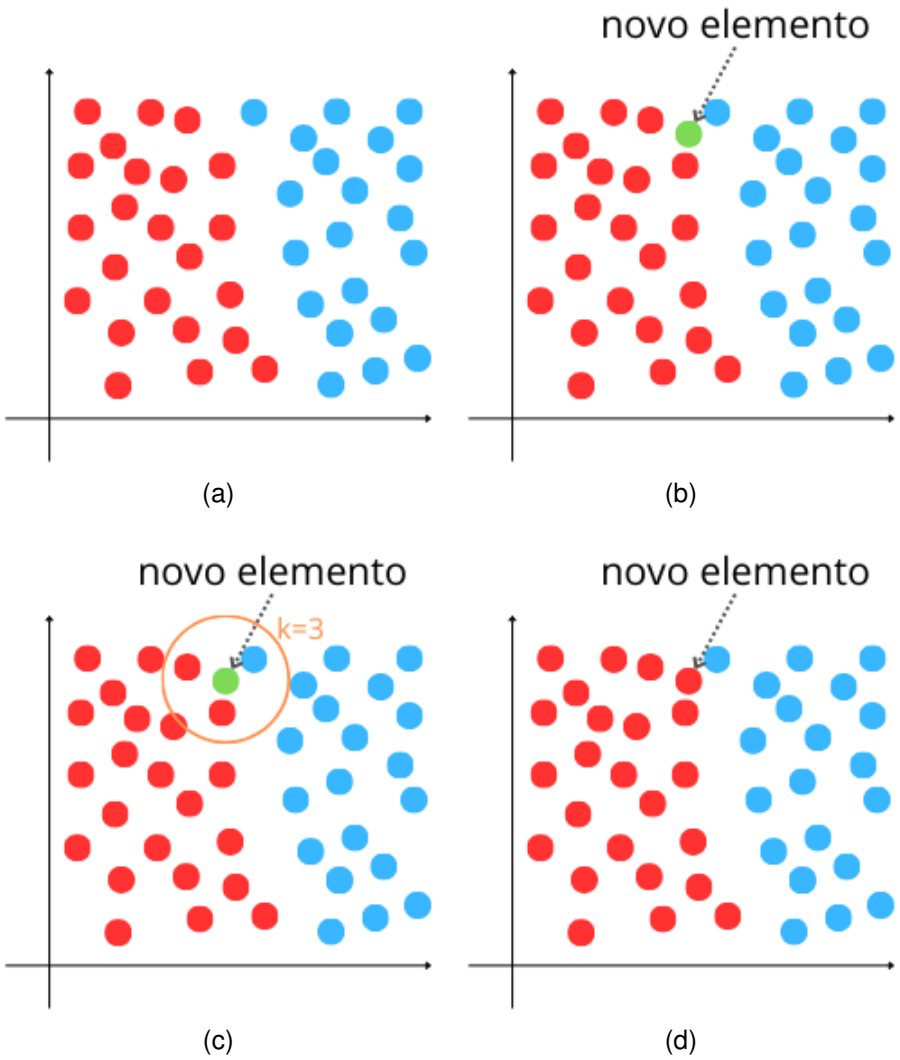


Figura 2.6 – Exemplo de funcionamento do KNN: (a) apresenta os elementos pertencentes as duas classes; na sequência, em (b), um novo elemento é adicionado para ser classificado; considerando a votação dos *k* vizinhos mais próximos em (c), a classe do novo elemento será vermelha (d).

2.3.2 *Naive Bayes*

Segundo Harrison [26], o *Naive Bayes* (NB) é um algoritmo probabilístico, baseado no teorema de *Bayes*. O *Naive Bayes* assume que os atributos dos dados são condicional-

mente independentes em relação à classe, o que simplifica os cálculos de probabilidade. Essa característica permite que o modelo seja eficiente em termos computacionais, podendo ser treinado com um conjunto relativamente pequeno de dados, ao mesmo tempo que consegue lidar com conjuntos de dados que possuem muitos atributos. O objetivo dele é simples, ele calcula a probabilidade de uma instância pertencer a determinada classe baseado em suas características. O Teorema de Bayes, que serve de base para o algoritmo, está representado na equação 2.1.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}, \tag{2.1}$$

onde:

- $P(A|B)$ é a probabilidade de A sendo que B ocorreu;
- $P(B|A)$ é a probabilidade de B sendo que A ocorreu;
- $P(A)$ e $P(B)$ são as probabilidades de A e B, respectivamente.

2.4 Redes Neurais Artificiais

Silva et al. [75] apresentam as *Artificial Neural Networks* (ANN), Redes Neurais Artificiais, como modelos computacionais criados a partir da ideia de funcionamento do sistema nervoso central de seres vivos. A Figura 2.7 ilustra o modelo de uma ANN. Nesta figura, podemos observar as informações de entrada (*x1*, *x2*, *x3*, *xn*), cada uma delas representa uma característica ou atributo dos dados. No exemplo, a rede possui duas camadas escondidas ou ocultas, e é nelas que ocorre a maior parte dos cálculos e a rede aprende a identificar padrões dos dados. Os resultados da rede neural são representados pelos valores (*y1*, *y2*, *yn*). Esta saída pode ser uma classificação ou regressão, de acordo com o resultado esperado.

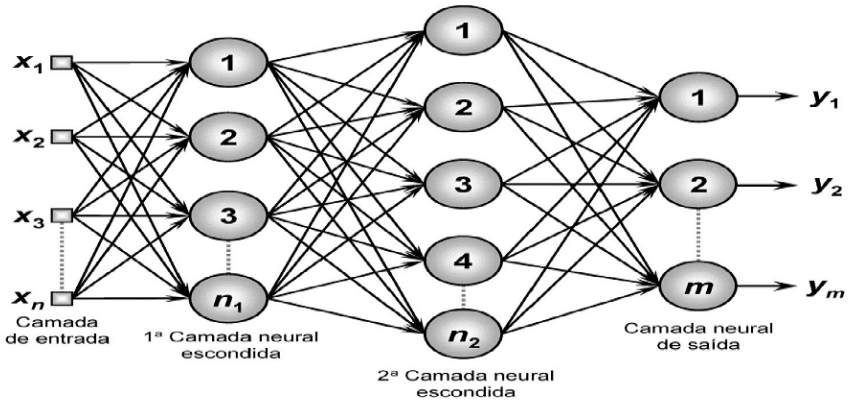


Figura 2.7 – A imagem apresenta uma *Multilayer Perceptron* (MLP), um exemplo de ANN [75].

Assim como nos seres vivos, as ANN são compostas por neurônios capazes de adquirir e manter conhecimento. Os neurônios artificiais de uma ANN são interconectados por meio de sinapses artificiais. A Figura 2.8 apresenta uma representação de um neurônio artificial. Nesta figura, pode-se observar os dados de entrada (x1, x2, xn) que são os dados que o neurônio recebe para processar. A importância de cada informação de entrada é determinada pelos pesos sinápticos (w1, w2, wn). O somatório representa uma soma ponderada dos sinais de entrada, ele soma os resultados da multiplicação de cada sinal pelo seu peso sináptico, processo chamado de combinação linear. O θ, é o limiar de ativação, ele determina o valor necessário para que o neurônio seja ativado. Se o resultado da soma for superior ao limiar de ativação, o neurônio será ativado. O resultado da soma ponderada dos sinais de entrada menos o limiar de ativação apresenta o potencial de ativação (u). A função de ativação (g(.)) determina a saída do neurônio (y) com base no potencial de ativação.

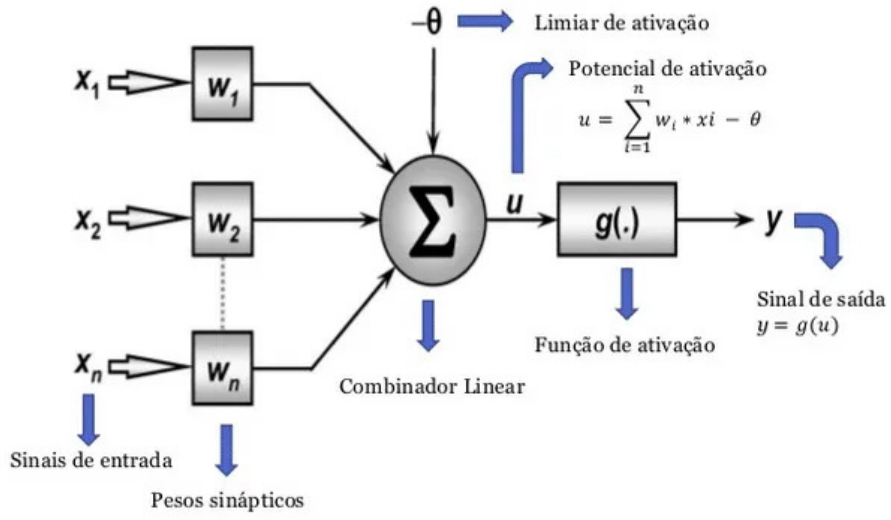


Figura 2.8 – Modelo de neurônio artificial [75].

Algumas das características das ANN são: a capacidade de se adaptar com a experiência, a capacidade de aprendizado e a habilidade de generalização, além de serem facilmente implementadas, conforme afirma Silva [75].

Entre os diferentes tipos de ANN, destaca-se a *Convolutional Neural Network* (CNN), ou Rede Neural Convolucional, que é aplicada em aprendizado profundo e é considerada muito bem-sucedida nos seus resultados. As CNNs foram desenvolvidas pelo cientista francês Yann LeCun junto com outros pesquisadores de destaque no final da década de 1980, a rede LeNet-5 foi proposta como uma solução para identificar caracteres em imagens [48]. Esse seria o início do desenvolvimento das redes convolucionais, que continuaram recebendo aperfeiçoamentos ao longo da década de 1990. As CNN podem ser aplicadas em muitas tarefas de ML, alguns exemplos de uso destas redes são reconhecimento de imagens, decifrar manuscritos e identificar objetos em imagens [61].

As CNNs foram desenvolvidas a partir da observação do processo da visão. Um experimento que foi fundamental para o desenvolvimento deste tipo de rede é o dos vencedores do Nobel David Hunter Hubel e Torsten Wiesel, que demonstraram que apenas alguns neurônios são ativados no cérebro quando o olho observa algum padrão específico como bordas orientadas em diferentes sentidos. O experimento demonstrou que os neurônios se organizam em uma hierarquia, o que aponta que a percepção visual é o resultado do trabalho conjunto de uma grande parcela de neurônios especializados [36].

Para reconhecer uma imagem, a CNN vai trabalhar com pequenas partes de cada vez, separando suas camadas *Red*, *Green* e *Blue* (RGB) em matrizes de pixels que terão valores entre 0 (completamente desligado) e 255 (completamente ligado). Um exemplo é apresentado na Figura 2.9

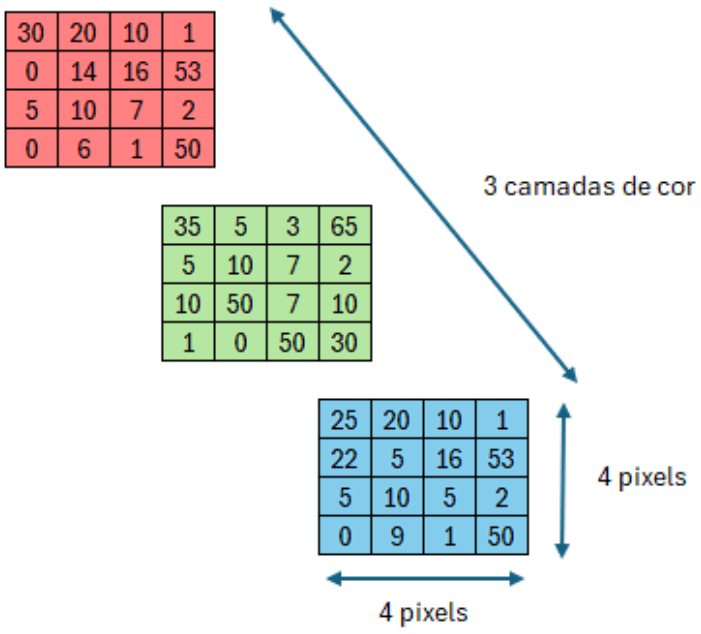


Figura 2.9 – Exemplo de entrada para uma CNN.

Em uma CNN existem basicamente três tipos de camadas: a *Convolutional layer* (CL), *Grouping layer* (GL) e *Fully Connected Layer* (FC). As primeiras camadas de uma CNN são do tipo CL, que podem ser sucedidas de outras camadas do tipo CL ou por camadas do tipo GL. A camada do tipo FC aparece no final da rede. A cada nova camada, a CNN ganha complexidade, permitindo identificar mais detalhes sobre a imagem que se está analisando, até chegar à conclusão sobre qual objeto está na imagem.

A camada *convolutional* age como um filtro, que observa pequenos quadros e percorre toda a imagem observando os traços mais marcantes, é nela que ocorrem a maioria dos cálculos. Recebe como entrada uma imagem como descrito na Figura 2.10. Essa entrada será percorrida por um detector de padrões, ele realiza um filtro observando uma região específica da imagem de entrada e determinando se o padrão está ou não presente.

Na Figura 2.10 é possível observar o funcionamento desta camada, que recebe uma entrada e um filtro (Figura 2.10(a)), varre toda a imagem aplicando o filtro (Figura 2.10(b)) e o resultado é armazenado em uma matriz de saída. Esse processo se repete até que o filtro tenha sido aplicado a toda a imagem como ilustra a Figura 2.10(c). No final, é gerada uma matriz de resultado. Os números no resultado indicam a semelhança entre o filtro e a entrada.

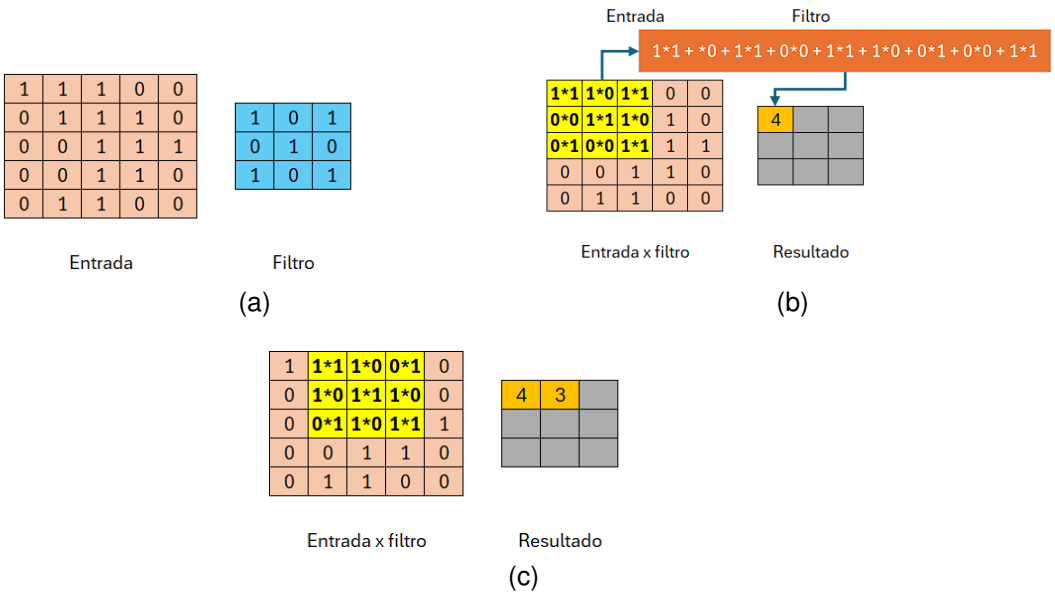


Figura 2.10 – Exemplo de processamento de uma CNN [16].

2.5 *Deep learning*

Mueller e Massaron [61] apresentam *Deep learning* (DP), Aprendizado Profundo, como uma subárea do aprendizado de máquina, que por sua vez é uma subárea da inteligência artificial. DP se caracteriza por um grande número de camadas de neurônios

artificiais definidas pelo usuário que está projetando a rede e dependendo do que se pretende reconhecer com a rede. As aplicações vão desde o reconhecimento de imagens até a análise de voz ou texto. O termo profundo, ou *deep*, está relacionado ao número de camadas que a rede possui.

Deep learning permite que os computadores possam aprender de forma independente e consigam realizar tarefas sem muita supervisão. Embora seja transparente para o usuário, é bastante provável que já tenha usado alguma ferramenta implementada usando DP. Um exemplo é a Alexa, assistente virtual da *Amazon*, que faz uso de *deep learning* para reconhecimento de voz [4].

2.6 Técnicas de validação dos resultados

Existem várias técnicas de validação de resultados que podem ser utilizadas para avaliar o desempenho de um modelo de aprendizado de máquina. Nesta seção, são apresentadas algumas das técnicas de validação dos resultados.

2.6.1 Matriz de confusão

Harrison [26] apresenta a matriz de confusão como uma ferramenta tabular que permite avaliar o desempenho de um classificador de forma detalhada. Ela apresenta uma visão dos resultados de uma classificação, destacando as diferentes categorias de classificação. Entre elas estão: os verdadeiros positivos (VP), que ocorrem quando o modelo classifica corretamente os casos positivos; os verdadeiros negativos (VN), que representam os casos negativos corretamente identificados pelo modelo; os falsos positivos (FP), que ocorrem quando o modelo erroneamente classifica um caso negativo como positivo; e os falsos negativos (FN), que ocorrem quando o modelo classifica erroneamente um caso positivo como negativo. A interpretação desses resultados pode ser realizada utilizando a matriz apresentada na Figura 2.11.

		Classificação real	
		P	**N**
Classificação prevista	**P**	**VP**	FP
	N	FN	**VN**

Figura 2.11 – Modelo de matriz de confusão.

Por exemplo, considere uma situação em que o algoritmo tem a tarefa de determinar se uma mulher está grávida ou não. A Figura 2.12 ilustra esse cenário, onde atribuímos o valor 1 ao estado “estar grávida” e o valor 0 ao estado “não estar grávida”. Os dados a serem analisados pelo algoritmo são mostrados na Figura 2.12(a), na qual a tabela está dividida em duas colunas: na primeira coluna temos os dados reais e na segunda coluna as respostas do algoritmo de classificação. A Figura 2.12(b) apresenta a matriz de confusão para esses dados, permitindo distinguir entre os acertos do algoritmo, verdadeiros positivos e verdadeiros negativos, e os erros, falsos positivos e falsos negativos. A construção da matriz de confusão é simples, basta contar e preencher nas colunas correspondentes. Por exemplo, ao analisar os dados, identificamos três casos em que a classe real é 1 e a classe prevista também é 1, esses são os três valores verdadeiro positivo.

Dados	
Classe real	Classe prevista
1	1
0	0
1	0
0	1
0	0
0	0
1	1
0	1
1	1
0	0

(a)

		Classificação real	
		P	N
Classificação prevista	P	3	1
	N	2	4

(b)

Figura 2.12 – Exemplo construção matriz de confusão, (a) apresenta uma tabela com os valores previstos pelo algoritmo e os valores reais, na (b), apresenta a matriz de confusão gerada com os resultados da tabela.

2.6.2 Métricas Acurácia, Precisão, *Recall* e F1-Score

A construção da matriz de confusão permite a avaliação do desempenho do classificador através de diversas métricas. Entre estas métricas, podemos destacar a acurácia (Equação 2.2), precisão (Equação 2.3), *recall* (Equação 2.4) e F1-Score (Equação 2.5). Essas medidas fornecem uma visão abrangente da eficácia do modelo em diferentes aspectos, como sua capacidade de classificar corretamente instâncias positivas e negativas, bem como sua habilidade de evitar falsos positivos e falsos negativos [26].

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN}, \tag{2.2}$$

$$\text{Precisão} = \frac{VP}{VP + FP}, \tag{2.3}$$

$$\text{Recall} = \frac{VP}{VP + FN}, \tag{2.4}$$

$$\text{F1-Score} = 2 * \frac{precisão * recall}{precisão + recall}, \tag{2.5}$$

onde:

- *VP* é a quantidade de verdadeiros positivos;
- *VN* é a quantidade de verdadeiros negativos;
- *FP* é a quantidade de falsos positivos;
- *FN* é a quantidade de falsos negativos.

Cada uma destas métricas permite avaliar o desempenho do classificador, oferecendo uma visão geral sobre os resultados. A acurácia, por exemplo, indica a performance geral do modelo, representando a proporção de todas as classificações realizadas que foram corretas. A precisão, por sua vez, revela o percentual de predições positivas que estão corretas. O *recall* mostra a porcentagem de valores positivos que foram classificados corretamente pelo modelo. Quanto ao *F1-Score*, ele proporciona uma medida de harmonia entre precisão e *recall*, sendo especialmente útil na avaliação de conjuntos de dados desbalanceados. Essas métricas são importantes para uma avaliação abrangente do desempenho do classificador [70].

2.6.3 Validação cruzada

Na validação cruzada, o conjunto de instâncias é dividido em um determinado número de subconjuntos de tamanho aproximadamente igual. Um desses subconjuntos é reservado para testar o modelo, enquanto os demais são utilizados para o seu treinamento. Esse processo é repetido, alterando qual subconjunto é reservado para teste, até que todos os subconjuntos tenham sido usados para testar o modelo, conforme ilustrado na Figura 2.13. O desempenho final do modelo é a média dos desempenhos obtidos em todas as rodadas de teste e treinamento [17].

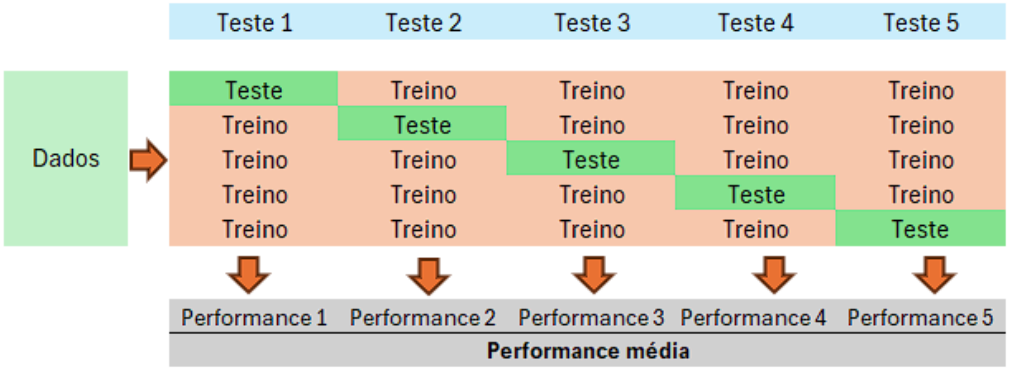


Figura 2.13 – Exemplo de funcionamento do processo de validação cruzada.

