UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS / FACULDADE DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL

Daniela Schimitz de Carvalho

Avanços em Modelos Computacionais para Predição da Sobrevida para o Câncer de Mama Feminino

Juiz de Fora

2024

Daniela Schimitz de Carvalho

Avanços em Modelos Computacionais para Predição da Sobrevida para o Câncer de Mama Feminino

Tese apresentada ao Programa de Pós-Graduação em Modelagem Computacionalda Universidade Federal de Juiz de Fora como requisito parcial à obtenção do grau de Doutor em Modelagem Computacional.

Orientadora: Profª. Drª. Priscila Vanessa Zabala Capriles Goliatt

Coorientador: Prof. Dr. Marco Paulo Lages Parente

Juiz de Fora

2024

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

DE CARVALHO, DANIELA.
Avanços em Modelos Computacionais para Predição da Sobrevida para o Câncer de Mama Feminino / DANIELA DE CARVALHO. -- 2024.
191 f. : il.

Orientadora: Priscila Vanessa Zabala Capriles Goliatt
Coorientador: Marco Paulo Lages Parente
Tese (doutorado) - Universidade Federal de Juiz de Fora, ICE/Engenharia. Programa de Pós-Graduação em Modelagem Computacional, 2024.

1. Análise de Sobrevida. 2. Aprendizado de Máquina. 3. Neoplasia da Mama. I. Zabala Capriles Goliatt, Priscila Vanessa, orient. II. Lages Parente, Marco Paulo, coorient. III. Título.

Daniela Schimitz de Carvalho

Avanços em Modelos Computacionais para Predição da Sobrevida para o Câncer de Mama Feminino

Tese apresentada ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Doutora em Modelagem Computacional. Área de concentração: Modelagem Computacional.

Aprovada em 20 de dezembro de 2024.

BANCA EXAMINADORA

Prof.ª Dr.ª. Priscila Vanessa Zabala Capriles Goliatt - Orientadora

Universidade Federal de Juiz de Fora

Prof. Dr. Marco Paulo Lages Parente - Coorientador

Universidade do Porto

Prof. Dr. Carlos Cristiano Hasenclever Borges

Universidade Federal de Juiz de Fora

Prof.ª Dr.ª Maria Teresa Bustamante Teixeira

Universidade Federal de Juiz de Fora

Prof. Dr. Eduardo Krempser da Silva

Fundação Oswaldo Cruz

Prof.ª Dr.ª. Inês Campos Monteiro Sabino Domingues

Research Centre of the Portuguese Institute of Oncology of Porto

Juiz de Fora, 06/12/2024.

Documento assinado eletronicamente por **Priscila Vanessa Zabala Capriles Goliatt**, **Professor(a)**, em 20/12/2024, às 17:42, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Maria Teresa Bustamante Teixeira**, **Coordenador(a)**, em 20/12/2024, às 20:40, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Eduardo Krempser da Silva**, **Usuário Externo**, em 23/12/2024, às 08:47, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Carlos Cristiano Hasenclever Borges**, **Professor(a)**, em 23/12/2024, às 15:33, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Marco Paulo Lages Parente**, **Usuário Externo**, em 30/01/2025, às 15:17, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Inês Domingues**, **Usuário Externo**, em 31/01/2025, às 07:05, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no Portal do SEI-Ufjf (www2.ufjf.br/SEI) através do ícone Conferência de Documentos, informando o código verificador **2143887** e o código CRC **7D3F0566**.

Dedico este trabalho às mulheres que me inspiram diariamente com sua força e resiliência.

AGRADECIMENTOS

Gostaria de expressar meus mais sinceros agradecimentos a todos que contribuíram para a realização desta tese. Primeiramente, agradeço a Deus, cuja orientação e fé foram fundamentais ao longo desta jornada. Sou profundamente grata à minha família — meus pais, esposo, filhos, irmãs, tios, primos e sogros — pelo amor incondicional e apoio constante.

Um agradecimento especial aos meus pais, José Maria e Regina, pelo exemplo e pelos valores que sempre me guiaram; ao meu esposo, Caio, pela cumplicidade e pelo suporte incondicional. Agradeço sinceramente aos meus filhos, Giulia e Guilherme, pelo carinho e pela assistência técnica e científica, e aos meus filhos menores, Alexandre e Giovanna, pela compreensão durante minha ausência. Minhas irmãs, Débora, Denise e Jenniffer, meu agradecimento pelo apoio contínuo.

Agradeço aos meus amigos pelos conselhos e alegrias compartilhadas. Com um carinho especial, agradeço a Jesuliana, Karla e Érica pela parceria durante a vida acadêmica e pessoal; a Debora, Luciana e aos amigos da Ensin.E pelo incentivo e contribuição ao meu crescimento profissional; e aos amigos do doutorado da UFJF — Gisely, Gustavos, Socorro e Thallys — pelo apoio e companheirismo. Um agradecimento especial vai para Renata, Maíra e Reginaldo pelos momentos de descontração durante os cafés e pela ajuda constante. Também sou grata aos novos amigos portugueses do INEGI (Instituto de Ciência e Inovação em Engenharia Mecânica e Engenharia Industrial) pelo acolhimento e pelas discussões valiosas sobre a progressão dos trabalhos de doutoramento.

Meu sincero agradecimento aos meus mestres pelos ensinamentos valiosos e pelos exemplos inspiradores. Em especial, agradeço a Max e à Jane por me incentivarem e ajudarem no propósito deste trabalho. Meu profundo reconhecimento vai aos meus orientadores: à minha orientadora Priscila, pela amizade, por acreditar no meu potencial e por sempre me incentivar; e ao meu coorientador Marco Parente, por me recepcionar e oferecer novas oportunidades em Portugal. Agradeço aos membros da banca de qualificação e de tese, Carlos Cristiano, Eduardo, Inês, José Karam e Maria Tereza por todas as considerações realizadas.

Agradeço à CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pelo apoio financeiro, à UFJF (Universidade Federal de Juiz de Fora) e ao PGMC (Programa de Pós-Graduação em Modelagem Computacional) pela infraestrutura e suporte acadêmico. Sou igualmente grata à Universidade do Porto, à FEUP (Faculdade de Engenharia da Universidade do Porto) e ao INEGI pelo suporte acadêmico durante o doutorado sanduíche. Finalmente, agradeço a todos que cruzaram meu caminho e contribuíram com exemplo, apoio e amizade ao longo desta trajetória.

“Mesmo quando tudo parece desabar, cabe a mim decidir entre rir ou chorar, ir ou ficar, desistir ou lutar; porque descobri, no caminho incerto da vida, que o mais importante é o decidir.”

Cora Coralina

RESUMO

O Câncer de Mama (CM) é uma das principais causas de morte entre mulheres, sendo frequentemente diagnosticado e tratado tardiamente, o que contribui para as altas taxas de mortalidade. A aplicação de métodos de Aprendizado de Máquina (AM) tem demonstrado grande potencial na predição de desfechos em doenças oncológicas. Este estudo teve como objetivo identificar e avaliar os atributos prognósticos para a predição da sobrevida de pacientes com CM feminino, por meio de modelos computacionais. A pesquisa envolveu a construção e análise de um Banco de Dados (BD) clínico contendo informações de pacientes da Zona da Mata Mineira. Após a obtenção dos dados, foram realizadas etapas de pré-processamento, com uma primeira fase de imputação simples e uma segunda fase de inferência dos valores ausentes, considerando a análise descritiva, a correlação entre os dados e a significância clínica de cada variável. As variáveis correlacionadas acima de 70% foram analisadas novamente, levando em conta sua relevância clínica, e, a partir disso, foram selecionadas as variáveis mais significativas. Essas variáveis passaram então por métodos de seleção de atributos, etapas essenciais para aprimorar a precisão dos modelos e identificar fatores críticos na predição da sobrevida. Com base nessa base de dados, foram avaliados diferentes métodos AM, incluindo modelos lineares, como a Regressão de *Cox Proportional-Hazards* (CPH), tanto não penalizada quanto penalizada (*Lasso* e *Elastic Net*), além de *Survival Support Vector Machine* (SSVM). Também foram testados modelos não lineares, como *Random Survival Forest* (RSF), *Gradient Boosting Survival* (GBS) e *Kernel Survival Support Vector Machine* (KSSVM). A avaliação do desempenho, utilizando métricas específicas para análise de sobrevida, indicou que o RSF obteve o melhor desempenho entre os modelos avaliados.O estudo também destacou a importância do pré-processamento dos dados e da aplicação dos métodos de seleção de atributos, que foram fundamentais para a identificação de variáveis prognósticas relevantes para a prática clínica. Além disso, foram discutidos os impactos dessa abordagem computacional na tomada de decisões clínicas. Os resultados evidenciam o potencial dos modelos computacionais na predição da sobrevida de pacientes com CM feminino, e essa pesquisa contribui para o avanço da oncologia computacional, com a possibilidade de melhorar prognósticos e a qualidade de vida das pacientes. .

Palavras-chave: Análise de Sobrevida. Aprendizado de Máquina. Neoplasia da Mama.

ABSTRACT

Breast Cancer is one of the leading causes of death among women, often diagnosed and treated at advanced stages, which contributes to high mortality rates. The application of Machine Learning methods has shown great potential in predicting outcomes in oncological diseases. This study aimed to identify and evaluate prognostic attributes for predicting the survival of female BC patients through computational models. The research involved the construction and analysis of a clinical Database containing information from patients in the Zona da Mata region of Minas Gerais. After data collection, preprocessing steps were carried out, beginning with a simple imputation phase followed by a second phase of inferring missing values, considering descriptive analysis, data correlation, and the clinical significance of each variable. Variables with a correlation above 70% were reanalyzed, taking into account their clinical relevance, and the most significant variables were selected. These variables were then subjected to feature selection methods, which were essential to enhance model accuracy and identify critical factors in survival prediction. Based on this database, different ML methods were evaluated, including linear models, such as Cox Proportional-Hazards Regression, both non-penalized and penalized (Lasso and Elastic Net), as well as Survival Support Vector Machine. Non-linear models, such as Random Survival Forest, Gradient Boosting Survival, and Kernel Survival Support Vector Machine, were also tested. Performance evaluation, using specific survival analysis metrics, showed that Random Survival Forest outperformed the other models. The study also highlighted the importance of data preprocessing and the application of feature selection methods, which were crucial in identifying relevant prognostic variables for clinical practice. Additionally, the impacts of this computational approach on clinical decision-making were discussed. The results underscore the potential of computational models in predicting the survival of female BC patients, and this research contributes to the advancement of computational oncology, with the potential to improve prognoses and patients' quality of life.

Keywords: Survival Analysis. Machine Learning. Breast Neoplasm.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ACC	*Accuracy*
AdaBoost	*Adaptive Boosting*
AJCC	*American Joint Committee on Cancer*
AM	Aprendizado de Máquina
ANN	*Artificial Neural Networks*
AUC	*Area Under the Curve*
BD	Banco de Dados
BS	*Brier Score*
C	Comparação
CC	*Confusion Matrix*
CDI	Carcinoma Ductal Infiltrante
CI	Coeficiente de Incerteza
CLI	Carcinoma Lobular Infiltrante
CM	Câncer de Mama
CP	Coeficiente de *Pearson*
CI5	*Cancer Incidence in Five Continents*
CID-10	Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde 10ª Revisão
C-Index	*Concordance Index*
CPH	*Cox Proportional-Hazards*
CPH-EN	CPH Penalizado *Elastic Net*
CPH-L	CPH Penalizado *Lasso*
CPH-R	CPH Penalizado *Ridge*
CPHM	Modificação do CPH
cTNM	Estadiamento Clínico do TNM
DATASUS	Departamento de Informática do Sistema Único de Saúde
DBSCAN	*Density-Based Spatial Clustering of Applications with Noise*
DCA	Análise de Curvas de Decisão
DCNT	Doenças Crônicas Não Transmissíveis
DP	Desvio Padrão
DT	*Decision Trees*
EACCD	*Ensemble Algorithm for Clustering Cancer Data*
EL	*Ensemble Learning*
EUA	Estados Unidos da América
F1	*F1-Score*
FA	Frequência Absoluta
FR	Frequência Relativa
GCO	*Global Cancer Observatory*
GB	*Gradient Boosting*
GBC	*Gradient Boosting Classifier*

GBDT	Gradient Boosting Decision Tree
GBS	*Gradient Boosting Survival*
GICR	*Global Initiative for Cancer Registry Development*
GLM	*Generalized Linear Model*
HER2	Receptor do Fator de Crescimento Humano Epidérmico-2
HER2-	HER2 Negativo (-)
HER2+	HER2 Positivo (+)
I	Intervenção
IA	Inteligência Artificial
IARC	*International Agency for Research on Cancer*
IBS	*Integrated Brier Score*
IDH	Índice de Desenvolvimento Humano
INCA	Instituto Nacional de Câncer
Index-C	Índice de Concordância de *Harrell*
K-Fold	*K-Fold Cross-Validation*
KNN	*K-Nearest Neighbor*
K-Statistic	*Kappa Statistic*
LightGBM	*Light Gradient Boosting Machine*
LIME	*Local Interpretable Model-agnostic Explanations*
LR	*Logistic Regression*
M	Presença de Metástases
MAE	*Mean Absolute Error*
MCC	*Matthews Correlation Coefficient*
MeSH	*Medical Subject Headings*
MLP	*Multilayer Perceptron*
MLP-ANN	*MLP-based ANN*
MO	Metástase Óssea
MS	Ministéro da Saúde
MTLSA	*Multi-Task Learning Model for Survival Analysis*
N	Comprometimento dos Linfonodos
NB	*Naïve Bayes*
NM	Não Metastático
NPV	*Negative Predictive Value*
O	Desfecho (*Outcomes*)
OMS	Organização Mundial da Saúde
P	População
PICO	Acrônimo que representa os elementos essenciais de uma pergunta norteadora de pesquisa
PPV	*Positive Predictive Value*
PREC	*Precision*
PRISMA	*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*
pTNM	Estadiamento Patológico do TNM

RBF	*Radial Basis Function Kernel*
RCBP	Registros de Câncer de Base Populacional
RE	Receptor de Estrogênio
RE-	RE Negativo (-)
RE+	RE Positivo (+)
REC	*Recall*
RF	*Random Forest*
RHC	Registros Hospitalares de Câncer
RIPPER	*Repeated Incremental Pruning to Produce Error Reduction*
RMSE	*Root Mean Squared Error*
ROC	*Receiver Operating Characteristic*
RP	Receptor de Progesterona
RP-	RP Negativo (-)
RP+	RP Positivo (+)
RS	Revisão Sistemática
RSF	*Random Survival Forest*
SBOC	Sociedade Brasileira de Oncologia Clínica
SEER	*Surveillance, Epidemiology, and End Results*
SEN	*Sensitivity*
SGB	*Stochastic Gradient Boosting*
SHAP	*SHapley Additive exPlanations*
SIHSUS	Morbidade Hospitalar do SUS
SIM	Sistema Nacional de Mortalidade
SISCAN	Sistema de Informação do Câncer
SISCOLO	Sistema de Informação do Câncer do Colo do Útero
SISMAMA	Sistema de Informação do Câncer de Mama
SOM	*Self-Organizing Map*
SPEC	*Specificity*
KSSVM	*Kernel Survival Support Vector Machine*
SSVM	*Survival Support Vector Machine*
Stan	*Bayesian estimation regression model*
Student-T	*Student-T test* com *Bonferroni correction*
SVM	*Support Vector Machine*
T	Tamanho do Tumor
TNM	Classificação de Estadiamento do Câncer com base em três parâmetros: T, N e M
UFJF	Universidade Federal de Juiz de Fora
VC	*Voting Classifier*
Vmin	Valor Mínimo
Vmax	Valor Máximo
XGBoost	*Extreme Gradient Boosting*

LISTA DE SÍMBOLOS

$\mathbb{R}$	Conjunto dos números Reais
$\forall$	Para todo
$\in$	Pertence
$\sum$	Somatório
$\prod$	Produtório
$\wedge$	Produto exterior
$\mid \beta_i \mid$	Valor absoluto dos coeficientes β_i
$\parallel w \parallel^2$	Norma $l2$ do vetor w
i	Número de interações
j	Número de interações
α	Parâmetro de regularização do CPH penalizados
β	Coeficiente fixo do CPH
β_i	Coeficientes estimados do CPH
δ_i	Indicador de eventos do CPH para sobrevida
ϵ	Intervalo de confiança entre as métricas C-Index
$\hat{\epsilon}$	Intervalo de confiança entre as métricas C-Index com estimação generalizada
γ_i	Coeficientes de expansão de GB e GBS
ι	Função de similaridade Kernel
κ	Similaridade entre as variáveis de entrada
σ	Parâmetro livre da largura Kernel
λ	Número de regularizações do SSVM
π_C	Probabilidade de concordância
π_D	Probabilidade de discordância
$\hat{\pi}(t \mid X_i)$	Probabilidade prevista de um indivíduo i permanecer livre de eventos até o tempo t
θ_i	Parâmetros de aprendizado do GB
ρ	Taxa de aprendizado do GB
τ	Ponto de tempo pré-definido
ξ	Variável de folga do SVM
A	Número de árvores preditoras da RF
a_i	Árvores preditoras da RF
b	Termo de polarização do SVM
C	Parâmetro de regularização (custo) do SVM
C_H	C-Index proposto por Harrel (1996)
$\hat{C}_H$	Estimação generalidada do C-Index proposto por Harrel (1996)
C_P	C-Index proposto por Pencina e D'Agostino (2004)
$\hat{C}_P$	Estimação generalidada do C-Index proposto por Pencina e D'Agostino (2004)
d_i	Número de eventos ocorridos (óbitos)

E_ϵ	Estimador de distribuição assintótica do C-Index
exp	Função exponencial natural (base é o número de Euler)
$F_{GB}(x)$	Função aditiva do GB
$F_{GBS}(X)$	Função aditiva do GBS
$\hat{F}_{RF}(x)$	Função de regressão da RF
$F_{SSVM}(W)$	Função objetivo o do SSVM
$F_{SVM}(x)$	Função do hiperplano de separação do SVM
$g(x, \theta_i$	Função do aprendizado de base do GB
$g(X, \theta_i$	Função do aprendizado de base do GBS
$G\hat{(Y)}$	Estimador de *Kaplan-Meier* para o C-Index
$h(t)$	Função de risco
$h\hat{(t)}$	Estimador de *Nelson-Aalen* para a função de sobrevida
$h_0(t)$	Função de risco de referência
I	Função indicadora de concordância
K	Número de tempos únicos do evento de interesse para sobrevida
$L(\beta)$	Função de verossimilhança parcial do CPH para sobrevida
M	Número total de interações de aprendizado do GB e GBS
n	Número da amostra
p	Número da covariáveis
P	Probabilidade
q	Parâmetro de regularização do SSVM
r	Hiperplanos discriminantes paralelos do SVM
$S(t)$	Função de sobrevida
$S\hat{(t)}$	Estimador de *Kaplan-Meier* para função de sobrevida
t_i	Tempos de observação da sobrevida
T	Tempo específico para a probabilidade de sobrevida
x_i	Vetor de covariáveis de entrada
X_i	Vetor de covariáveis (características clínicas) do paciente i
y_i	Vetor de saída
Y_i	Vetor do último tempo t observado do paciente i
$\hat{Y}$	Probabilidade prevista de sobrevida
w	Vetor de pesos perpendicular ao hiperplano de separação do SVM
$W(t)$	Função de ponderação do IBS

SUMÁRIO

1 INTRODUÇÃO

As Doenças Crônicas Não Transmissíveis (DCNT) desempenham um papel de destaque entre as principais causas de mortalidade em todo o mundo (BRAY et al., 2024). Dentro desse grupo, o câncer, uma das DCNT mais expressivas, se destaca como a principal causa de óbito. O câncer é caracterizado pela desregulação do crescimento celular, resistência à apoptose (morte celular programada) e a capacidade de invadir tecidos adjacentes, o câncer pode levar à formação de metástases, que representam sua forma mais agressiva e a principal causa de morte associada à doença (BRAY et al., 2018; FERLAY et al., 2021; WHO, 2022).

Embora o câncer seja uma doença desafiadora, oferece perspectivas de cura quando diagnosticado precocemente e tratado de maneira adequada. As neoplasias, que afetam uma em cada cinco pessoas ao longo da vida, representam um problema significativo para a saúde global. No século XXI, a prevenção do câncer tornou-se uma prioridade crescente, capaz de reduzir até 40% dos casos por meio de medidas preventivas primárias e minimizar a mortalidade por meio de diagnósticos precoces (IARC, 2023; FERLAY et al., 2021; WHO, 2022).

Projeções indicam que o impacto global do câncer poderá alcançar 35 milhões de casos até 2050, um aumento de 77% em relação a 2022. Esse crescimento é impulsionado por mudanças demográficas e pelo aumento dos fatores de risco associados à globalização e ao crescimento econômico (BRAY et al., 2024; SUNG et al., 2021). Em 2022, a *International Agency for Research on Cancer* (IARC) e a Organização Mundial da Saúde (OMS) estimaram aproximadamente 20 milhões de novos casos e 10 milhões de mortes atribuídas ao câncer. A prevalência acumulada de cinco anos foi de 53,5 milhões de casos, incluindo câncer de pele não melanoma (WHO, 2024; IARC, 2024).

Tanto globalmente quanto no Brasil, o controle do câncer é um processo contínuo que abrange desde a prevenção até os cuidados paliativos, exigindo planejamento e monitoramento rigorosos (SANTOS et al., 2023; BRASIL, 2022). Para 2023, o Instituto Nacional de Câncer (INCA) estimou cerca de 704.080 novos casos de câncer, destacando a magnitude e o impacto da doença no país. Os dados de mortalidade, extraídos do Sistema Nacional de Mortalidade (SIM), apontaram 244.009 óbitos em 2022 relacionados ao câncer (INCA, 2023; BRASIL, 2022; SANTOS et al., 2023; BRASIL, 2022).

Entre os tipos de neoplasias, o CM feminino é o mais diagnosticado, representando a forma mais prevalente tanto no Brasil (com uma incidência de 30,1%) quanto no cenário global (com 23,8%) (IARC, 2024; WHO, 2024; INCA, 2023). A pandemia de COVID-19, entretanto, trouxe incertezas quanto às estimativas de câncer, afetando tanto o diagnóstico quanto os tratamentos devido às restrições impostas. Esses efeitos ainda não foram plenamente considerados nas projeções atuais (BRASIL, 2022; IARC, 2023; SANTOS et

al., 2023; SIEGEL et al., 2023).

Esse cenário pandêmico também ressaltou a relevância da Inteligência Artificial (IA) na área da saúde, acelerando seu desenvolvimento e aplicação. A IA tem demonstrado seu potencial para otimizar a alocação de recursos, personalizar tratamentos e guiar políticas de saúde mais eficazes (INCA, 2023; SIEGEL et al., 2023; SUNG et al., 2021; SCHAAR et al., 2021). A crescente expansão dos dados médicos, impulsionada pela tecnologia e pelos métodos de Aprendizado de Máquina (AM), tornou a análise desses dados essencial para a pesquisa biomédica. Apesar dos desafios em compreender completamente os algoritmos, esses métodos têm permitido a identificação de padrões e relações até então desconhecidos (DENG et al., 2021; LI et al., 2021; SCHAAR et al., 2021).

A avaliação da sobrevida em DCNT, como o CM femino, é um fator fundamental para melhorar a qualidade de vida e aumentar a expectativa de vida após o diagnóstico. A pesquisa atual busca não apenas melhorar as taxas de sobrevivência, mas também reduzir o sofrimento dos pacientes (LI et al., 2021; MIN et al., 2021). Nesse contexto, os modelos prognósticos desempenham um papel crucial ao correlacionar fatores de risco com a probabilidade de resposta clínica. A aplicação de ciência de dados e modelos computacionais possibilita predições mais acuradas para diagnóstico, tratamento e prognóstico do câncer, viabilizando um planejamento terapêutico personalizado e contribuindo para melhores desfechos clínicos e uma maior taxa de sobrevida (CARVALHO et al., 2023; LI et al., 2021; MIN et al., 2021; MONCADA-TORRES et al., 2021; XIAO et al., 2022).

1.1 MOTIVAÇÃO

O CM é uma das principais causas de mortalidade entre mulheres em todo o mundo, independentemente de fatores econômicos e sociais (BRAY et al., 2024; IARC, 2024). Esse tipo de tumor maligno invasivo afeta predominantemente mulheres e apresenta alta heterogeneidade em suas características biológicas, resposta ao tratamento e prognóstico. A complexidade e variabilidade do CM, associadas à necessidade crescente de prognósticos individualizados, representam desafios substanciais para a saúde pública voltada às mulheres (BRASIL, 2014; TENG et al., 2019; OKAGBUE et al., 2021).

Para mitigar os riscos associados ao CM, estratégias de triagem e programas de conscientização são fundamentais. Intervenções terapêuticas, como cirurgias, radioterapia e tratamentos sistêmicos, desempenham papéis essenciais no combate à doença, sendo a detecção precoce um fator crucial para aumentar as taxas de sobrevida (HAQUE et al., 2022; OKAGBUE et al., 2021). Além disso, o avanço na predição precisa da sobrevida dos pacientes tem se tornado um campo promissor, com grande potencial para apoiar a tomada de decisões clínicas personalizadas, otimizar o uso de recursos financeiros e beneficiar pacientes, profissionais de saúde e a formulação de políticas públicas (LI et al., 2021; MIN et al., 2021; OKAGBUE et al., 2021; TENG et al., 2019).