O uso da validação cruzada é importante para identificar a ocorrência de *overfitting*, pois avalia o modelo com diferentes subconjuntos dos dados, garantindo que o desempenho não seja dependente de um único subconjunto específico. Além disso, esse processo melhora a capacidade de generalização do modelo de ML, permitindo uma estimativa mais confiável de como ele deve se comportar com novos dados [17].

2.6.4 Balanceamento

O balanceamento de dados é uma etapa fundamental em problemas de classificação, especialmente quando as classes não têm uma distribuição uniforme nos conjuntos de dados. Isso pode influenciar o desempenho do modelo, fazendo com que ele fique tendencioso para a classe majoritária e tenha dificuldade em prever corretamente as classes minoritárias [26].

Para lidar com essa questão, é possível realizar a manipulação da amostragem do conjunto de dados. As duas principais técnicas que permitem modificar a quantidade de instâncias do conjunto são o *undersampling* e o *oversampling*. No *undersampling*, o número de instâncias da classe majoritária é reduzido até que fique equilibrado com o da classe minoritária. O percentual de ajuste pode ser configurado no momento da implementação do algoritmo. Por outro lado, o *oversampling* cria instâncias sintéticas para a classe minoritária até que as classes estejam balanceadas [8].

É possível combinar essas duas estratégias para obter melhores resultados, reduzindo uma parte das instâncias da classe majoritária e gerando instâncias sintéticas para a classe minoritária, de modo a equilibrar as classes. O resultado do processo de balanceamento de classes pode ser visualizado em um gráfico de barras, como no exemplo apresentado na Figura 2.14. Nesse exemplo, foi aplicado *undersampling* para remover 70% do excesso de instâncias da classe majoritária e, em seguida, *oversampling* para gerar 30% de instâncias sintéticas na classe minoritária, tornando-as balanceadas.

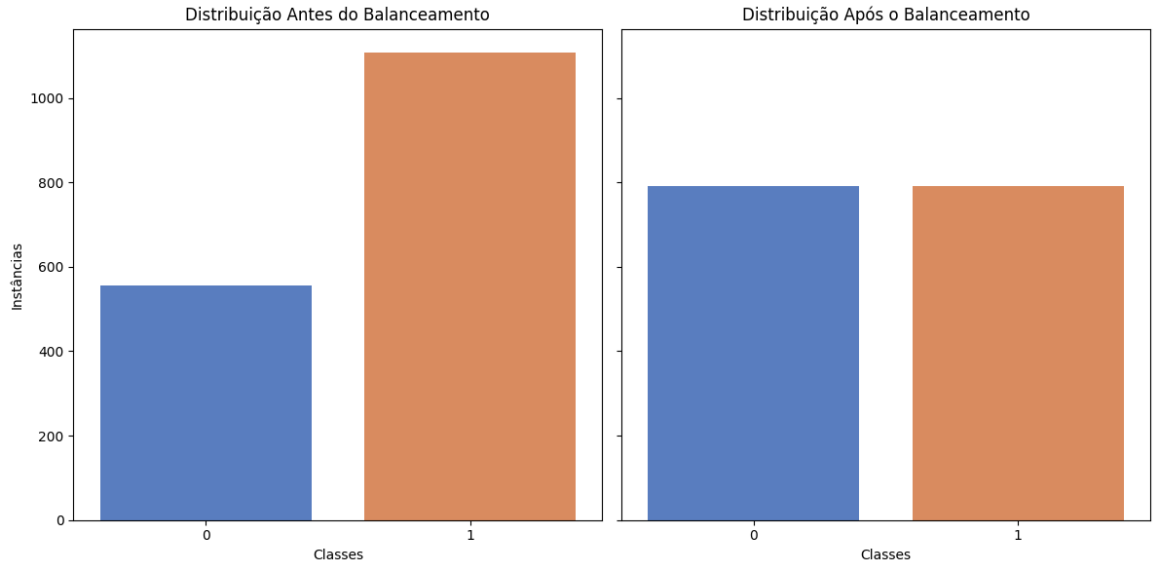


Figura 2.14 – Gráfico ilustrando o resultado do processo de balanceamento, aplicando simultaneamente *undersampling* e *oversampling*.

2.6.5 ROC e AUC

A curva ROC (*Receiver Operating Characteristic*) é uma ferramenta amplamente utilizada para avaliar o desempenho de modelos de classificação, sendo especialmente relevante em cenários nos quais o balanceamento de classes é uma preocupação [18]. Essa curva permite visualizar como o modelo equilibra a taxa de verdadeiros positivos (TPR) e a taxa de falsos positivos (FPR).

A Figura 2.15 apresenta um exemplo de curva ROC. No eixo y, está representada a taxa de verdadeiros positivos, e no eixo x, a taxa de falsos positivos. A linha vermelha indica os resultados de um classificador aleatório. Um aspecto fundamental a ser analisado nos resultados apresentados é o AUC (*Area Under the Curve*). Valores de AUC próximos a 1 indicam um desempenho excelente, enquanto um AUC de 0,5 corresponde a uma classificação aleatória. Valores abaixo de 0,5 indicam um desempenho inferior ao de um classificador aleatório, sendo, portanto, insatisfatórios [25].

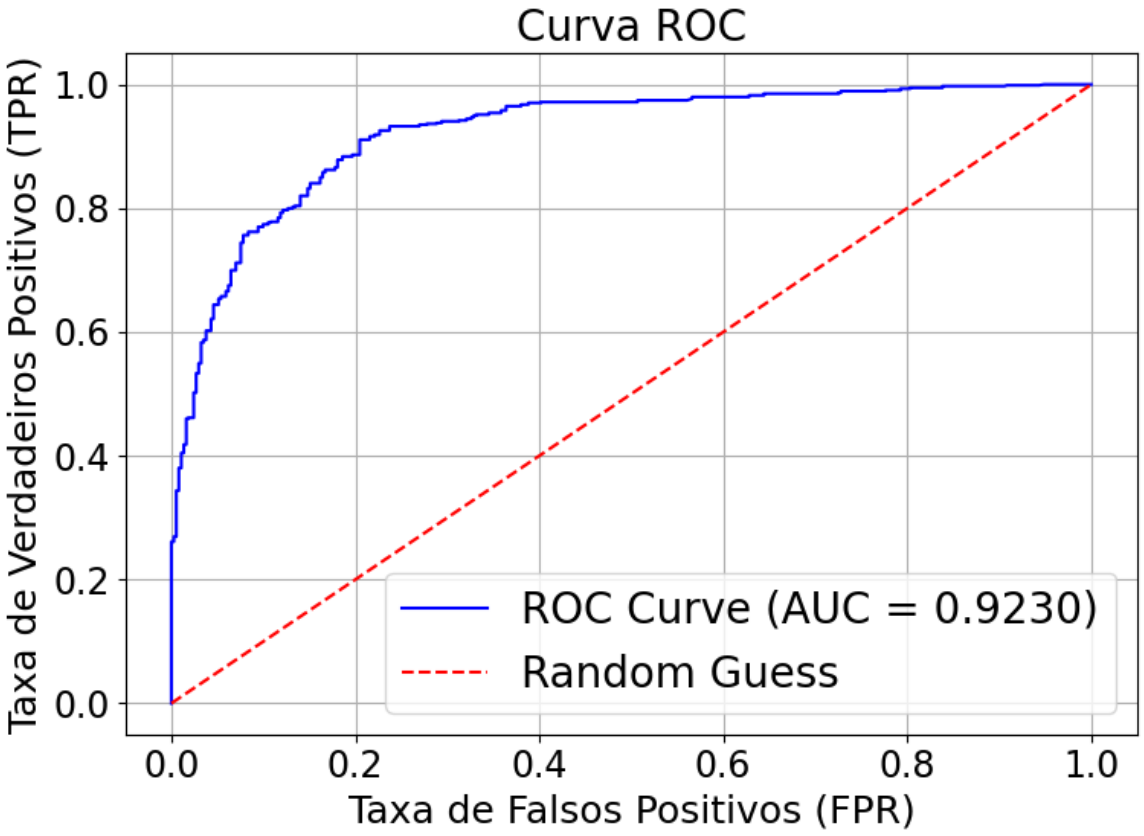


Figura 2.15 – Exemplo de curva ROC, comparando com um classificador aleatório e cálculo do AUC.

2.6.6 SHAP

Um tema que tem tido uma crescente importância em ML é a necessidade de entender o comportamento dos modelos implementados. Para isso, uma das alternativas é o uso do *SHapley Additive exPlanations* (SHAP). O cálculo do Shapley aponta o impacto de cada atributo sobre o resultado do modelo, além da importância de determinado valor em comparação aos demais [56]. Esses resultados podem ser apresentados em gráficos, fornecendo uma melhor interpretação dos resultados.

A Figura 2.16 apresenta um sumário dos resultados calculados pelo SHAP para um algoritmo *Random Forest*, treinado para classificar pacientes com câncer em sobrevida longa e curta usando dados clínicos dos pacientes. No eixo y, as variáveis estão dispostas em ordem decrescente de importância. Cada ponto ao longo do eixo x representa uma instância do *dataset*, e sua posição indica o impacto do atributo na decisão da classe para essa instância. Pontos localizados à direita (valores positivos) indicam que o atributo teve maior contribuição para a decisão na direção positiva da classe analisada. Além disso, a cor dos pontos representa o valor do atributo para a instância: vermelho indica valores altos, enquanto azul indica valores baixos [66].

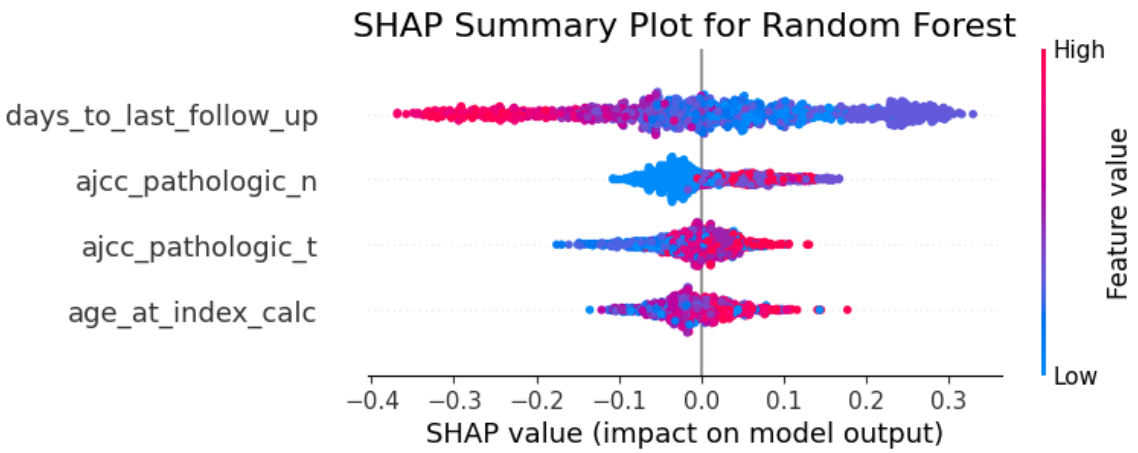


Figura 2.16 – Exemplo de *summary plot*.

Outra opção de visualização que pode ser gerada com os valores SHAP é o gráfico de importâncias. Um exemplo é apresentado na Figura 2.17. Neste gráfico, é possível identificar o impacto médio de cada atributo nas decisões do modelo. No eixo y, as variáveis são ordenadas da mais importante para a menos importante, enquanto no eixo x é apresentada a média dos valores absolutos de SHAP, que representa o impacto médio de cada variável nas previsões [66].

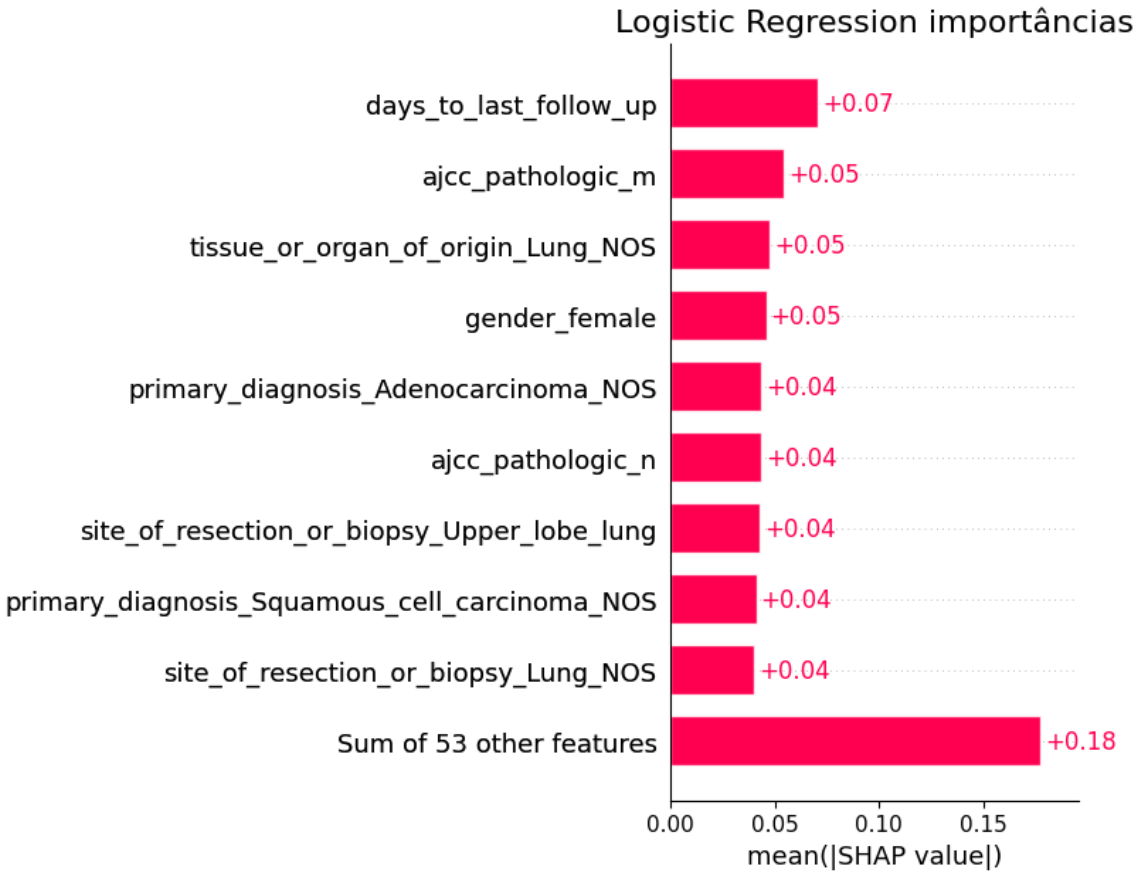


Figura 2.17 – Exemplo de gráfico de importâncias das variáveis no resultado do algoritmo.

3. TRABALHOS RELACIONADOS

Este capítulo apresenta como foi realizada a seleção e análise de trabalhos relacionados. A apresentação das informações e percepções obtidas está organizada de acordo com alguns critérios para melhorar a compreensão do contexto da pesquisa.

3.1 Metodologia de Pesquisa

Nesta seção, é apresentada a metodologia adotada para a busca de estudos relevantes, incluindo as questões de pesquisa, *string* de busca e os critérios de inclusão e exclusão. O primeiro passo para encontrar os estudos foi a aplicação de uma revisão sistemática da literatura apresentada por Kitchenham [43], que está ilustrada na Figura 3.1.

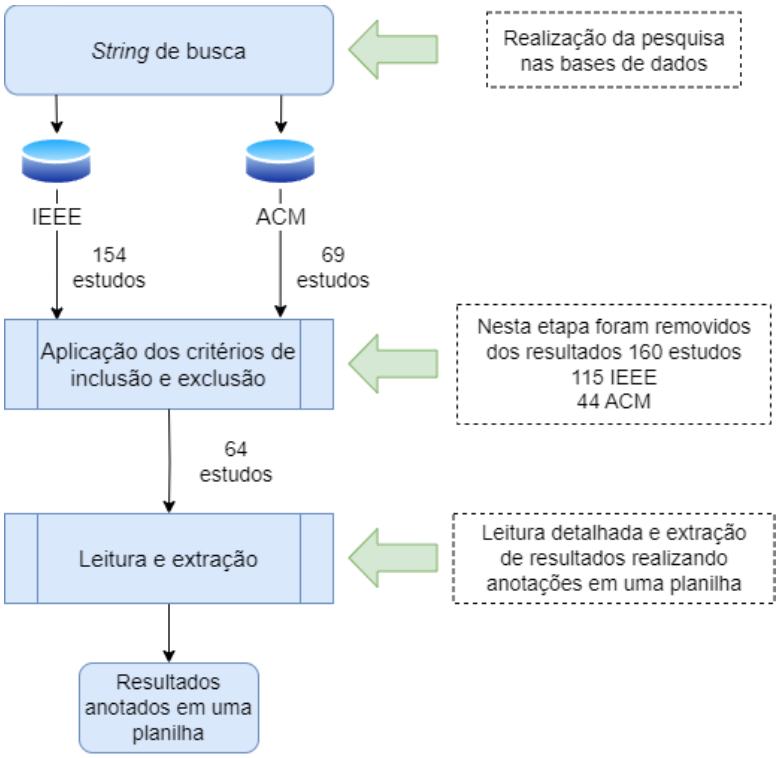


Figura 3.1 – Fluxo do processo de seleção dos estudos.

3.1.1 Questões de pesquisa

A seguir são listadas as questões que norteiam a pesquisa:

QP1 Quais técnicas têm sido aplicadas para prever o tempo de sobrevida em pacientes com câncer?

QP2 Quais as diferenças das técnicas usadas para prever o tempo de sobrevida em pacientes com câncer de pulmão em relação a outros órgãos?

QP3 Quais conjuntos de dados são utilizados para prever o tempo de sobrevida em pacientes com câncer?

QP4 Como as técnicas de ML tem sido usadas para chegar a tempos mais precisos?

QP5 Quais métricas tem sido usadas para avaliar o desempenho dos modelos de ML usados para prever o tempo de sobrevida de pacientes com câncer?

3.1.2 *String* de busca

Para o realizar o levantamento de conteúdo sobre o tema, foram adotados os seguintes termos de busca: *survival time*, *cancer* e *prediction*. A pesquisa foi conduzida em duas bases de dados: *IEEE Xplore Digital Library*[1] e *ACM Digital Library*[2], que, segundo Galvão [20], são referências especialmente na área de ciência da computação. Nessas duas bases, os termos de pesquisa mencionados foram aplicados utilizando *strings* de busca ajustadas para cada base, conforme apresentado na Tabela 3.1. Sobre os resultados obtidos usando a *string* de busca, foram aplicados os critérios de inclusão e exclusão apresentados na Seção 3.1.3.

Tabela 3.1 – *String* de busca utilizada.

Base	***String* de busca**
IEEE	survival time AND cancer AND prediction
ACM	[Abstract: "survival time"] AND [Abstract: cancer] AND [Abstract: prediction]

A escolha dos termos de pesquisa permite obter uma visão geral das diversas técnicas, com ou sem o uso de modelos de ML, utilizadas no processo de avaliação do tempo de sobrevida de pacientes com câncer. Mesmo que o principal interesse desta pesquisa seja pelo câncer de pulmão, não se restringiu a este tipo específico na pesquisa para podermos observar uma visão geral sobre as várias técnicas aplicadas no processo de avaliação do tempo de sobrevivência em pacientes com câncer.

[1] https://ieeexplore.ieee.org
[2] https://dl.acm.org

3.1.3 Critérios de inclusão e exclusão

Foram adotados os seguintes critérios de inclusão durante a seleção de artigos para compor a base referência desta pesquisa:

CI1 Estar disponível nas bibliotecas digitais que temos acesso.

CI2 Apresentar as palavras chave usadas para busca no resumo.

CI3 Estar escrito em Inglês.

CI4 Ter pelo menos 4 páginas.

Os artigos foram excluídos de acordo com os seguintes critérios:

CE1 Trabalhos duplicados.

CE2 Não responder a nenhuma das questões de pesquisa.

3.1.4 Resultados por base

Após realizar a busca, foram selecionados 223 trabalhos, sendo 154 da base de pesquisa IEEE e 69 da ACM. Na próxima etapa, os estudos foram analisados de acordo com os critérios de inclusão e exclusão estabelecidos, verificando-se o alinhamento do conteúdo com as questões a serem respondidas.

Os artigos resultantes da avaliação dos critérios de inclusão e exclusão foram lidos para extração de informações. Essas informações foram anotadas em uma tabela para análise. A listagem das quantidades de artigos pode ser observada na Tabela 3.2 e a lista completa dos artigos, com seus autores e anos de publicação, é apresentada na Tabela 3.3.

Tabela 3.2 – Detalhamento das quantidades de artigos encontrados.

Base	**Quantidade de estudos**			
	encontrados	**removidos**	**aproveitados**	**total**
IEEE	154	115	39	**64**
ACM	69	44	25	

Tabela 3.3 – Relação dos artigos, autores e anos de publicação.

Ano	Autor(es)	Título
2004	Li e Li [49]	Dimension reduction methods for microarrays with application to censored survival data
2008	Liu [54]	Cox's Proportional Hazards Model with Lp Penalty for Biomarker Identification and Survival Prediction
2009	Chen et al. [14]	Artificial Neural Network Prediction for Cancer Survival Time by Gene Expression Data
2009	Phong et al. [65]	Hedge Algebra Based Type-2 Fuzzy Logic System and its Application to Predict Survival Time of Myeloma Patients
2009	Ture et al. [82]	Using Kaplan-Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients
2010	Dekker et al. [15]	Survival Prediction in Lung Cancer Treated with Radiotherapy: Bayesian Networks vs. Support Vector Machines in Handling Missing Data
2010	Hayward et al. [29]	Machine learning of clinical performance in a pancreatic cancer database
2011	Chen et al. [13]	Prediction of survival in patients with liver cancer using artificial neural networks and classification and regression trees
2011	Chan et al. [12]	Machine learning of patient similarity: A case study on predicting survival in cancer patient after locoregional chemotherapy
2011	Kolasa et al. [45]	Optimization of hidden layer in a neural network used to predict bladder-cancer patient-survival
2011	Yu et al. [94]	Learning patient-specific cancer survival distributions as a sequence of dependent regressors
2013	Zhou et al. [98]	A Texture Feature Ranking Model for Predicting Survival Time of Brain Tumor Patients
2014	Zhou et al. [97]	Exploring Brain Tumor Heterogeneity for Survival Time Prediction
2014	Hawkins et al. [28]	Predicting Outcomes of Nonsmall Cell Lung Cancer Using CT Image Features
2014	Pu et al. [68]	Application of artificial neural network and multiple linear regression models for predicting survival time of patients with non-small cell cancer using multiple prognostic factors including FDG-PET measurements
2015	Gan et al. [21]	A survey of pattern classification-based methods for predicting survival time of lung cancer patients
2015	Chai et al. [11]	The L1/2 regularization approach for survival analysis in the accelerated failure time model
2015	Kim et al. [42]	Predicting censored survival data based on the interactions between meta-dimensional omics data in breast cancer
2016	Chaddad et al. [10]	Radiomic analysis of multi-contrast brain MRI for the prediction of survival in patients with glioblastoma multiforme
2016	Liu et al. [52]	Outcome Prediction for Patient with High-Grade Gliomas from Brain Functional and Structural Networks
2016	Zhang et al. [96]	Improve Glioblastoma Multiforme Prognosis Prediction by Using Feature Selection and Multiple Kernel Learning
2016	Malhotra et al. [58]	Constraint based temporal event sequence mining for Glioblastoma survival prediction
2016	Zhang et al. [95]	The modularity and dynamicity of miRNA–mRNA interactions in high-grade serous ovarian carcinomas and the prognostic implication
2017	Ferdinand Christ et al. [19]	SurvivalNet: Predicting patient survival from diffusion weighted magnetic resonance images using cascaded fully convolutional and 3D Convolutional Neural Networks
2017	Tyuryumina e Neznanov [83]	On Consolidated Predictive Model of the Natural History of Breast Cancer Considering Primary Tumor and Primary Distant Metastases Growth
2017	Isik e Ercan [39]	Integration of RNA-Seq and RPPA data for survival time prediction in cancer patients
2017	Kuo et al. [47]	Prognostic value of tumor volume for patients with advanced lung cancer treated with chemotherapy
2018	Bartholomai e Frieboes [6]	Lung Cancer Survival Prediction via Machine Learning Regression, Classification, and Statistical Techniques
2018	Wang et al. [87]	Unsupervised Deep Learning Features for Lung Cancer Overall Survival Analysis
2019	Ahmed et al. [1]	Neuroimaging Based Survival Time Prediction of GBM Patients Using CNNs from Small Data
2019	Jayashanka et al. [41]	Machine Learning Approach to Predict the Survival Time of Childhood Acute Lymphoblastic Leukemia Patients
2019	Cai et al. [9]	The Early Stage Lung Cancer Prognosis Prediction Model based on Support Vector Machine

Continua na próxima página.

Ano	Autor(es)	Título
2019	Huang e Liang [35]	A Novel Cox Proportional Hazards Model for High-Dimensional Genomic Data in Cancer Prognosis
2019	Wang et al. [88]	A novel Log penalty in a path seeking scheme for biomarker selection
2019	Li et al. [51]	Multi-task learning based survival analysis for multi-source block-wise missing data
2019	Hossain et al. [31]	Machine learning and bioinformatics models to identify gene expression patterns of ovarian cancer associated with disease progression and mortality
2020	Stepanek et al. [78]	A Machine-learning Approach to Survival Time-event Predicting: Initial Analyses using Stomach Cancer Data
2020	Lu et al. [55]	An integrated model of clinical information and gene expression for prediction of survival in breast cancer patients
2020	Nanda e Duraipandian [62]	Prediction of Survival Rate from Non-Small Cell Lung Cancer using Improved Random Forest
2020	Aljouie e Roshan [2]	Multi-path convolutional neural network for glioblastoma survival group prediction with point mutations and demographic features
2020	Wang et al. [86]	Cluster-Boosted Multi-Task Learning Framework for Survival Analysis
2020	Sun et al. [79]	Survival Risk Prediction of Esophageal Cancer Based on Self-Organizing Maps Clustering and Support Vector Machine Ensembles
2020	Aljouie et al. [3]	Challenges in predicting glioma survival time in multi-modal deep networks
2020	Wei et al. [90]	Using Multiple Machine Learning Algorithms for Cancer Prognosis in Lung Adenocarcinoma
2021	Wu et al. [92]	DeepMMSA: A Novel Multimodal Deep Learning Method for Non-small Cell Lung Cancer Survival Analysis
2021	Liu et al. [53]	Survival Time Prediction of Breast Cancer Patients Using Feature Selection Algorithm Crystall
2021	Wu et al. [91]	A generative adversarial network-based CT image standardization model for predicting progression-free survival of lung cancer
2021	Baker et al. [5]	Predicting Lung Cancer Survival Time Using Deep Learning Techniques
2021	Hossain et al. [32]	Machine learning and network-based models to identify genetic risk factors to the progression and survival of colorectal cancer
2022	Scalco et al. [71]	Automatic Feature Construction Based on Genetic Programming for Survival Prediction in Lung Cancer Using CT Images
2022	Wang et al. [85]	Data-driven intelligent decision for multimedia medical management
2022	Sharma et al. [74]	A deep learning-based integrative model for survival time prediction of head and neck squamous cell carcinoma patients
2022	Ma et al. [57]	Optimizing the Prognostic Model of Cervical Cancer Based on Artificial Intelligence Algorithm and Data Mining Technology
2022	Yang et al. [93]	A multi-omics machine learning framework in predicting the survival of colorectal cancer patients
2022	Wang et al. [89]	CondiS: A conditional survival distribution-based method for censored data imputation overcoming the hurdle in machine learning-based survival analysis
2023	Ghazipour et al. [23]	Survival Outcome Prediction for Stereotactic Body Radiation Therapy of Lung Cancer from Post-RT Ct Images with RNN/CNN Deep Learning
2023	Naser et al. [63]	Prediction Model of Breast Cancer Survival Months: A Machine Learning Approach
2023	Timilsina et al. [81]	Machine Learning Survival Models for Relapse Prediction in a Early Stage Lung Cancer Patient
2023	Shao et al. [73]	FAM3L: Feature-Aware Multi-Modal Metric Learning for Integrative Survival Analysis of Human Cancers
2023	Hou et al. [33]	Deep Clustering Survival Machines with Interpretable Expert Distributions
2023	Li et al. [50]	Causally-Aware Intraoperative Imputation for Overall Survival Time Prediction
2023	SShakir et al. [72]	A deep learning-based cancer survival time classifier for small datasets
2023	Kukreja et al. [46]	A Heuristic Machine Learning-Based Optimization Technique to Predict Lung Cancer Patient Survival

3.2 Aspectos gerais dos estudos

Após a seleção dos artigos, conforme descrito na Seção 3.1, eles foram analisados para a extração de informações, que foram organizadas em uma tabela, possibilitando uma visão comparativa entre os estudos.

Os 64 artigos selecionados foram publicados entre 2004 e 2023. A distribuição do número de estudos anual pode ser vista na Figura 3.2. O gráfico aponta um aumento nas publicações relacionadas ao tema da predição de sobrevida ao longo dos anos, indicando sua crescente relevância entre os pesquisadores. As seções a seguir apresentam uma análise destes trabalhos em diferentes pontos de vista.

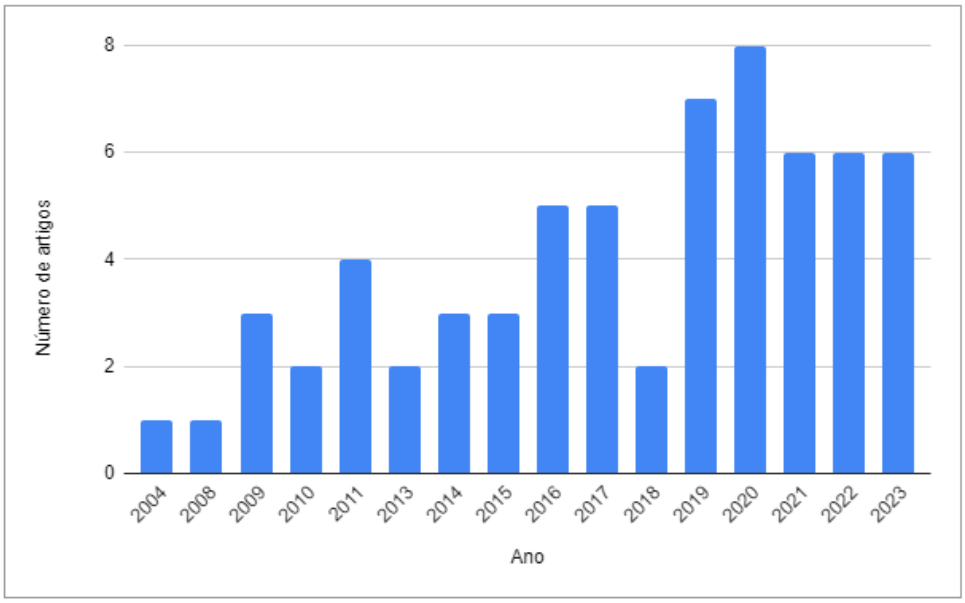


Figura 3.2 – Total de trabalhos por ano de publicação.

Nas seções a seguir, são apresentados os resultados dessa análise, incluindo informações como os anos de publicação, os algoritmos utilizados, as métricas de validação, entre outros aspectos.

3.2.1 Quanto ao órgão, tecido ou sistema afetado

Analisando os estudos pelo órgão, sistema ou tecido afetado pela doença é interessante destacar que 22 deles abordaram câncer de pulmão. O segundo tipo mais prevalente nos artigos é o câncer de mama, seguido do câncer de cérebro. Isto é justificável porque o câncer de pulmão é considerado um dos tipos mais letais da doença, de acordo com dados mundiais, segundo aponta o Instituto Nacional do Câncer (INCA) [38]. Alguns dos estudos não especificam o tipo de câncer analisado, dizendo apenas que utilizaram

uma base de dados com as informações sobre pacientes com a doença. Nestes casos, eles foram classificados como “não especificado”. A Figura 3.3 apresenta uma relação entre a quantidade de estudos em relação ao órgão, tecido ou sistema afetado.

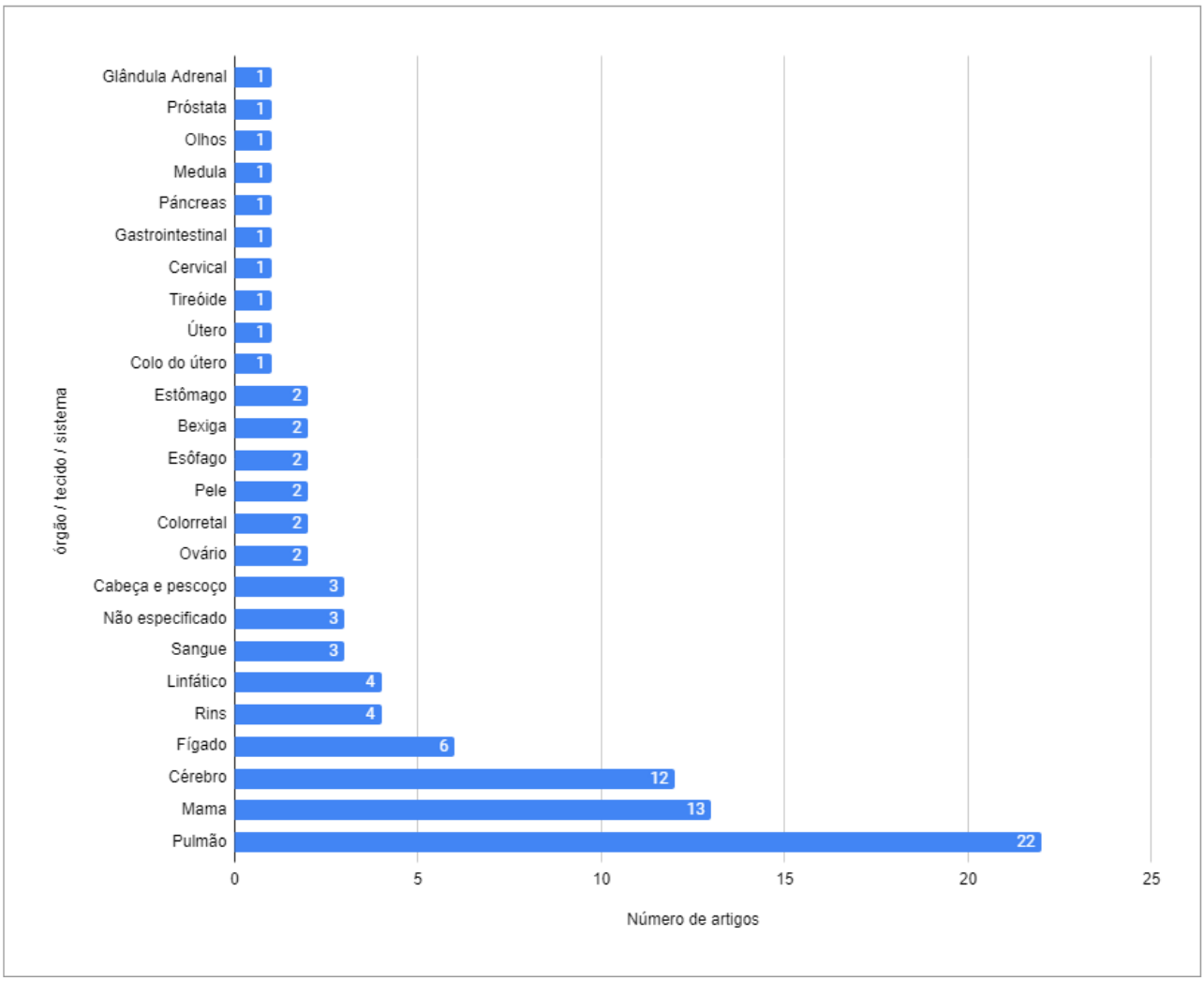


Figura 3.3 – Órgão, sistema ou tecido e a quantidade de estudos que os avaliam.

3.2.2 Quanto a aplicação de *Machine Learning*

O uso de inteligência artificial como ferramenta de apoio ao usuário em diversos tipos de tarefas e softwares de uso cotidiano tem aumentado, segundo aponta a *Microsoft* [60]. Na área da saúde também é cada vez mais comum o emprego de algoritmos de aprendizado de máquina, principalmente no auxílio de diagnósticos e prognósticos de doenças [24]. Mesmo não tendo sido um dos termos de pesquisa de seleção de artigos, a maioria dos estudos encontrados usa alguma técnica de ML. Do total de 65 artigos, 51 deles mencionam ter aplicado alguma técnica de AI no desenvolvimento da pesquisa, mesmo que nem todos detalhem especificamente o algoritmo aplicado. Modelos de ML têm a capaci-

dade de analisar grandes volumes de dados, tarefa que seria muito demorada com técnicas estatísticas tradicionais. Além disso, eles permitem realizar experimentos de forma mais ágil, o que contribui para o aumento do número de pesquisas que utilizam algoritmos de ML para obter resultados.

Tabela 3.4 – Relação dos algoritmos e artigos onde foram citados.

Algoritmo	**Citado em**
Support Vector Machine	[63, 98, 15, 79, 81, 5, 21, 12, 97, 9, 28, 52, 90, 93, 89]
(não especificado)	[78, 53, 73, 54, 10, 83, 49, 74, 88, 58, 57, 32, 31, 95]
Convolutional Neural Networks	[23, 2, 92, 91, 1, 33, 3, 50, 19, 85]
Random Forest	[63, 6, 62, 81, 21, 39, 93, 89]
K-Nearest Neighbors	[63, 21, 97, 29, 90, 46, 89]
Deep Convolutional Neural Networks	[63, 21, 97, 29, 90, 46, 89]
Decision Tree	[63, 28, 29, 85]
3D Convolutional Neural Network	[92, 91, 19]
Classification and Regression Trees	[13, 90, 82]
Naive Bayes	[28, 46, 93]
Bayesian Networks	[15, 29]
Genetic Algorithm	[79, 65]
Gradient Boosted Machines	[6, 89]
Least Absolute Shrinkage and Selection Operator	[55, 11]
Multi-Task Learning	[86, 51]
Multilayer Perceptron	[5, 45]
Recurrent Neural Networks	[23, 85]
Ridge regression	[11, 89]
Ada Boost Regressor	[5]
Adaptive Moment Estimation	[5]
Artificial Bee Colony for Support Vector Machine	[79]
Backpropagation	[14]
Borderline-SMOTE	[9]
C4.5	[82]
Chi-squared Automatic Interaction Detection	[82]
Convolutional Autoencoders	[87]
Elastic net	[11]
General Linear Regression	[6]
Generative Adversarial Networks	[91]
Genetic Algorithm for Support Vector Machine	[79]
Genetic Programming	[71]
Gradient Boosting Decision Tree	[85]
Grammatical Evolution Neural Network	[42]
Hierarchical Clustering	[86]
Improved Random Forest	[62]
Long Short-Term Memory	[85]
Multi-Path convolutional neural network	[2]
Multiple Kernel Learning	[96]
Multitask Logistic Regression	[94]
Neural Feedforward Network	[72]
Neural Networks With Three Hidden Layers	[68]
Particle Swarm Optimization for Support Vector Machine	[79]
Quick, Unbiased, Efficient Statistical Tree	[82]
Regularizing support vector machine	[11]
Self-Organizing Maps	[79]
Self-Paced Learning	[35]
Smoothly Clipped Absolute Deviation	[11]
Support Vector Regression	[41]
Survival Clustering Analysis	[33]
Weighted Fuzzy C-Means	[47]

A Tabela 3.4 apresenta a relação entre os estudos e os algoritmos de ML aplicados. Podemos observar que uma técnica muito popular nos estudos é o *Support Vector Machine*, seguido de *Convolutional Neural Networks*, algoritmos que podem ser aplicados em diferentes tarefas de classificação ou regressão, de acordo com objetivo da pesquisa.

3.2.3 Quanto ao tipo de tarefa

A Tabela 3.5 apresenta o total de estudos de aprendizado de máquina de acordo com a classificação dos tipos de tarefa executada. Alguns estudos, embora se refiram a técnicas de regressão, não aplicaram nenhuma técnica de aprendizado de máquina, limitando-se a análises estatísticas. Por esse motivo, não foram considerados nesta contagem.

Tabela 3.5 – Relação dos tipos de tarefas e quantidades de artigos.

Tipo de tarefa	**Quantidade de estudos**
Classificação	31
Regressão	18
Classificação e Regressão	5

Uma das aplicações simultâneas de tarefas de classificação e regressão realiza inicialmente uma separação das instâncias em duas ou mais classes por meio da classificação. Em seguida, sobre cada um desses subconjuntos menores de dados, aplica-se uma regressão, obtendo valores mais específicos. Esse foi o trabalho realizado por Bartholomai [6], que dividiu o conjunto de dados em três classes de sobrevida: menor ou igual a 6 meses; entre 7 e 24 meses; e maior que 24 meses. Para cada uma dessas classes, foi aplicado um modelo de regressão, permitindo obter resultados mais precisos.

Outra possibilidade é a adoção de tarefas paralelas de classificação e regressão, que embora estejam citadas como resultados do mesmo estudo, foram realizadas de forma independente. Este é o caso citado em Liu et al. [53], no qual os autores aplicaram modelos de classificação e regressão de forma separada. A classificação foi usada para distinguir pacientes que sobreviveram menos de cinco anos dos que sobreviveram mais de cinco anos. Já a tarefa de regressão foi aplicada a todo o conjunto de dados como uma análise separada.

Esses exemplos ilustram as diversas possibilidades de aplicação das tarefas de aprendizado de máquina com o objetivo de prever o tempo de sobrevida de pacientes com câncer.

3.3 Exploração das perspectivas futuras e limitações apontadas

Além dos resultados, é possível direcionar um olhar mais detalhado para as limitações e trabalhos futuros apontadas pelos estudos. Para facilitar esta análise, baseado nos textos dos artigos, as sugestões de melhorias foram agrupadas em categorias. A Tabela 3.6 apresenta uma relação das categorias de trabalhos futuros, quantidade de indicações nos artigos e em quais foram indicados.

Tabela 3.6 – Relação dos artigos e indicações de trabalhos futuros.

Trabalho futuro	**Nº indicações**	**Indicação**
Validar o modelo com mais dados	31	[78, 23, 63, 15, 5, 1, 12, 41, 3, 45, 50, 10, 68, 49, 90, 88, 55, 6, 14, 62, 2, 53, 71, 81, 83, 35, 94, 74, 96, 57, 51]
Inclusão de mais variáveis	14	[6, 14, 2, 86, 92, 79, 1, 12, 33, 97, 3, 74, 82, 42]
Aplicar outras técnicas de ML	11	[6, 62, 86, 98, 87, 3, 50, 28, 72, 29, 46]
Aplicação do modelo em outras áreas da oncologia	12	[91, 73, 12, 87, 39, 11, 51, 47, 89, 95, 42]
Incluir um modelo para pré-processamento dos dados	2	[41, 68]

Os itens apresentados na Tabela 3.6, são descritos abaixo:

- **Validar o modelo com mais dados**: validar o modelo usando um conjunto de dados maior que pode ser obtido de hospitais ou clínicas, possibilitando sua avaliação em um ambiente real.
- **Inclusão de mais variáveis**: avaliar qual o impacto que a inclusão de outras variáveis, como dados clínicos ou imagens, pode ter nos resultados obtidos originalmente;
- **Aplicar outras técnicas de ML**, avaliar se algoritmos de aprendizado de máquina mais modernos poderiam obter resultados melhores;
- **Aplicação do modelo em outras áreas da oncologia**: sugere que o modelo gerado poderia ser aplicado em outras áreas da medicina para verificar seu desempenho preditivo;

- **Incluir um modelo para pré-processamento dos dados**: inclusão de algum modelo que possa realizar automaticamente o pré-processamento dos dados, resolvendo questões como dados faltantes, por exemplo.

Faceli et al. [17] cita que, embora seja possível conseguir bases de dados para analisar, na maior parte das vezes esses dados não estão prontos para serem processados por um algoritmo de ML. Sendo assim, é preciso realizar um pré-processamento. Dependendo do volume de dados a ser analisado, esse processo pode se tornar inviável para ser realizado de forma manual. Neste ponto, a inclusão de algum algoritmo de pré-processamento dos dados, além de tornar o processo mais dinâmico, pode trazer eficiência ao resultado final do modelo.

Avaliar um grande conjunto de informações pode ser computacionalmente caro e demorado, desta forma determinar quais as variáveis têm maior importância no resultado do algoritmo de predição é uma das tarefas que podem ser realizadas. Os estudos sugerem a inclusão de novas variáveis para determinar seu impacto no resultado do algoritmo. Embora alguns dos trabalhos tenham realizado esta tarefa, nenhum dos estudos da literatura menciona ter feito esse processo de validação de variáveis para o câncer de pulmão.