A taxa de sobrevida é um importante indicador da eficácia do cuidado terapêutico em câncer. Estas probabilidades refletem e avaliam os avanços diagnósticos e terapêuticos, assim como a eficiência global do sistema de saúde (ALLEMANI et al., 2018; BUSTAMANTE-TEIXEIRA et al., 2002). Como por exemplo, o estudo CONCORD-3, que monitora a sobrevida global do câncer em um período de 15 anos (2000-2014), revela que, para mulheres com CM, a sobrevida de cinco anos é de cerca de 90% em países desenvolvidos. Contudo, essa taxa pode ser tão baixa quanto 40% em alguns países em desenvolvimento, evidenciando disparidades internacionais consideráveis e a importância de melhorar a expectativa de sobrevida e a qualidade do atendimento clínico no CM (ALLEMANI et al., 2018).

Na oncologia, uma ampla gama de modelos computacionais vem sendo aplicada, esses modelos são derivados de estruturas matemáticas, posteriormente discretizados e implementados para predizer o curso clínico-terapêutico (KöHN-LUQUE et al., 2020; LAI et al., 2019; LI et al., 2021; NAVE, 2020). Especificamente no contexto do CM, modelos que utilizam dados clínicos frequentemente registrados na prática médica têm o potencial de aprimorar o diagnóstico, tratamento, acompanhamento e predição de sobrevida, resultando em uma significativa redução no sofrimento dos pacientes e em uma melhoria na expectativa e qualidade de vida (LI et al., 2021; MCKENNA et al., 2018).

Portanto, a integração entre Ciência de Dados e a expertise de profissionais da saúde mostra-se, portanto, essencial para o desenvolvimento de tecnologias, como os métodos de AM, que aprimoram a precisão das predições de sobrevida, como de diagnósticos precoces e da eficácia da resposta terapêutica (LI et al., 2021; OKAGBUE et al., 2021). Para sua efetividade e aplicabilidade na prática clínica, necessita-se de repositórios de dados clínicos robustos para validação e aprimoramento contínuo destes modelos. Embora significativos avanços tenham sido alcançados, ainda há espaço para a inclusão e o refinamento de modelos adicionais que potencializem o uso da IA na melhoria do diagnóstico por imagem, do planejamento de tratamento e do prognóstico em saúde (LI et al., 2021; MIN et al., 2021; SEDIGHI-MAMAN; MONDELLO, 2021).

Além dos benefícios diretos à saúde pública, abordar a carga do câncer em mulheres também reconhece e enfatiza o papel fundamental delas como participantes ativas na sociedade, tanto no âmbito social quanto econômico. Além disso, valoriza sua significativa contribuição como cuidadoras familiares e aborda a complexa questão da desigualdade de gênero (TORRE et al., 2017).

1.2 JUSTIFICATIVA

O CM feminino foi o tipo mais incidente globalmente em 2020, com 2,26 milhões de casos registrados, superando o câncer de pulmão, que teve 2,21 milhões de casos (FERLAY et al., 2021; SIEGEL et al., 2023; WHO, 2022). No entanto, em 2022, apesar de um

aumento no número de novos casos, que ultrapassou 2,31 milhões, o CM passou a ser o segundo câncer mais comum, atrás do câncer de pulmão, que registrou 2,48 milhões de novos casos (IARC, 2024; WHO, 2024). Vale destacar que esta doença afeta predominantemente mulheres e é a principal causa de mortalidade relacionada ao câncer no sexo feminino. Segundo as estimativas da OMS e análises da IARC em 2022, aproximadamente 2,30 milhões de novos casos e 665.684 mortes globais foram atribuídas ao CM entre mulheres (IARC, 2024; BRAY et al., 2024; WHO, 2024). No Brasil, os números refletem essa tendência global, com o INCA estimando cerca de 73.610 novos casos em 2023 e o SIM registrando 19.363 óbitos relacionados em 2022 (INCA, 2023; BRASIL, 2021b).

Dado o impacto global e local do CM, sua análise torna-se uma questão crítica de saúde pública, sendo essencial a investigação detalhada de dados clínicos para melhorar as perspectivas de sobrevida das pacientes (LI et al., 2021). Para superar as limitações dos modelos estatísticos tradicionais, como o modelo estatístico tradicional, *Cox Proportional-Hazards* (CPH), o uso de técnicas de AM tem mostrado grande eficácia, proporcionando maior flexibilidade e desempenho superior na análise de sobrevida, principalmente devido às questões de dimensionalidade e não linearidades dos dados (MONCADA-TORRES et al., 2021; LIU et al., 2020; FANIZZI et al., 2023). As técnicas de IA complementam esses métodos tradicionais, visando aprimorar o entendimento do curso clínico e terapêutico do CM. Contudo, ainda há uma limitação importante: como a presença de lacunas nos dados. A imputação múltipla tem se mostrado uma estratégia eficaz para lidar com dados ausentes, otimizando o uso dos métodos de AM na predição da sobrevida de pacientes com CM (LIU et al., 2020; MAABREH et al., 2021; AFSHAR et al., 2021).

No campo da oncologia, um dos principais desafios ao trabalhar com grandes volumes de dados clínicos é a ocorrência de dados ausentes, o que dificulta tanto a análise descritiva quanto a modelagem preditiva. O pré-processamento dos dados, essencial para o tratamento de dados faltantes, torna-se especialmente desafiador devido às imprecisões, inconsistências e valores ausentes, que afetam negativamente o desempenho dos modelos preditivos e comprometem a precisão dos estudos (CURIOSO et al., 2023; KALAFI et al., 2019; LIU et al., 2020). Além das melhorias no pré-processamento, a seleção adequada de variáveis é crucial para a construção de modelos preditivos mais eficazes, utilizando apenas as variáveis mais relevantes. Isso não apenas aprimora a precisão das predições, mas também reduz a complexidade e melhora a interpretabilidade dos modelos, que podem se tornar difíceis de entender quando envolvem um grande número de variáveis, muitas vezes irrelevantes para os resultados desejados (LIU et al., 2020; KALAFI et al., 2019; XIAO et al., 2022).

O controle do CM feminino exige uma abordagem integrada que inclua prevenção, triagem, diagnóstico e tratamento precoce. A adaptação dessas intervenções às necessidades locais é decisiva para reduzir o impacto do câncer e melhorar a saúde das mulheres em todo o mundo. Além disso, a pesquisa atual se concentra na predisão da sobrevida para o

CM, o que é fundamental para orientar decisões clínicas e estratégias terapêuticas mais eficazes (TORRE et al., 2017; LI et al., 2021).

1.3 OBJETIVOS

1.3.1 OBJETIVO GERAL

O objetivo principal deste estudo é identificar e avaliar os atributos prognósticos via modelos computacionais para a predição da sobrevida em pacientes com Câncer de Mama, aplicando Métodos de Aprendizagem de Máquina a dados clínicos.

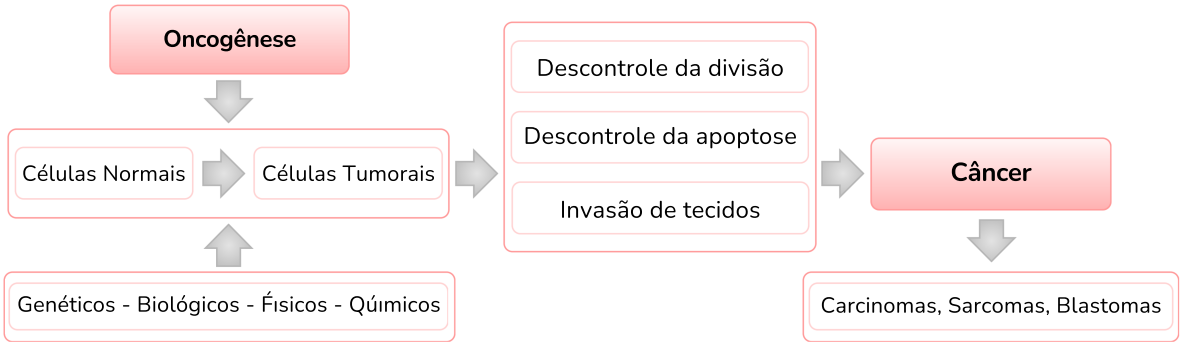
1.3.2 OBJETIVOS ESPECÍFICOS

1. Realizar uma Revisão Sistemática (RS) para identificar métodos de AM para predição da sobrevida para o CM validados com dados clínicos.
2. Promover o pré-processamento e análise dos dados do Banco de Dados (BD) clínicos.
3. Implementar, simular e validar os modelos selecionados, avaliando o desempenho.
4. Aplicar seleção de atributos visando aprimorar os resultados encontrados.
5. Comparar os resultados alcançados com os da literatura.

2 REFERENCIAL TEÓRICO

A oncogênese, processo pelo qual células normais transformam-se progressivamente em células tumorais, é um fenômeno complexo resultante de alterações genéticas e epigenéticas, conforme ilustrado na Figura 1. Esse processo envolve tanto fatores hereditários, como histórico familiar, quanto fatores ambientais, biológicos (infecções virais e bacterianas), físicos (exposição à radiação e traumas) e químicos (exposição a carcinógenos e poluição (CARVALHO, 2016; WEINBERG; WEINBERG, 2006). Os tumores, formados a partir desse processo, são classificados em benignos e malignos: os benignos possuem crescimento lento e bem diferenciado, enquanto os malignos exibem crescimento rápido, capacidade invasiva e potencial de metástase. O termo "câncer"refere-se geralmente aos tumores malignos, os quais podem ser classificados conforme sua origem embrionária: carcinomas (de origem epitelial), sarcomas (de tecidos conjuntivos) e blastomas (de células imaturas) (HAQUE et al., 2022; WEINBERG; WEINBERG, 2006; WHO, 2022).

Figura 1 – Fluxograma do processo de Oncogênese.



Fonte: Adaptado de Carvalho (2016).

Na oncologia, a implementação de protocolos rigorosos é essencial para orientar as etapas de prevenção, detecção, tratamento ativo e cuidados paliativos. Entre os tipos de câncer, o CM é o mais comumente diagnosticado em mulheres em todo o mundo e apresenta uma crescente incidência, atribuída a fatores genéticos e mudanças nos estilos de vida modernos (BRASIL, 2014; BRAY et al., 2024). A detecção precoce é fundamental: diagnósticos em estágios iniciais aumentam significativamente as chances de cura, enquanto diagnósticos em estágios avançados comprometem as taxas de sobrevida. O curso clínico do CM feminino é influenciado por fatores como idade, histórico familiar, características genéticas e fatores reprodutivos, conforme será descrito na Subseção 2.1.2 (BRASIL, 2014; CINTRA, 2012; LI et al., 2021).

O CM masculino, embora raro, é frequentemente diagnosticado em estágios avançados, o que reduz as chances de um prognóstico positivo. Os fatores de risco para o CM em homens incluem idade avançada, mutações genéticas, condições testiculares, obesidade e

exposição à radiação. Clinicamente, manifesta-se como nódulos indolores e, em casos mais avançados, com ulcerações e retrações cutâneas. A mamografia é particularmente eficaz para detectar lesões subareolares em homens e representa o principal método diagnóstico. Apesar da similaridade entre os tratamentos para CM masculino e feminino, a detecção tardia em homens contribui para menores taxas de sobrevida, destacando a necessidade de ampliar a conscientização sobre a doença no público masculino (GOMES et al., 2022).

Considerando que o CM é a forma mais comum de neoplasia entre mulheres em nível global (BRAY et al., 2024), observa-se um significativo avanço tecnológico na coleta e análise de dados clínicos dessa patologia. Essa evolução possibilita o desenvolvimento de modelos computacionais para a predição da sobrevida, utilizando métodos de AM para tarefas de classificação, predição e estimação. Tais abordagens têm demonstrado grande potencial para auxiliar o planejamento da prática clínica em pacientes com CM, oferecendo perspectivas promissoras na melhoria do atendimento (LI et al., 2021; DENG et al., 2021). Dada a relevância do CM feminino, este trabalho se concentra em investigar as características biológicas, clínicas e terapêuticas dessa patologia, com a realização de estudos aprofundados sobre o tema.

A análise da sobrevida, como nos casos de CM, é fundamental para compreender os fatores críticos associados à progressão da doença, além de contribuir para uma condução clínica mais eficaz e a melhoria da qualidade de vida dos pacientes (ALLEMANI et al., 2018; BUSTAMANTE-TEIXEIRA et al., 2002; LI et al., 2021). Nos últimos anos, os métodos de AM têm ganhado destaque na predição de desfechos em CM, superando abordagens estatísticas como o modelo de Regressão de Cox. Esses métodos se diferenciam por sua capacidade de processar grandes volumes de dados com alta dimensionalidade, lidar com informações censuradas e modelar relações não-lineares, eliminando suposições restritivas. Além disso, são capazes de identificar padrões ocultos que podem aprimorar significativamente os prognósticos (LIU et al., 2020; COX, 1972; MONCADA-TORRES et al., 2021). Essas abordagens são particularmente valiosas na identificação de fatores de risco específicos e na definição de estratégias terapêuticas mais precisas. O fortalecimento da confiabilidade desses métodos é crucial para a prática clínica oncológica, promovendo decisões mais informadas e tratamentos mais eficazes para pacientes com CM (LI et al., 2021). Estudos recentes têm corroborado esses avanços, evidenciando o potencial do AM na melhoria dos desfechos clínicos (CARVALHO et al., 2023; FANIZZI et al., 2023; LIU et al., 2020; MONCADA-TORRES et al., 2021; PINHEIRO et al., 2022; TIZI; BERRADO, 2023; XIAO et al., 2022).

Nesse contexto, a avaliação da sobrevida desempenha um papel vital na orientação de decisões clínicas e terapêuticas, contribuindo para a promoção da saúde das mulheres (CARVALHO et al., 2019; LI et al., 2021). Paralelamente, o desenvolvimento de métodos de predição prognóstica é igualmente importante, pois atua na redução da mortalidade e na melhoria da qualidade de vida das pacientes, ao considerar uma ampla gama de fatores de

risco, incluindo o estadiamento da doença, bem como características moleculares, clínicas, terapêuticas, reprodutivas, hormonais e comportamentais (RUBINGER et al., 2022; SIEGEL et al., 2023). A aplicação de modelos computacionais possibilita análises mais detalhadas e precisas, fornecendo *insights* que podem ser decisivos para o desenvolvimento de estratégias de tratamento mais eficazes e personalizadas (LI et al., 2021).

2.1 DESAFIOS E INOVAÇÕES NA ONCOLOGIA

O controle do câncer é atualmente compreendido como um contínuo de ações que começa pela gestão de fatores de risco e segue com a detecção precoce da doença, estendendo-se aos cuidados paliativos. Estes cuidados abrangem todas as etapas da jornada do paciente, desde o diagnóstico e tratamento até o acompanhamento durante a sobrevivência e, quando necessário, cuidados de fim de vida para aqueles que não obtêm cura ou controle da doença. Para assegurar a integralidade do cuidado em todas essas fases, torna-se essencial o planejamento meticuloso, uma organização eficiente dos serviços de saúde e a constante monitorização das ações de controle e promoção da saúde, viabilizada por indicadores específicos. Embora existam desafios persistentes nessas estratégias de controle, as estimativas atualizadas e os dados epidemiológicos são fundamentais para promover a avaliação e o aperfeiçoamento contínuos dessas iniciativas (INCA, 2023).

Apesar dos avanços, a luta contra o câncer ainda enfrenta barreiras substanciais, especialmente em países de baixa e média renda, onde o acesso a diagnósticos e tratamentos adequados é frequentemente limitado. Nesse cenário, estratégias de saúde pública e iniciativas de conscientização desempenham papéis cruciais no enfrentamento dessa doença devastadora em escala global, destacando a importância de políticas de saúde adaptadas às necessidades locais e globais (OPAS, 2020).

2.1.1 A IMPORTÂNCIA DA VIGILÂNCIA ONCOLÓGICA E DADOS EPIDEMIOLÓGICOS

O câncer é um dos maiores desafios para a saúde pública em escala global, sendo uma das principais causas de mortalidade e uma barreira significativa para o aumento da expectativa de vida em todo o mundo (BRAY et al., 2018; BRAY et al., 2024). O impacto crescente da incidência e mortalidade por câncer é particularmente evidente em países emergentes, que estão passando por uma transição epidemiológica similar aos países desenvolvidos. Este aumento é impulsionado por uma série de fatores de risco associados ao desenvolvimento socioeconômico dessas nações, destacando a urgência de investimentos adequados em recursos para tratamento, controle e prevenção primária da doença. Além disso, o aumento contínuo das taxas de mortalidade relacionadas ao câncer, quando comparado a outras condições, como acidente vascular cerebral e doença cardíaca coronária, o câncer continua a apresentar taxas crescentes de mortalidade, refletindo

diretamente as mudanças demográficas, envelhecimento populacional e transformações nos fatores de risco associados ao desenvolvimento socioeconômico globais (BRAY et al., 2018; SUNG et al., 2021; SIEGEL et al., 2023).

2.1.1.1 Vigilância Oncológica

A Vigilância Oncológica desempenha um papel fundamental no desenvolvimento de estratégias de controle do câncer, baseando-se em dados provenientes dos Registros de Câncer de Base Populacional (RCBP), que são coletados individualmente por cada país. A colaboração internacional também é crucial nesse contexto, como exemplificado pelo projeto Cancer *Cancer Incidence in Five Continents* (CI5), que, a cada cinco anos, oferece estatísticas comparáveis de alta qualidade sobre a incidência de câncer em todo o mundo (FERLAY et al., 2019; IARC, 2023; IARC, 2021). Outra iniciativa significativa é o *Global Initiative for Cancer Registry Development* (GICR), que visa enfrentar as desigualdades socioeconômicas e auxiliar na prevenção do câncer, bem como na melhoria dos resultados em países com menos recursos. A parceria entre o GICR e a IARC fortalece a capacidade de coleta, análise e disseminação de dados sobre o câncer, contribuindo diretamente para a missão de controle global da doença (IARC, 2024; SUNG et al., 2021; FERLAY et al., 2018; FERLAY et al., 2021).

Uma ferramenta essencial para a avaliação global do câncer é o GLOBOCAN, que fornece informações fundamentais para o planejamento de políticas de saúde e prevenção eficazes em nível mundial. As atualizações periódicas deste Banco de Dados (BD) são cruciais para garantir estimativas precisas sobre a incidência e mortalidade do câncer em escala global (FERLAY et al., 2019; IARC, 2024). O GLOBOCAN abrange 36 tipos de câncer, conforme a Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde 10ª Revisão (CID-10), fornecendo dados para 185 países ou territórios, segmentados por sexo e 18 faixas etárias. Esses dados destacam a necessidade de adaptar os esforços de controle do câncer a níveis regionais e nacionais, levando em consideração os padrões específicos de cada região (IARC, 2024; FERLAY et al., 2021; SAÚDE, 1994; SUNG et al., 2021).

As estimativas do GLOBOCAN, compiladas pela IARC, estão disponíveis por meio do *Global Cancer Observatory* (GCO), que oferece dados confiáveis sobre a incidência e mortalidade do câncer. Esses dados são coletados a partir dos RCBP e do BD do CI5, enquanto os dados de óbito são registrados em um BD de mortalidade por câncer compilado pela OMS (IARC, 2024; IARC, 2021; WHO, 2022). O GCO fornece uma visão abrangente da carga global do câncer, segmentada por idade, sexo e Índice de Desenvolvimento Humano (IDH), permitindo comparações entre diferentes populações e refletindo os níveis socioeconômicos das regiões analisadas (FERLAY et al., 2021; SUNG et al., 2021; IARC, 2021).

No Brasil, as informações sobre câncer também são obtidas por meio dos RCBP, apoiados pelos Registros Hospitalares de Câncer (RHC) e o SIM. Esses dados são essenciais para o desenvolvimento de ações de prevenção e controle do câncer no país. O INCA e o Departamento de Informática do Sistema Único de Saúde (DATASUS) desempenham papel fundamental na coleta e disponibilização dessas informações (INCA, 2023; BRASIL, 2021a). O INCA consolida os dados dos RCBP e dos RHC, sendo que os RCBP visam informar sobre a incidência de câncer, sua distribuição geográfica e tendências temporais. Já os dados dos RHC fornecem informações clínicas detalhadas, como sexo, idade, topografia, morfologia, extensão e meio de diagnóstico, sendo essenciais para o acompanhamento da qualidade do tratamento de câncer nos hospitais (INCA, 2023; INCA, 2022a; INCA, 2022b).

O DATASUS, por meio do TABNET, disponibiliza informações que orientam o planejamento baseado em evidências para programas de saúde, incluindo registros de mortalidade, morbidade, nascidos vivos e dados epidemiológicos. Esses dados categorizados são fundamentais para entender os determinantes de saúde da população e a eficácia das intervenções de saúde pública. No contexto do câncer, como ilustrado na Figura 2, o TABNET oferece informações gerais sobre morbidade pelo Sistema de Informação Hospilar do SUS (SIH/SUS)e Sistema de Informação Ambulatorial (SIA-SUS), e especificamente do câncer pelo Sistema de Informação do Câncer de Colo de Útero e Mama (SISCOLO/SISMAMA), Sistema de Informação do Câncer (SISCAN), tempo até o início do tratamento oncológico – PAINEL ONCOLÓGICO, entre outros. Contudo, esses dados não contêm informações detalhadas sobre o curso clínico e terapêutico e a sobrevivência dos pacientes (INCA, 2022a; INCA, 2022b; BRASIL, 2022; BRASIL, 2021a).

Figura 2 – Informações de Câncer disponibilizadas no TABNET.

Fonte: Adaptado do TABNET, Brasil (2021).

As informações nacionais sobre câncer são essenciais para o estabelecimento de ações estratégicas de Vigilância Oncológica, que é uma peça-chave na para a promoção da saúde e controle do câncer. Esses dados, abrangendo uma variedade de tipos de câncer e variações temporais, desempenham um papel fundamental na formulação de políticas públicas e na alocação eficaz de recursos para o combate à doença. Os RCBP, RHC e o SIM fornecem dados abrangentes que são essenciais para enfrentar o desafio do câncer

no Brasil, além de oferecerem insights para gestores de saúde e orientações para futuras pesquisas (BRASIL, 2022; INCA, 2023; SANTOS et al., 2023).

2.1.1.2 Estimativas Epidemiológicas

Embora os registros de câncer, tanto em nível global quanto nacional, apresentem algumas limitações, eles fornecem informações essenciais sobre a incidência, prevalência e mortalidade da doença. No caso da mortalidade, as informações sobre as causas de óbito oferecem uma visão detalhada da distribuição dos principais tipos de câncer, além de enriquecerem a coleta contínua de dados a partir dos RCBP e registros de mortalidade. Esses dados são essenciais para compreender a magnitude e o impacto do câncer no Brasil e no mundo. No contexto brasileiro, o aumento da incidência e mortalidade por câncer, fenômeno observado globalmente, é particularmente influenciado pelo envelhecimento da população e pelo crescimento populacional. Esses fatores, combinados com mudanças comportamentais e ambientais, como a mobilidade, estilo de vida, dieta e exposição a poluentes, têm impacto direto na dinâmica do câncer no país (SANTOS et al., 2023; BRASIL, 2022; SIEGEL et al., 2023; WHO, 2022).

A **Incidência do Câncer**, medida pelo número de novos casos diagnosticados em uma região durante um determinado período de tempo, pode ser expressa em valores absolutos ou em taxas por 100.000 pessoas por ano. Esses dados são essenciais para a comparação de riscos entre países e regiões (BRAY et al., 2018; IARC, 2024). Provenientes dos RCBP nacionais ou subnacionais, as taxas de incidência desempenham um papel crucial no controle da doença, mesmo em contextos com recursos limitados. Isso ocorre porque as taxas de incidência estão diretamente relacionadas à capacidade de diagnosticar novos casos, o que, por sua vez, depende do acesso aos serviços de diagnóstico. Estratégias eficazes de detecção precoce aumentam a capacidade de diagnóstico e aprimoram a atenção primária aos serviços oncológicos. Portanto, as taxas de incidência oferecem uma estimativa do risco médio de desenvolvimento da doença, servindo como base para o planejamento de estratégias locais e o desenvolvimento de estimativas em níveis nacional e global (BRAY et al., 2018; FERLAY et al., 2019; IARC, 2024; BRASIL, 2022).

O CM feminino liderou como o tipo mais comum de neoplasia diagnosticada globalmente em 2020, em ambos os sexos, com cerca de 2,3 milhões de novos casos, representando 11,7% de todos os casos de câncer, superando o câncer de pulmão em termos de incidência, que teve 2,21 milhões de novos casos e representou 11,4% de todos os casos. No entanto, em 2022, o câncer de pulmão se tornou o segundo tipo mais comum diagnosticado, e entre os cinco tipos de câncer mais frequentes nesse ano estão o câncer de pulmão (2,48 milhões - 12,4%), CM (2,30 milhões - 11,6%), cólon e reto (1,93 milhão - 9,6%), próstata (1,47 milhão - 7,3%) e estômago (0,97 milhão - 4,8%), excluindo o câncer de pele (não melanoma) (1,24 milhão - 6,2%). Especificamente entre as mulheres, o CM

representa 1 em cada 4 casos de câncer, sendo o tipo mais incidente na grande maioria dos países do GLOBOCAN (IARC, 2024; WHO, 2024; SUNG et al., 2021; SIEGEL et al., 2023; BRAY et al., 2024).