Alguns dos estudos sugerem que um modelo desenvolvido para prever o tempo de sobrevida de um determinado tipo de câncer poderia ser aplicado em outros tipos de câncer. Neste sentido, entende-se que uma série de adaptações teria que ser realizada no modelo, considerando as peculiaridades de cada tipo de câncer.

As pesquisas na área de ML estão em constante evolução, a escolha por um algoritmo ou uma variação de um algoritmo pode ser guiada pelo estado da arte atual. No entanto, com o processo de evolução, pode ser que estejam disponíveis aprimoramentos dos algoritmos já aplicados. Desta forma, os estudos sugerem que a pesquisa seja repetida com algoritmos mais modernos para verificar se os resultados obtidos seriam diferentes.

É possível combinar diversos dados para aperfeiçoar a precisão do modelo desenvolvido. Dessa forma, estudos poderiam incluir informações variadas, como dados clínicos, imagens de TC, ou ainda adicionar novas variáveis no contexto investigado para avaliar o impacto dessas variáveis sobre o desempenho do modelo.

A sugestão de melhoria mais apontada nos estudos é testar o desempenho do modelo com dados diferentes, seja com um volume maior de dados ou com dados obtidos de um ambiente real. Embora haja bancos de dados disponíveis, nem sempre as informações são adequadas, ou nem sempre o volume de informações é suficiente para treinar e avaliar um modelo de ML. Outro aspecto a se observar é que, mesmo apresentando o número de pacientes considerados para desenvolvimento dos modelos, nenhum artigo aponta a quantidade mínima que seria necessária para treinar o mesmo modelo, o que poderia ser uma informação interessante para dar direcionamento aos futuros estudos e dar uma dimensão de população necessária para aplicar o modelo.

3.4 Respondendo as questões de pesquisa

Uma análise de todos os resultados e estudos relacionados permitiu responder às questões de pesquisa elaboradas, QP1 a QP5, que foram apresentadas na Seção 3.1.1.

QP1: *"Quais técnicas têm sido aplicadas para prever o tempo de sobrevida em pacientes com câncer?"*

São aplicados diferentes algoritmos de ML. A Tabela 3.4 apresenta as ferramentas e em quais estudos foram aplicadas. As técnicas mais referenciadas pelos estudos são *Support Vector Machine* e *Convolutional Neural Networks*.

QP2: *"Quais são as diferenças das técnicas usadas para prever o tempo de sobrevida em pacientes com câncer de pulmão em relação a outros órgãos?"*

Não se identificam diferenças significativas nas técnicas de ML aplicadas à predição do tempo de sobrevida para câncer de pulmão em comparação com outros órgãos. Conforme evidenciam os estudos, as diferentes aplicações de algoritmos de ML estão mais relacionadas ao tipo de informação analisada do que propriamente à doença.

QP3: *"Quais conjuntos de dados são utilizados para prever o tempo de sobrevida em pacientes com câncer?"*

Podemos citar diferentes conjuntos de dados, como o *The Cancer Genome Atlas* (TCGA)[3], *Cancer Browser*[4], *Surveillance, Epidemiology, and End Results* (SEER)[5], *The Cancer Imaging Archive* (TCIA)[6], entre outros. Esses conjuntos de dados contêm diferentes tipos de informações, como resultados de exames de imagem, dados clínicos, dados de expressão gênica, malignidade do tumor, taxas de sobrevida e, em alguns casos, informações sobre tratamentos. A Tabela 3.8 apresenta a relação entre as informações analisadas e o tipo de câncer referido.

QP4: *"Como as técnicas de ML tem sido usadas para chegar a tempos mais precisos?"*

[3] www.cancer.gov/ccg/research/genome-sequencing/tcga
[4] xena.ucsc.edu
[5] seer.cancer.gov
[6] www.cancerimagingarchive.net

As técnicas de ML têm sido amplamente utilizadas para prever tempos de sobrevida de pacientes em diversos contextos médicos, especialmente em oncologia. Essas técnicas oferecem vantagens sobre os métodos estatísticos tradicionais, pois conseguem lidar melhor com grandes volumes de dados e captar padrões complexos. Outra vantagem do uso de *Machine Learning* é a capacidade de lidar com dados de grande dimensionalidade, como dados de expressão gênica e imagens médicas; os modelos de ML também dispõem de recursos que permitem tratar dados censurados; além de poderem passar por customizações de parâmetros, permitindo obter resultados mais precisos.

QP5: *"Quais métricas tem sido usadas para avaliar o desempenho dos modelos de ML usados para prever o tempo de sobrevida de pacientes com câncer?"*

A métrica mais comumente utilizada para apresentar o desempenho de modelos de ML é a acurácia. Porém, em alguns casos, ela pode não ser suficiente para demonstrar os resultados, então é preciso um olhar mais detalhado para outras métricas como Precisão, *Recall* e *F1-Score* para ter uma visão geral do real desempenho do algoritmo. Outras formas visuais podem ser aplicadas na avaliação de algoritmos, como matrizes de confusão, que podem ser representadas em gráficos (a forma mais comum) ou tabelas. Outra abordagem visual é o gráfico da curva ROC, que avalia o desempenho do algoritmo por meio da *Area Under the Curve* (AUC). Quanto mais próximo de 1 for o valor da AUC, maior será a eficiência do algoritmo avaliado.

3.5 Discussão

Mesmo tendo objetivos semelhantes, que é prever o tempo de sobrevida de pacientes com câncer, a análise dos trabalhos relacionados oferece um panorama abrangente sobre a área de pesquisa, revelando as diversas abordagens adotadas para alcançar resultados similares. A Tabela 3.7 apresenta uma seleção desses trabalhos ordenados por ano de publicação (do mais antigo para o mais recente), destacando diferentes abordagens sobre o tema. Dos 64 estudos relacionados, foram selecionados nove para apresentar na tabela. Os artigos foram selecionados com base nas diferentes abordagens adotadas, visando demonstrar as múltiplas possibilidades de pesquisa.

Essas abordagens variam de acordo com: o tipo de tarefa realizada, que pode envolver classificação, regressão ou a aplicação simultânea de ambos; o tipo de algoritmo de aprendizado de máquina aplicado, evidenciando os principais algoritmos e sua adaptabilidade, que permite sua utilização em diversas áreas de pesquisa; e o tipo de informação analisada, onde a análise de diferentes tipos de dados, como dados clínicos ou imagens, pode conduzir a resultados similares, ou ainda, a combinação dessas informações pode

potencializar os resultados em uma pesquisa. Outro ponto a se considerar é que a amplitude temporal dos estudos apresentados na tabela, cobrindo artigos publicados entre 2007 e 2023, proporciona uma compreensão mais profunda da evolução e das tendências na pesquisa sobre a previsão de sobrevida em pacientes com câncer.

Tabela 3.7 – Estudos selecionados, apresentando diferentes abordagens para prever o tempo de sobrevida de pacientes com câncer.

Artigo	Ano	Objetivos	Algoritmo	Resultados
[45]	2007	Determinar o número ótimo de neurônios na camada oculta de uma ANN para prever o tempo de sobrevida de pacientes com câncer de bexiga pós-operatório analisando dados clínicos.	MLP	O estudo chegou a um número ótimo de 13 neurônios na camada oculta da ANN. Com esta configuração o modelo obteve uma precisão de 80,6%.
[12]	2010	Desenvolver um algoritmo de similaridade de pacientes, para classificar pacientes com carcinoma hepatocelular submetidos à quimioterapia em duas classes com base em 14 medidas de similaridade.	SVM	O desenvolvimento de uma ferramenta capaz de identificar o tempo de sobrevida baseado em similaridade.
[21]	2014	Prever o tempo de sobrevida de pacientes com câncer de pulmão utilizando a abordagem de reconhecimento de padrões. A intenção é classificar o tempo de sobrevida em três intervalos menor que 3 anos, entre 3 a 5 anos e maior que 5 anos.	KNN SVM RF	Os autores verificaram que a normalização dos dados resulta em melhores resultados.
[52]	2016	Classificar o tempo de sobrevida de pacientes com glioblastoma em bom (+ de 650 dias) e ruim (até 650 dias), considerando a conectividade cerebral (estrutural e funcional) e não só a região afetada pelo tumor.	SVM	Usando apenas dados clínicos o resultado de precisão de previsão ficou em 63,2%, valor que sobe para 72% considerando a conectividade funcional. Combinando as caraterísticas estruturais e funcionais da rede a taxa de precisão aumenta para 75%.
[6]	2018	Explorar a capacidade de prever o tempo de sobrevida de pacientes com câncer de pulmão. Realiza uma classificação e regressão, primeiro separa as instâncias em classes e depois sobre essas classes aplica um algoritmo de regressão.	RF GL GBM	O modelo preditivo funcionou com precisão para tempos de sobrevida curtos de 6 meses; no entanto, a precisão é reduzida à medida que o modelo tenta prever tempos de sobrevida mais elevados.
[90]	2020	Classificar o tempo de sobrevida de pacientes com câncer de pulmão em duas classes, menos de 3 anos e 3 anos ou mais, usando dados de expressão gênica.	SVM CART KNN	O SVM e CART, demonstraram maior poder preditivo na identificação de pacientes com maior ou menor probabilidade de sobreviver por mais de 3 anos após o diagnóstico.
[92]	2021	Aplicar técnicas de *deep learning* em dados clínicos e imagens de TC para prever o tempo de sobrevida de pacientes com câncer de pulmão.	CNN	O modelo proposto obteve resultados superiores a outras técnicas, principalmente por combinar informações de TC com dados clínicos.
[50]	2023	Desenvolver um modelo que leve em consideração indicadores intraoperatórios para classificar pacientes com carcinoma hepatocelular em 4 intervalos curto prazo (menos de 36 meses), médio-curto prazo (entre 36 e 72 meses), médio longo prazo (entre 72 e 108 meses) e longo prazo (mais de 108 meses).	CNN	O modelo proposto obteve uma precisão média de 85,36% na classificação do tempo de sobrevida dos pacientes.
[63]	2023	Desenvolver um modelo capaz de prever meses de sobrevida de pacientes com câncer de mama usando dados clínicos, classificando em intervalos [0 2), [2 4), [4 6), [6 8) e [8 10], além de identificar variáveis com maior impacto no resultado.	DT RF SVM KNN	Os classificadores DT e RF obtiveram resultados acima de 68%. A idade da paciente tem grande impacto na precisão do modelo, que é reduzida em até 17% se essa informação for desconsiderada.

Os estudos trazem diversas estratégias para abordar o tempo de sobrevida de pacientes com diferentes tipos de câncer, oferecendo contribuições à área médica. Eles desenvolvem modelos de predição que podem auxiliar os profissionais na tomada de decisões relacionadas ao tratamento dos pacientes. As pesquisas também podem identificar

quais variáveis ou informações tem maior impacto no tempo de sobrevida de pacientes. Além disso, os estudos demonstram que abordagens baseadas em aprendizado de máquina tendem a ser mais precisas do que os modelos estatísticos convencionais na predição do tempo de sobrevida. Muitos desses estudos conduzem experimentos utilizando técnicas de aprendizado de máquina, buscando determinar quais destas técnicas atingem os melhores resultados, indicando como trabalhos futuros as técnicas de ML mais eficazes na classificação do tempo de sobrevida.

Entre os artigos selecionados, destacam-se os 22 que estão relacionados com câncer de pulmão, um dos tipos mais prevalentes. Podemos ressaltar as diferentes abordagens adotadas pelos autores para alcançar os resultados. Alguns estudos realizaram avaliações de imagens de ressonância magnética, dados genômicos, dados clínicos ou uma combinação de diferentes tipos de informações. Além disso, as análises empregam uma gama variada de algoritmos de aprendizado de máquina, obtendo, em geral, precisões semelhantes nos resultados.

Em Wei et al. [90], os autores usaram o algoritmo SVM para analisar a expressão do RNA sobre o tempo de sobrevida em pacientes com câncer de bexiga e Timilsina et al. [81] que analisam dados clínicos de pacientes com câncer de pulmão, aplicando o mesmo algoritmo. Em ambos os casos o SVM, apresentou resultados satisfatórios na predição do tempo de sobrevida de pacientes, o que demonstra a versatilidade dos algoritmos de aprendizado de máquina.

Em relação ao uso de múltiplos tipos de informações para prever o tempo de sobrevida, o trabalho de Wu et al. [92], que prevê o tempo de sobrevida de pacientes com câncer de pulmão combinando dados de imagens obtidos de tomografia computadorizada e dados clínicos, apresentou bons resultados. Outro estudo combina dados gênicos e clínicos para prever o tempo de sobrevida em pacientes com câncer de mama [53].

Considerando o volume de informações disponíveis, a quantidade média de instâncias analisadas pelos estudos foi de 442, com os *datasets* variando de 28 (o menor) a 4024 (o maior) linhas. No estudo dos autores Prakash et al. [67], cujo objetivo é entender os impactos do tamanho da amostra nos resultados do modelo, mostrou que a redução do tamanho da amostra pode impactar significativamente o desempenho do modelo desenvolvido. Isso demonstra que os resultados podem ser influenciados pelo conjunto de dados disponível para o treinamento e teste do modelo.

A Tabela 3.8 apresenta de uma forma resumida a lista de artigos, tipo de câncer abordado e quais informações foram utilizadas para realizar a predição do tempo de sobrevida. A tabela mostra que os dados mais usados são os dados clínicos, genéticos e imagens. Para apresentação na tabela, alguns tipos de câncer foram agrupados, como, por exemplo: gastrointestinal, que engloba câncer de esôfago e estômago; sangue, que engloba leucemia, linfoma e mieloma; pele, que inclui melanoma e outros tipos de câncer de pele; e útero, que inclui câncer de útero e do colo do útero.

Em relação aos dados analisados, alguns agrupamentos foram realizados: dados genéticos, que abrangem informações como metilação do DNA, expressão de miRNA, expressão de mRNA; dados de imagens que incluem resultados de diferentes exames de imagens como tomografia computadorizada e ressonância magnética. Estes dados são apresentados na Tabela 3.8, que aponta que os estudos realizados utilizaram diferentes tipos de dados, sem uma relação específica entre o tipo de dado avaliado e o tipo de doença. Um exemplo é o câncer de pulmão, listado em 22 dos estudos, e analisado utilizando dados clínicos, imagens e dados genéticos em diferentes pesquisas. No entanto, observa-se que exames de laboratório, estadiamento TNM [76], tamanho do tumor ou taxa de sobrevivência não foram aplicados, o que poderia melhorar os resultados.

A pesquisa permitiu obter um panorama geral dos estudos relacionados ao tema, tempo de sobrevida de pacientes em tratamento de câncer, permitindo identificar qual a principal limitação encontrada pelos pesquisadores está relacionada a obtenção de conjuntos de dados que permitam otimizar o treinamento e validação do modelo desenvolvido.

Tabela 3.8 – Relação dos estudos, órgão estudado e dados analisados.

Estudos		Órgão ou sistema afetado																				Dados analisados														
Artigo	Ano	Não especificado	Bexiga	Cabeça e pescoço	Cérebro	Cervical	Colorretal	Fígado	Gastrointestinal	Glândula adrenal	Mama	Olhos	Ovário	Pele	Pâncreas	Próstata	Pulmão	Rins	Sangue	Tireoide	Útero	Não especificado	Caract. das células cancerosas	Clínicos	Demográficos	Genéticos	Proteínas	Imagens	Laboratoriais	Envolvimento da vasculatura	Escores ECOG	Estadiamento TNM	Histologia do tumor	Malignidade	Tamanho do tumor	Taxas de sobrevivência
Li e Li [49]	2004																		X							X										
Liu [54]	2008																		X							X										
Chen et al. [14]	2009																		X							X										
Phong et al. [65]	2009																		X					X												
Ture et al. [82]	2009										X													X												
Hayward et al. [29]	2010								X																				X	X	X	X	X	X	X	X
Dekker et al. [15]	2010																X							X												
Yu et al. [94]	2011	X																						X												
Chan et al. [12]	2011							X																X												
Kolasa et al. [45]	2011		X																					X												
Chen et al. [13]	2011							X																X												
Zhou et al. [98]	2013				X																							X								
Pu et al. [68]	2014																X									X		X								
Zhou et al. [97]	2014				X																							X								
Hawkins et al. [28]	2014																X											X								
Gan et al. [21]	2015										X						X									X										
Kim et al. [42]	2015										X																X									
Chai et al. [11]	2015	X																																		
Malhotra et al. [58]	2016				X																	X														
Zhang et al. [96]	2016				X																			X		X										
Liu et al. [52]	2016				X																							X								
Chaddad et al. [10]	2016				X																							X								
Zhang et al. [95]	2016												X											X		X										
Isik e Ercan [39]	2017				X												X	X								X										
Tyuryumina e Neznanov [83]	2017										X											X														
Kuo et al. [47]	2017																X											X								
Ferdinand Christ et al. [19]	2017							X																				X								
Bartholomai e Frieboes [6]	2018																X											X								
Wang et al. [87]	2018																X											X								
Huang e Liang [35]	2019										X			X	X		X		X							X										
Wang et al. [88]	2019										X															X										
Hossain et al. [31]	2019												X											X		X										
Jayashanka et al. [41]	2019																		X						X	X			X							
Li et al. [51]	2019			X	X												X			X	X					X										
Ahmed et al. [1]	2019				X																							X								
Cai et al. [9]	2019																X											X								
Stepanek et al. [78]	2020								X															X												
Lu et al. [55]	2020										X													X												
Aljouie et al. [3]	2020				X																					X		X								
Wang et al. [86]	2020	X	X	X	X			X	X	X	X	X		X		X	X	X			X			X		X										
Aljouie e Roshan [2]	2020				X																				X	X										
Nanda e Duraipandian [62]	2020																X							X												
Sun et al. [79]	2020								X																			X								
Wei et al. [90]	2020																X									X										
Wu et al. [91]	2021																X											X								
Wu et al. [92]	2021																X							X				X								
Hossain et al. [32]	2021						X															X														
Baker et al. [5]	2021																X							X				X								
Liu et al. [53]	2021										X																									
Sharma et al. [74]	2022			X																				X		X										
Yang et al. [93]	2022						X																			X										
Scalco et al. [71]	2022																X											X								
Wang et al. [89]	2022										X								X			X														
Wang et al. [85]	2022	X																						X	X											
Ma et al. [57]	2022					X																		X		X										
Shakir et al. [72]	2023																X											X								
Kukreja et al. [46]	2023																X						X	X		X										
Li et al. [50]	2023							X																X				X								
Hou et al. [33]	2023	X																				X														
Shao et al. [73]	2023										X							X								X		X								
Timilsina et al. [81]	2023																X							X												
Naser et al. [63]	2023										X													X												
Ghazipour et al. [23]	2023																X											X								

4. SOLUÇÃO PROPOSTA

A Figura 4.1 ilustra as atividades desenvolvidas ao longo da pesquisa. Iniciando com a definição do tema, são elaboradas as questões a serem respondidas, as quais guiaram a seleção de literatura e o progresso do estudo. A busca por dados desempenha um papel importante e é um processo recorrente, o qual será detalhado na Seção 4.1.

Com o conhecimento dos dados, foi possível desenvolver um modelo de aprendizado de máquina que é responsável pela classificação dos pacientes em sobrevida longa e curta. O processo de elaboração do modelo é detalhado na Seção 4.2. Finalizado o modelo, tivemos que avaliar seu desempenho, compilar e analisar os resultados obtidos, além de apontar as dificuldades encontradas ao longo do processo.

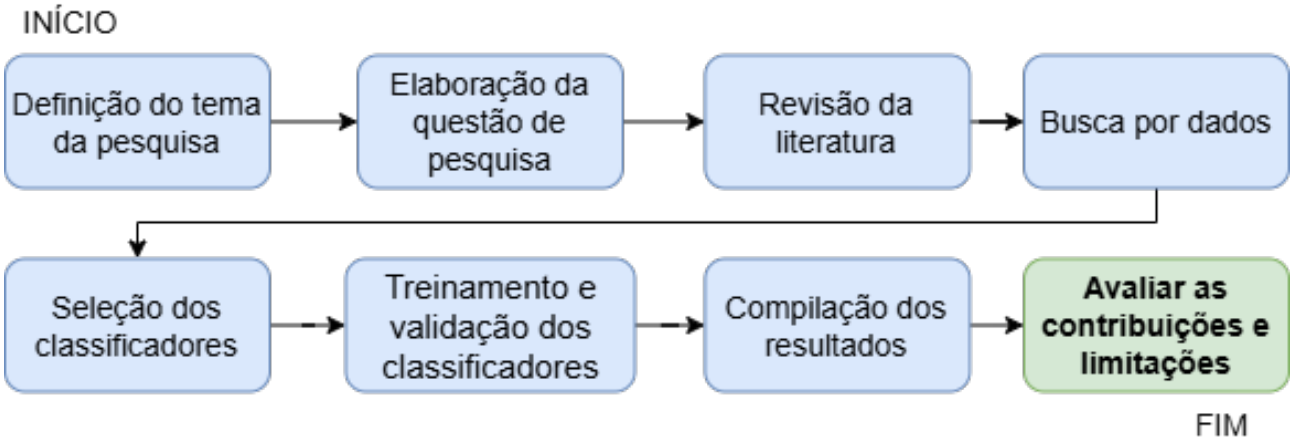


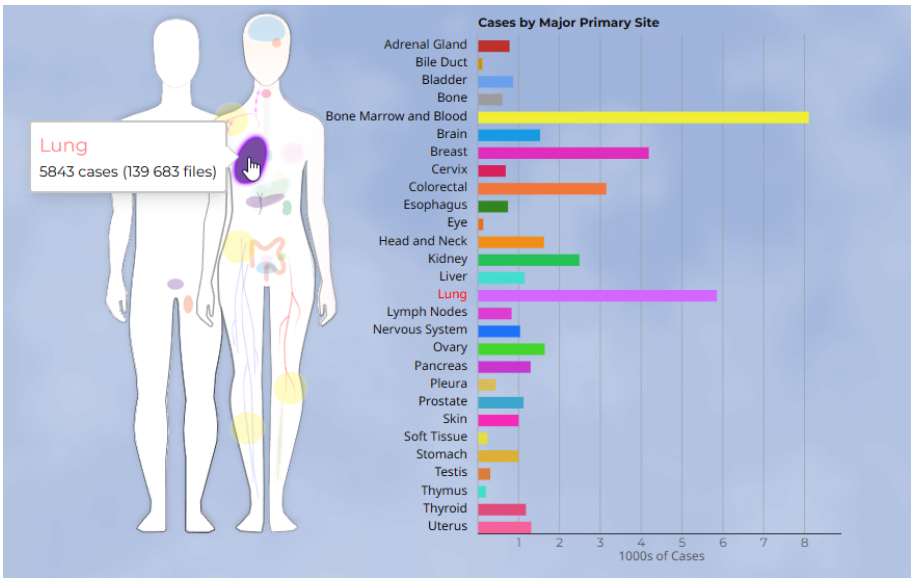
Figura 4.1 – Estrutura da pesquisa: fluxo das atividades principais.

4.1 Dados de entrada

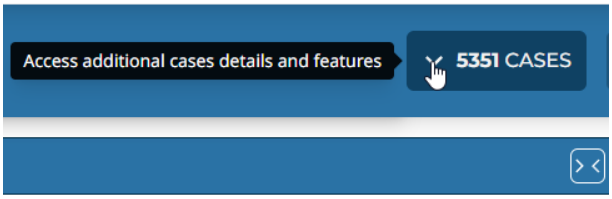
O conjunto de dados utilizado foi obtido do TCGA [80], que é um consórcio de pesquisa mantido pelo *National Cancer Institute* (NCI)[1] e pelo *National Human Genome Research Institute* (NHGRI)[2] dos Estados Unidos. Os dados são armazenados pelo *Genomic Data Commons* (GDC)[3] e podem ser baixados livremente. Ao entrar na plataforma, é necessário ler e aceitar os termos clicando no botão *Accept* em uma janela *modal* que será apresentada. Em seguida, deve ser selecionada a região desejada no mapa anatômico interativo para visualizar os dados correspondentes, neste caso *Lung*, como mostra a Figura (a). Após a confirmação, clicando em *Yes* na *modal* que é apresentada, na sequência,

[1] https://www.cancer.gov/
[2] https://www.coriell.org/
[3] https://portal.gdc.cancer.gov/

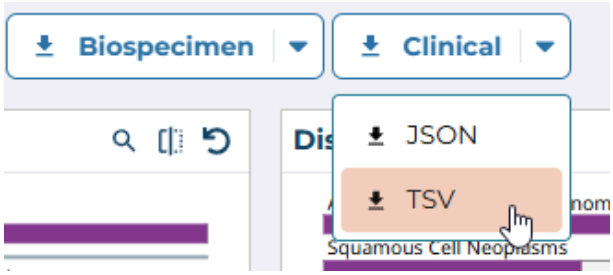
deve-se clicar sobre o botão com o número de casos apresentado no canto superior direito da página, como apresentado na Figura (b). Para efetuar o download dos dados clínicos, basta clicar no botão *clinical* ilustrado na Figura (c), e realizar o *download* dos dados em TSV (*text separated value*), formato usado neste estudo.



(a)



(b)



(c)

Figura 4.2 – Processo realizado para obter os dados do GDC: (a) mapa anatômico para selecionar o tipo de câncer; (b) botão para expandir as opções disponíveis; (c) botão para realizar o *download* em formato TSV.

O conjunto de dados contém informações sobre pacientes com câncer de pulmão, apresentando 56 colunas, com informações como: gênero, etnia, raça, idade, ano de nascimento, ano de diagnóstico, estadiamento TNM clínico e patológico, classificação do tumor, método de diagnóstico, diagnóstico primário, dados sobre a existência de câncer anterior, progressão da doença, se existe doença residual após o tratamento, posição do tumor, presença ou não de outros tipos de tumor ao mesmo tempo e tipo de tratamento.

O conjunto de dados obtido do TCGA possui muitas colunas com muitos valores faltantes, o que exigiu um amplo trabalho de pré-processamento. A Figura 4.3 apresenta o fluxo do processamento realizado usando a ferramenta Tableau Prep[4], através da qual foram removidas colunas que continham apenas valores nulos e foi feita a mesclagem de

[4]www.tableau.com

colunas que continham informações duplicadas. A última etapa realizada dentro do Tableau Prep foi a criação da coluna alvo, baseada na coluna tempo de sobrevida, considerando pessoas que sobreviveram mais de 5 anos como sobrevida longa e pessoas com menos de 5 anos de sobrevida como sobrevida curta. A *American Cancer Society* (ACS) [77] aponta que 5 anos é o tempo de sobrevida usado como referência para câncer de pulmão. Essa coluna será posteriormente usada como coluna alvo para os algoritmos de classificação aplicados.

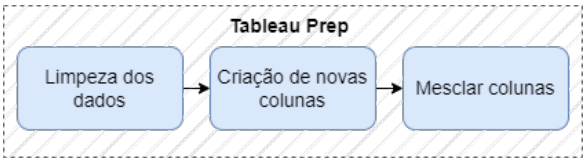


Figura 4.3 – Etapa de pré-processamento usando Tableau Prep.

O resultado gerado pelo Tableau Prep foi compilado em uma planilha e importado para um código Python[5] utilizando a biblioteca Pandas[6]. O conjunto de dados passou por mais uma etapa de pré-processamento para gerenciar os dados ausentes, que foram preenchidos de acordo com o tipo de informação presente em cada coluna. A Tabela 4.1 lista as colunas que receberam preenchimento e os valores usados para substituir os valores nulos. Esse processo foi realizado por meio de uma substituição simples dos valores, utilizando funções disponíveis na biblioteca Pandas. O fluxo do processo de importação, preenchimento e tratamento das colunas é ilustrado pela Figura 4.4.

Tabela 4.1 – Colunas e seus respectivos valores de preenchimento.

Atributo	Valor atribuído
gender	not_reported
race	not_reported
ajcc_pathologic_t	not_reported
ajcc_pathologic_n	not_reported
ajcc_pathologic_m	not_reported
site_of_resection_or_biopsy	not_reported
tissue_or_organ_of_origin	not_reported
age_at_index_calc	valor médio
days_to_last_follow_up	valor médio
synchronous_malignancy	No

Outro tratamento necessário antes de realizar o treinamento e avaliação dos modelos de classificação é a separação das colunas categóricas de acordo com os critérios:

- Categóricas nominais, são separadas em múltiplas colunas, processo também chamado de *one-hot encoding* [7]. Neste processo, cada valor presente na coluna vira

[5]www.python.org
[6]pandas.pydata.org/

uma nova coluna do conjunto de dados. Cada instância recebe 1 na coluna correspondente ao seu valor na coluna original e 0 nas demais. Por exemplo, uma coluna gênero com valores M e F, resultaria em duas colunas, uma “genero_M” e outra “genero_F”. Para cada instância com valor F na coluna “genero”, a coluna “genero_F” receberia o valor 1 e para as demais instâncias, que possuem o valor M na coluna “genero”, receberia o valor 0 na coluna “genero_F”.

- Categóricas ordinais, possuem valores com uma ordem natural que são substituídos por números, mantendo o peso de cada categoria, mas de uma forma numérica. Por exemplo uma coluna “nível_de_escolaridade”, com os valores fundamental, médio, superior e pós-graduação, poderia receber um valor numérico para cada um dos valores. Como resultado, os valores manteriam sua ordem, mas agora de forma numérica, ou seja, 1 para fundamental, 2 para médio, 3 para superior e 4 para pós-graduação.

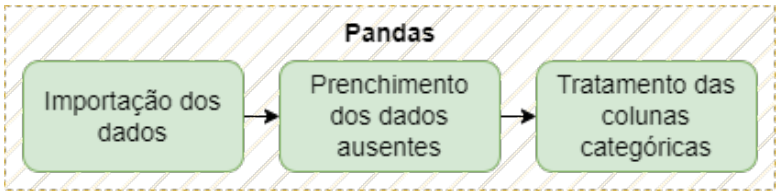


Figura 4.4 – Etapa de importação dos dados usando Pandas.

Para evitar que os algoritmos de classificação fiquem tendenciosos a uma das classes, elas precisam ser balanceadas. O balanceamento das classes foi realizado usando uma junção de duas técnicas. Inicialmente, foi aplicado um algoritmo de *undersampling*, neste caso a classe majoritária teve o número de instâncias reduzidas até estar 30% maior que a classe minoritária. No próximo passo, foi aplicada uma técnica de *oversampling*, para gerar 30% de instâncias sintéticas para a classe minoritária, fazendo com que as classes ficassem balanceadas.

O *undersampling* foi realizado utilizando o *Cluster Centroids*, um algoritmo que reduz o número de instâncias da classe majoritária por meio de clusterização, aplicando o *k-means*. Nesse processo, os dados da classe majoritária são agrupados em *clusters*, onde cada *cluster* representa um subconjunto de exemplos com características semelhantes. Em seguida, o centroide de cada *cluster*, calculado como a média dos exemplos que o compõem, é usado para substituir os exemplos do *cluster* no conjunto de dados. Essa abordagem visa reduzir a classe majoritária, mantendo sua representatividade estatística [34].

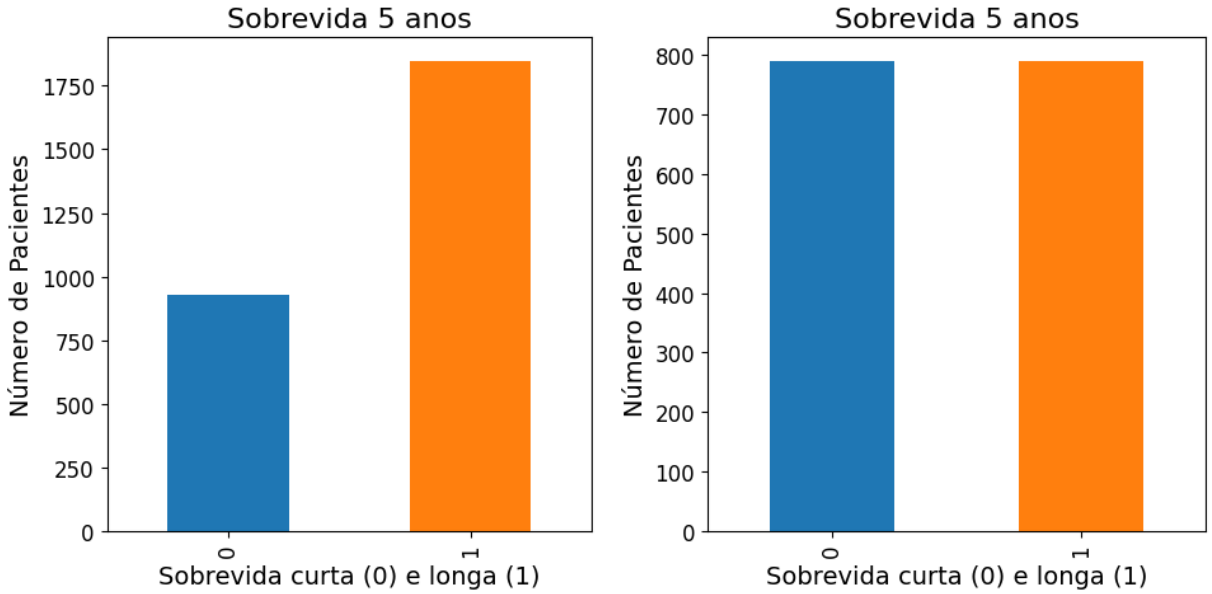
Para o *oversampling*, foi aplicada uma técnica chamada *Synthetic Minority Oversampling Technique* (SMOTE), que identifica qual classe tem menos instâncias e aplica o algoritmo *K-nearest neighbors* (KNN) para criar novas amostras sintéticas baseadas nos vizinhos mais próximos para equilibrar as duas classes. O SMOTE calcula a distância euclidiana entre os k vizinhos mais próximos, e o valor para o dado sintético é a multiplicação

desta distância por um valor aleatório entre 0 e 1. Esse processo se repete até que as duas classes estejam com 50% das instâncias [37].

A Tabela 4.2 apresenta as quantidades de instâncias resultantes para cada passo do processo de balanceamento. Os resultados da aplicação das técnicas de *undersampling* e *oversampling* também podem ser visualizados na Figura 4.5, na qual são apresentados os dados antes (Figura 4.5(a)) e depois (Figura 4.5(b)) da execução do balanceamento. Os gráficos mostram as duas classes de sobrevida, longa (1) e curta (0), em relação ao número de instâncias observadas.

Tabela 4.2 – Quantidades de instâncias em cada momento do balanceamento, treinamento e teste.

Dados		**Treinamento**		**Balanceado**		**Teste**	
2773		1663		1584		1110	
curta	**longa**	**curta**	**longa**	**curta**	**longa**	**curta**	**longa**
926	1847	555	1108	792		371	739



(a) (b)

Figura 4.5 – Resultado do balanceamento de classes.

Os dados balanceados passam, então, para o processo de normalização, que é importante para que os modelos tenham um melhor desempenho [26]. A Figura 4.6, ilustra o fluxo do processo de balanceamento e normalização dos dados. Para executar a normalização dos dados, foi aplicado o *MinMax Scaler* [25], que transforma os dados de forma linear aplicando a Equação 4.1.

$$X_{\text{scaled}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}, \tag{4.1}$$

onde:

- X é o valor original do dado,
- X_{min} é o menor valor do dado no conjunto de dados,
- X_{max} é o maior valor do dado no conjunto de dados,
- X_{scaled} é o valor transformado (normalizado).

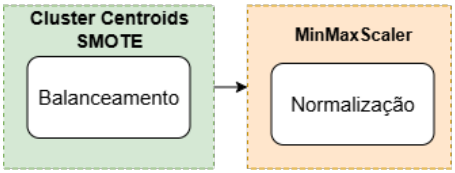


Figura 4.6 – Etapa de balanceamento e normalização dos dados.

4.2 Implementação dos modelos

Com todas as etapas de processamento realizadas, é o momento de separar o conjunto de dados em treino e teste, deixando 60% do conjunto de instâncias para treino e os 40% restantes para teste dos algoritmos e validação dos resultados.

Para o experimento foram selecionados cinco classificadores, *Random Forest* (RF), *Logistic Regression* (LR), KNN, *Decision Tree* (DT) e *Support Vector Classifier* (SVC), que permitem ter exemplos de diferentes abordagens de classificação. Sendo DT e RF baseados em árvores, KNN baseado em distâncias entre instâncias, LR e SVC baseados em hiperplanos.

A Figura 4.7(a) apresenta o fluxo do processo de treinamento e teste realizado com cada um dos algoritmos. O treinamento é realizado utilizando o conjunto de dados destinado a essa etapa, sendo importante salientar que essa parte dos dados contém as respostas; dessa forma, o algoritmo pode aprender os padrões dos dados que levam a determinada classe de saída. O modelo treinado pode ser testado com o conjunto de dados de teste. Neste momento, sem conhecer as respostas, ele prevê a qual classe cada instância de teste pertence. Por fim, podemos verificar qual o grau de precisão do algoritmo comparando as respostas que ele gerou com as respostas reais do conjunto de dados de teste. Essa comparação permite calcular métricas para avaliar o desempenho do algoritmo. Em um segundo processo, ilustrado na Figura 4.7(b), o procedimento é repetido, porém, os melhores parâmetros de configuração para cada algoritmo são determinados utilizando o *Grid Search Cross-Validation*. Essa técnica realiza uma busca exaustiva por combinações de parâmetros, simulando múltiplos ciclos de treinamento e teste, e identifica as configurações que proporcionam o melhor desempenho para os dados disponíveis.

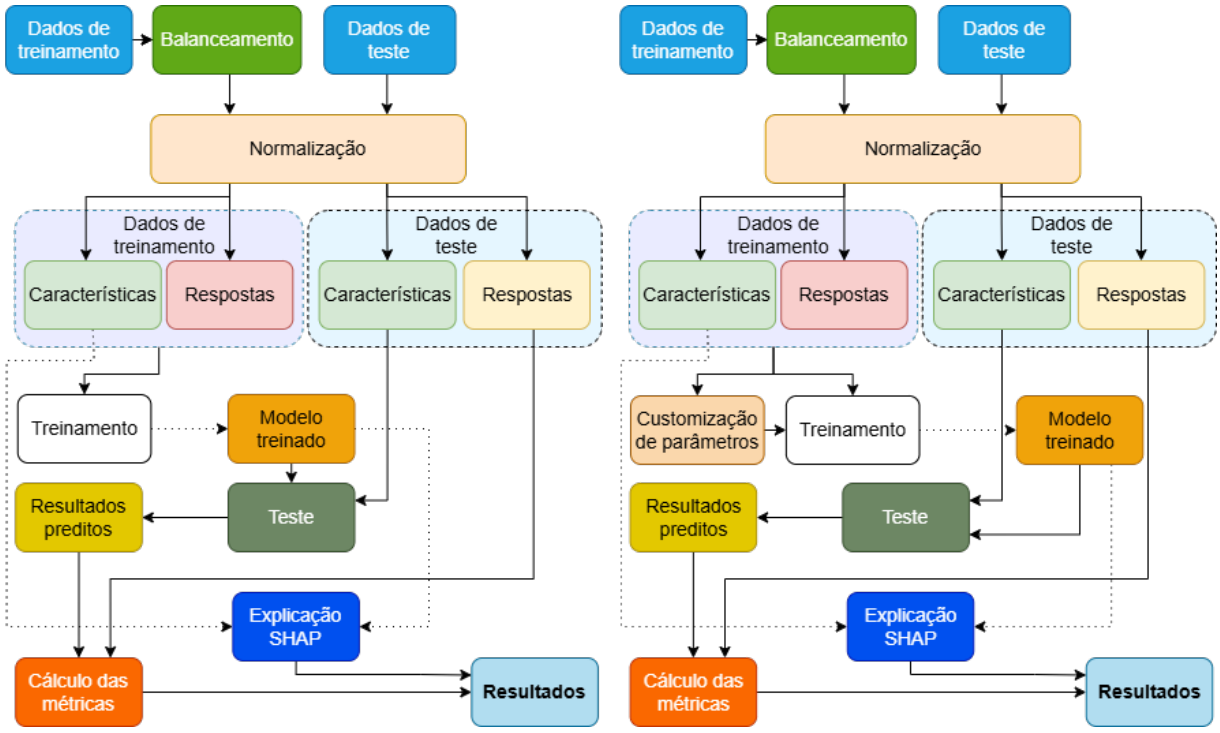


Figura 4.7 – Fluxo de treinamento dos algoritmos: (a) apresentação da implementação padrão dos modelos; (b) apresentação da adição do processo de customização dos parâmetros por meio do *Grid Search Cross-Validation*.

A Figura 4.8 apresenta o fluxo completo da metodologia descrita neste capítulo, oferecendo uma visão geral dos processos e do fluxo de informações. Esse fluxo abrange desde o pré-processamento, incluindo etapas de balanceamento e normalização, até a geração dos resultados finais. Os resultados são obtidos com a implementação, a verificação do desempenho dos modelos através das métricas de avaliação (Acurácia, Precisão e *F1-Score*) e a explicação dos modelos usando o SHAP, que permite entender a tomada de decisão dos algoritmos. O processo de customização dos parâmetros é apresentado como uma etapa opcional, para avaliar os impactos dessa modificação nos resultados dos classificadores.

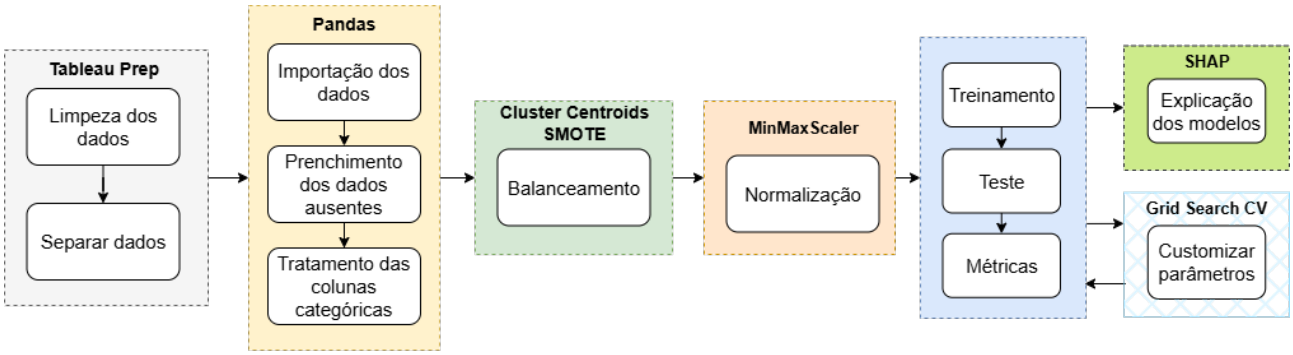


Figura 4.8 – Processo de implementação completo.

5. RESULTADOS E DISCUSSÃO

Concluído o processo de treinamento e teste com cada um dos algoritmos, foram calculadas as métricas de avaliação de desempenho acurácia, precisão, *recall* e *f1-score*. Além destas métricas, foram geradas as matrizes de confusão, que permitem ter uma visualização dos resultados dos algoritmos. Este capítulo apresenta os resultados obtidos ao longo do processo, separando os valores obtidos com e sem a customização de parâmetros dos algoritmos. O código-fonte desenvolvido nesta pesquisa está disponível em um repositório público no GitHub[1]. Os resultados deste trabalho foram publicados na conferência MedInfo 2025 [30].

As tabelas de resultados (Tabela 5.1 e Tabela 5.2) apresentam a métrica de acurácia na primeira coluna. As métricas de precisão, *recall* e *f1-score* são exibidas em três colunas que correspondem, respectivamente, às classes de sobrevida curta, longa e à média ponderada, que considera o número de instâncias de cada classe no cálculo.

5.1 Implementação com a configuração padrão

Em um primeiro momento, os algoritmos foram inicializados usando os parâmetros de configuração padrão, ou seja, não foram passados parâmetros exceto os essenciais para que o modelo disponibilizasse as informações necessárias para calcular determinadas métricas. Por exemplo, esse é o caso do *Support Vector Classifier*, que não disponibiliza as probabilidades normalmente, apenas a classe, e como as probabilidades são necessárias para calcular a curva ROC do algoritmo, precisamos informar o parâmetro de configuração *probability* como ligado.

5.1.1 Métricas

A Tabela 5.1 apresenta os resultados obtidos com a implementação padrão dos algoritmos, apenas com os parâmetros mínimos necessários para calcular as métricas. A acurácia média dos 5 modelos neste cenário foi de 76,6%.

Dois algoritmos merecem destaque, o *Decision Tree* e o *Random Forest*, que obtiveram os melhores resultados de acurácia, 86,9% e 86,8% respectivamente. O pior desempenho foi do algoritmo *K Nearest Neighbors*, que conseguiu prever corretamente 68,5% dos casos. No entanto, a acurácia não pode ser analisada separadamente, em especial em

[1] https://github.com/DAVINTLAB/lung-cancer-survival-prediction

Tabela 5.1 – Métricas dos algoritmos implementados sem customização de parâmetros.