No Brasil, os dados de incidência são obtidos por meio dos 30 RCBP ativos em todo o território nacional. Para 2023, as estimativas indicam 220 mil novos casos de câncer, representando 31,3% do total de neoplasias. Dentre esses, destacam-se o CM (73.610 casos - 10,5%), seguido pelo câncer de próstata (71.730 - 10,2%), cólon e reto (45.630 - 6,5%), pulmão (32.560 - 4,6%) e estômago (21.480 - 3,1%). No que diz respeito à distribuição por sexo, 50,5% dos casos de câncer no Brasil ocorrem em mulheres, sendo o CM o mais diagnosticado em todo o país, correspondendo a 41,9% de todos os casos, excluindo o câncer de pele não melanoma. Na região Sudeste, essa taxa de incidência é ainda mais expressiva, alcançando 52,8% dos casos registrados (INCA, 2023; BRASIL, 2022; SANTOS et al., 2023).

As taxas de incidência do CM estão em ascensão, influenciadas por mudanças no estilo de vida, fatores socioculturais e o processo de urbanização, impulsionado pelo crescimento econômico e pela maior participação das mulheres no mercado de trabalho. Além disso, fatores como a redução na taxa de fertilidade e o aumento da obesidade também contribuem para esse cenário. A implementação de programas de prevenção primária para o CM é de suma importância. Programas focados na redução do excesso de peso, no consumo controlado de álcool, na promoção de atividade física regular e na incentivação da amamentação podem ter um impacto significativo na redução da incidência dessa doença em todo o mundo (BRAY et al., 2024; SUNG et al., 2021; SIEGEL et al., 2023; TORRE et al., 2017).

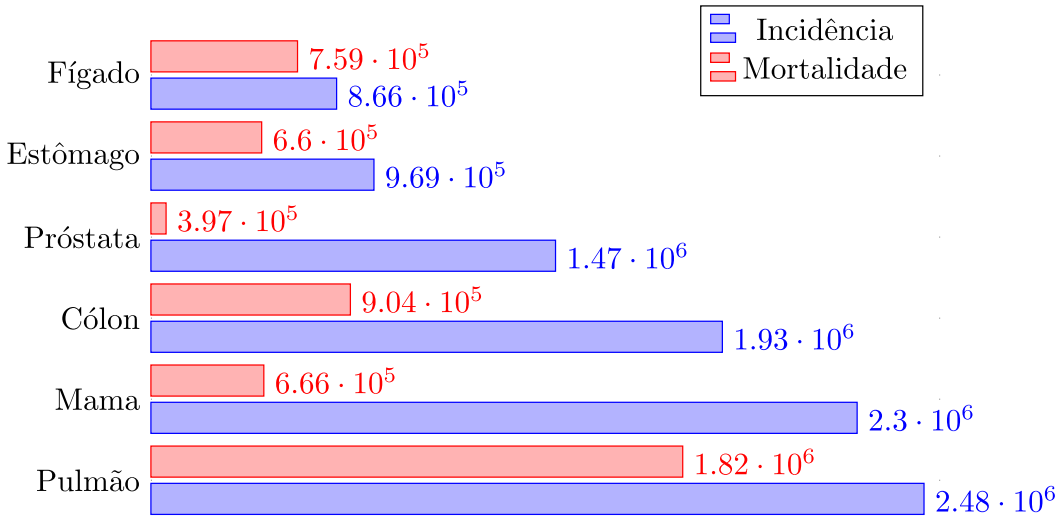
A **Prevalência de Câncer** refere-se ao número de indivíduos diagnosticados com a doença, que permanecem vivos com o câncer por um período específico, sem a cura. Essa métrica é particularmente útil para avaliar a carga do câncer em uma população. No cenário global, a prevalência de cinco anos para o CM, em 2022, foi de 8,18 milhões de casos (15,3%), seguida pelo câncer de cólon e reto (5,77 milhões - 10,8%), próstata (5,03 milhões - 9,4%), pulmão (3,22 milhões - 6,0%) e tireoide (2,91 milhões - 5,4%) (IARC, 2024).

A **Mortalidade por Câncer** refere-se ao número de óbitos causados pela doença em uma região durante um determinado período. As taxas de mortalidade são frequentemente expressas como o número de mortes por 100.000 pessoas por ano (BRAY et al., 2018; IARC, 2023). A qualidade e abrangência desses registros podem variar entre os países, impactando a precisão das informações sobre as causas de óbito. Aproximadamente 1 em cada 5 países possui registros de óbitos confiáveis. As taxas de mortalidade são usadas para estimar o risco da doença em diferentes grupos, mas podem ser menos precisas para cânceres com prognósticos favoráveis, especialmente quando há melhorias na detecção

precoce e no tratamento eficaz, além das disparidades no acesso ao tratamento em países com menos recursos (BRAY et al., 2018; FERLAY et al., 2019; WHO, 2022).

Em 2022, as principais causas de mortalidade por câncer em ambos os sexos foram: câncer de pulmão (1.817.469 óbitos), seguido pelo câncer de cólon e reto (904.019 óbitos), fígado (758.725 óbitos), CM (666.103 óbitos) e estômago (660.175 óbitos) (IARC, 2024; WHO, 2024; SUNG et al., 2021; BRAY et al., 2024). No Brasil, os dados de mortalidade são extraídos dos registros do SIM e do Atlas de Mortalidade do INCA. Em 2022, as principais causas de morte por câncer no país foram: pulmão (29.576 óbitos), cólon e reto (21.255 óbitos), CM (19.363 óbitos), próstata (16.429 óbitos) e estômago (14.340 óbitos) (BRASIL, 2022; INCA, 2022). Dessa forma, o câncer se configura como um dos maiores problemas sociais, de saúde pública e econômicos do século XXI, responsável por quase 16,8% das mortes globais e 22,8% das mortes por DCNT (BRAY et al., 2024).

Figura 3 – Estimativas de 2022 do GLOBOCAN - Câncer Mundial.

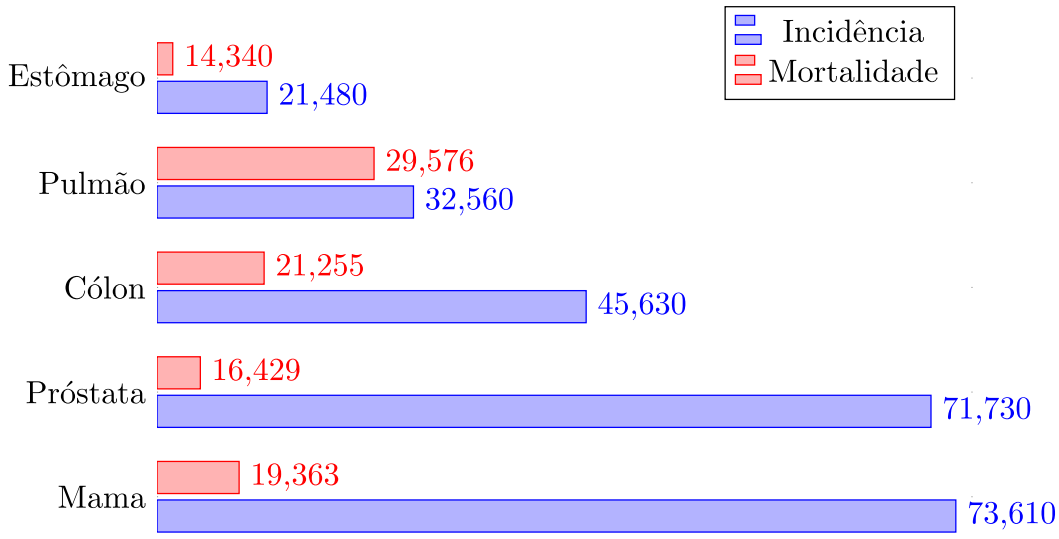


Fonte: Adaptado IARC (2024).

As Figuras 3 e 4 mostram as estimativas globais e nacionais, respectivamente, da incidência e das causas mais frequentes de mortalidade por câncer (BRASIL, 2022; BRASIL, 2022; IARC, 2024). Embora o CM ocupe a quarta posição como causa de morte por câncer globalmente e a terceira no Brasil, quando considerados ambos os sexos, sua taxa de mortalidade pode ser impactada pelas futuras evoluções nas taxas de incidência e prevalência. Entre as mulheres, o CM é responsável por 1 em cada 4 casos de câncer e 1 em cada 6 mortes por câncer, sendo o tipo mais incidente na grande maioria dos países (SUNG et al., 2021; SIEGEL et al., 2023; BRAY et al., 2024).

Esses dados destacam a gravidade do CM no contexto da saúde pública global,

Figura 4 – Estimativas de 2023/2022 do INCA e SIM- Câncer Nacional.



Fonte: Adaptado INCA (2024).

especialmente entre as mulheres, devido às elevadas taxas de incidência, prevalência e mortalidade associadas a essa doença, com variações significativas entre as pacientes (CARVALHO et al., 2023; CINTRA et al., 2012). A complexidade do CM reforça a necessidade de ações contínuas de prevenção, detecção precoce e tratamento eficaz, com o objetivo de reduzir seu impacto na saúde das mulheres (SANTOS et al., 2023; BRASIL, 2022; SIEGEL et al., 2023; WHO, 2022).

Um exemplo importante dessas ações são os programas de rastreamento do CM, cujo principal objetivo é reduzir a mortalidade por meio da detecção precoce e subsequente tratamento. A OMS recomenda triagens bienais para mulheres entre 50 e 69 anos, em áreas com infraestrutura adequada, enquanto a **American Cancer Society** sugere triagens anuais a partir dos 45 anos. No Brasil, as diretrizes sugerem o início do rastreamento aos 40 anos, conforme orientação da Sociedade Brasileira de Oncologia, e aos 50 anos, de acordo com o Ministério da Saúde (MS) (BRASIL, 2014; SUNG et al., 2021).

O enfrentamento do câncer exige esforços colaborativos amplos, envolvendo governos, organizações e indivíduos, a fim de minimizar seu impacto global. No entanto, é essencial lembrar que as estimativas pontuais não substituem a coleta contínua de dados por meio dos RCBP, que desempenham um papel fundamental na avaliação e no monitoramento da doença em nível local. Dessa forma, a colaboração internacional é fundamental para superar os desafios operacionais e garantir a qualidade e comparabilidade dos dados em escala global (FERLAY et al., 2021; SUNG et al., 2021).

As estimativas de câncer no Brasil, publicadas pelo INCA e baseadas na metodologia da IARC/OMS, possibilitam análises abrangentes sobre fatores de risco e disparidades temporais e regionais. Essas informações são valiosas para gestores de saúde, profissionais da área, pesquisadores e a sociedade, destacando a importância de decisões e ações passadas e incentivando melhorias nas estratégias de controle do câncer. Importante ressaltar que os dados apresentados não abordam completamente os impactos da pandemia de COVID-19, que provavelmente causou atrasos no diagnóstico e na notificação de novos casos de câncer (BRASIL, 2022; IARC, 2023; WHO, 2022).

2.1.2 CÂNCER DE MAMA: ABORDAGEM DO CURSO CLÍNICO-TERAPÊUTICO E IMPACTO NA SOBREVIDA

O CM feminino é um dos maiores desafios em saúde pública, sendo o tipo de câncer mais diagnosticado entre as mulheres globalmente e com uma história que remonta a 1600 a.C., no Egito. Este câncer tem sido estudado amplamente devido às suas graves implicações para a saúde feminina, o CM continua a impulsionar inovações nas estratégias de manejo clínico e terapêutico (HAQUE et al., 2022; TORRE et al., 2017).

No contexto da saúde da mulher, o aumento nas taxas de incidência, prevalência e mortalidade por CM é preocupante, sendo influenciado por fatores de risco como história reprodutiva, histórico familiar, alta densidade mamária e idade (com maior incidência após os 50 anos no Brasil). As altas taxas de mortalidade são agravadas pelo diagnóstico tardio e pelo limitado acesso a tratamentos de adequados (ARNOLD et al., 2022; BRASIL, 2014).

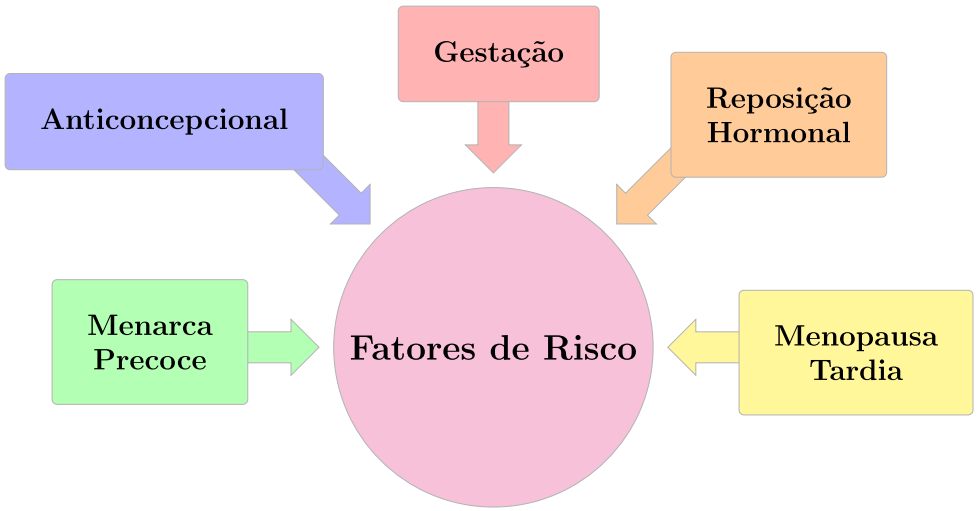
A Iniciativa Global para o Controle do CM visa promover a saúde, a detecção precoce e o tratamento abrangente como pilares centrais na redução da mortalidade. A detecção tardia e a metástase, associadas a fatores de risco, elevam significativamente a taxa de mortalidade e afetam negativamente a qualidade de vida e a sobrevida das pacientes (BRAY et al., 2024; CINTRA, 2012; HAQUE et al., 2022; TORRE et al., 2017).

Em termos globais, o CM feminino é o tipo de câncer mais comum, com exceção dos cânceres de pele não melanoma, como apontam as estimativas do IARC e do INCA (IARC, 2024; INCA, 2023). Esta neoplasia é amplamente diagnosticada e uma das principais causas de mortalidade por câncer entre as mulheres, com a probabilidade de sobrevida intrinsecamente relacionada à detecção precoce e à implementação de terapias eficazes (IARC, 2023; NICOLÒ et al., 2020; OKAGBUE et al., 2021).

Embora o CM feminino seja uma das neoplasias mais prevalentes, o prognóstico é favorável quando detectado precocemente, o que reforça a importância das estratégias de prevenção e diagnóstico precoce. Dentre as medidas de prevenção, como controle de peso, redução do consumo de álcool e estímulo à prática de atividade física, contribuem para a diminuição da incidência de CM, conforme mostrado na Figura 5 os fatores de risco desta

doença (BRAY et al., 2024; LIU et al., 2020; TORRE et al., 2017).

Figura 5 – Diagrama dos fatores de risco do CM femino.



Fonte: Adaptado BRAY *et al.* (2024).

O CM feminino é um tumor maligno invasivo, conhecido por sua notável heterogeneidade, tanto na composição celular quanto nas características moleculares, nas respostas terapêuticas e nos desfechos clínicos. Tal como ocorre com outros tipos de câncer, o CM apresenta uma desregulação na proliferação celular e na apoptose, manifestando uma grande diversidade celular. Essa diversidade se assemelha à composição das células das glândulas mamárias saudáveis, incluindo células mioepiteliais, epiteliais do ducto e epiteliais alveolares (CARVALHO, 2016; SOUZA, 2018).

2.1.2.1 Classificação e Estadiamento do Câncer de Mama

A heterogeneidade do CM feminino é uma das características mais marcantes dessa doença e representa um grande desafio para sua abordagem clínica. Essa variabilidade se reflete nas características morfológicas, histopatológicas, fenotípicas e genéticas das células neoplásicas da mama. A diversidade celular encontrada no CM contribui para as variações nas respostas terapêuticas e nos desfechos clínicos, tornando esta neoplasia uma doença complexa, com prognóstico individualizado e um importante problema de saúde pública (CINTRA, 2012; TENG et al., 2019).

O diagnóstico histopatológico do CM é fundamental para sua classificação e é realizado por meio de exames citológicos. O CM é classicamente subdividido em dois tipos principais: o carcinoma ductal e o carcinoma lobular, que podem ser *in situ* ou infiltrantes. O adenocarcinoma do tipo carcinoma ductal infiltrante (CDI) é o mais comum, representando cerca de 90% dos casos, seguido pelo carcinoma lobular infiltrante (CLI), que corresponde de 5 a 10% dos casos (BRASIL, 2014; CINTRA, 2012; INCA, 2023).

Para estabelecer um tratamento oncológico personalizado, são utilizados biomarcadores como marcadores citológicos, histopatológicos, perfil molecular e imuno-histoquímicos, os quais atribuem a cada paciente um subtipo da doença, orientando assim a escolha do tratamento mais adequado (LAI et al., 2019; KöHN-LUQUE et al., 2020).

O CM é categorizado em subtipos intrínsecos (Luminal A, Luminal B, Basal e superexpressão HER2) com base em marcadores moleculares imuno-histoquímicos como Receptores de Estrogênio (RE), Receptores de Progesterona (RP), o receptor do fator de crescimento epidérmico humano 2 (HER2) e Ki-67 (CIRQUEIRA et al., 2011). Esta classificação é fundamental para direcionar terapias específicas e prever o prognóstico, como descrito no Quadro 1. Tumores luminais, que expressam RE/RP e não expressam HER2, frequentemente têm prognóstico mais favorável, enquanto tumores triplo negativos (sem expressão de RE/RP/HER2) e tumores com superexpressão de HER2 têm prognóstico menos otimista (CINTRA, 2012; CIRQUEIRA et al., 2011; NAVE, 2020).

Quadro 1 – Classificação molecular do perfil imuno-histoquímico.

Subtipo	**Perfil Imuno-histoquímicos**
Luminal A	RE+ e/ou RP+, HER2- e Ki-67 $< 14\%$
Luminal B	RE+ e/ou RP+, HER2- e Ki-67 $\geq 14\%$
	RE+ e/ou RP+, HER2+
Superexpressão de HER2	RE-, RP- e HER2+
Triplo Negativo	RE-, RP-e HER2-

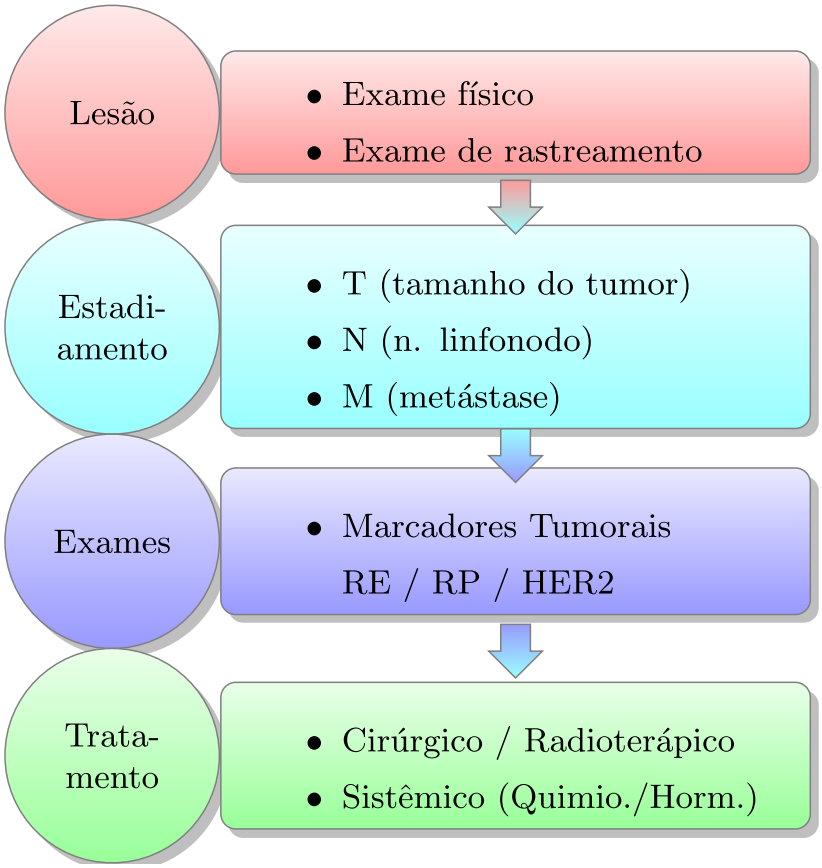
Fonte: Adaptado Subtipos Moleculares CM (Cirqueira et. al. 2011).

A subclassificação do CM, baseada em exames moleculares e imuno-histoquímicos que avaliam a resposta endócrina (RE e RP) e fatores moleculares reguladores do crescimento (HER2, P53, Ki-67), desempenha um papel decisivo na orientação das diretrizes clínico-terapêuticas específicas para cada subtipo. Com o avanço contínuo das opções de tratamento, a escolha da terapia mais adequada para cada subtipo de CM tornou-se um desafio essencial na gestão clínica dessa doença (CINTRA et al., 2012; MONCADA-TORRES et al., 2021; MIN et al., 2021; NAVE, 2020).

A detecção precoce do CM envolve a identificação de nódulos por meio de exames físicos e mamografias, como ilustrado na Figura 6, que demonstra a sequência do curso clínico-terapêutico (CARVALHO, 2016; BRASIL, 2014). No Brasil, o rastreamento mamográfico é recomendado a partir dos 40 anos pela Sociedade Brasileira de Oncologia Clínica (SBOC) (ou a partir dos 50 anos pelo Ministério da Saúde - MS), o que contribui significativamente para a redução da mortalidade e melhora da sobrevida, resultando em diminuição do sofrimento dos pacientes e melhoria da qualidade de vida (XIAO et al., 2022; CARVALHO, 2016; BRASIL, 2014).

Após a identificação de um nódulo, é essencial realizar uma biópsia para determinar

Figura 6 – Fluxograma do curso clínico-terapêutico do CM feminino.



Fonte: Adaptado Carvalho (2016).

se é maligno ou benigno. O diagnóstico do CM é baseado no exame clínico combinado com estudos de imagem e confirmado por avaliação histopatológica. Em seguida, é estabelecido o estadiamento da doença, utilizando o sistema TNM, e a categorização dos subtipos intrínsecos, orientando as decisões do curso clínico terapêutico (BRASIL, 2014; BARRIOS et al., 2022). Durante a anamnese, são avaliados fatores como o estado menopausal, histórico familiar de CM e ovário, comorbidades e outros fatores de risco, para orientar o plano de tratamento (BRASIL, 2014; BARRIOS et al., 2022; HUEMAN et al., 2018; XIAO et al., 2022).

O estadiamento do CM desempenha um papel crucial na avaliação prognóstica, sendo realizado por meio do sistema TNM (T: tamanho do tumor; N: número de linfonodos comprometidos; M: presença de metástases à distância). Desde sua primeira publicação em 1976, o Manual de Estadiamento do Câncer passou por oito revisões, que incorporaram novos fatores prognósticos, aprimorando a estratificação dos pacientes e a precisão no planejamento terapêutico, além de melhorar a predição dos desfechos clínicos (AMIN et al., 2017; BARRIOS et al., 2022; HUEMAN et al., 2018).

No Brasil, os Protocolos Clínicos e Diretrizes Terapêuticas em Oncologia categori-

zam o CM pelo estadiamento clínico (cTNM), baseado em exames físicos, diagnóstico por imagem, endoscopia, cirurgia e outros exames pertinentes, e pelo estadiamento patológico (pTNM), que incorpora achados cirúrgicos e histopatológicos (BRASIL, 2014; BARRIOS et al., 2022; CINTRA, 2012).

O sistema TNM classifica o estágio do CM com base nas características do tumor (T), no acometimento linfonodal (N) e na presença de metástases à distância (M). Para T, existem categorias como Tis (carcinoma *in situ*), T1 (T <= 2,0 cm), T2 (2,0 < T <= 5,0 cm), T3 (T > 5,0 cm) e T4 (qualquer T com extensão nos tecidos adjacentes). Quanto ao N, inclui Nx (linfonodo não avaliável), N0 (sem metástase em linfonodos regionais) e N1-N3 (metástases em linfonodos regionais). Para M, temos M0 (sem metástase) e M1 (com metástase). Esta classificação, conhecida como Estadiamento Anatômico, é realizada antes da cirurgia, conforme apresentado no Quadro 2 (BARRIOS et al., 2022; BRASIL, 2014; CINTRA, 2012; SOBIN et al., 2011).

Quadro 2 – Estadiamento Anatômico baseado no Sistema TNM.

Estadiamento	**Agrupamentos (TNM)**
0	(TisN0M0)
IA	(T1N0M0)
IB	(T0N1miMO, T1N1mi M0)
IIA	(T0N1MO, T1N1M0, T2N0M0)
IIB	(T2N1M0, T3N0M0)
IIIA	(T0N2M0, T1N2M0, T2N2M0, T3N1M0, T3N2M0)
IIIB	(T4N0M0, T4N1M0, T4N2M0)
IIIC	(Qualquer T N3M0)
IV	(Qualquer T Qualquer N M1)

Fonte: Adaptado Diretrizes SBOC (Barrios et. al. 2022).

Após o tratamento cirúrgico, o estadiamento TNM é precedido pela letra "p", e para o acometimento linfonodal, incluem-se informações sobre imuno-histoquímica positiva (i+) ou negativa (i-), histologia ou molecular positiva (m+) ou negativa (m-), micrometástase linfonodal (mi), e o número de linfonodos afetados(1- 1 a 3; 2- 4 a 9; e 3- >= 10). Para metástases, cM0 (i+) indica células tumorais ou depósitos de até 0,2 mm, e cM1 indica metástases distantes detectadas clinicamente ou por imagem, enquanto pM1 indica metástases à distância confirmadas histopatologicamente. A 8ª edição do Manual de Estadiamento do Câncer da *American Joint Committee on Cancer* (AJCC) orienta o estadiamento clínico e patológico, fornecendo uma base atualizada para a avaliação e gestão do CM (AMIN et al., 2017; BARRIOS et al., 2022; SAúDE, 2019; ROSEN; SAPRA, 2023).