Alg.	Acurácia	Precisão			Recall			F1-score		
		curta	longa	média	curta	longa	média	curta	longa	média
RF	**0,8676**	0,7995	0,9022	**0,8678**	0,8059	0,8985	**0,8676**	0,8027	0,9003	**0,8677**
LR	**0,6955**	0,5401	0,7868	**0,7044**	0,5984	0,7442	**0,6955**	0,5678	0,7650	**0,6990**
KNN	**0,6847**	0,5221	0,8063	**0,7113**	0,6685	0,6928	**0,6847**	0,5863	0,7453	**0,6921**
DT	**0,8694**	0,8005	0,9046	**0,8698**	0,8113	0,8985	**0,8694**	0,8059	0,9016	**0,8696**
SVC	**0,7108**	0,5644	0,7895	**0,7143**	0,5903	0,7713	**0,7108**	0,5771	0,7803	**0,7124**

casos como este, em que a classe de sobrevida longa tem mais instâncias (para o conjunto de teste). Neste caso, é necessário atentar para as outras métricas que juntas podem ajudar a entender melhor os modelos. No caso da precisão, em todos os modelos ficou maior para a classe longa, indicando que todos têm uma tendência a gerar menos falsos positivos na classe de sobrevida longa. Observando o panorama geral, os algoritmos RF e DT têm o melhor desempenho também nesta métrica. O bom equilíbrio entre as métricas precisão, *recall* e *F1-Score* demonstra que os modelos estão equilibrados, embora o desempenho menor dos modelos para a classe de sobrevida curta pode estar associado ao desbalanceamento do conjunto de teste.

5.1.2 Curva ROC

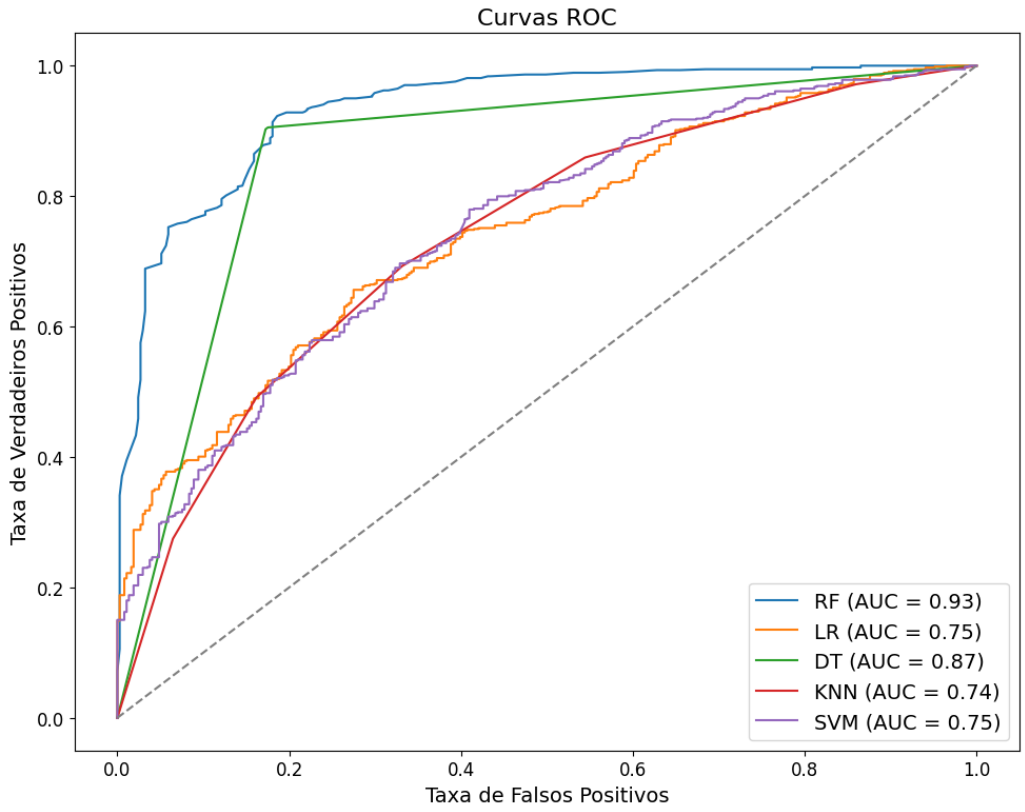


Figura 5.1 – Curva ROC comparando o desempenho dos 5 algoritmos testados.

A Figura 5.1 apresenta as curvas ROC para os cinco algoritmos implementados. Os valores de AUC indicam a capacidade preditiva geral de cada modelo, com o *Random Forest* destacando-se como o mais eficaz (AUC = 0,93), seguido pela *Decision Tree* (AUC = 0,87). Por outro lado, os demais algoritmos *Logistic Regression*, *k-nearest neighbors* e *Support Vector Machine*, apresentaram desempenhos similares, com valores de AUC em torno de 0,74-0,75. Esse desempenho é relativamente baixo, aproximando-se do comportamento de uma classificação aleatória (AUC = 0,50).

5.1.3 Matrizes de confusão

A Figura 5.2 apresenta as matrizes de confusão geradas para os cinco algoritmos de aprendizado de máquina implementados. Os algoritmos *Decision Tree* (Figura 5.2(d)) e *Random Forest* (Figura 5.2 (a)) foram os mais assertivos do conjunto, com o DT se saindo melhor com a classe de sobrevida curta, acertando 304 das 371 instâncias desta classe presentes no conjunto de teste e o algoritmo RF com melhor desempenho para a classe longa, acertando 664 das 739 instâncias de sobrevida longa. Os dois algoritmos têm um bom desempenho de classificação, reafirmando as métricas apresentadas.

O desempenho dos algoritmos *Logistic Regression* (Figura 5.2(b)) e *Support Vector Classifier* (Figura 5.2(e)) foi semelhante, enquanto o pior desempenho foi apresentado pelo *K-Nearest Neighbor* (Figura 5.2(c)), que realizou 512 predições corretas para a classe de sobrevida longa e 248 para a classe de sobrevida curta.

5.1.4 Análise SHAP

A análise SHAP permite entender o impacto de cada variável na resposta do algoritmo. A Figura 5.3 apresenta os *summary plot* dos algoritmos que obtiveram os melhores desempenhos de classificação, *Random Forest* (Figura 5.3 (a)) e *Decision tree* (Figura 5.3 (b)), com os cinco principais atributos na ordem de impacto para o resultado da classificação. Este gráfico permite entender melhor como os algoritmos chegaram aos resultados apresentados. No eixo y, são apresentadas as variáveis organizadas em ordem decrescente de importância. Cada ponto ao longo do eixo x representa uma instância do *dataset*, e sua posição indica o impacto do atributo na decisão da classe para aquela instância específica. Pontos localizados à direita (com valores positivos) indicam que o atributo contribuiu de forma mais significativa para uma decisão favorável à classe analisada. Além disso, a cor dos pontos reflete o valor do atributo na instância: tons de vermelho indicam valores altos, enquanto tons de azul indicam valores baixos [66].

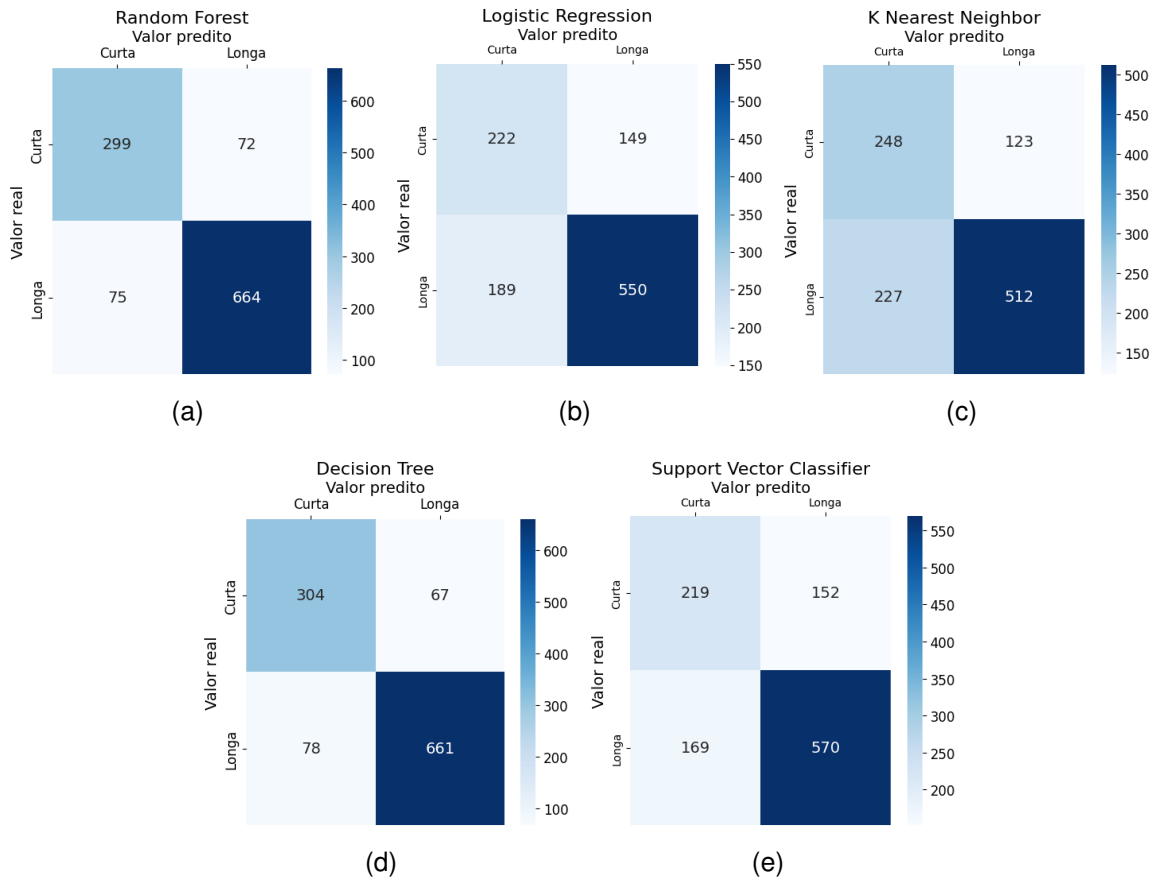
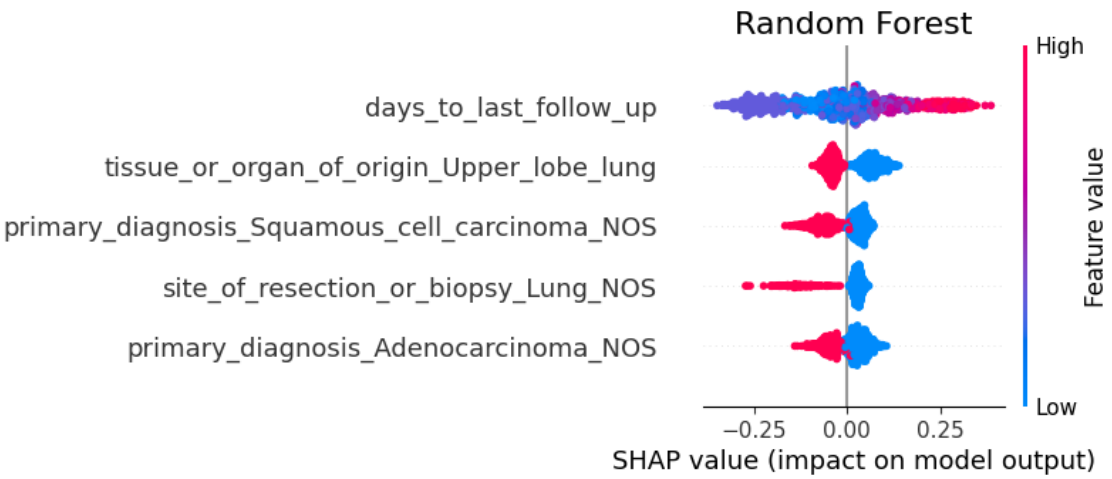


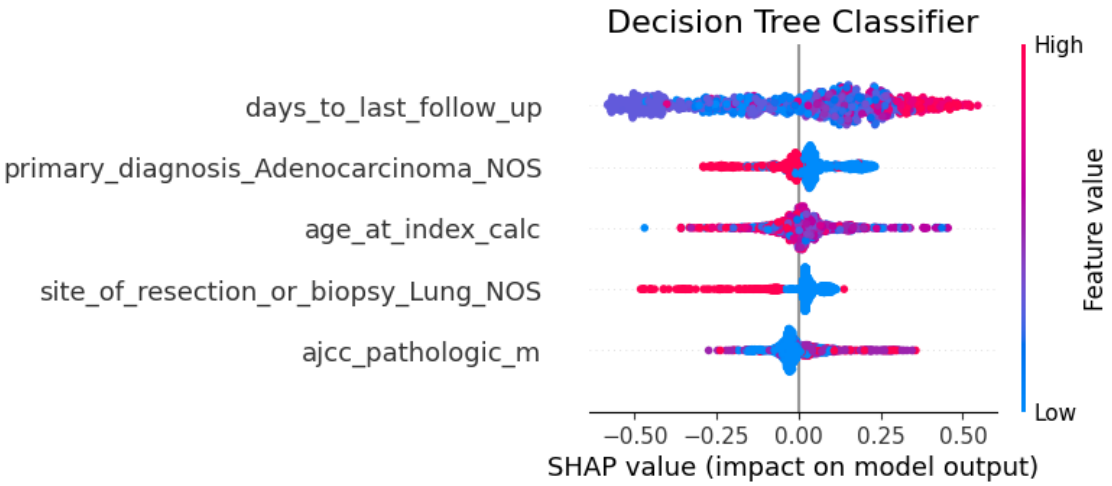
Figura 5.2 – Matrizes de confusão para os algoritmos de ML. RF (a), LR (b), KNN (c), DT (d) e SVC (e).

Em ambos os algoritmos, a variável “days_to_last_follow_up” é a mais importante, e valores altos para esta variável têm impacto positivo na predição. Variáveis como “tissue_or_organ_of_origin_Upper_lobe_Lung”, sugerem que pacientes cuja origem do tumor seja o lobo superior do pulmão, têm um impacto negativo na sobrevida longa. A variável “site_of_resection_or_biopsy_Lung_NOS” indica que a ressecção ou biópsia foi realizada no pulmão, sem especificação exata do local. Pacientes que não têm especificação do local exato da realização do procedimento estão associados a uma classificação de sobrevida curta. O atributo “primary_diagnosis_Adenocarcinoma_NOS” tem um impacto negativo para a classe positivo, ou seja, valores baixos nestas variáveis favorecem a classe de sobrevida longa. O adenocarcinoma é um tipo de câncer que se origina em células glandulares epiteliais. No contexto do câncer de pulmão, ele costuma se desenvolver nas regiões periféricas dos pulmões e é um dos subtipos mais comuns dessa doença [69].

Para o algoritmo *Decision tree*, embora as instâncias estejam dispersas ao longo da distribuição, a variável “age_at_index_calc” indica que pacientes mais jovens tendem a ter uma classificação positiva para sobrevida longa. Além da idade e dos demais atributos já detalhados, para o algoritmo DT, o atributo “ajcc_pathologic_m” indica que não ter a presença de metástase em órgãos distantes favorece a classe de sobrevida longa.



(a)



(b)

Figura 5.3 – *Shap Summary Plot* para os dois algoritmos com o melhor desempenho, *Random Forest* (a) e *Decision tree* (b).

5.2 Implementação com otimização de parâmetros

Buscando identificar os melhores parâmetros para os algoritmos, foi aplicado o *Grid Search Cross Validation*, que explora uma lista de possíveis combinações de parâmetros e retorna a configuração ideal de parâmetros para o algoritmo e conjunto de dados analisado. Esse processo foi repetido para cada um dos cinco algoritmos. Para avaliar o impacto dessa otimização nos resultados, todo o processo de treinamento foi realizado novamente com o mesmo conjunto de dados.

5.2.1 Métricas

As métricas obtidas com a nova configuração estão apresentadas na Tabela 5.2.

Tabela 5.2 – Métricas dos algoritmos implementados com customização de parâmetros.

Alg.	Acurácia	Precisão			Recall			F1-score		
		curta	longa	média	curta	longa	média	curta	longa	média
RF	**0,8712**	0,8016	0,9071	**0,8718**	0,8167	0,8985	**0,8712**	0,8091	0,9028	**0,8715**
LR	**0,6784**	0,5165	0,7784	**0,6909**	0,5903	0,7226	**0,6784**	0,5509	0,7495	**0,6831**
KNN	**0,8279**	0,7103	0,9018	**0,8378**	0,8194	0,8322	**0,8279**	0,7610	0,8656	**0,8306**
DT	**0,8126**	0,7162	0,8622	**0,8134**	0,7278	0,8552	**0,8126**	0,7219	0,8587	**0,8130**
SVC	**0,7595**	0,6166	0,8554	**0,7756**	0,7412	0,7686	**0,7595**	0,6732	0,8097	**0,7641**

Com o resultado do processo de customização de parâmetros dos algoritmos, três dos cinco modelos apresentaram melhorias nos resultados: *Random Forest*, *K-Nearest Neighbor* (KNN) e o *Support Vector Classifier* (SVC). O algoritmo que mais se beneficiou da customização foi o KNN, cuja acurácia aumentou de 68,5% para 82,8%, uma melhoria significativa, considerando que ele foi o pior classificador nos testes realizados anteriormente, quando foi executado sem a configuração de nenhum parâmetro e usando os valores pré-definidos pela biblioteca. Mesmo após a customização de parâmetros, o algoritmo RF se manteve com os melhores resultados.

Por outro lado, os algoritmos *Logistic Regression* e *Decision Tree* tiveram uma piora no desempenho após a customização, sendo que o *Decision Tree* foi o modelo que mais perdeu desempenho. Já o *Support Vector Classifier* registrou um incremento sutil, com a acurácia subindo de 71% para 75,9%, além de apresentar métricas mais equilibradas em comparação à configuração anterior.

Os resultados indicam que a customização de parâmetros por meio do *Grid Search Cross Validation* pode melhorar o desempenho dos algoritmos. No entanto, é essencial avaliar cuidadosamente o impacto dessas alterações, pois, em alguns casos, a redução do desempenho pode indicar que o modelo se tornou mais ajustado às particularidades do conjunto de dados utilizado, não representando necessariamente um problema.

5.2.2 Curva ROC

A Figura 5.4 apresenta as curvas ROC para os cinco algoritmos de aprendizado de máquina implementados, após a customização de parâmetros.

Ao comparar os modelos com e sem ajuste de parâmetros, observamos uma melhora no desempenho do *K-Nearest Neighbor* (AUC = 0,92) e *Support Vector Machine* (AUC = 0,80), *Random Forest* (AUC = 0,94) e *Logistic Regression* (AUC = 0,75) mantiveram pra-

ticamente o mesmo desempenho obtido sem a customização de parâmetros, com o RF passando de uma AUC de 0,93 para 0,94 e o LR se mantendo no AUC de 0,75. O *Decision Tress* apresentou uma pequena piora no valor de AUC, passando de 0,87 para 0,85 com a customização de parâmetros.

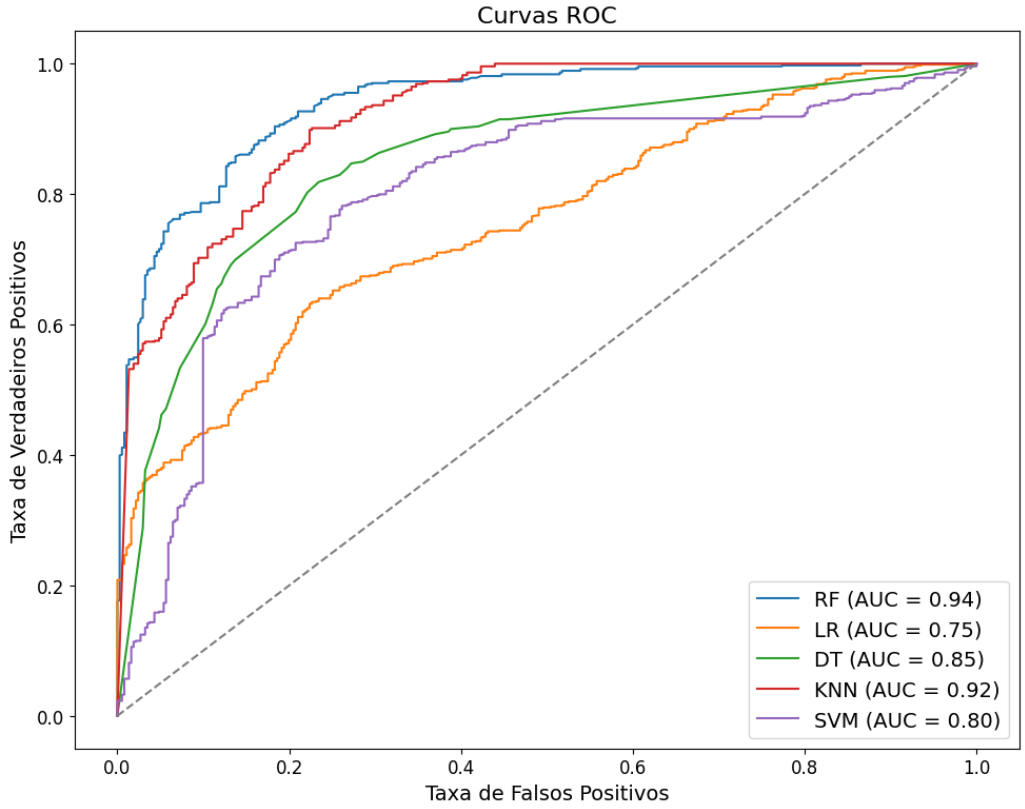


Figura 5.4 – Curva ROC comparando o desempenho dos 5 algoritmos testados após a otimização de parâmetros.

5.2.3 Matrizes de confusão

Na Figura 5.5 são apresentadas as matrizes de confusão para os modelos após a customização de parâmetros.

Assim como quando implementado sem parâmetros customizados, o algoritmo RF (Figura 5.5(a)) manteve o bom resultado apresentado, gerando uma pequena melhora no acerto para a classe de sobrevida curta, passando de 299 para 303 instâncias classificadas corretamente, e mantendo o número de acertos em 664 instâncias para a classe de sobrevida longa. KNN (Figura 5.5(c) melhorou o número de acertos de 248 para 303 (classe de sobrevida curta), e de 512 para 615 (classe de sobrevida longa), uma melhora que evidencia as métricas apresentadas na Tabela 5.3. O algoritmo *Support Vector Classifier* (Figura 5.5(e)) passou a errar menos para a classe de sobrevida curta, mas se manteve estável em relação à classe de sobrevida longa.