2.1.2.2 Curso terapêutico e e perspectivas para a Sobrevida

Adotar uma abordagem abrangente no tratamento do CM, que inclua a detecção precoce e o tratamento imediato, tem se mostrado promissora para melhorar o prognóstico da doença e, consequentemente, a sobrevida. A triagem e a conscientização sobre a importância do diagnóstico precoce são fundamentais para a redução dos riscos, enquanto modalidades terapêuticas como cirurgia, radioterapia e quimioterapia continuam sendo as mais utilizadas (BRASIL, 2014; LI et al., 2021). Nos últimos anos, a hormonioterapia e a imunoterapia emergiram como tratamentos mais específicos e modernos, oferecendo novas perspectivas para o tratamento do CM. O processo diagnóstico, por sua vez, evoluiu de simples exames físicos para o uso de biópsias e métodos de imagem sofisticados, como mamografia, ultrassonografia e ressonância magnética (BRASIL, 2014; LI et al., 2021; OKAGBUE et al., 2021; SHUKLA et al., 2018).

Após a classificação e o estadiamento do tumor, é iniciado um plano terapêutico individualizado, sendo a cirurgia a intervenção primária. Esta pode ser complementada por terapias locais, como a radioterapia, e tratamentos sistêmicos, administrados de forma neoadjuvante (antes da cirurgia) ou adjuvante (após a cirurgia) (CARVALHO, 2016; CINTRA, 2012; BRASIL, 2014). Uma abordagem multidisciplinar, que combina cirurgias curativas ou paliativas com a avaliação do envolvimento linfonodal, radioterapia local e tratamentos sistêmicos com quimioterapia e/ou hormonioterapia, visa não apenas a cura, mas também a prolongação da vida e a melhoria da qualidade de vida dos pacientes. Além disso, os cuidados paliativos desempenham um papel fundamental em estágios avançados da doença, aliviando sintomas e proporcionando um melhor conforto aos pacientes (CARVALHO, 2016; BRASIL, 2014; CINTRA, 2012; SHUKLA et al., 2018).

O tratamento medicamentoso é orientado pela avaliação do risco, considerando fatores como o estadiamento da doença, os resultados histopatológicos e imuno-histoquímicos. Com base nesses critérios, são selecionados fármacos quimioterápicos ou hormonioterápicos, que podem ser aplicados de forma profilática, citorredutora ou paliativa, dependendo das necessidades e características do caso (BARRIOS et al., 2022; BRASIL, 2014).

No século XXI, o tratamento do CM passou por avanços significativos, com destaque para o desenvolvimento de medicamentos direcionados, como o trastuzumabe, para pacientes com superexpressão do HER2, e a hormonioterapia adjuvante para aquelas com receptores hormonais positivos. Estes avanços têm contribuído para a eficácia do tratamento e para a melhoria dos desfechos clínicos. No entanto, é essencial que os pacientes, bem como seus responsáveis legais, sejam informados sobre os potenciais riscos, benefícios e efeitos adversos associados ao tratamento, especialmente no que diz respeito ao uso de agentes antineoplásicos (BARRIOS et al., 2022; BRASIL, 2014; XIN et al., 2022).

O prognóstico do CM é mais favorável quando a doença é diagnosticada e tratada precocemente. No entanto, cerca de 20% a 30% das pacientes ainda enfrentam recorrência

com metástases distantes após a cirurgia, o que sugere a presença de micrometástases clinicamente ocultas. A avaliação dos fatores prognósticos, como o estadiamento, os resultados moleculares e as características histopatológicas, é fundamental para estimar o risco de recorrência e disseminação da doença, que pode ser locorregional ou sistêmica. A recorrência do CM é caracterizada pelo reaparecimento da doença após um período assintomático, geralmente após o tratamento inicial (BRASIL, 2014; CINTRA, 2012; BOERI et al., 2020).

O curso clínico do CM é determinado por diversos fatores, incluindo marcadores tumorais, histórico médico, estadiamento, tipo histológico e a sobrevida. Com base nessa análise, uma estratégia terapêutica personalizada é delineada e monitorada ao longo do tratamento, permitindo ajustes conforme a resposta da paciente. Compreender esses fatores é essencial para o planejamento, a avaliação e o acompanhamento adequado do CM, de modo a otimizar os resultados terapêuticos (BRASIL, 2014; CARVALHO et al., 2018; HOWARD et al., 2018; LAI et al., 2019).

A análise da sobrevida, que se refere ao tempo de vida após o diagnóstico do câncer, é um elemento essencial para orientar o manejo clínico-terapêutico e aprimorar os desfechos clínicos. Esse parâmetro desempenha um papel fundamental na redução da mortalidade por CM e na promoção da saúde das mulheres, ao permitir uma compreensão mais aprofundada dos fatores críticos relacionados ao diagnóstico, tratamento e acesso ao sistema de saúde (ALLEMANI et al., 2018; BUSTAMANTE-TEIXEIRA et al., 2002).

A importância dessa abordagem pode ser ilustrada pelos primeiros modelos estatísticos prognósticos baseados em variáveis clínicas, como o "Nottingham Prognostic Index"(D'EREDITA et al., 2001), o "Adjuvant! Online"(HESS, 2008) e o "Predict"(ENGLAND, 2022), conforme descrito por Min et al. (2021) . Esses modelos foram fundamentais para a estratificação do risco e a definição de condutas terapêuticas mais adequadas (MIN et al., 2021)

Os avanços na modelagem computacional têm permitido predições mais precisas sobre o curso da doença e auxiliando na tomada de decisões clínicas. Como demonstrado na revisão sistemática conduzida por Li et al. (2021) (LI et al., 2021), que avaliou métodos de AM aplicados à predição da sobrevida em cinco anos para o CM, esses modelos oferecem um potencial significativo para otimizar o tratamento e aprimorar os prognósticos (LI et al., 2021). Além de preencher lacunas na prática médica atual, as abordagens preditivas computacionais fornecem informações valiosas que podem contribuir para tratamentos mais eficazes e melhores perspectivas de sobrevida das pacientes (LI et al., 2021; MCKENNA et al., 2018; MONCADA-TORRES et al., 2021; NAVE, 2020).

2.1.3 A PANDEMIA DE COVID-19: DESAFIOS E OPORTUNIDADES NA ONCOLOGIA

A pandemia de COVID-19, iniciada em 2019, teve um impacto profundo na detecção e no tratamento do câncer. No Brasil, onde o primeiro caso foi registrado em fevereiro de 2020, o sistema de saúde enfrentou desafios complexos, agravados pela sobrecarga hospitalar e pelas medidas de isolamento, que dificultaram o acesso a cuidados médicos. O fechamento temporário de unidades de saúde, interrupções laborais, dificuldades no acesso a planos de saúde e o receio da exposição ao vírus comprometeram a continuidade dos tratamentos oncológicos (RIBEIRO et al., 2022; SIEGEL et al., 2023; SUNG et al., 2021).

Embora os efeitos mais imediatos tenham sido sentidos em meados de 2020, a recuperação dos serviços de saúde ainda está em andamento. Na oncologia, a crise resultou em atrasos no diagnóstico e no tratamento, aumentando o número de casos detectados em estágios avançados e, potencialmente, a taxa de mortalidade. Além disso, os impactos psicológicos, sociais e nas bases de dados de saúde continuarão a ser avaliados nos próximos anos (RIBEIRO et al., 2022; FERLAY et al., 2021; SIEGEL et al., 2023; SUNG et al., 2021).

Em 2020, o Brasil experimentou uma redução acentuada acentuada em procedimentos oncológicos em comparação ao ano anterior, com quedas significativas nos exames citopatológicos (44,6%), mamografias (42,6%), biópsias (35,3%), cirurgias oncológicas (15,7%) e procedimentos de radioterapia (0,7%). Surpreendentemente, a quimioterapia foi uma exceção, mantendo ou até aumentando sua produção. Essas reduções foram observadas em todo o país, variando de acordo com a incidência da COVID-19 e as medidas restritivas adotadas pelas autoridades locais e estaduais (RIBEIRO et al., 2022).

A pandemia destacou a necessidade de fortalecer os sistemas de registro de saúde e aprimorar a infraestrutura para enfrentar crises sanitárias futuras. A IA, impulsionada pela urgência em conter o vírus e desenvolver vacinas, tornou-se uma ferramenta essencial para otimizar recursos, personalizar tratamentos, apoiar políticas públicas e acelerar ensaios clínicos (SCHAAR et al., 2021; DICUONZO et al., 2023). A aplicação da IA tem transformado fluxos de trabalho administrativos e clínicos, mas seu uso pleno ainda é limitado por desafios práticos e éticos, como privacidade e integridade dos dados (SCHAAR et al., 2021; DICUONZO et al., 2023; DAVENPORT; KALAKOTA, 2019).

De acordo com Schaar et al. (2021), a IA pode ajudar a enfrentar cinco desafios destacados pela pandemia da COVID-19:

- Otimização de Recursos: A escassez global de recursos críticos, como testes e leitos hospitalares, é enfrentada com a IA, que ajuda a identificar indivíduos de alto risco e alocar recursos financeiros de maneira eficaz.

- Personalização de Tratamentos: A IA aborda a diversidade de sintomas e progressão

da doença, permitindo terapias personalizadas baseadas em dados individuais.

- Coordenação de Resposta: A IA auxilia na identificação de melhores práticas, alinhamento de políticas e promoção da colaboração, enfrentando os desafios de coordenação da resposta à pandemia.
- Melhoria dos Ensaios Clínicos: Os tradicionais ensaios clínicos limitados são ampliados pela IA, considerando eficientemente subgrupos e gerando resultados mais robustos.
- Adaptação Contínua: A evolução da pandemia exige pesquisa em aprendizado de transferência, aplicando insights a novos cenários. A IA também mede a incerteza nas recomendações, tornando as decisões mais confiáveis (SCHAAR et al., 2021).

A IA se mostrou promissora na resposta à pandemia e na oncologia, mas sua implementação ainda enfrenta barreiras técnicas, regulatórias e éticas. Segundo van Dicuonzo et al. (2023), a tecnologia tem se mostrado eficaz em áreas como diagnóstico, terapia, intercâmbio de informações, monitoramento, coleta de dados e até cirurgia remota. No entanto, seu uso também apresenta desafios, especialmente no que diz respeito à integração de dados, segurança, privacidade e conformidade com regulamentações rigorosas. O estudo analisou 132 publicações acadêmicas sobre o impacto da IA na saúde, destacando tanto as vantagens quanto as limitações da tecnologia e enfatizando a necessidade de uma implementação responsável. Para que a IA contribua de forma eficaz na tomada de decisões clínicas e no aprimoramento do atendimento ao paciente, é crucial que futuros pesquisadores avaliem criticamente obstáculos éticos e técnicos, garantindo uma gestão de dados otimizada e alinhada às exigências regulatórias (DICUONZO et al., 2023).

O futuro da IA na pesquisa médica é promissor, mas exige um equilíbrio entre inovação e responsabilidade ética (DAVENPORT; KALAKOTA, 2019; DICUONZO et al., 2023). A transparência e a interpretabilidade dos sistemas de IA são essenciais para garantir sua aplicação segura e ética. Os modelos computacionais, ao serem aplicados aos registros médicos eletrônicos, podem transformar dados complexos em *insights* úteis, melhorando a tomada de decisões clínicas. A colaboração entre médicos e desenvolvedores de IA será crucial para otimizar essas abordagens e garantir sua aplicação ética e eficaz na saúde (DAVENPORT; KALAKOTA, 2019; RUBINGER et al., 2022).

2.2 AVANÇOS E APLICAÇÕES DE MODELOS COMPUTACIONAIS NA SAÚDE

A modelagem na área da saúde é um campo de estudo antigo, com ampla aplicação em pesquisas oncológicas. De forma geral, as abordagens de modelagem podem ser classificadas em três grandes áreas no estudo do câncer: estatística, matemática e computacional

(WODARZ; KOMAROVA, 2005; ASSIS et al., 2019; CARVALHO, 2016; CARVALHO et al., 2011).

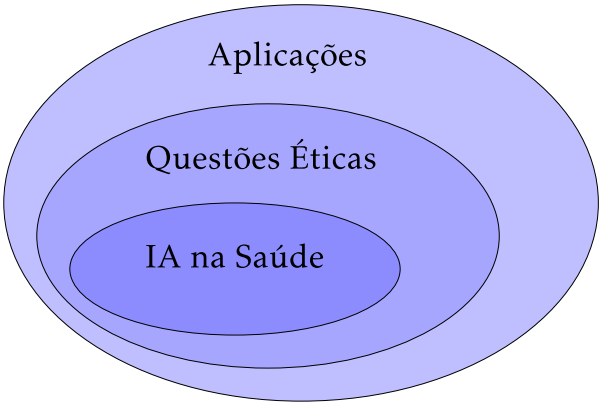
A modelagem estatística é uma das metodologias mais antigas e bem-sucedidas no estudo do câncer, utilizando dados epidemiológicos de registros de saúde para criar modelos descritivos que sintetizam essas informações. Esses modelos são essenciais para a análise de dados relacionados à saúde pública e ao estudo de doenças em populações específicas. Por meio de métodos estatísticos, é possível descrever características demográficas, incidência, prevalência, fatores de risco, eficácia de intervenções e outras variáveis relevantes, embasando decisões em saúde pública e orientando políticas de prevenção e intervenção (ASSIS et al., 2019).

Por sua vez, a modelagem matemática tem se mostrado crucial na compreensão e predição do comportamento do câncer, sendo subdividida em modelos populacionais e mecanicistas (CARVALHO, 2016). Os modelos populacionais, como o de Gompertz, utilizam equações para descrever o crescimento tumoral ao longo do tempo, proporcionando uma visão geral do crescimento e disseminação metastática. Por outro lado, os modelos mecanicistas baseiam-se nas relações causais entre processos biológicos nos tecidos neoplásicos (BRITTON; BRITTON, 2003; CARVALHO, 2016; PREZIOSI, 2003). Um exemplo importante é o modelo de Lotka-Volterra, que descreve interações entre diferentes espécies (BRITTON; BRITTON, 2003), e o modelo de Nicolò et al. (2020), que inclui aspectos biológicos, como taxa de crescimento e probabilidade de disseminação metastática. Esses modelos permitem predições personalizadas do curso da doença, auxiliando na escolha de estratégias terapêuticas mais adequadas (NICOLÒ et al., 2020).

Nos últimos anos, a modelagem computacional, especialmente com o uso de IA e suas técnicas de AM, tem ganhado destaque devido aos benefícios no setor da saúde (CARVALHO et al., 2011; KITSIOS et al., 2023). A IA oferece soluções inovadoras que atendem a diversas necessidades, mas sua aplicação ainda enfrenta desafios, como integração de dados, privacidade do paciente, questões legais e segurança. Os métodos de AM combinam abordagens estatísticas e matemáticas, gerando *insights* valiosos por meio de técnicas como classificação, regressão e clusterização. Esses modelos são treinados com grandes volumes de dados, permitindo uma análise mais eficiente e personalizada dos dados de saúde (CARVALHO et al., 2011; PANCH et al., 2018; KITSIOS et al., 2023).

Conforme ressaltado por Kitsios et al. (2023), e ilustrado na Figura 7, a IA pode desempenhar diversas funções no setor da saúde, abrangendo diagnóstico, terapia, troca de informações, proteção, consulta, monitoramento, coleta de dados e até cirurgia remota. Essas aplicações demonstram a versatilidade da IA na melhoria dos cuidados com a saúde, desde a prevenção até o tratamento (KITSIOS et al., 2023). Além disso, os métodos de AM são amplamente utilizados na pesquisa médica, oferecendo uma abordagem eficaz e econômica para analisar conjuntos de dados complexos. No entanto, é fundamental

Figura 7 – Diagrama das implicações da IA na Saúde.



Fonte: Adaptado Kitsios et al. (2023).

equilibrar os avanços tecnológicos com a consideração de questões éticas, garantindo que os sistemas baseados em IA sejam transparentes e interpretáveis. Embora o uso dessas técnicas ainda esteja em desenvolvimento, elas estão se tornando uma realidade cada vez mais presente na prática médica (PAIXÃO et al., 2022; RUBINGER et al., 2022).

2.2.1 APLICAÇÃO DA INTELIGÊNCIA ARTIFICIAL NA MEDICINA

A pandemia de COVID-19 videnciou a urgência da incorporação de tecnologias inovadoras na saúde, impulsionando a adoção da IA. Essas inovações frequentemente superam as capacidades humanas na análise de grandes volumes de dados, tornando-se ferramentas poderosas para classificação, predição e tomada de decisões clínicas (MAABREH et al., 2021; RUBINGER et al., 2022). No entanto, sua implementação enfrenta desafios éticos e práticos, especialmente no que se refere à interpretabilidade e transparência, essenciais para garantir resultados confiáveis e seguros. Nesse contexto, a colaboração multidisciplinar é essencial para o desenvolvimento de modelos computacionais robustos, que, quando alimentados por dados clínicos adequados, demonstram grande potencial na prática médica (DAVENPORT; KALAKOTA, 2019; LI et al., 2021).

O avanço tecnológico atual pode ser comparado a uma nova revolução industrial, com a IA como protagonista. As máquinas não se limitam mais a tarefas manuais, expandindo sua atuação para atividades que exigem inteligência racional. Os métodos de AM, por exemplo, otimizam o desempenho de sistemas computacionais por meio da experiência acumulada (LUDERMIR, 2021; MITCHELL; LEARNING, 1997).

A IA, especialmente em aplicações clínicas, tem avançado rapidamente, muitas vezes superando a capacidade humana em diversas tarefas médicas (MAABREH et al., 2021). Paralelamente, os métodos de AM vêm se consolidando no desenvolvimento de algoritmos que capacitam os computadores a aprender com dados, gerando modelos preditivos e classificatórios. À medida que esses métodos se aprimoram, surgem debates sobre como

essas abordagens se comparam às estatísticas tradicionais, bem como sobre o seu potencial de conflito ou complementaridade com essas técnicas mais convencionais (SHUKLA et al., 2018; AIVALIOTIS et al., 2021).

No amplo campo da IA, destacam-se os métodos de AM e as redes neurais profundas, inspiradas na biologia neural. Segundo Mitchell (1997), os MAM podem ser definidos como "a capacidade de melhorar o desempenho de uma tarefa por meio da experiência"(CARVALHO et al., 2011; MITCHELL; LEARNING, 1997; LUDERMIR, 2021). Esses métodos têm demonstrado grande potencial na saúde, permitindo diagnósticos mais rápidos e precisos (DICUONZO et al., 2023; KITSIOS et al., 2023).

Como ressalta Panch (2018), os métodos de AM simulam habilidades cognitivas humanas, desenvolvendo algoritmos que aprendem a partir de grandes volumes de dados. Esses modelos são usados para prever, classificar ou detectar padrões, proporcionando soluções eficazes na medicina (PANCH et al., 2018; RUBINGER et al., 2022; PAIXÃO et al., 2022).

Dentro do AM, destacam-se três abordagens principais: supervisionado, não supervisionado e por reforço. No aprendizado supervisionado, os algoritmos são treinados com dados rotulados para mapear entradas e saídas, sendo amplamente utilizados em problemas de classificação e regressão. No aprendizado não supervisionado, os algoritmos identificam padrões em dados não rotulados, possibilitando a formação de clusters e a redução de dimensionalidade. Já no aprendizado por reforço, o sistema aprende por meio de recompensas e punições, sendo aplicado em áreas como jogos e robótica. Em todas essas abordagens, o reconhecimento de padrões é essencial para resolver problemas complexos de classificação e aumentar a flexibilidade dos modelos (LUDERMIR, 2021; PAPA; FALCAO, 2010; PANCH et al., 2018; PÖLSTERL, 2020).

A aplicação do AM na saúde traz grandes oportunidades para o desenvolvimento de modelos voltados à prevenção, diagnóstico precoce e prognóstico de doenças. No entanto, desafios como o pré-processamento de dados, a otimização de parâmetros e a seleção de variáveis ainda precisam ser superados para aumentar a eficácia desses modelos (LI et al., 2021; PAIXÃO et al., 2022; SEDIGHI-MAMAN; MONDELLO, 2021). Além disso, o crescimento exponencial dos dados biomédicos exige algoritmos mais precisos e robustos para lidar com sua alta dimensionalidade, exigindo cautela na interpretação dos resultados e na validação desses modelos em populações mais amplas (DENG et al., 2021; PAIXÃO et al., 2022; VERÍSSIMO et al., 2016).

Os modelos de AM voltados à predição temporal têm demonstrado eficácia comparável ou superior aos métodos estatísticos tradicionais. No entanto, sua complexidade pode dificultar a interpretação humana (JANSEN et al., 2020; KRZYZIŃSKI et al., 2023). A seleção criteriosa de variáveis é essencial para a precisão dos modelos preditivos baseados em dados históricos. Apesar dos avanços, há uma demanda crescente por mais pesquisas,

incluindo a validação de modelos existentes e o desenvolvimento de novas abordagens que integrem dados clínicos mais completos (MIN et al., 2021).

No contexto da saúde, a seleção de características e a avaliação da importância das variáveis são estratégias essenciais para otimizar os modelos de AM, especialmente na redução de dimensionalidade e na identificação de variáveis relevantes do problema. A escolha dessas abordagens impacta diretamente a eficácia e interpretabilidade dos modelos, sendo crucial para o sucesso das aplicações na prática clínica (PÖLSTERL, 2020; VERÍSSIMO et al., 2016; KRZYZIŃSKI et al., 2023).

A seleção de características visa identificar um subconjunto relevante de variáveis, mantendo ou melhorando o desempenho preditivo (PÖLSTERL, 2020; VERÍSSIMO et al., 2016; SIMON et al., 2011). Já a avaliação da importância das variáveis determina sua contribuição para a predição, sendo frequentemente realizada com técnicas como árvores de decisão e algoritmos baseados em permutação (PÖLSTERL, 2020; ISHWARAN et al., 2008; KRZYZIŃSKI et al., 2023).

A pesquisa multidisciplinar, aliada à melhoria da interpretabilidade dos modelos de AM, é essencial para sua adoção na prática clínica. A colaboração entre diferentes áreas do conhecimento pode viabilizar o desenvolvimento de modelos mais eficazes e confiáveis. No entanto, a contínua adaptação dessas tecnologias aos diversos contextos da saúde é fundamental para garantir sua aplicabilidade (MIN et al., 2021; LI et al., 2021; OKAGBUE et al., 2021).

2.2.2 MODELAGEM COMPUTACIONAL NA ONCOLOGIA

A crescente utilização da IA tem sido fundamental para a compreensão e resolução de problemas em diversas áreas, incluindo a saúde. Na oncologia, os métodos de AM se destacam ao fornecer informações preditivas valiosas para a prática clínica, impulsionando o diagnóstico, prognóstico e tratamento do câncer (LAI et al., 2019; LI et al., 2021; KöHN-LUQUE et al., 2020; NAVE, 2020). O avanço tecnológico, aliado ao desenvolvimento terapêutico, favorece a criação de modelos prognósticos, que se mostram promissores na medicina (LI et al., 2021; OKAGBUE et al., 2021; TAPAK et al., 2019).

A modelagem tumoral tem sido essencial para esclarecer fatores como dados epidemiológicos, progressão e microambiente tumoral, utilizando análises estatísticas e modelos populacionais e mecanicistas (CARVALHO, 2016; WODARZ; KOMAROVA, 2005). No entanto, com a evolução clínica e terapêutica, há uma demanda crescente por modelos computacionais que permitam prever e personalizar terapias anticâncer com maior precisão (LAI et al., 2019; MCKENNA et al., 2018).

Na oncologia moderna, modelos computacionais baseados em princípios matemáticos lidam com a complexidade do câncer, utilizando métodos de AM para prever o curso terapêutico da doença (LAI et al., 2019; MCKENNA et al., 2018; NAVE, 2020). Esses

modelos também auxiliam na análise de aspectos como a heterogeneidade tumoral e a resposta aos tratamentos (HOWARD et al., 2018; KöHN-LUQUE et al., 2020; LAI et al., 2019).

Além de contribuírem para o desenvolvimento de tratamentos personalizados, desde análises estatísticas a modelos mecanicistas, os métodos de AM apoiam a predição do curso clínico e da sobrevida dos pacientes. Essas abordagens baseadas em ciência de dados facilitam a interpretação dos complexos dados oncológicos (KöHN-LUQUE et al., 2020; VERÍSSIMO et al., 2016).

No contexto oncológico, os modelos matemáticos de AM utilizam dados clínicos para prever o prognóstico do paciente, aprendendo com padrões extraídos de registros históricos. A eficácia desses modelos depende de teorias robustas de probabilidade, estatística, ciência da computação e outras áreas. Com alto poder computacional, eles conseguem analisar grandes volumes de dados e extrair informações úteis para diversas aplicações clínicas (MIN et al., 2021; XIN et al., 2022).

Os modelos matemáticos oferecem uma linguagem precisa para representar fenômenos complexos. Uma estratégia comum é a simplificação do modelo sem comprometer seu objetivo, embora esse processo seja desafiador, especialmente na formulação de equações que representem interações biológicas e clínicas, exigindo adaptação contínua e refinamento (CARVALHO, 2016; EDELSTEIN-KESHET, 2005; VELTEN, 2009).

Apesar do potencial dos modelos matemáticos para prever o curso terapêutico do câncer, ainda há desafios na representação das interações celulares e na validação com dados clínicos robustos. Um exemplo promissor é o estudo de Nave et al. (2020), que combinou um modelo matemático com técnicas de AM (regressão linear e redes neurais) para prever o tamanho tumoral. Utilizando dados de 1869 mulheres diagnosticadas com CM, o estudo analisou detecção, progressão e recorrência após cirurgia conservadora, demonstrando o potencial dessas abordagens para a medicina personalizada (NAVE, 2020).

Atualmente, os métodos de AM vêm sendo aplicados no desenvolvimento de protocolos terapêuticos que fornecem informações prognósticas para orientar decisões clínicas (NAVE, 2020; KöHN-LUQUE et al., 2020; LI et al., 2021; TAHMASSEBI et al., 2019). Essa abordagem visa otimizar os resultados e melhorar a qualidade de vida dos pacientes (LAI et al., 2019; KöHN-LUQUE et al., 2020). Este trabalho tem como objetivo identificar métodos de AM específicos para a predição da sobrevida de pacientes diagnosticadas com CM, tema que será abordado na próxima seção. Para isso, são considerados modelos baseados em abordagens tradicionais de AM, ainda não adaptadas para dados de sobrevida. Entre essas abordagens, destacam-se modelos matemáticos amplamente utilizados em diversas áreas, como o modelo de CPH (COX, 1972), além dos modelos dos algoritmos de AM, como *Gradient Boosting* (GB) (FRIEDMAN, 2001; FRIEDMAN, 2002), *Random Forest* (RF) (BREIMAN, 2001), *Support Vector Machine*

(SVM) (CORTES; VAPNIK, 1995; BURGES, 1998), os quais serão detalhados a seguir.

2.2.2.1 *Cox Proportional Hazards*

O modelo de Regressão de Cox, também conhecido como Modelo de Cox ou CPH, é uma ferramenta estatística amplamente utilizada na análise de dados epidemiológicos e de sobrevida. Desenvolvido por Sir David Cox em 1972, seu objetivo é avaliar a relação entre covariáveis e a função de risco, combinando essas características (variáveis independentes) de forma linear (COX, 1972; SIMON et al., 2011). A função de risco do modelo é definida por:

$$h(t) = h_0(t) \exp^{x_i^T \beta}, \tag{2.1}$$

onde $h(t) > 0$ representa a função de risco de um indivíduo no instante t, $h_0(t) > 0$ é a função de risco de referência, x_i é o vetor de covariáveis e β corresponde aos coeficientes estimados a partir das características preditoras de x_i (COX, 1972; SIMON et al., 2011).

O modelo de Cox é classificado como semi-paramétrico, pois não faz suposições específicas sobre a forma da função de risco de referência, concentrando-se exclusivamente no efeito das covariáveis sobre a razão de riscos. A premissa fundamental do CPH é que a razão de riscos permanece constante ao longo do tempo, ou seja, a relação entre as taxas de risco de quaisquer dois indivíduos é constante e independente do tempo. Matematicamente, isso é expresso como:

$$\frac{h(t)}{h_0(t)} = \exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p), \tag{2.2}$$

onde $\beta_1, \beta_2, \ldots, \beta_p$ são os coeficientes das covariáveis $x_1, x_2, \ldots, x_p$, sendo p o número total de variáveis consideradas no modelo. Essas covariáveis representam fatores explicativos que podem influenciar o tempo até a ocorrência do evento de interesse, permitindo interpretar seu impacto na taxa de risco ao longo do tempo (COX, 1972; SIMON et al., 2011).

2.2.2.2 *Gradient Boosting*

O GB é um métodos de AM supervisionado utilizado tanto para tarefas de regressão quanto de classificação. Esse *Ensemble Learning* (EL - método de conjunto) é baseado em um processo iterativo aditivo, combinando múltiplos aprendizes fracos (geralmente árvores de decisão) para formar um aprendiz forte (FRIEDMAN, 2001; FRIEDMAN, 2002). A matemática por trás do GB envolve a otimização de uma função de perda por meio do gradiente descendente (FRIEDMAN, 2001; FRIEDMAN, 2002; HASTIE et al., 2009; HOTHORN et al., 2006).

De forma simplificada, a formulação matemática do modelo de regressão aditiva proposto por Friedman (2001) é construída iterativamente, ajustando sequencialmente

uma função parametrizada simples aos pseudo-resíduos das iterações anteriores. O objetivo é minimizar a função de perda, que quantifica a discrepância entre as predições do modelo e os valores reais. A função aditiva do GB é dada por:

$$F_{GB}(x) = \sum_{i=0}^{M} \gamma_i g(x; \theta_i), \tag{2.3}$$

onde $F_{GB}(x)$ representa a função aditiva, γ_i os coeficientes de expansão, $g(x; \theta_i)$ corresponde à função do aprendiz base (g) parametrizado pelo vetor $\theta = \theta_1, \theta_2, \ldots$, que modela o aprendizado entre a relação das variáveis de entrada (x) e saída (y); e M é o número total de iterações (aprendizado) no processo de *boosting* do modelo, com $i = 1, 2, \ldots, M$. Um valor elevado de M pode levar a um ajuste excessivo, enquanto um valor muito baixo pode resultar em subajuste (FRIEDMAN, 2001; FRIEDMAN, 2002).

O *Stochastic Gradient Boosting* (SGB), proposto por Friedman (2002), introduz randomização no processo de treinamento, selecionando aleatoriamente subamostras dos dados de treinamento em cada iteração. Essa abordagem aumenta a robustez do modelo, reduzindo a correlação entre as iterações e melhorando sua generalização. A atualização do modelo é dada por:

$$F_{GB\ (i)}(x) = F_{GB\ (i-1)}(x) + \gamma_i\ g(x; \theta_{(i)}), \tag{2.4}$$

onde $F_{GB\ (i)}(x)$ corresponde à função aditiva na i-ésima iteração, $F_{GB\ (i-1)}(x)$ é a função do modelo da iteração anterior, γ_i representa os coeficientes de expansão (FRIEDMAN, 2001; FRIEDMAN, 2002).

Além disso, a regularização da taxa de aprendizagem do modelo, ou seja, a magnitude com que as predições são ajustadas em cada iteração aditiva, pode ser aplicada para evitar o sobreajuste, incluindo um termo de aprendizagem $0 < \rho \leq 1$. O modelo matemático regularizado é dado por:

$$F_{GB\ (i)}(x) = F_{GB\ (i-1)}(x) + \rho\ \gamma_i\ g(x; \theta_{(i)}). \tag{2.5}$$

Dessa maneira, temos dois parâmetros de regularização: a taxa de aprendizado ρ e o número de iterações M, onde cada um pode controlar o grau de ajuste e, portanto, afetar o melhor valor para o outro. Com taxas de aprendizado menores, pode-se melhorar a capacidade de generalização do modelo, porém, isso requer um maior número de iterações, o que gera um maior custo computacional (FRIEDMAN, 2001; HASTIE et al., 2009).

2.2.2.3 *Random Forest*

Os modelos RF são amplamente reconhecidos como ferramentas eficazes de predição, que empregam o princípio de EN de árvores de decisão durante o treinamento de forma

randomizada, ou seja, aleatória. Essa aleatoriedade é introduzida de duas maneiras principais: primeiro, uma amostra de dados *bootstrap* é utilizada para cultivar cada árvore; segundo, durante o crescimento de cada nó da árvore, um subconjunto de variáveis é selecionado como candidatos à divisão. O objetivo dessa aleatoriedade é reduzir a variância dos dados em cada árvore, enquanto a profundidade fixa das árvores ajuda a diminuir o viés (BREIMAN, 2001; ISHWARAN; KOGALUR, 2007; ISHWARAN et al., 2008; ISHWARAN; KOGALUR, 2023).

Para ilustrar, considere um conjunto de aprendizes $a_1, a_2, \ldots, a_T$ treinados no mesmo conjunto de dados de aprendizado $(x_1, y_1), \ldots, (x_n, y_n)$, onde x corresponde aos dados de entrada e y às saídas. Para cada árvore a_i na floresta, com $i = 1, 2, \ldots, A$, uma amostra *bootstrap* é inicialmente selecionada dos dados de treinamento x. O processo de construção da árvore inclui então a seleção aleatória de algumas variáveis dentre as disponíveis, seguida pela escolha do melhor ponto de divisão entre essas variáveis selecionadas e a divisão do nó em dois nós filhos (BREIMAN, 2001; ISHWARAN; KOGALUR, 2023).

Assim, a função de regressão da RF $\hat{F}_{RF}$ é obtida pela média das predições de todas as árvores da floresta, expressa matematicamente pela equação:

$$\hat{F}_{RF}(x) = \frac{1}{A} \sum_{i=1}^{A} a_i(x), \tag{2.6}$$

onde $\hat{F}_{RF}(x)$ é a função de regressão da floresta para predição dos dados x, A é o número de árvores na floresta e $a_i(x)$ é a predição da i-ésima árvore da floresta A para os dados x. Além desses parâmetros, o modelo inclui a definição do número de amostras consideradas em cada divisão do nó, bem como o critério de divisão de cada nó. Ajustar esses parâmetros corretamente é essencial para otimizar o desempenho do RF (BREIMAN, 2001; ISHWARAN; KOGALUR, 2023).

Portanto, a construção de cada árvore a partir de diferentes amostras bootstrap e subconjuntos de características selecionadas aleatoriamente leva à predição da Floresta Aleatória $\hat{F}_{RF}(x)$ para os dados de entrada x. Esse processo cria um modelo robusto que reduz o sobreajuste (*overfitting*) e melhora a generalização, combinando as predições de várias árvores construídas de forma randomizada e profunda. Ao combinar as predições de todas essas árvores, o RF fornece predições mais precisas do que qualquer árvore individualmente (BREIMAN, 2001; ISHWARAN; KOGALUR, 2007; ISHWARAN et al., 2008; ISHWARAN; KOGALUR, 2023).

2.2.2.4 *Support Vector Machine*

Os métodos de SVM são algoritmos de AM supervisionados, amplamente utilizados em tarefas de classificação e regressão, com o objetivo de encontrar um hiperplano ótimo para separar duas classes de dados (CORTES; VAPNIK, 1995; BURGES, 1998).

Introduzidos por Cortes e Vapnik em 1995, esses modelos revolucionaram o problema de classificação ao utilizar uma abordagem baseada na transformação dos vetores de entrada em um espaço de características de alta dimensão. Nesse espaço, a construção de uma superfície de decisão linear permite uma excelente capacidade de generalização (CORTES; VAPNIK, 1995). O principal objetivo do SVM é maximizar a margem entre os pontos de dados mais próximos de cada classe, garantindo um melhor desempenho na generalização para novos dados (CORTES; VAPNIK, 1995; BURGES, 1998).

Matematicamente, a função do SVM (F_{SVM}) pode ser expressa como:

$$F_{SVM}(x) = w^T x + b \tag{2.7}$$

onde w é o vetor de pesos perpendicular ao hiperplano de separação, x o vetor de características (entrada) e b é o termo de polarização (CORTES; VAPNIK, 1995; SMOLA; SCHÖLKOPF, 2004).

A otimização do SVM envolve maximizar a margem entre os vetores de suporte e a superfície de decisão, um problema de otimização convexa que pode ser formulado como:

$$\min \ \frac{1}{2} \parallel w \parallel^2, \tag{2.8}$$

sujeito à restrição:

$$y_i(w^T x_i + b) \geq 1, \quad \text{para todo } y_i = +1, \quad \text{ou} \quad y_i(w^T x_i + b) \leq 1, \quad \text{para todo } y_i = -1,$$

garantindo que todos os pontos sejam corretamente classificados (CORTES; VAPNIK, 1995; BURGES, 1998).

Em problemas de classificação linear, considerando um conjunto de treinamento composto por n pares $(x_1, y_1), \ldots, (x_n, y_n)$, onde x_i é o vetor de características e y_i representa a classe associada ao exemplo i (tipicamente -1 ou $+1$ para classificação binária), a solução pode ser obtida utilizando multiplicadores de Lagrange. A formulação resultante é a seguinte:

$$\min_{w,b} \frac{1}{2} \parallel w \parallel^2 + C \sum_{i=1}^{n} \xi_i, \tag{2.9}$$

sujeito às condições:

$$\xi_i \geq 0, \quad \text{e} \quad y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \forall i,$$

onde w e b definem o hiperplano separador, ξ_i são as variáveis de folga que permitem que alguns pontos fiquem dentro da margem ou do lado incorreto do hiperplano, e C é o parâmetro de regularização que controla o equilíbrio entre maximizar a margem e

minimizar os erros de classificação (BURGES, 1998; BELLE et al., 2011; HASTIE et al., 2009).

Para problemas de regressão linear, em que temos um conjunto de treinamento de n pares $(x_1, y_1), \ldots, (x_n, y_n)$, com x_i representando o vetor de características e y_i o valor contínuo associado à regressão, o vetor de suporte w tenta encontrar a direção ótima de mapeamento para $r-1$, que corresponde aos hiperplanos discriminantes paralelos para as r classificações. A formulação é dada por:

$$\min_{w,b,\xi,\xi^*} \frac{1}{2} \parallel w \parallel^2 + C \sum_{i=1}^{r-1} \left(\sum_{i=1}^{n^j} \xi_{(j)(i)} + \sum_{i=1}^{n^{j+1}} \xi^*_{(j+1)(i)} \right), \tag{2.10}$$

sujeito às condições:

$$w^T x_i - b_j \leq -1 + \xi_{(j)(i)}, \quad \xi_{(j)(i)} \geq 0, \quad \forall i = 1, 2, \ldots, n^j,$$

$$w^T x_i - b_j \geq 1 - \xi^*_{(j+1)(i)}, \quad \xi^*_{(j+1)(i)} \geq 0, \quad \forall i = 1, 2, \ldots, n^{j+1},$$

$$b_j \leq b_{j+1}, \quad \text{para} \quad j = 1, \ldots, r-2, \quad \text{e} \quad C > 0.$$

Para cada limiar b_j, são consideradas duas categorias adjacentes, j e $j+1$, para erros empíricos, ou seja, cada amostra na j-ésima categoria deve ter um valor de função que seja menor que a margem inferior $b_j - 1$, caso contrário, o erro é denotado como $\xi^*_{(j)(i)}$; da mesma forma, cada amostra da categoria $j+1$ deve ter um valor de função que seja maior que a margem superior $b_j + 1$, caso contrário o erro é denotado como $\xi^*_{(j+1)(i)}$. Os parâmetros w e b representam o hiperplano de regressão, $\xi_{(j)(i)}$ e $\xi^*_{(j+1)(i)}$ são as variáveis de folga que permitem que alguns pontos de treinamento estejam fora da faixa de tolerância, e C é o parâmetro de regularização (custo), que controla a largura da faixa de tolerância e a quantidade de violações. Um valor maior de C penaliza mais as violações, enquanto um valor menor de C permite mais violações (BELLE et al., 2011; CHU; KEERTHI, 2007).

A modelagem computacional, baseada em princípios matemáticos, desempenha um papel fundamental na compreensão de uma vasta gama de problemas, incluindo o estudo do câncer (NAVE, 2020; NICOLÒ et al., 2020; LAI et al., 2019). Essa abordagem permite a captura precisa de microeventos e processos críticos que influenciam a progressão dessa doença complexa. Muitos desses aspectos são intrínsecos ao câncer e não podem ser facilmente observados por métodos clínicos convencionais, o que ressalta a importância dos modelos computacionais na investigação dessa patologia (HAQUE et al., 2022; LAI et al., 2019). Embora tais modelos não necessitem abranger todos os fenômenos relacionados à doença, suas predições podem fornecer contribuições valiosas tanto para avanços práticos

quanto teóricos, impulsionando o progresso na pesquisa e na prática clínica (CARVALHO, 2016; NAVE, 2020).

2.2.3 MODELAGEM COMPUTACIONAL PARA PREDIÇÃO DE SOBREVIDA

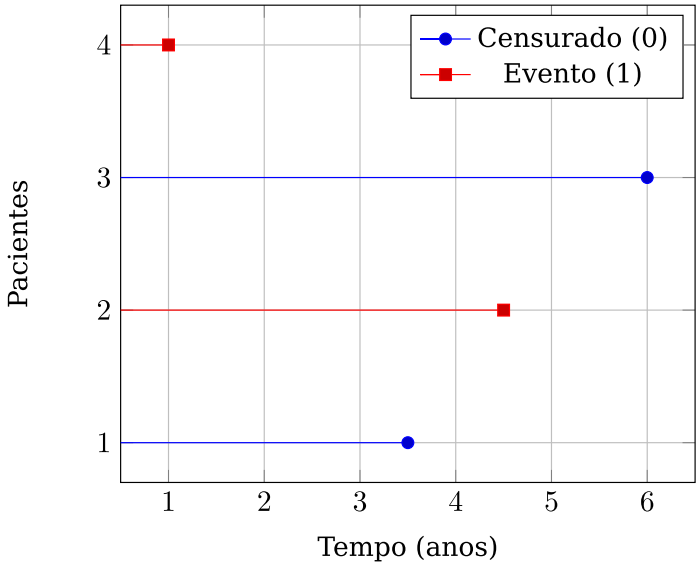
Nos estudos oncológicos, o registro detalhado de informações clínicas, como diagnóstico, remissão, recorrência e óbito, é essencial para a análise da sobrevida e a avaliação do prognóstico dos pacientes. Essa análise permite estimar desfechos, identificar pacientes de alto risco e otimizar a alocação de recursos médicos (TENG et al., 2019; KALAFI et al., 2019). Geralmente conduzida em períodos de acompanhamento de cinco anos, a análise de sobrevida avalia a capacidade de sobrevivência de uma *coorte* e padroniza a atribuição temporal de rótulos aos registros clínicos. Estudos clínicos e de acompanhamento de pacientes com CM são fundamentais para o desenvolvimento de modelos prognósticos validados, contribuindo para a melhoria das taxas de sobrevida (PINHEIRO et al., 2022; LIU et al., 2020).

A análise de sobrevida é essencial na medicina, pois permite avaliar a progressão das doenças e estimar a probabilidade de sobrevivência em períodos específicos. Essa análise, ao considerar múltiplos fatores, oferece estimativas da função de sobrevida e do risco associado (CARVALHO et al., 2011; LI et al., 2021). A sobrevida, definida como o período após o diagnóstico em que o paciente permanece vivo, é um indicador amplamente utilizado para avaliar a evolução clínica, ou seja, a capacidade de sobrevivência, refletindo avanços no diagnóstico e tratamento, além de auxiliar na identificação de fatores prognósticos (BUSTAMANTE-TEIXEIRA et al., 2002; LI et al., 2021). No entanto, essa análise frequentemente lida com dados censurados, uma situação comum nestes estudos clínicos, quando as informações sobre o tempo de sobrevivência dos pacientes podem ser incompletas, como ilustrado na Figura 8 (CARVALHO et al., 2011; LI et al., 2021; KRZYZIŃSKI et al., 2023; LIU et al., 2020).

Os métodos de AM têm se destacado na oncologia ao utilizar BD clínicos para aprimorar a predição da sobrevida em pacientes com CM e outras doenças (AIVALIOTIS et al., 2021; LI et al., 2021). Essas técnicas, que combinam múltiplas variáveis, oferecem predições mais precisas, identificam fatores prognósticos e vêm ganhando relevância na pesquisa em IA, impulsionadas por avanços tecnológicos e suas aplicações práticas. Assim como, desempenham um papel crucial na estimativa de desfechos clínicos (AIVALIOTIS et al., 2021; LI et al., 2021; OKAGBUE et al., 2021; TENG et al., 2019).

Nos últimos anos, os métodos de AM supervisionados têm superado abordagens estatísticas tradicionais, como o modelo de Regressão de Cox (CPH), na predição de desfechos oncológicos (LIU et al., 2020; MONCADA-TORRES et al., 2021; FANIZZI et al., 2023). A capacidade desses métodos de lidar com dados complexos e de alta dimensionalidade, incluindo informações censuradas, elimina suposições restritivas, modela

Figura 8 – Gráfico de Análise de Sobrevida: dados censurados (azul) e dados não censurados (vermelho).



Fonte: Adaptado PÖLSTERL (2023).

não-linearidades e identifica padrões ocultos, resultando em uma análise de sobrevida mais robusta (LIU et al., 2020; MONCADA-TORRES et al., 2021; XIAO et al., 2022). (LIU et al., 2020; MONCADA-TORRES et al., 2021; XIAO et al., 2022). Além disso, esses métodos exploram conexões sutis entre padrões clínicos e respostas ao tratamento, possibilitando prognósticos mais precisos. Estudos recentes reforçam essa tendência e consolidam sua importância (CARVALHO et al., 2023; CARVALHO et al., 2024; FANIZZI et al., 2023; LIU et al., 2020; MONCADA-TORRES et al., 2021; PINHEIRO et al., 2022; XIAO et al., 2022). Essa abordagem é fundamental para a identificação de fatores prognósticos de risco e para o aprimoramento das estratégias diagnósticas e terapêuticas, aumentando a confiabilidade das decisões clínicas e permitindo tratamentos mais adequados para pacientes com CM (BURGES, 1998; OKAGBUE et al., 2021).

Os métodos de AM aplicados à análise de sobrevida visam estabelecer relações entre as covariáveis e o momento da ocorrência do evento, aprendendo com os dados para prever a probabilidade desse evento acontecer (PÖLSTERL, 2023). Uma diferença fundamental em relação aos métodos AM tradicionais é a capacidade desses modelos de lidar com dados censurados, ou seja, situações em que a observação dos dados é parcial. Assim como nos métodos tradicionais, os modelos de AM para análise de sobrevida recebem como entrada as variáveis a serem analisadas, mas sua saída vai além da predição do tempo de sobrevida, incluindo também a ocorrência ou não do evento de interesse (PÖLSTERL, 2023). Esses modelos são adaptados especificamente para o problema em questão, exceto o modelo de Regressão de Cox (CPH), em que a avaliação da correlação entre o risco predito e o risco

observado (SIMON et al., 2011; PÖLSTERL, 2023).

A biblioteca *Scikit-Survival*, um pacote de código aberto desenvolvido em *Python*, oferece ferramentas para análise de sobrevida, incluindo métodos de predição, avaliação de desempenho, otimização e seleção de atributos (PÖLSTERL, 2020; PÖLSTERL, 2023). Os modelos dessa biblioteca, detalhados na próxima seção, estão alinhados aos objetivos deste trabalho e serão aplicados seguindo suas recomendações (PÖLSTERL, 2023).

2.2.3.1 *Cox Proportional Hazards - Survival Analysis*

A análise de sobrevida busca relacionar covariáveis ao tempo de ocorrência de um evento, sendo amplamente utilizada na pesquisa clínica para prever sua incidência (CARVALHO et al., 2011). Os dados de sobrevivência consistem em registros de informações dos pacientes associados às covariáveis, ao indicador de ocorrência ou não do evento de interesse e aos tempos de eventos ou censura. Os eventos podem ser censurados, tornando partes dos dados apenas parcialmente observáveis (PÖLSTERL, 2023; SIMON et al., 2011).

Função de Sobrevida

Os dados de sobrevida são representados como $(X_1, Y_1, \delta_1), \ldots, (X_n, Y_n, \delta_n)$, onde n corresponde ao número de pacientes; $X_i \in \mathbb{R}^m$, sendo X_i um vetor de covariáveis do paciente i e m a dimensão das covariáveis; $Y_i \in \mathbb{R}^+$, sendo Y_i o último tempo observado do paciente i, ou seja, o último tempo de acompanhamento t_i; e $\delta_i = 1$ indica a ocorrência do evento de interesse, enquanto $\delta_i = 0$ dindica censura (LIU et al., 2020; SIMON et al., 2011).

A função de sobrevida $S(t)$ expressa a probabilidade de um paciente sobreviver além de um determinado tempo t, ou seja, por pelo menos esse período (PÖLSTERL, 2023; CARVALHO et al., 2011). Matematicamente, é definida como:

$$S(t) = P(T > t), \tag{2.11}$$

onde T é uma variável aleatória contínua e não negativa, $T \in \mathbb{R}^+$, representando o tempo específico para o qual se calcula a probabilidade de sobrevida, e t corresponde ao tempo de sobrevivência do paciente i (PÖLSTERL, 2023; CARVALHO et al., 2011).

Agora, lembrando que a função de distribuição acumulada da variável T, definida estatísticamente como a probabilidade do evento de interesse ocorrer até um tempo t, representada por:

$$f(t) = P(T \leq t), \tag{2.12}$$

onde $f(t)$ é a função de distribuição acumulada de T (CARVALHO et al., 2011). Assim, $S(t)$ é seu complemento:

$$S(t) = 1 - f(t). \tag{2.13}$$

A função de risco, que expressa a taxa instantânea de ocorrência do evento em t, relaciona-se com a função de sobrevida pela equação:

$$h(t) = \frac{f(t)}{S(t)}. \tag{2.14}$$

Observa-se que a função de risco $h(t)$ e a função de sobrevida $S(t)$ são inversamente proporcionais: à medida que o risco aumenta, a probabilidade de sobrevivência diminui, e vice-versa (PÖLSTERL, 2023; LIU et al., 2020; SIMON et al., 2011; CARVALHO et al., 2011).

Cox Proportional Hazards

O modelo de CPH, ou Regressão de Cox, é amplamente utilizado na análise de sobrevida em oncologia, especialmente por sua interpretabilidade. Como descrito na Equação 2.2, ele avalia o impacto das covariáveis nas taxas de risco de eventos, como o óbito, permitindo a identificação de fatores prognósticos relevantes (LIU et al., 2020; COX, 1972; MONCADA-TORRES et al., 2021). O modelo assume que o risco de um evento é proporcional ao longo do tempo e estima os coeficientes das covariáveis por meio da maximização da verossimilhança parcial. Para tratar empates nos tempos de eventos, utiliza-se a aproximação de *Breslow* (PÖLSTERL, 2023; SIMON et al., 2011).