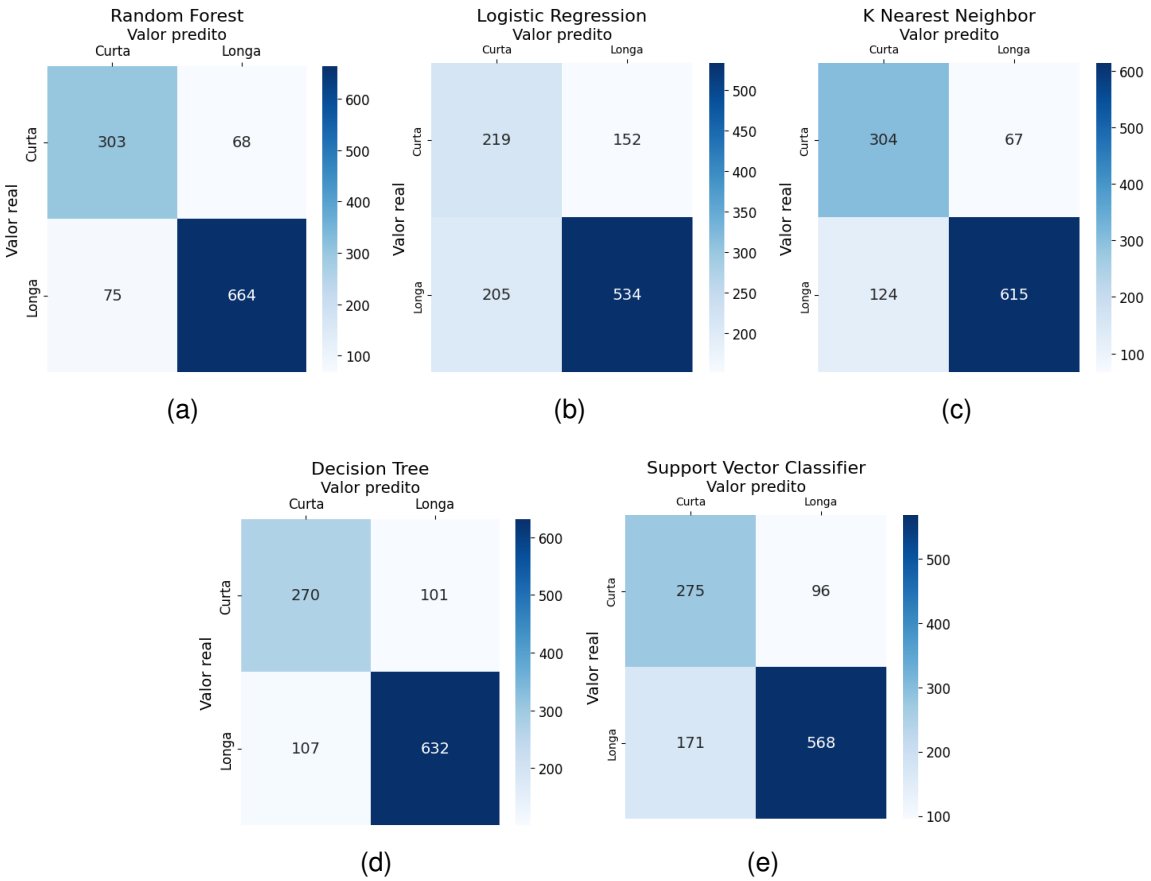
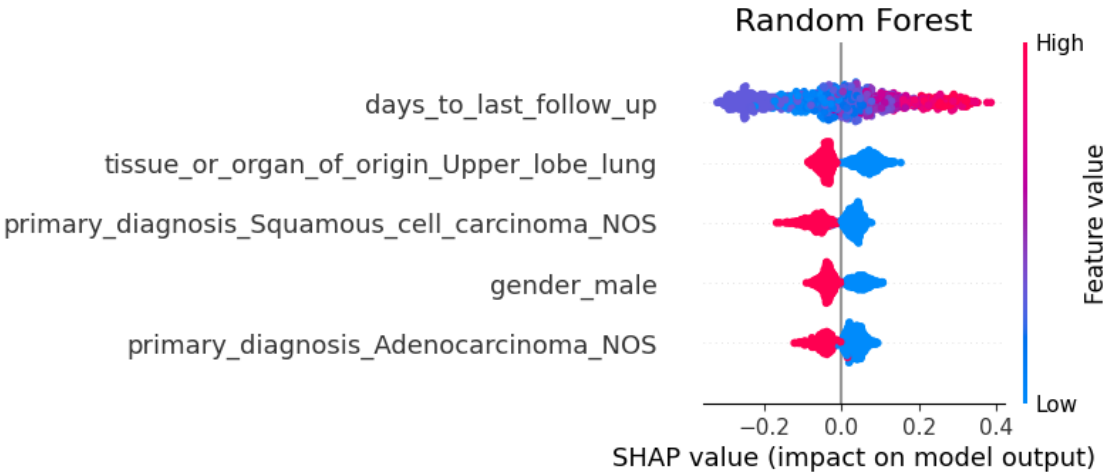


Figura 5.5 – Matrizes de confusão para os algoritmos de ML usando a customização de parâmetros. RF (a), LR (b), KNN (c), DT (d) e SVC (e).

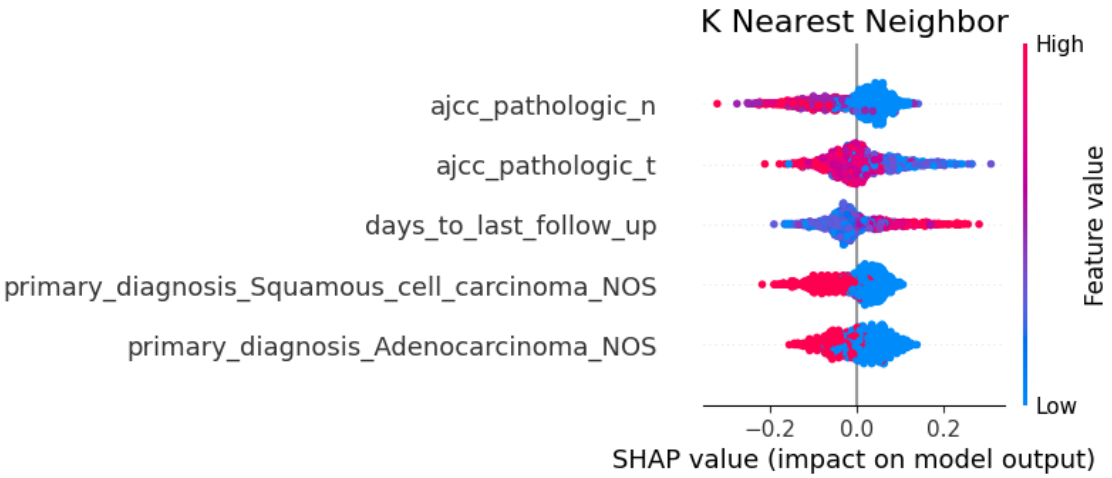
5.2.4 SHAP

A Figura 5.6 apresenta o *summary plot* gerado pelo SHAP para dois algoritmos que obtiveram bons resultados de classificação. O objetivo é explorar o comportamento desses algoritmos, destacando a seleção de atributos e o impacto de cada um nos resultados obtidos.

Ao analisar os resultados do SHAP para os algoritmos que apresentaram melhor desempenho com a customização de parâmetros, *Random Forest* (Figura 5.6(a)) e *K-Nearest Neighbors* (Figura 5.6(b)), observa-se que para ambos os algoritmos, o atributo “days_to_last_follow_up” é listado entre os cinco mais importantes, assim como na implementação sem customização, evidenciando a grande relevância desta variável para os resultados. O atributo “primary_diagnosis_Squamous cell_carcinoma_nos” está relacionado ao diagnóstico primário do paciente. Quando esse diagnóstico tem origem em células escamosas, há uma tendência à sobrevida curta. Essas células revestem diversas partes do corpo, incluindo o trato respiratório [59]. Além disso, ao observar o atributo “gender_male”, nota-se que homens apresentam maior tendência à sobrevida curta.



(a)



(b)

Figura 5.6 – *Shap Summary Plot* para os dois algoritmos com o melhor desempenho obtido com a customização de parâmetros, RF (a) e KNN (b).

Observando os atributos importantes para o algoritmo *K Nearest Neighbor*, a variável “ajcc_pathologic_n” indica que não ter presença de câncer nos linfonodos próximos favorece a classe de sobrevida longa. Já a variável “ajcc_pathologic_t”, que também faz parte do estadiamento TNM, mostra que tumores menores estão associados a uma maior sobrevida.

5.3 Comparando os resultados entre os modos de implementação

Para facilitar a comparação, a Tabela 5.3 apresenta as métricas de acurácia, e as médias de precisão, *Recall* e *F1-Score* para os cinco modelos implementados antes e após a customização dos parâmetros, permitindo uma comparação dos resultados. Observando as métricas dispostas lado a lado, podemos notar uma melhora nos resultados, especi-

almente no KNN, que apresentou um avanço significativo em todas as métricas. Apesar dessa melhora, o algoritmo RF manteve o melhor resultado. Considerando os dados apresentados, o algoritmo DT apresentou os melhores resultados nas métricas de acurácia, recall e F1-Score quando implementado sem customização de parâmetros. Por outro lado, com a customização, o algoritmo RF obteve os melhores desempenhos em acurácia, precisão, recall e F1-Score.

Tabela 5.3 – Comparativo das métricas dos algoritmos de ML, apresentando a acurácia e os valores médios de precisão, *recall* e *F1-Score* para os dois modos de implementação.

Alg.	**Acurácia**		**Precisão**		**Recall**		**F1-score**	
	norm.	**cust.**	**norm.**	**cust.**	**norm.**	**cust.**	**norm.**	**cust.**
RF	0,8676	**0,8712**	0,8678	**0,8718**	0,8676	**0,8712**	0,8677	**0,8715**
LR	0,6955	0,6784	0,7044	0,6909	0,6955	0,6784	0,6990	0,6831
KNN	0,6847	0,8279	0,7113	0,8378	0,6847	0,8279	0,6921	0,8306
DT	**0,8694**	0,8126	**0,8698**	0,8134	**0,8694**	0,8126	**0,8696**	0,8130
SVC	0,7108	0,7595	0,7143	0,7756	0,7108	0,7595	0,7124	0,7641

5.4 Validação cruzada

A Tabela 5.4 apresenta os resultados da validação cruzada que foi configurada com 10 divisões, ou seja, o conjunto de dados de treinamento foi dividido em 10 partes iguais. Em cada iteração, 9 partes foram utilizadas para o treinamento do modelo, enquanto a parte restante foi usada para teste. Na tabela, é apresentada a média das acurácias obtidas ao longo das 10 iterações e o desvio padrão, tanto para a implementação padrão quanto para a versão com customização de parâmetros. A escolha do número de divisões para a validação cruzada foi baseada no estudo de Kohavi [44], no qual o autor aponta bons resultados para 10 divisões.

Valores baixos para o desvio padrão indicam que o modelo apresentou desempenho consistente, independentemente da divisão do conjunto de dados. Os melhores resultados foram alcançados pelos algoritmos *Random Forest* e *Decision Tree*. Outro ponto a se destacar é a melhora dos valores após a customização de parâmetros.

5.5 Discussão com base em estudos Anteriores

Os resultados apresentados permitem avaliar os diferentes comportamentos dos algoritmos e qual tem o melhor desempenho com o conjunto de dados testado. Podemos observar resultados semelhantes em outros estudos que aplicaram os mesmos algoritmos,

Tabela 5.4 – Apresentação dos resultados da validação cruzada realizada em 10 divisões para cada algoritmo.

Alg.	Forma de implementação			
	normal		usando customização	
	Acurácia média	Desvio padrão	Acurácia média	Desvio padrão
RF	0,8725	$\pm 0,0257$	0,8755	$\pm 0,0220$
LR	0,7150	$\pm 0,0408$	0,6465	$\pm 0,0347$
KNN	0,6928	$\pm 0,0425$	0,8425	$\pm 0,0342$
DT	0,8569	$\pm 0,0282$	0,7956	$\pm 0,0262$
SVC	0,6843	$\pm 0,0371$	0,7727	$\pm 0,0300$

porém em aplicações diferentes. Nesta seção, são abordados alguns desses estudos a fim de estabelecer um comparativo.

Em um estudo semelhante, Naser et al. [63] utilizaram os algoritmos DT, RF e KNN para realizar a classificação de pacientes com câncer de mama em cinco classes de sobrevida [0, 2), [2-4), [4-6), [6-8) e [8-10), analisando dados clínicos dos pacientes. Os resultados apresentados apontam que RF e DT obtiveram os melhores resultados no estudo, com acurácia de 69,7% para o RF e 68,8% para o DT. Resultado semelhante aos obtidos por este estudo, onde o RF e DT obtiveram o melhor desempenho entre os 5 classificadores.

Em Gan et al. [21], os autores avaliam a capacidade dos algoritmos KNN, SVM e RF na tarefa de agrupar pacientes em três classes de sobrevida, até 3 anos, entre 3 e 5 anos e mais de 5 anos, analisando dados de expressão gênica. Foram estudados diferentes cenários de aplicação dos vários algoritmos, e o melhor resultado foi apresentado pelo algoritmo KNN com uma acurácia de 70%, resultado alinhado com a implementação do KNN obtido por este estudo que foi de 68,5% sem customização e 82,8% com customização de parâmetros. Em Liu et al. [52], os autores apresentam um exemplo de aplicação de imagens para classificar pacientes tratados com glioblastoma pós-cirúrgicos, em sobrevida de até 650 dias ou mais de 650 dias. O estudo avaliou as conexões estruturais (como as regiões do cérebro se conectam entre si fisicamente) e funcionais (baseado na inter-relação das diferentes regiões do cérebro, entendendo o impacto que ocorre nas diferentes regiões avaliadas). Os melhores resultados foram obtidos ao analisar os dois conjuntos de informações ao mesmo tempo, sendo que o algoritmo SVM obteve uma acurácia de 75%, semelhante ao obtido por este estudo pelo mesmo algoritmo quando implementado aplicando a customização de parâmetros (75,9%).

Esses relatos apontam que os resultados obtidos por este estudo estão alinhados com os resultados obtidos pela literatura. As diferentes aplicações dos algoritmos de ML apresentam resultados semelhantes. Outro ponto importante a se destacar é que a escolha do algoritmo de ML está muito relacionada ao tipo e quantidade de informações que se pretende avaliar. Um resumo desta análise é apresentado na Tabela 5.5.

Tabela 5.5 – Resumo da comparação deste estudo com alguns dos trabalhos relacionados.

Estudo	Algoritmos Avaliados	Tipo de câncer	Tipo de Dados	Melhor Alg.	Acurácia
Este estudo	RF, LR, DT, SVM, KNN	pulmão	Dados clínicos	RF	RF: 87,1% (cust.) KNN: 82,8% (cust.) DT: 86,9% (norm.)
Naser et al. [63]	DT, RF, SVM, KNN	mama	Dados clínicos	RF	RF: 69,7% DT: 68,8%
Gan et al. [21]	KNN, SVM, RF	pulmão	Expressão gênica	KNN	70%
Liu et al. [52]	SVM	glioblastoma	Imagens cerebrais	SVM	75%

5.6 Limitações e trabalhos futuros

Outro aspecto que poderia ser avaliado neste estudo é a aplicação dos modelos de ML usando outras bases de dados. Embora o conjunto de dados utilizado tenha 2773 instâncias, um volume maior de dados poderia contribuir para mudanças nos resultados. Como não se tem uma padronização entre as colunas das diferentes fontes de dados disponibilizadas, a junção destas bases em um único conjunto de dados é muito complexa. Uma possibilidade de estudo futuro seria desenvolver um modelo capaz de mapear as diferentes colunas, buscando encontrar semelhanças entre as bases de dados, fazendo, assim, com que o volume e a heterogeneidade do conjunto de dados sejam ampliados.

Outro ponto ainda relacionado aos dados é o desbalanceamento natural dos conjuntos de dados disponibilizados. Embora faça parte do processo, para o treinamento dos modelos, esse desbalanceamento não é interessante. Então, outra possibilidade de estudo é explorar outras formas de balanceamento de dados, com outras estratégias, além do processo híbrido usando *oversampling* e *undersampling* ao mesmo tempo, que foi aplicado neste estudo. A escolha de uma técnica inadequada pode gerar impactos negativos nos resultados obtidos, tornando o modelo tendencioso, em especial para dados que tendem a ter um desbalanceamento natural, como é o caso da sobrevida curta e longa. Uma das questões que se pode avaliar neste ponto é o quanto os dados sintéticos gerados pelas técnicas de *oversampling* podem impactar nos resultados dos algoritmos, avaliando se essas instâncias podem deixar os modelos tendenciosos para uma das classes.

Além do aspecto relacionado ao balanceamento dos dados, o estudo poderia ser ampliado, tomando como base os resultados de importância dos atributos apresentados pelo SHAP, realizando outros arranjos de atributos, para avaliar os impactos nos resultados. Um exemplo seria remover uma variável com grande importância nos resultados e comparar o comportamento dos algoritmos em relação aos demais atributos.

O acompanhamento do projeto por profissionais de saúde, em especial da área oncológica, poderia contribuir para melhorias nos resultados. Assim, poderia haver um melhor entendimento sobre o câncer de pulmão, além de verificar a possibilidade de integração

com sistemas já existentes nos hospitais para permitir a avaliação dos modelos com outros dados. Com isso, seria possível avaliar a aplicação dos modelos na prática clínica.

Este estudo comparou o desempenho de cinco algoritmos de ML (RF, LR, DT, SVM e KNN) na tarefa de classificar pacientes em sobrevida curta e longa. Porém, existem outros classificadores que também poderiam ser avaliados para estabelecer um comparativo com os resultados obtidos.

6. CONCLUSÃO

Pesquisas relacionadas ao tempo de sobrevida em pacientes com câncer indicam seu potencial para apoiar equipes médicas na decisão sobre o melhor caminho de tratamento em cada caso. Com o objetivo de buscar estudos sobre a previsão do tempo de sobrevida em pacientes com câncer, foi aplicada uma metodologia de Revisão Sistemática da Literatura (RSL) para avaliar como estão sendo realizadas as pesquisas na área. O processo iniciou com uma busca nas bases IEEE e ACM e terminou com a seleção de 64 estudos que foram analisados para a extração de informações que permitiram responder às questões de pesquisa.

Assim, foi possível entender como são desenvolvidos os projetos na área, identificar quais são os principais modelos de ML usados nas predições do tempo de sobrevida de pacientes e obter sugestões de fontes de dados confiáveis para posterior aplicação no desenvolvimento da pesquisa. O desenvolvimento foi iniciado com uma busca e pré-processamento dos dados obtidos do TCGA. Este processo foi realizado usando a ferramenta Tableau Prep e, posteriormente, as bibliotecas do Python. Na sequência, os modelos foram desenvolvidos e analisados a fim de identificar qual tem o melhor desempenho de classificação. Os algoritmos foram implementados em dois formatos: usando os algoritmos com suas configurações padrão e usando uma ferramenta para identificar os melhores parâmetros. Em ambas as implementações, o desempenho do algoritmo *Random Forest* merece destaque, com uma acurácia de 87%, seguido pelo *Decision Tree* com uma acurácia de 86% quando implementado sem customização de parâmetros. Algo importante a se destacar é o desempenho do algoritmo *K Nearest Neighbors*, que obteve uma melhora significativa no processo de classificação, quando aplicada a customização, chegando a uma acurácia de 82%.

Os resultados apresentados respondem à questão inicial da pesquisa: “Qual modelo de aprendizado de máquina apresenta o melhor desempenho na tarefa de prever o tempo de sobrevida de pacientes com câncer de pulmão?”. Os algoritmos com melhor desempenho na tarefa de classificação de pacientes foram o *Random Forest* e o *Decision Tree*, que tiveram bons resultados nos testes. Também é importante ressaltar que a configuração customizada dos parâmetros dos algoritmos pode impactar os resultados dos modelos, já que serão escolhidos parâmetros específicos para maximizar o desempenho para o conjunto de dados trabalhado. Esses resultados se destacam como as principais contribuições do estudo, auxiliando na escolha de algoritmos para a tarefa de classificação de pacientes em tratamento de câncer, além de demonstrar que a escolha dos melhores parâmetros para a configuração dos algoritmos de ML pode trazer importantes melhorias nos resultados.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Ahmed, K. B.; Hall, L. O.; Liu, R.; Gatenby, R. A.; Goldgof, D. B. “Neuroimaging Based Survival Time Prediction of GBM Patients Using CNNs from Small Data”. In: International Conference on Systems, Man and Cybernetics (SMC), 2019, pp. 1331–1335.

[2] Aljouie, A.; Roshan, U. “Multi-path convolutional neural network for glioblastoma survival group prediction with point mutations and demographic features”. In: International Conference on Bioinformatics and Biomedicine (BIBM), 2019, pp. 1274–1279.

[3] Aljouie, A.; Xue, Y.; Xie, M.; Roshan, U. “Challenges in predicting glioma survival time in multi-modal deep networks”. In: International Conference on Bioinformatics and Biomedicine (BIBM), 2020, pp. 2558–2562.

[4] Amazon, W. S. “Aprendizado profundo na AWS”. Source: https://aws.amazon.com/pt/deep-learning/, 2024-02-20.

[5] Baker, Q. B.; Gharaibeh, M.; Al-Harahsheh, Y. “Predicting Lung Cancer Survival Time Using Deep Learning Techniques”. In: International Conference on Information and Communication Systems (ICICS), 2021, pp. 177–181.

[6] Bartholomai, J. A.; Frieboes, H. B. “Lung Cancer Survival Prediction via Machine Learning Regression, Classification, and Statistical Techniques”. In: International Symposium on Signal Processing and Information Technology (ISSPIT), 2018, pp. 632–637.

[7] Brownlee, J. “Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python”. Machine Learning Mastery, 2020, 398p.

[8] Bruce, P. C.; Bruce, A.; Gedeck, P. “Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python”. Beijing; Boston; Farnham; Sebastopol; Tokyo: O’Reilly Media, 2020, 2 ed., 363p.

[9] Cai, Z.; Yu, Z.; Zhou, H.; Gu, Z. “The Early Stage Lung Cancer Prognosis Prediction Model based on Support Vector Machine”. In: International Conference on Digital Signal Processing (DSP), 2018, pp. 1–4.

[10] Chaddad, A.; Desrosiers, C.; Toews, M. “Radiomic analysis of multi-contrast brain MRI for the prediction of survival in patients with glioblastoma multiforme”. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2016, pp. 4035–4038.

[11] Chai, H.; Liang, Y.; Liu, X.-Y. “The L1/2 regularization approach for survival analysis in the accelerated failure time model”, *Computers in Biology and Medicine*, vol. 64, September 2015, pp. 283–290.