Na regressão de Cox, a inferência é conduzida com base na verossimilhança parcial, utilizada para estimar os coeficientes do modelo e avaliar a significância estatística das covariáveis (PÖLSTERL, 2023; SIMON et al., 2011). Considere um conjunto de dados de sobrevivência representado por $(X_1, Y_1, \delta_1), \ldots, (X_n, Y_n, \delta_n)$, onde X_i é o vetor de características do paciente i, Y_i corresponde ao último tempo observado, e δ_i indica a ocorrência do óbito ($\delta_i = 1$) ou censura ($\delta_i = 0$). Os tempos de ocorrência dos eventos são organizados em ordem crescente $t_1 < t_2, \ldots, t_k$, onde k é o número de tempos únicos do evento, d representa o número de óbitos, e $j(i)$ denota o índice da observação dos eventos em t_i (LIU et al., 2020; SIMON et al., 2011).

A função de verossimilhança parcial para o modelo de Cox, conforme descrita por Simon et al. (2011), é dada por:

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp(X_{j(i)}^T \beta)}{\sum_{j \in R(t_i)} \exp(X_j^T \beta)}, \tag{2.15}$$

onde $L(\beta)$ representa a verossimilhança parcial, X_i é o vetor de covariáveis do indivíduo i, e $R(t_i)$ é o conjunto de indivíduos em risco no tempo t_i. A maximização dessa função

permite estimar os coeficientes β do modelo, ignorando $h_0(t)$, conforme descrito na Eq. 2.1 (SIMON et al., 2011).

Apesar de sua ampla aplicação, o CPH apresenta limitações, como dificuldades em lidar com dados de alta dimensão, suposições restritivas e baixa capacidade de modelar não-linearidades e interações complexas entre variáveis. Portanto, ao aplicá-lo em análises mais complexas, é fundamental considerar essas restrições (LIU et al., 2020; COX, 1972; MONCADA-TORRES et al., 2021).

Cox Proportional Hazards Penalizados

Os modelos de Cox penalizados, descritos por Polsterl (2023) e Simon (2011) (PÖLSTERL, 2023; SIMON et al., 2011), incluem *Ridge*, *Lasso* e *Elastic Net*, que introduzem penalizações nos coeficientes para evitar *overfitting* e melhorar a generalização. Essas penalizações reduzem a complexidade do modelo ao regularizar os coeficientes de acordo com critérios como tamanho ou esparsidade (PÖLSTERL, 2023).

O CPH penalizado *Ridge* (CPH-R) adiciona um termo de penalidade quadrática, $\frac{1}{2}(1-\alpha)\sum_{i=1}^{p}\beta_i^2$, aos coeficientes na função de verossimilhança parcial (2.15). A formulação matemática é dada por:

$$\hat{\beta} = \arg\max_{\beta} \left[\frac{2}{n} \left(\sum_{i=1}^{k} X_{j(i)}^T \beta - log \left(\sum_{j \in R_i} \exp^{(X_j^T \beta)} \right) \right) - \frac{1}{2}(1-\alpha) \sum_{i=1}^{p} \beta_i^2 \right], \quad (2.16)$$

onde $\hat{\beta}$ são os coeficientes estimados e $\alpha \geq 0$ controla a regularização. A penalização *Ridge* (norma l_2) reduz os coeficientes proporcionalmente, sem eliminá-los, permitindo a incorporação de todas as variáveis no modelo (PÖLSTERL, 2023; SIMON et al., 2011).

2- Lasso:

O CPH penalizado *Lasso* (CPH-L) adiciona um termo de penalidade de valor absoluto, $\alpha \sum_{i=1}^{p} \mid \beta_i \mid$, aos coeficientes na função de verossimilhança parcial (2.15). A formulação matemática é semelhante à de *Ridge*, mas a penalidade é dada por:

$$\hat{\beta} = \arg\max_{\beta} \left[\frac{2}{n} \left(\sum_{i=1}^{k} X_{j(i)}^T \beta - \log \left(\sum_{j \in R_i} \exp^{(X_j^T \beta)} \right) \right) - \alpha \sum_{i=1}^{p} \mid \beta_i \mid \right], \quad (2.17)$$

onde $\mid \beta_i \mid$ representa o valor absoluto dos coeficientes. A penalização *Lasso* (norma l_1) reduz alguns coeficientes a zero, promovendo a eliminação de algumas variáveis do modelo (PÖLSTERL, 2023; SIMON et al., 2011).

3- Elastic Net:

O CPH penalizado *Elastic Net* (CPH-EN) combina as penalidades de *Ridge* e *Lasso* aos coeficientes na função de verossimilhança parcial (Eq. 2.15). A formulação matemática é dada por:

$$\hat{\beta} = \arg\max_{\beta} \left[\frac{2}{n} \left(\sum_{i=1}^{k} X_{j(i)}^{T}\beta - \log \left(\sum_{j \in R_i} \exp^{(X_j^T \beta)} \right) \right) - \left(\alpha \sum_{i=1}^{p} \mid \beta_i \mid + \frac{1}{2}(1-\alpha) \sum_{i=1}^{p} \beta_i^2 \right) \right], \tag{2.18}$$

onde $\alpha \in [0, 1]$ define a combinação das normas l_1 e l_2. Esse método permite maior flexibilidade na seleção de variáveis e robustez na modelagem (PÖLSTERL, 2023; SIMON et al., 2011).

Essas formulações são amplamente utilizadas na análise de sobrevivência para ajustar modelos CPH penalizados e selecionar um conjunto relevante de características para prever a sobrevida (PÖLSTERL, 2020; PÖLSTERL, 2023; SIMON et al., 2011; XIAO et al., 2022).

2.2.3.2 *Gradient Boosting Survival*

O GBS não se refere a um modelo específico, mas sim a uma estrutura altamente versátil para otimizar diversas funções de perda. Trata-se de um métodos de AM formulado como um problema de otimização, no qual uma função de perda é continuamente minimizada pela inclusão iterativa de aprendizes fracos, como nos métodos de árvores de decisão. Em cada iteração, uma nova árvore de decisão é gerada para aprender os resíduos do modelo anterior. Esse processo de adição de árvores continua até que não haja mais melhorias substanciais, e a predição final é obtida pela combinação ponderada das predições de cada árvore (LIU et al., 2020; MONCADA-TORRES et al., 2021; HOTHORN et al., 2006).

As predições são combinadas de forma aditiva, onde a adição de cada modelo base melhora (ou "impulsiona") o modelo geral. Matematicamente, seguindo o modelo geral da equação aditiva 2.3, para dados de sobrevida, a função aditiva é dada por:

$$F_{GBS}(X) = \sum_{i=1}^{M} \gamma_i g(X, \theta_i), \tag{2.19}$$

onde $F_{GBS}(X)$ é a função de "impulso"para dados de sobrevida, $M > 0$ denota o número de aprendizes base, $\gamma_i \in \mathbb{R}$ representa o termo de expansão, e g o aprendiz da base parametrizado pelo vetor θ. Os aprendizes de base individuais diferem na configuração de seus parâmetros θ_i, o que é indicado por um indice i (PÖLSTERL, 2023).

Aplicando a função de "impulso"(2.19) ao log da função de verossimilhança parcial do modelo CPH (2.15), obtemos a seguinte expressão:

$$\arg\min_{f} \sum_{i=1}^{n} \delta_i [F_{GBS}(X_i) - log(\sum_{j \in R_i} e^{F_{GBS}(X_j)})], \tag{2.20}$$

substituindo o termo $X^T\beta$ da Eq. 2.3 pela função aditiva $F_{BGS}(X)$ (PÖLSTERL, 2023).

O GBS foi descrito conforme na biblioteca "*Scikit-survival*"(PÖLSTERL, 2023), que se baseia no GB desenvolvido de forma sequencial e gananciosa (FRIEDMAN, 2001; FRIEDMAN, 2002). A função de verossimilhança parcial é otimizada conforme descrito por Ridgeway (1999) (RIDGEWAY, 1999), com a ponderação dos tempos de falha e censura otimizada conforme Hothorn et al. (2006) (HOTHORN et al., 2006). Além disso, a taxa de desistência não nula, onde a regularização é aplicada durante o treinamento, é utilizada conforme descrito por Vinayak e Gilard-Bachrach (2015) (VINAYAK; GILAD-BACHRACH, 2015).

2.2.3.3 *Random Survival Forest*

O RSF é uma abordagem baseada no método RF, proposto por Breiman em 2001, que utiliza o algoritmo de *bootstrap* para desenvolver múltiplas árvores de decisão, onde os nós são divididos com base em uma seleção aleatória de características. A predição final da floresta é a média das predições de cada árvore individual (BREIMAN, 2001; ISHWARAN et al., 2008).

Desenvolvido por Ishwaran et al. (ISHWARAN et al., 2008), o RSF é um conjunto de modelos baseados em árvores, projetado especificamente para a análise de dados de sobrevivência censurados à direita. Ele incorpora informações de censura diretamente nas regras de divisão das árvores. Durante a construção das árvores, a aleatoriedade é promovida ao dividir os nós usando um critério de sobrevivência que considera tanto o tempo de sobrevivência quanto o indicador de ocorrência do evento de interesse. O teste de *log-rank* é utilizado para dividir as árvores, visando maximizar a diferença de sobrevivência entre os nós. Quanto maior o valor do teste, melhor a divisão (LIU et al., 2020; MONCADA-TORRES et al., 2021; XIAO et al., 2022).

O RSF estima a função de sobrevida dos pacientes com o estimador de *Kaplan-Meier* e calcula a função de risco acumulado associada a cada nó terminal com o estimador de *Nelson-Aalen* (LIU et al., 2020; PÖLSTERL, 2020; PÖLSTERL, 2023). O estimador de *Kaplan-Meier* (KAPLAN; MEIER, 1958), denotado por $\hat{S}(t)$, é uma estimativa não paramétrica da função de sobrevida $S(t)$ (2.11) no tempo t. Matematicamente, o estimador de *Kaplan-Meier* é expresso como:

$$\hat{S}(t) = \prod_{t_i \leq T} \left(1 - \frac{d_i}{n_i} \right), \tag{2.21}$$

onde t é o tempo do evento observado em ordem crescente, d_i é o número de eventos no tempo t_i, e n_i é o número de indivíduos em risco nesse tempo. O estimador calcula a probabilidade de sobrevivência como o produto das probabilidades de sobrevivência em cada ponto de tempo observado até o tempo especificado T (KAPLAN; MEIER, 1958; PÖLSTERL, 2020).

A função de *Nelson-Aalen* é uma estimativa não paramétrica da função de risco cumulativo do CPH (2.1) (ISHWARAN; KOGALUR, 2007; ISHWARAN et al., 2008; PÖLSTERL, 2023). Matematicamente, o estimador de *Nelson-Aalen* $\hat{h}(t)$ em um dado tempo t é dado por:

$$\hat{h}(t) = \sum_{t_i \leq t} \frac{d_i}{n_i}, \tag{2.22}$$

onde d_i é o número de eventos ocorridos no tempo t_i, e n_i é o número de indivíduos em risco nesse tempo. O estimador calcula a função de risco cumulativo como a soma das razões entre o número de eventos e o total de indivíduos em risco em cada ponto de tempo (ISHWARAN et al., 2008; ISHWARAN; KOGALUR, 2007; PÖLSTERL, 2023). Na biblioteca "*Scikit-survival*"(PÖLSTERL, 2023), a função de risco cumulativo da RSF $\hat{h}$ é dada por:

$$\sum_{i=1}^{d} \hat{h}(t_i \in X), \tag{2.23}$$

d denota o número total de eventos ocorridos nos dados de treinamento (PÖLSTERL, 2020; PÖLSTERL, 2023).

A formulação matemática do RSF envolve a construção de árvores de sobrevivência usando *bootstrap* e a combinação dessas árvores para formar o modelo de floresta. Ela inclui critérios de divisão com o teste de *log-rank* e a agregação dos resultados das árvores usando os estimadores de *Kaplan-Meier* e *Nelson-Aalen*. Essa abordagem é descrita nos trabalhos de Ishwaran et al. (ISHWARAN et al., 2008; ISHWARAN; KOGALUR, 2007; ISHWARAN; KOGALUR, 2023). O resultado final é obtido pela média das predições de cada árvore, oferecendo uma estimativa robusta da função de sobrevida (LIU et al., 2020; MONCADA-TORRES et al., 2021; XIAO et al., 2022; PÖLSTERL, 2023).

2.2.3.4 *Survival Support Vector Machine*

O SSVM é uma extensão do SVM, tradicionalmente usado em tarefas de classificação e regressão para encontrar o hiperplano ótimo, conforme definido na Equação 2.7. Adaptado para dados de sobrevida, o SSVM emprega iterações para encontrar um hiperplano que minimize o erro e maximize a margem entre eventos e censura. O diferencial do SSVM reside na sua capacidade de ranquear instâncias com base no número de observações, permitindo estimar as probabilidades de sobrevida e predizer o risco de eventos para os pacientes (MONCADA-TORRES et al., 2021; XIAO et al., 2022; PÖLSTERL et al., 2015).

Na análise de sobrevivência, os dados de treinamento consistem em n tríades (X_i, Y_i, δ_i), onde X_i é o vetor de características, Y_i é o vetor de tempos (t_i) de sobrevida

ou censura, e δ_i é o indicador binário do evento de interesse. O objetivo é minimizar a função do SSVM, $F_{SSVM}(w)$, definida por Pölsterl et al. (2015) como:

$$F_{SSVM}(w) = \frac{1}{2}w^T w + \frac{q}{2}\sum_{i,j\in P}\max(0, 1-(w^T X_i - w^T X_j))^2, \tag{2.24}$$

onde w é o vetor de coeficientes perpendicular ao hiperplano de separação ($w \in \mathbb{R}^d$), X é o vetor de características e $q > 0$ é o parâmetro de regularização. O conjunto $P = (i,j) \mid Y_i > Y_j \wedge \delta_j = 1$ define os pares de amostras comparáveis usados no treinamento (PÖLSTERL et al., 2015).

O SSVM para classificação e regressão linear na análise de sobrevivência, conforme descrito na biblioteca "*Scikit-survival*"(PÖLSTERL, 2023), é expresso como:

$$\arg\min_{w,b}\frac{1}{2}w^T w + \frac{\lambda}{2}[q\sum_{i,j\in P} max(0, 1-(w^T x_i - w^T x_j))^2 + (1-q)\sum_{i=1}^{n}(\xi_{w,b}(X_i, Y_i, \delta_i))^2], \tag{2.25}$$

sujeito a: se $\delta_i = 0$, $\xi_{w,b}(X_i, Y_i, \delta_i) = \max(0, Y_i - (w^T x_i - w^T x_j - b))$, e se $\delta_i = 1$, $\xi_{w,b}(X_i, Y_i, \delta_i) = \max(0, Y_i - (w^T X_i - b))$. O hiperparâmetro $\lambda > 0$ determina a quantidade de regularização, enquanto o hiperparâmetro $q \in [0; 1]$ determina o *trade-off* entre os objetivos de classificação e regressão (PÖLSTERL et al., 2015; PÖLSTERL, 2020; PÖLSTERL, 2023).

Para o modelo não linear, o *Kernel Survival Support Vector Machine* (KSSVM), usando a *Radial Basis Function Kernel* (RBF), emprega a distância euclidiana quadrática $|x - x'|^2$ entre dois vetores de características. A função RBF é descrita por:

$$\kappa(x - x') = \exp^{(-\iota\|x-x'\|^2)}, \tag{2.26}$$

onde κ representa a similaridade entre $(x - x')$ das variáveis de entrada, e ι é definido como $\iota = \frac{1}{2\sigma^2}$, sendo σ um parâmetro livre que controla a largura da similaridade Kernel (DUVENAUD, 2014; PÖLSTERL et al., 2016; PÖLSTERL, 2023).

A aplicação de métodos de AM na oncologia tem se mostrado fundamental para o desenvolvimento de modelos prognósticos, proporcionando uma melhor compreensão do processo evolutivo e da resposta terapêutica ao CM. Em resumo, cada métodos de AM explora aspectos distintos da análise de sobrevida: o modelo de Regressão de Cox avalia o impacto das covariáveis nas taxas de risco de ocorrência do evento de interesse, permitindo a identificação de fatores prognósticos (COX, 1972); o GBS, formulado como um problema de otimização, utiliza árvores de decisão iterativas para fornecer predições robustas por meio da combinação ponderada de cada árvore (LIU et al., 2020; MONCADA-TORRES et al., 2021); o RSF aplica o método de *bootstrap* e a seleção aleatória de características para desenvolver árvores de decisão, incorporando informações sobre censura nas regras

de divisão e fornecendo predições robustas por meio da média das árvores (BREIMAN, 2001; ISHWARAN et al., 2008); por fim, o SSVM linear e o KSSVM não linear encontram o hiperplano no espaço de características de modo a maximizar a margem entre eventos e censura (PÖLSTERL et al., 2015). Dessa forma, esses métodos se revelam valiosos para o acompanhamento do curso clínico-terapêutico e para a análise de sobrevida em pacientes com CM (CARVALHO et al., 2023; CARVALHO et al., 2024; LIU et al., 2020; MONCADA-TORRES et al., 2021; PINHEIRO et al., 2022; XIAO et al., 2022; TENG et al., 2019).

3 REVISÃO SISTEMÁTICA

O CM continua sendo uma das principais causas de mortalidade entre mulheres no mundo, ressaltando a importância de predições precisas de sobrevida para orientar tratamentos e otimizar desfechos clínicos. Nesse contexto, os métodos de AM surgem como ferramentas promissoras, aprimorando a acurácia das predições de sobrevida. Estudos recentes têm evidenciado o avanço significativo na aplicação de técnicias de AM para este propósito, com abordagens inovadoras que contribuem para a melhoria da predições no contexto do CM (LI et al., 2021; MIN et al., 2021).

Para esta RS, foram selecionados artigos que utilizaram métodos de AM na análise de sobrevida, seguindo uma metodologia rigorosa conduzida por um (1) único avaliador. Os critérios de inclusão e elegibilidade foram cuidadosamente definidos para assegurar a seleção de estudos que reportassem tanto o desempenho dos diferentes métodos de AM quanto as métricas de avaliação e validação em dados clínicos (BRASIL; CIêNCIA TECNOLOGIA, 2012; PAGE et al., 2023; PRISMA, 2021).

Os achados desta RS sugerem uma tendência positiva na utilização de métodos de AM para predições de sobrevida em CM. As abordagens supervisionadas demonstraram alta eficácia, enquanto as não supervisionadas também evidenciaram avanços relevantes (LI et al., 2021; MIN et al., 2021; SHUKLA et al., 2018; TENG et al., 2019). Além disso, a revisão aponta a necessidade de pesquisas adicionais para validar a generalização dos modelos, bem como aprimorar a sua aplicabilidade prática (HUANG et al., 2022; MONCADA-TORRES et al., 2021).

Assim, esta RS não só reforça a crescente relevância dos métodos de AM na predição de sobrevida para CM feminino, como também destaca a necessidade de investigações contínuas para integrar esses modelos de forma eficaz na prática clínica. A evolução e adaptação constante dos métodos de AM prometem aprimorar progressivamente a precisão das predições, trazendo impactos diretos no manejo clínico das pacientes diagnosticadas com CM.

3.1 TENDÊNCIAS E INOVAÇÕES DA MODELAGEM COMPUTACIONAL PARA O CÂNCER DE MAMA

As tendências e inovações na modelagem computacional para o CM refletem avanços significativos na predição da sobrevida e no direcionamento do tratamento. Com o advento dos MAM, há um foco crescente na individualização do cuidado, abordando a complexidade biológica da doença em cada paciente (GU et al., 2020; KALAFI et al., 2019; LI et al., 2022; BOERI et al., 2020). Esses métodos proporcionam uma análise abrangente e precisa, integrando múltiplas variáveis clínicas e biológicas para prever desfechos clínicos com maior confiabilidade (ZHOU et al., 2021; XIN et al., 2022; DENG et al., 2021; KLEINLEIN;

RIAÑO, 2019; HAQUE et al., 2022; GANGGAYAH et al., 2019; AFSHAR et al., 2021; HUANG et al., 2022). Além disso, a aplicação de técnicas não supervisionadas tem contribuído para identificar padrões ocultos nos dados, ampliando nosso entendimento sobre a progressão da doença e suas interações (AFSHAR et al., 2021; SHUKLA et al., 2018).

A pesquisa sobre o CM, a neoplasia mais prevalente entre mulheres, é complexa devido à sua relação com uma variedade de fatores biológicos e seu estadiamento, que impactam diretamente na evolução da doença e na sobrevida das pacientes, como discutido na Subseção 2.1.2. A personalização dos tratamentos, adaptando-os ao perfil individual de cada paciente, exige uma abordagem multidisciplinar que permita identificar esses fatores de risco com precisão. Neste sentido, diversas técnicas, desde métodos estatísticos convencionais até MAM, têm sido empregadas para prever a sobrevida, com destaque para a análise de regressão de Cox, amplamente aceita na na oncologia. Contudo, os métodos de AM têm ganhado espaço recentemente, oferecendo vantagens significativas (BOERI et al., 2020; KALAFI et al., 2019; MONCADA-TORRES et al., 2021; SHUKLA et al., 2018).

Apesar dos desafios contínuos na detecção precoce e no tratamento eficaz do CM, a tecnologia desempenha um papel significativo na coleta e análise análise de grandes volumes de dados. Com o crescimento da quantidade e complexidade dos dados disponíveis, surge a necessidade de algoritmos mais robustos, como os MAM, que conseguem classificar, prever e estimar desfechos clínicos de forma eficaz, contribuindo para o prognóstico e tratamento de pacientes com CM (HAQUE et al., 2022; DENG et al., 2021). A aplicação desses métodos na análise de dados da patologia e na identificação de fatores prognósticos apresenta uma abordagem promissora, permitindo intervenções mais precisas, inclusive na identificação de pacientes com alto risco de recorrência pós-operatória (GANGGAYAH et al., 2019; HUANG et al., 2022; SEDIGHI-MAMAN; MONDELLO, 2021; ZHOU et al., 2021).

A expansão do uso da IA está transformando a resolução de problemas em diversas áreas, incluindo a oncologia. Neste cenário, os métodos de AM têm se destacado por fornecer *insights* preditivos que auxiliam na tomada de decisões clínicas para o CM (LAI et al., 2019; LI et al., 2021; KöHN-LUQUE et al., 2020; NAVE, 2020). Esses avanços fomentam o desenvolvimento de modelos computacionais preditivos de sobrevida, com grande potencial e aplicabilidade na prática médica (LI et al., 2021; OKAGBUE et al., 2021; TAPAK et al., 2019).

O tratamento personalizado do CM é viabilizado pela identificação de biomarcadores específicos, que categorizam os pacientes em subtipos e orientam as escolhas terapêuticas individualizadas. Estudos comprovam a eficácia das terapias personalizadas, nas quais modelos computacionais simulam e avaliam os efeitos de diferentes intervenções terapêuticas, permitindo decisões mais precisas para cada paciente (KöHN-LUQUE et al.,

2020; LAI et al., 2019). Em vista da relevância do CM e da importância da análise de sobrevida no delineamento do curso clínico-terapêutico, observa-se um aumento na adoção dos MAM, que utilizam combinações de variáveis para gerar prognósticos personalizados (OKAGBUE et al., 2021).

A análise de sobrevida atrai crescente interesse na comunidade de IA, impulsionada pelos avanços tecnológicos e suas aplicações clínicas. Embora os métodos de AM ofereçam previsões mais precisas que abordagens estatísticas convencionais, facilitando diagnóstico, tratamento e prognóstico do câncer, ainda há desafios, como a transparência dos modelos, que podem ser vistos como "caixas-pretas", o que pode afetar a confiança de médicos e pacientes nesses métodos (TENG et al., 2019; XIAO et al., 2022; MONCADA-TORRES et al., 2021).

Em uma busca sistemática nas bases de dados *Google Scholar*, *Medline* (*PubMed*) e *Scopus*, foram identificados artigos de revisão relevantes ao tema. Após análise criteriosa, destacaram-se os seguintes estudos:

1. **Min et al. (2021):** Este estudo explora modelos prognósticos para CM, analisando aspectos clinicopatológicos, de expressão gênica e MAM. Os autores identificaram variáveis comuns a cada tipo de modelo, como idade, comprometimento linfonodal, grau e tamanho do tumor, e status do receptor hormonal. A estratificação baseada em informações genômicas e clínicas permite identificar riscos e evita tratamentos desnecessários, enquanto a análise de imagens aprimora a previsão do prognóstico (MIN et al., 2021).

2. **Li et al. (2021):** Focado no desenvolvimento e validação de métodos de AM para previsão de sobrevida em CM, este estudo utiliza diretrizes PRISMA e analisa artigos de alta qualidade selecionados nas bases *PubMed*, *Embase* e *Web of Science Core*. Os métodos de AM mostraram-se promissores na predição de sobrevida, embora os autores recomendem melhorias no pré-processamento de dados e padronização na seleção de variáveis. Pesquisas interdisciplinares podem fortalecer a eficácia desses modelos na prática clínica (LI et al., 2021).

Apesar dos avanços na área, ainda são escassos os estudos de revisão que oferecem uma análise abrangente sobre o uso de métodos de AM na predição de sobrevida em pacientes com CM. Nesse contexto, esta RS se torna relevante, pois busca contribuir tanto para a formulação de uma proposta de pesquisa da tese quanto para o aprimoramento da compreensão dessas técnicas no cenário do CM. O objetivo desta RS é avaliar a aplicação dos métodos de AM para predição de sobrevida em pacientes com CM, validados com dados clínicos. Para isso, foi realizada uma busca detalhada nas principais bases de dados, identificando estudos relevantes que aplicaram esses métodos à análise de sobrevida. Além de destacar o uso desses modelos, a RS visa mapear os principais métodos empregados,

comparar os estudos mais relevantes e identificar suas contribuições para a construção da pesquisa.

3.2 MÉTODOS DA REVISÃO SISTEMÁTICA

A RS é uma metodologia de síntese de evidências que permite a análise crítica e a interpretação das pesquisas disponíveis sobre uma questão específica, área do conhecimento ou fenômeno de interesse. Esse método distingue-se por sua abordagem explícita e sistemática para identificar, selecionar e avaliar a qualidade das evidências, assegurando confiabilidade, rigor e transparência nos resultados (BRASIL; CIêNCIA TECNOLOGIA, 2012).

Nesta RS, a metodologia foi desenvolvida em conformidade com as diretrizes estabelecidas pelo Ministério da Saúde (MS) para a elaboração de revisões sistemáticas e metanálises de ensaios clínicos randomizados (BRASIL; CIêNCIA TECNOLOGIA, 2012). Para assegurar a transparência e o rigor em todas as etapas do processo, serão seguidas as recomendações do protocolo PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*). Esse protocolo define detalhadamente as etapas de definição, identificação, seleção, elegibilidade e inclusão dos estudos, além de especificar os critérios correspondentes para cada fase (PRISMA, 2021).

Para otimizar a organização e análise das publicações em cada uma dessas etapas, foi utilizado o *Software Rayyan*, que proporciona assistência de qualidade e agilidade na elaboração e execução das etapas de uma RS (OUZZANI et al., 2016). Embora o protocolo PRISMA normalmente requeira múltiplos avaliadores para assegurar maior controle de qualidade, esta revisão foi conduzida por um único avaliador. A seguir, detalham-se as etapas metodológicas aplicadas, conforme o protocolo PRISMA (PRISMA, 2021).

3.2.1 DEFINIÇÕES

O acrônimo PICO é amplamente utilizado para estruturar os elementos essenciais de uma pergunta de pesquisa, abrangendo População (P), Intervenção (I), Comparação (C) e Desfecho (O - "*Outcomes*"). Esse modelo facilita a definição clara e objetiva dos objetivos da pesquisa, assegurando uma direção focada para a investigação (BRASIL; CIêNCIA TECNOLOGIA, 2012). Para esta RS, o PICO foi definido conforme descrito a seguir:

P **População:** Pacientes diagnosticadas com CM feminino.

I **Intervenção:** métodos de AM aplicados à predição de sobrevida em CM.

C **Comparação:** Validação dos modelos preditivos com dados clínicos.

O **Desfecho:** Eficácia dos métodos de AM no apoio à tomada de decisões clínicas.

A questão norteadora desta pesquisa, formulada a partir do modelo PICO, é: “Quais são os métodos de Aprendizado de Máquina aplicados à predição de sobrevida para o câncer de mama feminino, e como esses modelos têm sido validados com dados clínicos?”

As bases de dados utilizadas para a pesquisa foram *Google Scholar*, *Medline* (*PubMed*) e *Scopus*. As palavras-chave, definidas com base nos descritores MeSH (*Medical Subject Headings*) para refletir o PICO e a questão norteadora, incluem "*breast neoplasms*", "*machine learning*"e "*survival*". A pesquisa foi delimitada a artigos publicados entre 2018 e 2022, estabelecendo um período de interesse recente.

3.2.2 IDENTIFICAÇÃO

Na etapa de identificação, será implementada a estratégia de busca nas bases de dados selecionadas, com o objetivo de identificar artigos originais. As palavras-chave foram configuradas para refletir o PICO e a questão norteadora, incluindo descritores como "*breast neoplasms*", "*machine learning*", "*survival*"e "*not review*". A busca será restrita a artigos originais publicados entre 2018 e 2022. Os resultados de cada base serão compilados e importados para o *Software Rayyan*, que auxiliará na organização, análise e remoção de duplicados, facilitando as etapas subsequentes da RS.

3.2.3 SELEÇÃO

Nesta etapa, serão aplicados critérios rigorosos de inclusão (1) e exclusão (2) para garantir a relevância e a qualidade dos estudos selecionados. Os critérios são os seguintes:

1. **Inclusão:**
    - Estudos originais que utilizam métodos de AM para prever a sobrevida de pacientes diagnosticadas com CM.
    - Estudos que validam modelos preditivos com dados clínicos oncológicos usuais.
    - Publicações que apresentam informações completas sobre a população, intervenção, comparação e desfecho, conforme definido pelo PICO.
2. **Exclusão:**
    - Estudos que não têm como foco a predição de sobrevida de pacientes com CM.
    - Estudos não originais, como revisões, RS, meta-análises e literatura cinzenta (ex.: cartas editoriais, dissertações, teses).
    - Estudos que se concentram em variáveis genéticas ou moleculares, em vez de dados clínicos usuais da prática oncológica.
    - Estudos que não incluem pacientes diagnosticadas com CM como parte da população de estudo.

3.2.4 ELEGIBILIDADE

Na etapa de elegibilidade, os títulos e resumos dos artigos pré-selecionados serão avaliados para verificar sua relevância e aderência aos critérios de inclusão estabelecidos. Os critérios de elegibilidade são:

1. **Relevância do desfecho:** Serão excluídos estudos que não se concentram na predição da sobrevida de pacientes com CM, uma vez que este é o foco principal da RS (*Wrong outcome*).

2. **Desenho de estudo:** Serão excluídos estudos que utilizam dados genéticos como variáveis para predição prognóstica, pois esta RS tem como objetivo analisar especificamente a validação com variáveis clínicas usuais (*Wrong study design*).

3. **Tipo de publicação:** Serão excluídas revisões sistemáticas, meta-análises e literatura cinzenta, uma vez que o foco está em avaliar estudos primários que investigam diretamente a aplicação de métodos de AM na predição da sobrevida para o CM (*Wrong publication type*).

4. **População de estudo:** Serão excluídos estudos que não envolvam pacientes diagnosticadas com CM, já que os resultados desses estudos não são diretamente aplicáveis ao contexto desta RS (*Wrong population*).

Esses critérios serão rigorosamente aplicados para garantir que os artigos selecionados ofereçam evidências essenciais sobre a população de estudo, intervenção, comparação e desfecho, alinhando-se à questão norteadora e ao objetivo desta RS.

3.2.4.1 INCLUSÃO

Na etapa de inclusão, os textos completos dos artigos selecionados serão avaliados na íntegra, considerando sua contribuição para o objetivo da pesquisa. A análise detalhada determinará a elegibilidade final dos artigos para inclusão na RS. Para cada estudo incluído, serão extraídas as seguintes informações:

1. **Identificação dos artigos incluídos:** Para facilitar a identificação de cada estudo por autores e ano de publicação.

2. **Características da população de estudo:** PPara garantir que os estudos incluídos investigaram pacientes diagnosticadas com CM, conforme os critérios de inclusão.

3. **Métodos de AM utilizados nos artigos:** Para compreender os métodos de AM aplicados na predição da sobrevida.

4. **Métricas de desempenho e validação dos métodos de AM:**Para avaliar a robustez e a generalização dos modelos preditivos com dados clínicos.

5. **Principais resultados da aplicação dos métodos de AM:** Para analisar os desfechos em relação à predição da sobrevida no CM.

Os dados extraídos serão sintetizados quantitativa e qualitativamente para descrever a população do estudo, os métodos de AM utilizados, as métricas de desempenho reportadas e os principais resultados dos estudos incluídos. Vale destacar que, na maioria dos artigos analisados nesta RS, os métodos de seleção de atributos e otimização não foram explicitamente descritos e, em muitos casos, sequer mencionados, razão pela qual não foram abordados nesta revisão.

O processo metodológico será conduzido conforme as diretrizes do MS e do protocolo PRISMA, assegurando rigor e transparência em todas as etapas da RS. A adoção do PRISMA fortalece a robustez metodológica e assegura maior clareza, facilitando a tomada de decisões clínicas baseadas em evidências (BRASIL; CIêNCIA TECNOLOGIA, 2012; PAGE et al., 2023; PRISMA, 2021). Além disso, a análise dos dados extraídos incluirá uma revisão crítica dos estudos selecionados, destacando suas principais descobertas, limitações e as implicações clínicas dos métodos de AM na predição da sobrevida em pacientes com CM feminino.

3.3 RESULTADOS DA REVISÃO SISTEMÁTICA

Nesta seção, apresentamos os principais resultados dda RS sobre a aplicação de métodos de AM na predição da sobrevida no CM. O CM é uma doença prevalente e complexa, afetando predominantemente mulheres e representando uma das principais causas de mortalidade relacionada ao câncer em todo o mundo. A previsão precisa da sobrevida desempenha um papel fundamental na prática clínica oncológica, pois oferece *insights* essenciais para orientar decisões terapêuticas individualizadas e a prestação de cuidados paliativos. Dessa forma, a predição da sobrevida no CM constitui uma área de pesquisa de grande relevância, com potencial para impactar diretamente o curso clínico e os resultados das pacientes afetadas por essa condição devastadora (MIN et al., 2021; LI et al., 2021; GU et al., 2020; SEDIGHI-MAMAN; MONDELLO, 2021; TENG et al., 2019; SHUKLA et al., 2018; HUEMAN et al., 2018; MONCADA-TORRES et al., 2021; KALAFI et al., 2019; JANSEN et al., 2020; LI et al., 2022; BOERI et al., 2020; ZHOU et al., 2021; XIN et al., 2022; LIU et al., 2020; DENG et al., 2021; KLEINLEIN; RIAÑO, 2019; HAQUE et al., 2022; GANGGAYAH et al., 2019; AFSHAR et al., 2021; HUANG et al., 2022).

Os modelos prognósticos para o CM variam desde abordagens clinicopatológicas até ensaios de expressão gênica e MAM. Nos últimos anos, os métodos de AM têm atraído

crescente atenção devido ao seu grande potencial para fornecer informações precisas e personalizadas na previsão da sobrevida e na seleção de tratamentos adequados para pacientes com CM (MIN et al., 2021; LI et al., 2021). Além disso, novas abordagens têm sido desenvolvidas, como a integração de fatores prognósticos adicionais ao sistema de estadiamento do câncer, utilizando algoritmos de métodos de AM para construir modelos de prognóstico mais precisos (HUEMAN et al., 2018).

Embora os métodos de AM mostrem grande potencial, ainda existem desafios significativos a serem superados para sua plena adoção na prática clínica. Um dos principais obstáculos é a opacidade dos modelos, o que pode dificultar a sua interpretação e confiança por parte dos profissionais de saúde. Para superar essa limitação, métodos de explicabilidade, como o *Local Interpretable Model-agnostic Explanations* (LIME) e *SHapley Additive exPlanations* (SHAP), têm sido cada vez mais empregados para esclarecer as previsões dos modelos e aumentar sua confiabilidade (GU et al., 2020; JANSEN et al., 2020).

Apesar desses desafios, a aplicação de métodos de AM na predição da sobrevida do CM é promissora e tem o potencial de contribuir significativamente para a melhoria do cuidado ao paciente. O uso dessas tecnologias pode não apenas otimizar a escolha dos tratamentos, mas também fornecer uma visão mais precisa do prognóstico das pacientes, permitindo um acompanhamento mais eficaz e personalizado (SEDIGHI-MAMAN; MONDELLO, 2021; MONCADA-TORRES et al., 2021; SHUKLA et al., 2018; KALAFI et al., 2019; LIU et al., 2020; DENG et al., 2021; BOERI et al., 2020; GANGGAYAH et al., 2019; AFSHAR et al., 2021; HAQUE et al., 2022; HUANG et al., 2022; XIN et al., 2022).

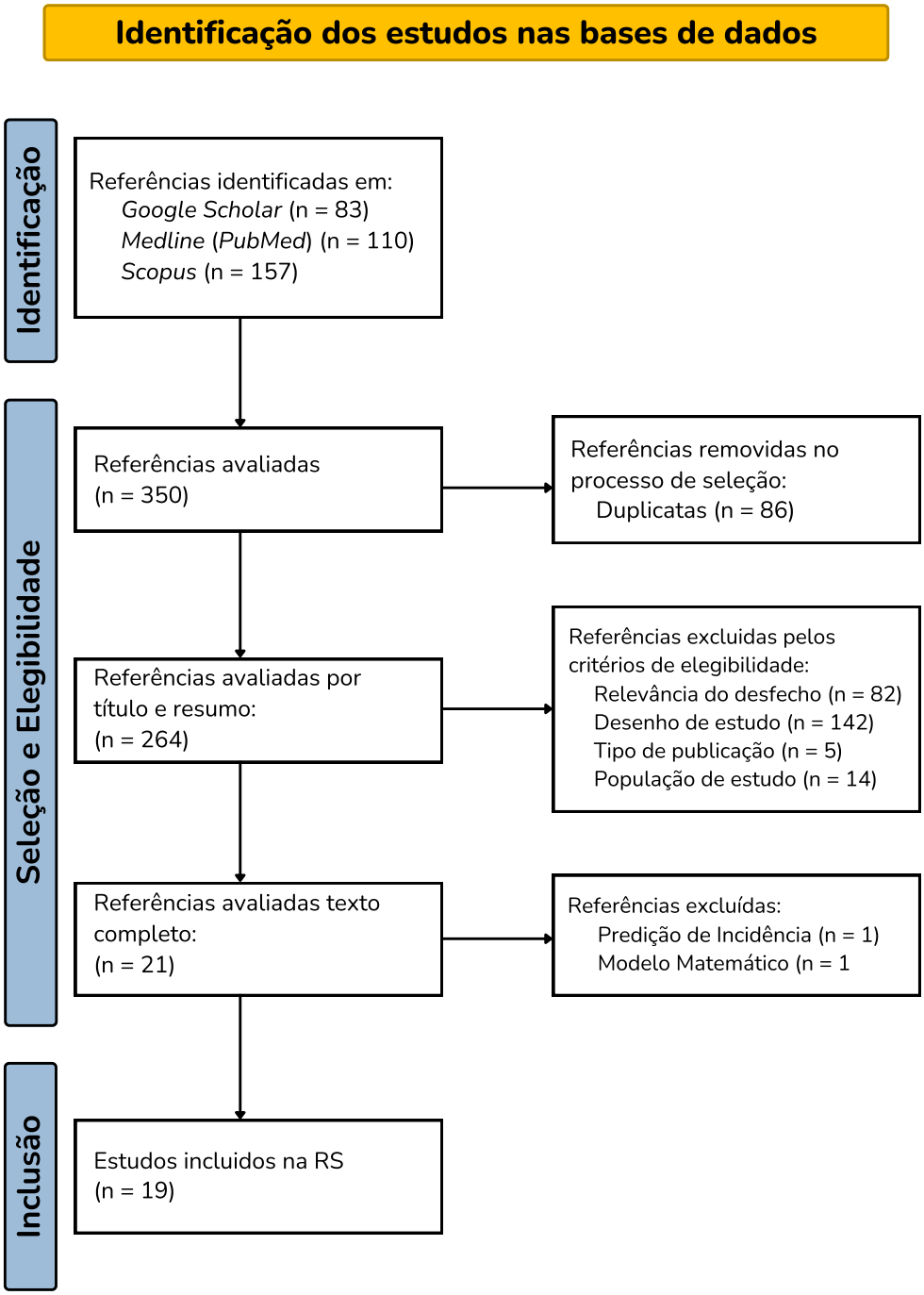
A Figura 9 resume as etapas metodológicas da revisão, abrangendo a identificação, seleção, elegibilidade, inclusão e análise dos artigos. Essa visão geral proporciona uma compreensão aprofundada da literatura existente sobre o uso de métodos de AM na predição da sobrevida no CM, evidenciando tanto os avanços alcançados quanto os desafios que ainda precisam ser enfrentados. Assim, destaca-se a importância contínua da pesquisa nesta área, que tem o potencial de transformar a forma como o prognóstico do CM é abordado na prática clínica.

3.3.1 IDENTIFICAÇÃO

Na etapa de identificação dos artigos, realizamos buscas nas bases de dados selecionadas, conforme ilustrado na Figura 8 com o objetivo de incluir apenas artigos originais. Para isso, foram empregadas palavras-chave e descritores MeSH específicos relacionados ao CM, MAM, sobrevida, excluindo artigos de revisões. As buscas foram realizadas nas seguintes bases de dados, resultando no número de artigos descrito abaixo:

1. *Google Scholar:* A busca foi realizada utilizando os descritores "*breast neoplasms*"OR

Figura 9 – Fluxograma PRISMA 2020.



Fonte: Adaptado de PRISMA (2020).

"*breast neoplasmmachine learning*"*survival -review*. Essa busca resultou na identificação de 83 trabalhos.

2. *Medline* (*PubMed*): Utilizou-se a combinação de descritores *breast neoplasms*"OR "*breast neoplasm*"AND "*machine learning*"AND *survival* NOT *review*, o que resultou na coleta de 110 publicações.

3. *Scopus*:A pesquisa foi conduzida com os descritores "*breast neoplasms*"OR "*breast neoplasm*"AND "*machine learning*"AND *survival* AND NOT *review*, identificando 157 trabalhos.

.

Ao final dessa etapa, um total de 350 artigos foi identificado. Esses documentos foram então compilados e importados para o *Software Rayyan*, onde foram organizados para análise nas etapas subsequentes RS.

3.3.2 SELEÇÃO & ELEGIBILIDADE

Após a importação dos arquivos das bases de dados para o *Software Rayyan*, foram removidos 86 artigos duplicados, restando 264 artigos para avaliação. A seleção inicial foi realizada com base nos critérios de inclusão, exclusão e elegibilidade, através da análise dos títulos e resumos dos artigos pré-selecionados. Para essa triagem, foram aplicados os seguintes identificadores no *Rayyan*:

- *Wrong outcome* (resultado incorreto)
- *Wrong study design* (design do estudo incorreto)
- *Wrong publication type* (tipo de publicação incorreto)
- *Wrong population* (população incorreta)

Na primeira etapa da elegibilidade, foram excluídos 243 artigos que não atendiam aos critérios definidos. Em seguida, os artigos pré-selecionados foram analisados na íntegra, resultando na exclusão adicional de dois artigos. A seguir, detalhamos os critérios de exclusão utilizados:

1. ***Wrong outcome***: Foram excluídos 84 artigos originais que não abordavam a questão central da revisão, focando em desfechos distintos. Os artigos excluídos abordavam áreas como diagnóstico (49), tratamento (19), biomarcadores (5), modelos matemáticos (4), fatores de risco e qualidade de vida (3), e incidência (2).

2. ***Wrong study design***: Nesta categoria, foram excluídos 142 artigos que utilizavam características genéticas, dos quais 56 correspondiam a doenças gerais e 86 especificamente ao CM.

3. ***Wrong publication type***: Foram removidos 5 artigos que não eram originais. Esses incluíam artigos de bibliometria e revisões (2) e literatura cinzenta (cartas editoriais (2) e uma dissertação (1)).

4. ***Wrong population***: Foram excluídos 14 estudos que envolviam uma população inadequada. Isso incluiu artigos que utilizavam dados sintéticos e simulados (2), dados de outros tipos de câncer (pulmão, fígado, ovário e nasofaríngeo) (4) e dados *in vitro* e *in vivo* (8).

A aplicação rigorosa desses critérios, como ilustrado na 9 permitiu a seleção dos artigos mais relevantes e adequados para a análise detalhada e subsequente inclusão na RS.

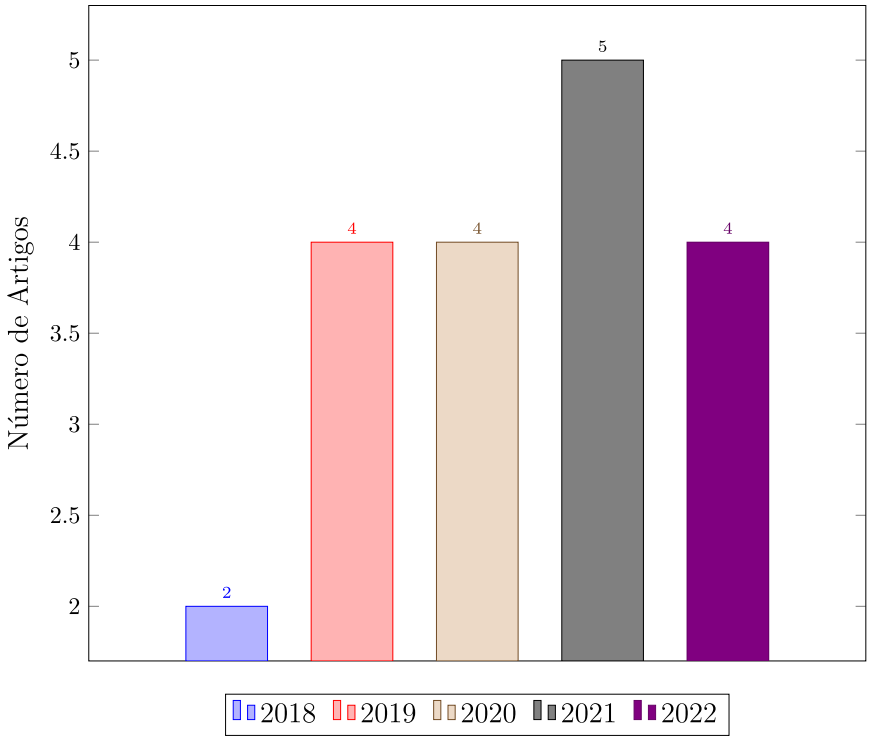
3.3.3 INCLUSÃO E ANÁLISE

Na etapa de inclusão, os textos completos dos 21 estudos previamente selecionados foram analisados para verificar sua relevância em relação à questão central da pesquisa, conforme ilustrado na Figura 9. Após uma análise criteriosa, foram excluídos 2 artigos, sendo um por focar na predição de incidência em vez de sobrevida e o outro por utilizar um modelo exclusivamente matemático. Assim, 19 artigos foram considerados adequados e incluídos na análise final da revisão sistemática, por atenderem aos critérios estabelecidos.

A Figura 10 apresenta a distribuição anual dos artigos incluídos nesta RS, distribuídos da seguinte forma: 2 artigos de 2018 (10,5%) (SHUKLA et al., 2018; HUEMAN et al., 2018), 4 de 2019 (21,1%) (TENG et al., 2019; KALAFI et al., 2019; KLEINLEIN; RIAÑO, 2019; GANGGAYAH et al., 2019), 4 de 2020 (21,1%) (GU et al., 2020; JANSEN et al., 2020; BOERI et al., 2020; LIU et al., 2020), 5 de 2021 (26,3%) (SEDIGHI-MAMAN; MONDELLO, 2021; MONCADA-TORRES et al., 2021; ZHOU et al., 2021; DENG et al., 2021; AFSHAR et al., 2021), e 4 de 2022 (21,1%) (XIN et al., 2022; HAQUE et al., 2022; HUANG et al., 2022).

Cada artigo incluído foi submetido a uma análise detalhada, considerando informações chave descritas no Quadro 3. A análise e discussão dos dados extraídos de cada estudo são apresentadas nas próximas subseções, oferecendo uma visão abrangente sobre a aplicação de métodos de AM na predição da sobrevida em casos de CM. Esta análise destaca tanto os avanços quanto as limitações identificadas na literatura, contribuindo para o desenvolvimento de modelos computacionais mais robustos e úteis na prática clínica oncológica.

Figura 9 – Gráfico do número de artigos incluídos na RS por ano de publicação..



Fonte: Elaborado pela autora (2024).

Quadro 3 – Informações extraídas de cada artigo incluído na RS.

Informações	**Descrição**
Características Populações	Informações sobre o país de origem e fatores clínicos relevantes que influenciam os resultados.
MAM	Descrição dos algoritmos e técnicas utilizados para a predição da sobrevida.
Desempenho e Validação	Estratégias para avaliar e validar o desempenho dos métodos de AM utilizando dados clínicos.
Principais Resultados	Análise da eficácia dos métodos de AM na predição da sobrevida de pacientes com CM.

Fonte: Elaborado pela autora(2024).

3.3.3.1 Características das Populações

A maioria dos estudos incluídos nesta RS foi conduzida em países como os Estados Unidos (EUA) e a China, abrangendo amostras variando de 210 a 510.000 pacientes. Esses estudos utilizaram, em sua maioria, dados clínicos secundários ou registros hospitalares de pacientes diagnosticadas com CM para a aplicação de métodos de AM na predição da sobrevida (SEDIGHI-MAMAN; MONDELLO, 2021; SHUKLA et al., 2018; KALAFI et al., 2019; GANGGAYAH et al., 2019; AFSHAR et al., 2021; MONCADA-TORRES et

al., 2021; JANSEN et al., 2020; KLEINLEIN; RIAÑO, 2019). As pesquisas abordaram diversos aspectos, incluindo prognóstico (HUEMAN et al., 2018; BOERI et al., 2020; TENG et al., 2019; LI et al., 2022; LIU et al., 2020), recorrência (GU et al., 2020; HAQUE et al., 2022; XIN et al., 2022), e mortalidade (ZHOU et al., 2021).