[12] Chan, L.; Chan, T.; Cheng, L.; Mak, W. “Machine learning of patient similarity: A case study on predicting survival in cancer patient after locoregional chemotherapy”. In: International Conference on Bioinformatics and Biomedicine Workshops (BIBMW), 2010, pp. 467–470.

[13] Chen, C.-M.; Hsu, C.-Y.; Chiu, H.-W.; Rau, H.-H. “Prediction of survival in patients with liver cancer using artificial neural networks and classification and regression trees”. In: Seventh International Conference on Natural Computation (ICNC), 2011, pp. 811–815.

[14] Chen, Y.-C.; Yang, W.-W.; Chiu, H.-W. “Artificial Neural Network Prediction for Cancer Survival Time by Gene Expression Data”. In: International Conference on Bioinformatics and Biomedical Engineering (ICBBE), 2009, pp. 1–4.

[15] Dekker, A.; Dehing-Oberije, C.; Ruysscher, D. D.; Lambin, P.; Komati, K.; Fung, G.; Yu, S.; Hope, A.; Neve, W. D.; Lievens, Y. “Survival Prediction in Lung Cancer Treated with Radiotherapy: Bayesian Networks vs. Support Vector Machines in Handling Missing Data”. In: International Conference on Machine Learning and Applications (ICMLA), 2009, pp. 494–497.

[16] Dertat, A. “Applied Deep Learning: Convolutional Neural Networks”. Source: https://tinyurl.com/yx2uank2, 2024-04-28.

[17] Faceli, K.; Lorena, A. C.; Gama, J.; Carvalho, A. C. P. D. L. F. D.; Almeida, T. A. D. “Inteligência Artificial-Uma Abordagem De Aprendizado De Máquina”. "Rio de Janeiro, RJ": Ltc-Livros Tecnicos e Cientificos Editora Lda, 2021, 2 ed., 395p.

[18] Fawcett, T. “An introduction to ROC analysis”, *Pattern Recognition Letters*, vol. 27–8, June 2006, pp. 861–874.

[19] Ferdinand Christ, P.; Ettlinger, F.; Kaissis, G.; Schlecht, S.; Ahmaddy, F.; Grun, F.; Valentinitsch, A.; Ahmadi, S.-A.; Braren, R.; Menze, B. “SurvivalNet: Predicting patient survival from diffusion weighted magnetic resonance images using cascaded fully convolutional and 3D Convolutional Neural Networks”. In: International Symposium on Biomedical Imaging (ISBI), 2017, pp. 839–843.

[20] Galvão, M. C. B.; Ricarte, I. L. M. “Revisão sistemática da literatura: conceituação, produção e publicação”, *Logeion: Filosofia da Informação*, vol. 6, September 2019, pp. 57–73.

[21] Gan, B.; Zheng, C.-H.; Wang, H.-Q. “A survey of pattern classification-based methods for predicting survival time of lung cancer patients”. In: International Conference on Bioinformatics and Biomedicine (BIBM), 2014, pp. 5–12.

[22] GCO. “Cancer Today”. Source: https://gco.iarc.who.int/today/, 2024-02-27.

[23] Ghazipour, A.; Settle, T.; Veasey, B.; Daugherty, E.; Keltner, S.; Kumar, N.; Ververs, J.; Farris, M.; Dunlap, N.; Amini, A. A. “Survival Outcome Prediction for Stereotactic Body Radiation Therapy of Lung Cancer from Post-RT Ct Images with RNN/CNN Deep Learning”. In: International Symposium on Biomedical Imaging (ISBI), 2023, pp. 1–4.

[24] Gonçalves, E. S.; Silva, A. L.; Ramalho, J. G. G.; Spricigo, T.; Castro, A. C. F. “O impacto da tecnologia de inteligência artificial na medicina diagnóstica”, *Revista ft*, vol. 28, December 2023.

[25] Géron, A. “Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems”. Beijing; Boston; Farnham; Sebastopol; Tokyo: O’Reilly, 2019, 2 ed., 851p.

[26] Harrison, M. “Machine Learning – Guia de Referência Rápida: Trabalhando com dados estruturados em Python”. São Paulo, SP: Novatec Editora, 2019, 272p.

[27] Hartshorn, S. “Machine Learning With Random Forests And Decision Trees: A Visual Guide For Beginners”. eBook Kindle, 2016, 82p.

[28] Hawkins, S. H.; Korecki, J. N.; Balagurunathan, Y.; Yuhua Gu; Kumar, V.; Basu, S.; Hall, L. O.; Goldgof, D. B.; Gatenby, R. A.; Gillies, R. J. “Predicting Outcomes of Nonsmall Cell Lung Cancer Using CT Image Features”, *IEEE Access*, vol. 2, November 2014, pp. 1418–1426.

[29] Hayward, J.; Alvarez, S. A.; Ruiz, C.; Sullivan, M.; Tseng, J.; Whalen, G. “Machine learning of clinical performance in a pancreatic cancer database”, *Artificial Intelligence in Medicine*, vol. 49, July 2010, pp. 187–195.

[30] Henrich, R.; Manssour, I. H.; Bordini, R. H. “Using machine learning techniques for lung cancer survival prediction”, *World Congress on Medical and Health Informatics (MedInfo)*, 2025.

[31] Hossain, M. A.; Saiful Islam, S. M.; Quinn, J. M.; Huq, F.; Moni, M. A. “Machine learning and bioinformatics models to identify gene expression patterns of ovarian cancer associated with disease progression and mortality”, *Journal of Biomedical Informatics*, vol. 100, December 2019, pp. 103313.

[32] Hossain, M. J.; Chowdhury, U. N.; Islam, M. B.; Uddin, S.; Ahmed, M. B.; Quinn, J. M.; Moni, M. A. “Machine learning and network-based models to identify genetic

risk factors to the progression and survival of colorectal cancer", *Computers in Biology and Medicine*, vol. 135, August 2021, pp. 104539.

[33] Hou, B.; Li, H.; Jiao, Z.; Zhou, Z.; Zheng, H.; Fan, Y. "Deep Clustering Survival Machines with Interpretable Expert Distributions". In: International Symposium on Biomedical Imaging (ISBI), 2023, pp. 1–4.

[34] Huang, C.-M.; Hung, C.-S.; Hsu, Y.-Y.; Zheng, Y.-C.; Yu, C.-H.; Lin, C.-H. R.; Chen, S.-H. "A K-means Clustering Based Under-Sampling Method for Imbalanced Dataset Classification". In: International Conference on Information Networking (ICOIN), 2024, pp. 708–713.

[35] Huang, H.-H.; Liang, Y. "A Novel Cox Proportional Hazards Model for High-Dimensional Genomic Data in Cancer Prognosis", *Transactions on Computational Biology and Bioinformatics*, vol. 18, September 2021, pp. 1821–1830.

[36] Hubel, D. H.; Wiesel, T. N. "Receptive fields of single neurones in the cat's striate cortex", *J Physiol*, vol. 148, October 1959, pp. 574–591.

[37] Imbalanced-learn developers. "Imblearn.over_sampling.SMOTE", 2024, imbalanced-learn 0.11.0 documentation, Source: https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html.

[38] INCA. "Câncer de pulmão". Source: https://www.gov.br/inca/pt-br/assuntos/cancer/tipos/pulmao, 2024-02-12.

[39] Isik, Z.; Ercan, M. E. "Integration of RNA-Seq and RPPA data for survival time prediction in cancer patients", *Computers in Biology and Medicine*, vol. 89, October 2017, pp. 397–404.

[40] Izbicki, R.; Santos, T. M. d. "Aprendizado de máquina: uma abordagem estatística". São Carlos, SP: UICLAP, 2020, 270p.

[41] Jayashanka, R.; Wijesinghe, C.; Weerasinghe, A.; Pieris, D. "Machine Learning Approach to Predict the Survival Time of Childhood Acute Lymphoblastic Leukemia Patients". In: International Conference on Advances in ICT for Emerging Regions (ICTer), 2018, pp. 426–432.

[42] Kim, D.; Li, R.; Dudek, S. M.; Ritchie, M. D. "Predicting censored survival data based on the interactions between meta-dimensional omics data in breast cancer", *Journal of Biomedical Informatics*, vol. 56, August 2015, pp. 220–228.

[43] Kitchenham, B. A. "Systematic review in software engineering: where we are and where we should be going". In: International workshop on Evidential Assessment of Software Technologies (EAST), 2012, pp. 1–2.

[44] Kohavi, R. “A study of cross-validation and bootstrap for accuracy estimation and model selection”. In: International Joint Conference on Artificial Intelligence (IJCAI), 1995, pp. 1137–1143.

[45] Kolasa, M.; Jozwicki, W.; Wojtyna, R.; Jarzemski, P. “Optimization of hidden layer in a neural network used to predict bladder-cancer patient-survival”. In: Signal Processing Algorithms, Architectures, Arrangements, and Applications SPA, 2007, pp. 69–74.

[46] Kukreja, S.; Sabharwal, M.; Shah, M. A.; Gill, D. S. “A Heuristic Machine Learning-Based Optimization Technique to Predict Lung Cancer Patient Survival”, *Computational Intelligence and Neuroscience*, vol. 2023, February 2023, pp. 1–9.

[47] Kuo, C.-F. J.; Ke, B.-H.; Wu, N.-Y.; Kuo, J.; Hsu, H.-H. “Prognostic value of tumor volume for patients with advanced lung cancer treated with chemotherapy”, *Computer Methods and Programs in Biomedicine*, vol. 144, June 2017, pp. 165–177.

[48] LeCun, Y.; Boser, B.; Denker, J. S.; Henderson, D.; Howard, R. E.; Hubbard, W.; Jackel, L. D. “Backpropagation applied to handwritten zip code recognition”, *Neural Computation*, vol. 1, December 1989, pp. 541–551.

[49] Li, L.; Li, H. “Dimension reduction methods for microarrays with application to censored survival data”, *Bioinformatics*, vol. 20, December 2004, pp. 3406–3412.

[50] Li, X.; Qian, X.; Liang, L.; Kong, L.; Dong, Q.; Chen, J.; Liu, D.; Yao, X.; Fu, Y. “Causally-Aware Intraoperative Imputation for Overall Survival Time Prediction”. In: Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15681–15690.

[51] Li, Y.; Wang, L.; Zhou, J.; Ye, J. “Multi-task learning based survival analysis for multi-source block-wise missing data”, *Neurocomputing*, vol. 364, October 2019, pp. 95–107.

[52] Liu, L.; Zhang, H.; Rekik, I.; Chen, X.; Wang, Q.; Shen, D. “Outcome Prediction for Patient with High-Grade Gliomas from Brain Functional and Structural Networks”. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI), Ourselin, S.; Joskowicz, L.; Sabuncu, M. R.; Unal, G.; Wells, W. (Editors), 2016, pp. 26–34.

[53] Liu, S.; Li, H.; Zheng, Q.; Yang, L.; Duan, M.; Feng, X.; Li, F.; Huang, L.; Zhou, F. “Survival Time Prediction of Breast Cancer Patients Using Feature Selection Algorithm Crystall”, *IEEE Access*, vol. 9, JAN 2021, pp. 24433–24445.

[54] Liu, Z. “Cox’s Proportional Hazards Model with Lp Penalty for Biomarker Identification and Survival Prediction”. In: International Conference on Machine Learning and Applications (ICMLA), 2007, pp. 624–628.

[55] Lu, W.; Guo, L.; Mao, L. “An integrated model of clinical information and gene expression for prediction of survival in breast cancer patients”. In: Chinese Control Conference (CCC), 2020, pp. 5873–5877.

[56] Lundberg, S. “An introduction to explainable AI with Shapley values — SHAP latest documentation”. Source: https://shap.readthedocs.io/en/latest/, 2024-02-29.

[57] Ma, Y.; Zhu, H.; Yang, Z.; Wang, D. “Optimizing the Prognostic Model of Cervical Cancer Based on Artificial Intelligence Algorithm and Data Mining Technology”, *Wireless Communications and Mobile Computing*, vol. 2022, August 2022, pp. 1–10.

[58] Malhotra, K.; Navathe, S. B.; Chau, D. H.; Hadjipanayis, C.; Sun, J. “Constraint based temporal event sequence mining for Glioblastoma survival prediction”, *Journal of Biomedical Informatics*, vol. 61, June 2016, pp. 267–275.

[59] Maluf, D. F. C. “Tudo sobre Câncer de Pulmão | Vencer o Câncer”. Running Time: 73 Section: Sem Categoria, Source: https://vencerocancer.org.br/tipos-de-cancer/cancer-de-pulmao-o-que-e/, 2025-01-21.

[60] Microsoft. “Liberar sua produtividade com IA e Microsoft Copilot - Suporte da Microsoft”. Source: https://tinyurl.com/4ebea46k, 2024-02-12.

[61] Mueller, J. P.; Massaron, L. “Aprendizado profundo”. Rio de Janeiro, RJ: Alta Books, 2020, 653p.

[62] Nanda, P.; Duraipandian, N. “Prediction of Survival Rate from Non-Small Cell Lung Cancer using Improved Random Forest”. In: International Conference on Inventive Computation Technologies (ICICT), 2020, pp. 93–97.

[63] Naser, M. Y. M.; Chambers, D.; Bhattacharya, S. “Prediction Model of Breast Cancer Survival Months: A Machine Learning Approach”. In: SoutheastCon, 2023, pp. 851–855.

[64] OPAS. “Câncer - OPAS/OMS | organização pan-americana da saúde”. Source: https://www.paho.org/pt/topicos/cancer, 2024-08-27.

[65] Phong, P. A.; Dong, D. K.; Khang, T. D. “Hedge Algebra Based Type-2 Fuzzy Logic System and its Application to Predict Survival Time of Myeloma Patients”. In: International Conference on Knowledge and Systems Engineering (KSE), 2009, pp. 13–18.

[66] Ponce-Bobadilla, A. V.; Schmitt, V.; Maier, C. S.; Mensing, S.; Stodtmann, S. “Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development”, *Clinical and Translational Science*, vol. 17, November 2024, pp. e70056.

[67] Prakash, T.; Dhamija, T.; Kumar, R.; Panda, J. “Leveraging Explainable Artificial Intelligence for Understanding the Effect of Model Capacity on Training Dataset Size”.

In: International Conference on Service Operations and Logistics, and Informatics (SOLI), 2022, pp. 1–6, iSSN: 2768-1890.

[68] Pu, Y.; Baad, M. J.; Jiang, Y.; Chen, Y. “Application of artificial neural network and multiple linear regression models for predicting survival time of patients with non-small cell cancer using multiple prognostic factors including FDG-PET measurements”. In: International Joint Conference on Neural Networks (IJCNN), 2014, pp. 225–230.

[69] Radiologico, C. “Adenocarcinoma pulmonar: saiba o que é, sintomas e como é feito o tratamento”. Source: https://tinyurl.com/5954uwt9, 2025-01-17.

[70] Rodrigues, V. “Métricas de Avaliação: acurácia, precisão, recall... quais as diferenças?” Source: https://tinyurl.com/5879xsps, 2024-02-29.

[71] Scalco, E.; Rizzo, G.; Gomez-Flores, W. “Automatic Feature Construction Based on Genetic Programming for Survival Prediction in Lung Cancer Using CT Images”. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2022, pp. 3797–3800.

[72] Shakir, H.; Aijaz, B.; Khan, T. M. R.; Hussain, M. “A deep learning-based cancer survival time classifier for small datasets”, *Computers in Biology and Medicine*, vol. 160, June 2023, pp. 106896.

[73] Shao, W.; Liu, J.; Zuo, Y.; Qi, S.; Hong, H.; Sheng, J.; Zhu, Q.; Zhang, D. “FAM3L: Feature-Aware Multi-Modal Metric Learning for Integrative Survival Analysis of Human Cancers”, *IEEE Transactions on Medical Imaging*, vol. 42, September 2023, pp. 2552–2565.

[74] Sharma, D.; Deepali; Garg, V. K.; Kashyap, D.; Goel, N. “A deep learning-based integrative model for survival time prediction of head and neck squamous cell carcinoma patients”, *Neural Comput & Applic*, vol. 34, December 2022, pp. 21353–21365.

[75] Silva, I. N. d.; Spatti, D. H.; Flauzino, R. A. “Redes Neurais Artificiais”. são Paulo, SP: Artliber, 2023, 2 ed., 432p.

[76] Sobin, L. H.; Wittekind, C. “TNM classification of malignant tumours”. New York: Wiley-Liss, 2002, 6 ed., 281p.

[77] Society, A. C. “Lung Cancer Survival Rates | 5-Year Survival Rates for Lung Cancer”. Source: https://www.cancer.org/cancer/types/lung-cancer/detection-diagnosis-staging/survival-rates.html, 2024-03-01.

[78] Stepanek, L.; Habarta, F.; Mala, I.; Marek, L.; Pazdirek, F. “A Machine-learning Approach to Survival Time-event Predicting: Initial Analyses using Stomach Cancer

Data". In: International Conference on e-Health and Bioengineering (EHB), 2020, pp. 1–4.

[79] Sun, J.; Yang, Y.; Wang, Y.; Wang, L.; Song, X.; Zhao, X. "Survival Risk Prediction of Esophageal Cancer Based on Self-Organizing Maps Clustering and Support Vector Machine Ensembles", *IEEE Access*, vol. 8, July 2020, pp. 131449–131460.

[80] The Cancer Genome Atlas Research Network; 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", *Nat Genet*, vol. 45, October 2013, pp. 1113–1120.

[81] Timilsina, M.; Buosi, S.; JANik, A.; Minervini, P.; Costabello, L.; Torrente, M.; Provencio, M.; Calvo, V.; Camps, C.; Ortega, A. L.; Massutí, B.; Campelo, M. G.; Del Barco, E.; Bosch-Barrera, J.; Novacek, V. "Machine Learning Survival Models for Relapse Prediction in a Early Stage Lung Cancer Patient". In: International Joint Conference on Neural Networks (IJCNN), 2023, pp. 1–8.

[82] Ture, M.; Tokatli, F.; Kurt, I. "Using Kaplan–Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients", *Expert Systems with Applications*, vol. 36, March 2009, pp. 2017–2026.

[83] Tyuryumina, E. Y.; Neznanov, A. A. "On Consolidated Predictive Model of the Natural History of Breast Cancer Considering Primary Tumor and Primary Distant Metastases Growth". In: International Conference on Healthcare Informatics (ICHI), 2017, pp. 484–489.

[84] Vapnik, V. N. "Statistical Learning Theory: 2". New York: Wiley-Interscience, 1998, 1 ed., 768p.

[85] Wang, J.; Wu, H.; Cheng, X.; Guo, Z.; Yu, K.; Shen, Y. "Data-driven intelligent decision for multimedia medical management", *Multimed Tools Appl*, vol. 81, December 2022, pp. 42023–42039.

[86] Wang, L.; Chignell, M.; Jiang, H.; Charoenkitkarn, N. "Cluster-Boosted Multi-Task Learning Framework for Survival Analysis". In: International Conference on Bioinformatics and Bioengineering (BIBE), 2020, pp. 255–262.

[87] Wang, S.; Liu, Z.; Chen, X.; Zhu, Y.; Zhou, H.; Tang, Z.; Wei, W.; Dong, D.; Wang, M.; Tian, J. "Unsupervised Deep Learning Features for Lung Cancer Overall Survival Analysis". In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2018, pp. 2583–2586.

[88] Wang, S.; Zhang, H.; Chai, H.; Liang, Y. “A novel Log penalty in a path seeking scheme for biomarker selection”, *Technology and Health Care*, vol. 27, June 2019, pp. 85–93.

[89] Wang, Y.; Flowers, C. R.; Li, Z.; Huang, X. “CondiS: A conditional survival distribution-based method for censored data imputation overcoming the hurdle in machine learning-based survival analysis”, *Journal of Biomedical Informatics*, vol. 131, July 2022, pp. 104117.

[90] Wei, L.; Wen, W.; Fang, Z. “Using Multiple Machine Learning Algorithms for Cancer Prognosis in Lung Adenocarcinoma”. In: International Conference on Bioscience, Biochemistry and Bioinformatics (ICBBB), 2020, pp. 52–55.

[91] Wu, Q.; Huang, W.; Wang, S.; Yu, H.; Wang, L.; Wu, Z.; Zhu, Y.; Liu, Z.; Ma, H.; Tian, J. “A generative adversarial network-based CT image standardization model for predicting progression-free survival of lung cancer”. In: Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2021, pp. 3411–3414.

[92] Wu, Y.; Ma, J.; Huang, X.; Ling, S. H.; Weidong Su, S. “DeepMMSA: A Novel Multimodal Deep Learning Method for Non-small Cell Lung Cancer Survival Analysis”. In: International Conference on Systems, Man, and Cybernetics (SMC), 2021, pp. 1468–1472.

[93] Yang, M.; Yang, H.; Ji, L.; Hu, X.; Tian, G.; Wang, B.; Yang, J. “A multi-omics machine learning framework in predicting the survival of colorectal cancer patients”, *Computers in Biology and Medicine*, vol. 146, July 2022.

[94] Yu, C.-N.; Greiner, R.; Lin, H.-C.; Baracos, V. “Learning Patient-Specific Cancer Survival Distributions as a Sequence of Dependent Regressors”. In: International Conference on Neural Information Processing Systems (NIPS), 2011, pp. 1845–1853.

[95] Zhang, W.; Edwards, A.; Fan, W.; Flemington, E. K.; Zhang, K. “The modularity and dynamicity of miRNA–mRNA interactions in high-grade serous ovarian carcinomas and the prognostic implication”, *Computational Biology and Chemistry*, vol. 63, August 2016, pp. 3–14.

[96] Zhang, Y.; Li, A.; Peng, C.; Wang, M. “Improve Glioblastoma Multiforme Prognosis Prediction by Using Feature Selection and Multiple Kernel Learning”, *Transactions on Computational Biology and Bioinformatics*, vol. 13, September 2016, pp. 825–835.

[97] Zhou, M.; Hall, L. O.; Goldgof, D. B. “Exploring Brain Tumor Heterogeneity for Survival Time Prediction”. In: International Conference on Pattern Recognition (ICPR), 2014, pp. 580–585.

[98] Zhou, M.; Hall, L. O.; Goldgof, D. B.; Gatenby, R. A.; Gillies, R. J. “A Texture Feature Ranking Model for Predicting Survival Time of Brain Tumor Patients”. In: International Conference on Systems, Man, and Cybernetics (SMC), 2013, pp. 4533–4538.

Pontifícia Universidade Católica do Rio Grande do Sul
Pró-Reitoria de Pesquisa e Pós-Graduação
Av. Ipiranga, 6681 – Prédio 1 – Térreo
Porto Alegre – RS – Brasil
Fone: (51) 3320-3513
E-mail: propesq@pucrs.br
Site: www.pucrs.br