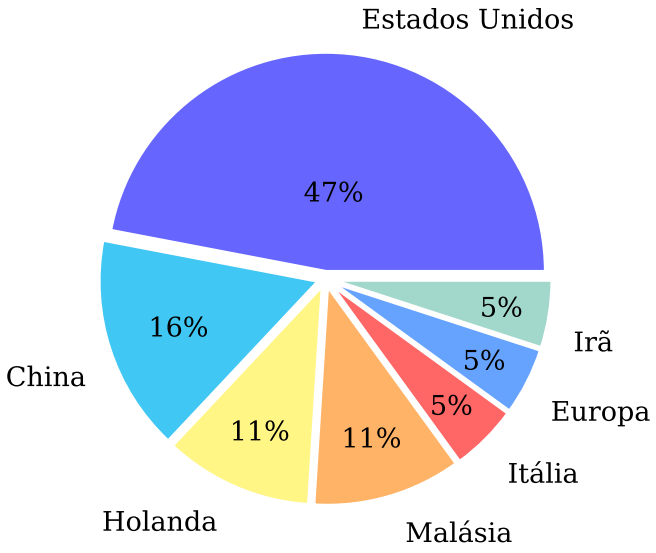
O Quadro 4 sintetiza as principais características extraídas de cada estudo, detalhando a composição das amostras, as características patológicas e a origem das populações estudadas. A maior parte dos estudos focou em informações clínicas de pacientes com diagnóstico de CM (GU et al., 2020; SEDIGHI-MAMAN; MONDELLO, 2021; SHUKLA et al., 2018; HUEMAN et al., 2018; KALAFI et al., 2019; BOERI et al., 2020; DENG et al., 2021; GANGGAYAH et al., 2019; AFSHAR et al., 2021). Alguns estudos, entretanto, concentraram-se em subpopulações específicas, como casos de CM CDI e CLI (TENG et al., 2019; HAQUE et al., 2022), pacientes com CM Não Metastático (NM) (MONCADA-TORRES et al., 2021; JANSEN et al., 2020; ZHOU et al., 2021), com Metástase Óssea (MO)(LI et al., 2022), com o receptor HER2 positivo (HER2+) (XIN et al., 2022), e estadiamentos I a III (I-III) (HUANG et al., 2022). . Adicionalmente, um dos estudos abrangeu toda a trajetória do CM, incluindo os estágios *in situ*, localizado, regional e distante (KLEINLEIN; RIAÑO, 2019).

Quadro 4 – Características clínicas e regionais das amostras dos 19 estudos analisados na RS.

Artigos	**Pacientes**	**Patologia**	**País**
Afshar et al. (2021)	856	CM	Irã
Boeri et al. (2020)	610	CM	Itália
Deng et al. (2021)	45.085	CM	EUA
Ganggayah et al. (2019)	8.066	CM	Malásia
Gu et al. (2020)	217	CM	China
Haque et al. (2022)	4.024	CM (CDI e CLI)	EUA
Huang et al. (2022)	4.696	CM (TN)	EUA
Hueman et al. (2018)	514.213	CM	EUA
Kalafi et al. (2019)	4.902	CM	Malásia
Kleinlein and Riaño (2019)	312.446	CM (Estágios)	EUA
Jansen et al. (2020)	46.284	CM (NM)	Holanda
Li et al. (2022)	15.129	CM (MO)	EUA
Liu et al. (2020)	6.012	CM (Estágio I-III)	China
Moncada-Torres et al. (2021)	36.658	CM (NM)	Holanda
Sedighi-Maman e Mondello (2021)	51.408	CM	EUA
Shukla et al. (2018)	85.189	CM	EUA
Teng et al. (2019)	5.279	CM (CDI e CLI)	EUA
Xin et al. (2022)	6.486	CM (HER2+)	China
Zhou et al. (2021)	1.661	CM (NM)	Europa

Fonte: Elaborada pela autora (2024).

Figura 11 – Gráfico da origem dos dados dos artigos incluídos na RS.



Fonte: Elaborado pela autora (2024).

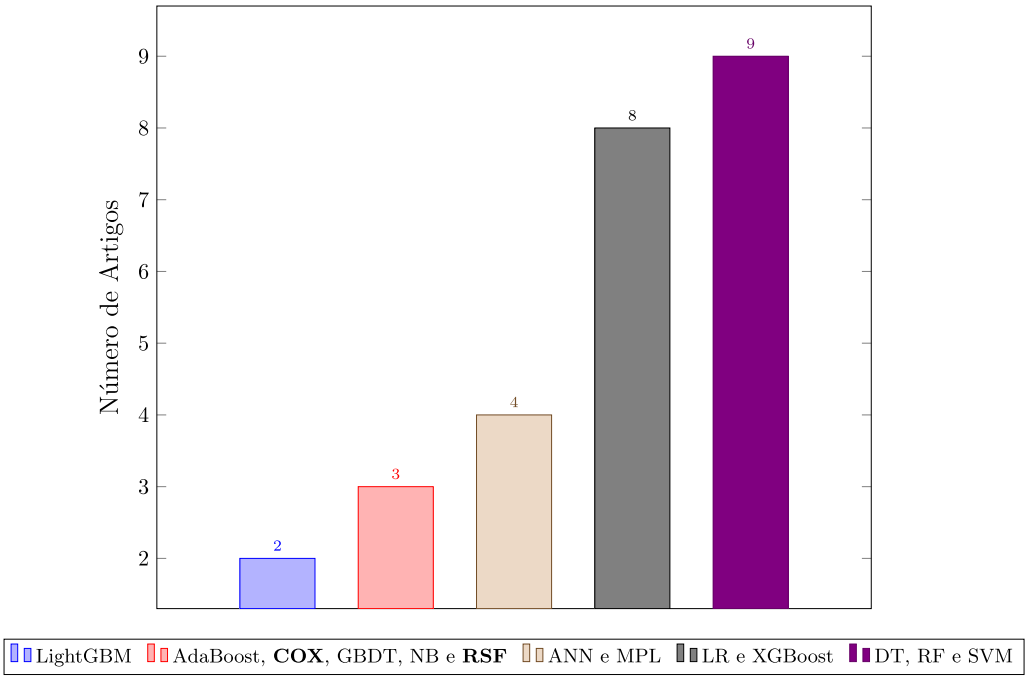
A Figura 11 apresenta a distribuição geográfica das populações de estudo nos artigos incluídos nesta revisão sistemática. Em termos de origem dos dados clínicos, quase metade dos estudos (47,4%)utilizou bases de dados norte-americanas, destacando-se a base *Surveillance, Epidemiology, and End Results* (SEER) (SEDIGHI-MAMAN; MONDELLO, 2021; TENG et al., 2019; SHUKLA et al., 2018; HUEMAN et al., 2018; LI et al., 2022; DENG et al., 2021; KLEINLEIN; RIAÑO, 2019; HAQUE et al., 2022; HUANG et al., 2022). Outros estudos recorreram a dados dos Registros de Hospitais Chineses (15,8%) (GU et al., 2020; XIN et al., 2022; LIU et al., 2020), Registros de Câncer da Holanda (10,5 %) (MONCADA-TORRES et al., 2021; JANSEN et al., 2020), Registros de CM da Malásia (10,5 %)(KALAFI et al., 2019; GANGGAYAH et al., 2019), Instituto de Câncer Italiano (5,3 %) (BOERI et al., 2020), Registros Médicos de Câncer da cidade de Urmia, no Irã (5,3 %) (AFSHAR et al., 2021), e a base de dados europeia *BioStudies* (5,3 %) (ZHOU et al., 2021).

3.3.3.2 Métodos de Aprendizado de Máquina

Os artigos selecionados nesta RS empregaram uma variedade de MAM, cuja distribuição é apresentada na Figura 12 , enquanto a correspondência entre métodos e estudos está detalhada no Quadro 5. A maioria dos trabalhos comparou o desempenho desses métodos na predição de sobrevida para o CM (GU et al., 2020; TENG et al., 2019; MONCADA-TORRES et al., 2021; KALAFI et al., 2019; LI et al., 2022; BOERI et al., 2020; ZHOU et al., 2021; XIN et al., 2022; LIU et al., 2020; DENG et al., 2021;

KLEINLEIN; RIAÑO, 2019; HAQUE et al., 2022; GANGGAYAH et al., 2019; AFSHAR et al., 2021; HUANG et al., 2022). Alguns autores optaram pela biblioteca "*Scikit-survival*", especializada em análise de sobrevida (MONCADA-TORRES et al., 2021; LIU et al., 2020; TENG et al., 2019)enquanto outros desenvolveram abordagens customizadas envolvendo etapas de tratamento, classificação e predição de dados (SEDIGHI-MAMAN; MONDELLO, 2021; SHUKLA et al., 2018). Adicionalmente, alguns estudos buscaram explicar as previsões geradas (JANSEN et al., 2020; GU et al., 2020; MONCADA-TORRES et al., 2021), enquanto um estudo explorou uma abordagem não supervisionada, utilizando um dendrograma para construção de um sistema prognóstico baseado em estadiamento (HUEMAN et al., 2018). Além disso, um estudo desenvolveu um modelo prognóstico (LIU et al., 2020).

Figura 12 – Gráfico dos métodos de AM utilizados nos artigos incluídos na RS.



Fonte: Elaborado pela autora (2024).

Tanto na Figura 12 como no Quadro 5 podemos observar que alguns métodos de métodos de AM são mais frequentes, como *Decision Trees* (DT), RF e SVM, presentes em 47, 37% dos artigos; seguidos pelos métodos *Logistic Regression* (LR) e *Extreme Gradient Boosting* (XGBoost), aplicados em 42, 11%; *K-Nearest Neighbor* (KNN) em 31, 58%; *Artificial Neural Networks* (ANN) e *Multilayer Perceptron* (MLP) em 21, 05%; *Adaptive Boosting* (AdaBoost), CPH, *Gradient Boosting Decision Tree* (GBDT), *Naïve Bayes* (NB) e RSF em 15, 79%; e, *Light Gradient Boosting Machine* (LightGBM) em 10, 53%. Os demais métodos de AM foram aplicados apenas em um estudo, como Modificação do CPH (CPHM), C5, *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN), *Ensemble*

Quadro 5 – Sumário dos métodos de AM aplicados nos 19 estudos da analisados na RS.

Artigo	**Métodos**
Afshar et al. (2021)	C5 e RIPPER
Boeri et al. (2020)	ANN e SVM
Deng et al. (2021)	ANN, DT, RF e SVM
Ganggayah et al. (2019)	DT, LR, MLP-ANN, RF, SVM e XGBoost
Gu et al. (2020)	DT, GBDT, KNN, LR, MLP, RF, SVM e XGBoost
Haque et al. (2022)	AdaBoost, DT, GBC, KNN, LR, RF, SVM e VC
Huang et al. (2022)	AdaBoost, ANN, DT, GBDT, KNN, LightGBM, LR, RF, SVM e XGBoost
Hueman et al. (2018)	EACCD
Kalafi et al. (2019)	DT, MLP, RS, e SVM
Kleinlein and Riaño (2019)	DT, LR e NB
Jansen et al. (2020)	ANN, KNN, LR, NB, RF, e XGBoost
Li et al. (2022)	DT, KNN, SVM e XGBoost
Liu et al. (2020)	CPH, EXSA, GB, RSF e XGBoost
Moncada-Torres et al. (2021)	CPH, RSF, SSVM e XGBoost
Sedighi-Maman e Mondello (2021)	GLM, MLP e XGBoost
Shukla et al. (2018)	DBSCAN, MLP e SOM
Teng et al. (2019)	CPH, CPHM, MTLSA, RSF e Stan
Xin et al. (2022)	AdaBoost, KNN, LR, NB, RF e SVM
Zhou et al. (2021)	DT, GBDT, LightGBM, LR e RF

Fonte: Elaborada pela autora (2024).

Algorithm for Clustering Cancer Data (EACCD), EXSA, *Gradient Boosting Classifier* (GBC), *Generalized Linear Model* (GLM), *Multi-Task Learning Model for Survival Analysis* (MTLSA), *Repeated Incremental Pruning to Produce Error Reduction* (RIPPER), *Bayesian Estimation Regression Model* (Stan), *Self-Organizing Map* (SOM), *Survival Support Vector Machine* (SSVM) e *Voting Classifier* (VC). Para organizar a análise, os métodos são agrupados a seguir em categorias como análise de sobrevida, classificação e regressão, clusterização, aprendizado de conjunto (EL), redes neurais e regras de decisão.

Análise de Sobrevida

Métodos específicos para análise de sobrevida, como CPH, Stan, EXSA, MTLSA, RSF e SSVM, foram empregados para estimar o tempo até a ocorrência de um evento, considerando a censura dos dados. O modelo CPH foi amplamente utilizado devido à sua relevância estatística na análise de risco ao longo do tempo (LIU et al., 2020; MONCADA-TORRES et al., 2021; TENG et al., 2019), enquanto o CPHM, uma versão modificada para estruturas temporais independentes, foi empregado em um estudo (TENG et al., 2019). O EXSA, por sua vez, combina a estrutura do XGBoost e o modelo CPH para predição de

sobrevida (LIU et al., 2020). , e o MTLSA, voltado para a previsão de múltiplos eventos de sobrevida, foi utilizado junto ao Stan, uma abordagem bayesiana para estimativa de parâmetros (TENG et al., 2019). O RSF estende o RF para problemas de análise de sobrevivência (LIU et al., 2020; MONCADA-TORRES et al., 2021; TENG et al., 2019), enquanto o SSVM adapta o SVM para esse contexto específico (MONCADA-TORRES et al., 2021).

Classificação e Regressão

Os métodos de classificação e regressão incluem técnicas supervisionadas que mapeiam dados de entrada para uma saída-alvo. Exemplos amplamente empregados são DT, GB, GBC, GLM, KNN, LR, NB e SVM, aplicados a problemas de classificação e/ou regressão. O DT, utilizado em estudos como técnica interpretável e eficaz (GU et al., 2020; KALAFI et al., 2019; LI et al., 2022; ZHOU et al., 2021; DENG et al., 2021; KLEINLEIN; RIAÑO, 2019; HAQUE et al., 2022; GANGGAYAH et al., 2019; HUANG et al., 2022). , destaca-se por sua simplicidade e robustez. LR, uma técnica para modelagem probabilística de eventos binários como sobrevida no CM, aparece em diversos estudos (GU et al., 2020; JANSEN et al., 2020; ZHOU et al., 2021; XIN et al., 2022; KLEINLEIN; RIAÑO, 2019; HAQUE et al., 2022; GANGGAYAH et al., 2019; HUANG et al., 2022) . O GBC, variante do GB para classificação, foi aplicado em um estudo (HAQUE et al., 2022), , enquanto o GB foi utilizado em outro (LIU et al., 2020). O GLM foi empregado para modelar relações flexíveis entre variáveis em um estudo específico (SEDIGHI-MAMAN; MONDELLO, 2021). O KNN, uma técnica baseada em proximidade, foi aplicado em vários estudos (GU et al., 2020; JANSEN et al., 2020; LI et al., 2022; XIN et al., 2022; HAQUE et al., 2022; HUANG et al., 2022). NB, um classificador probabilístico bayesiano, foi utilizado em estudos específicos (JANSEN et al., 2020; XIN et al., 2022; KLEINLEIN; RIAÑO, 2019), e o SVM, robusto para classificação e regressão, foi amplamente empregado (GU et al., 2020; LI et al., 2022; XIN et al., 2022; DENG et al., 2021; HAQUE et al., 2022; GANGGAYAH et al., 2019; HUANG et al., 2022).

Clusterização

Os métodos de *clusterização* ou agrupamento, foram utilizados para agrupar observações em grupos homogêneos, baseados em proximidade ou densidade dos dados. Em um estudo, SOM e DBSCAN foram aplicados conjuntamente para aprimorar a predição de MLP (SHUKLA et al., 2018).

Ensemble Learning

Os métodos de EL, como AdaBoost, EACCD, GBDT, LightGBM, RF, VC e XGBoost, combinaram diversos modelos para aumentar robustez e precisão. EACCD, voltado ao agrupamento de dados de câncer, foi aplicado em um estudo (HUEMAN et al., 2018). GBDT, por sua vez, foi empregado para construção de modelos sequenciais (ZHOU et al., 2021; HUANG et al., 2022), enquanto o LightGBM, reconhecido por sua

velocidade e eficiência, foi aplicado em outros estudos (ZHOU et al., 2021; HUANG et al., 2022). O RF, uma técnica amplamente utilizada para reduzir *overfitting* e melhorar a generalização, aparece em diversos estudos (GU et al., 2020; KALAFI et al., 2019; JANSEN et al., 2020; ZHOU et al., 2021; XIN et al., 2022; DENG et al., 2021; HAQUE et al., 2022; GANGGAYAH et al., 2019; HUANG et al., 2022). O VC, EL que combina previsões de diferentes classificadores, foi aplicado em um estudo (HAQUE et al., 2022), assim como o AdaBoost (XIN et al., 2022; HAQUE et al., 2022; HUANG et al., 2022). XGBoost destaca-se pela eficácia e popularidade em grandes conjuntos de dados e no manejo de regularização e valores ausentes (GU et al., 2020; MONCADA-TORRES et al., 2021; JANSEN et al., 2020; LI et al., 2022; LIU et al., 2020; HUANG et al., 2022).

Redes Neurais

Os métodos de redes neurais, como ANN, MLP e MLP-ANN, são baseados no funcionamento do cérebro humano, uma estrutura de camadas de processamento para aprendizado de padrões complexos. ANN foi amplamente empregada (JANSEN et al., 2020; BOERI et al., 2020; DENG et al., 2021; GANGGAYAH et al., 2019; HUANG et al., 2022), enquanto MLP foi utilizado em diversos estudos para capturar relações intrínsecas nos dados (GU et al., 2020; SEDIGHI-MAMAN; MONDELLO, 2021; SHUKLA et al., 2018; KALAFI et al., 2019). O MLP-ANN, uma variante de rede neural profunda, também foi empregado em um estudo (GANGGAYAH et al., 2019).

Regras de Decisão

Métodos de regras de decisão, como C5 e RIPPER, foram utilizados para gerar regras de classificação interpretáveis. O C5, em conjunto com RIPPER, foi aplicado em um estudo para extração de regras de decisão (AFSHAR et al., 2021).

3.3.3.3 Avaliação do Desempenho e Validação

O Quadro 6 apresenta uma visão geral das métricas de desempenho e dos métodos de validação adotados nos estudos revisados, permitindo uma análise detalhada da robustez e generalização dos métodos de AM aplicados à predição de sobrevida para o CM. As métricas de desempenho utilizadas incluem: *Accuracy* (ACC), *Recall* (REC), *Precision* (PREC), *F1-Score* (F1), *Trade-off*, *Sensitivity* (SEN), *Specificity* (SPEC), *Positive Predictive Value* (PPV), *Negative Predictive Value* (NPV), *Confusion Matrix* (CC), *Matthews Correlation Coefficient* (MCC), *Receiver Operating Characteristic* (ROC), *Area Under the Curve* (AUC), *Mean Absolute Error* (MAE), *Root Mean Squared Error* (RMSE), *Concordance Index* (C-Index), *Student-T test* com *Bonferroni correction* (Student-T), *Kappa Statistic* (K-Statistic), LIME, SHAP e *K-Fold Cross-Validation* (K-Fold)..

Entre as métricas mais recorrentes, a ACC é amplamente utilizada para medir a proporção de predições corretas realizadas pelo modelo, sendo mencionada em diversos estudos (GU et al., 2020; TENG et al., 2019; KALAFI et al., 2019; BOERI et al., 2020;

Quadro 6 – Resumo dos métodos de avaliação dos métodos de AM nos 19 estudos da analisados na RS.

Artigo	**Desempenho e Validação**
Afshar et al. (2021)	ACC, SEN, SPEC, K-Statistic e AUC
Boeri et al. (2020)	ACC, SEN, SPEC, AUC, ROC e K-Fold
Deng et al. (2021)	ACC, PREC, REC, AUC, ROC e F1 e K-Fold
Ganggayah et al. (2019)	ACC, SEN, SPEC, AUC, PREC, MCC e ROC
Gu et al. (2020)	ACC, REC, F1, AUC, ROC, SHAP e 10-Fold
Haque et al. (2022)	CC, ACC, PREC, REC, *Trade-off* e AUC
Huang et al. (2022)	ACC, PREC, SEN, F1, AUC e CC
Hueman et al. (2018)	C-Index
Kalafi et al. (2019)	ACC, SEN, SPEC, PREC, CC, F1, MCC e 10-Fold
Kleinlein and Riaño (2019)	ROC, AUC e 10-Fold
Jansen et al. (2020)	AUC, LIME, SHAP e 10-Fold
Li et al. (2022)	SEN, SPEC, ROC, AUC, e CC e 10-Fold
Liu et al. (2020)	C-Index, AUC, ROC e K-Fold
Moncada-Torres et al. (2021)	C-Index, Student-T, SHAP e 10-Fold
Sedighi-Maman e Mondello (2021)	RMSE, MAE e 5-Fold
Shukla et al. (2018)	ACC, 10-Fold
Teng et al. (2019)	C-Index, AUC e ROC
Xin et al. (2022)	SEN, PPV, SPEC, NPV, AUC e ROC
Zhou et al. (2021)	ACC, PREC, REC, AUC e F1

Fonte: Elaborada pela autora (2024).

ZHOU et al., 2021; DENG et al., 2021; HAQUE et al., 2022; GANGGAYAH et al., 2019; AFSHAR et al., 2021; HUANG et al., 2022). O REC mede a taxa de verdadeiros positivos, ou seja, a proporção de casos positivos corretamente identificados, e foi abordado em estudos como (GU et al., 2020; ZHOU et al., 2021; HAQUE et al., 2022). A PREC, que representa a precisão nas predições positivas, aparece em estudos focados na qualidade das predições em situações reais (KALAFI et al., 2019; ZHOU et al., 2021; HAQUE et al., 2022; GANGGAYAH et al., 2019; HUANG et al., 2022).

Outra métrica importante é o F1-Score, que combina PREC e REC em uma média harmônica, e é particularmente útil para dados desbalanceados; foi empregada em (GU et al., 2020; KALAFI et al., 2019; ZHOU et al., 2021; DENG et al., 2021; HUANG et al., 2022). O *Trade-off*, por sua vez, refere-se ao equilíbrio entre PREC e REC e foi analisado em (HAQUE et al., 2022).

As métricas SEN e SPEC, que medem a capacidade do modelo de identificar verdadeiros positivos e verdadeiros negativos, respectivamente, foram utilizadas em estudos como (KALAFI et al., 2019; LI et al., 2022; BOERI et al., 2020; XIN et al., 2022; GANGGAYAH et al., 2019; AFSHAR et al., 2021; HUANG et al., 2022). O PPV e o

NPV, que indicam a precisão das predições positivas e negativas, respectivamente, são apresentados em trabalhos como (XIN et al., 2022).

A CC que organiza os resultados de classificação em uma tabela de contingência, descrevendo o desempenho de um modelo em termos de verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos, aparece em diversos estudos (KALAFI et al., 2019; LI et al., 2022; HAQUE et al., 2022; HUANG et al., 2022). á o MCC, que oferece uma medida de correlação entre predições e valores reais e considera todos os quatro resultados possíveis, é abordado em (KALAFI et al., 2019).

Para avaliar o desempenho geral dos modelos em diferentes limiares, o ROC foi utilizado, especialmente nos estudos (GU et al., 2020; TENG et al., 2019; LI et al., 2022; BOERI et al., 2020; XIN et al., 2022; LIU et al., 2020; DENG et al., 2021; KLEINLEIN; RIAÑO, 2019; HAQUE et al., 2022; GANGGAYAH et al., 2019; AFSHAR et al., 2021; HUANG et al., 2022). A AUC, que mede a discriminação do modelo pela área sob a curva ROC, é apresentada em (GU et al., 2020; TENG et al., 2019; LI et al., 2022; BOERI et al., 2020; XIN et al., 2022; LIU et al., 2020; DENG et al., 2021; KLEINLEIN; RIAÑO, 2019; GANGGAYAH et al., 2019).

Métricas de erro absoluto, como o MAE e o RMSE, que medem a precisão das predições em relação aos valores reais, foram utilizadas em (SEDIGHI-MAMAN; MONDELLO, 2021).Já o C-Indexé particularmente relevante para análise de sobrevida, permitindo avaliar a capacidade preditiva dos modelos para a sequência de eventos de interesse e foi analisado em (TENG et al., 2019; HUEMAN et al., 2018; MONCADA-TORRES et al., 2021; LIU et al., 2020).

O Student-T com correção de Bonferroni, foi empregado para comparar os valores de C-Index entre modelos em (MONCADA-TORRES et al., 2021).Para verificar a concordância entre as predições e os resultados observados, o K-Statistic foi aplicado no estudo (AFSHAR et al., 2021).

Com relação à interpretação dos modelos, o LIME oferece explicações locais para as predições, enquanto o SHAP, baseado na teoria dos jogos, atribui pesos de importância às variáveis, fornecendo insights sobre sua contribuição para as predições. Essas técnicas de interpretabilidade foram destacadas em (GU et al., 2020; JANSEN et al., 2020; MONCADA-TORRES et al., 2021).

Para validação, a técnica de K-Fold Cross-Validation foi amplamente adotada.. Alguns estudos não especificaram o valor de K (BOERI et al., 2020; LIU et al., 2020; DENG et al., 2021), enquanto outros utilizaram 5-Fold (SEDIGHI-MAMAN; MONDELLO, 2021) e 10-Fold (GU et al., 2020; MONCADA-TORRES et al., 2021; KALAFI et al., 2019; JANSEN et al., 2020; LI et al., 2022; KLEINLEIN; RIAÑO, 2019). A K-Fold Cross-Validation permite avaliar a consistência do modelo em diferentes divisões do conjunto de dados, reforçando sua robustez.

3.3.3.4 Principais Resultados para Predição da Sobrevida

Os resultados desta RS revelaram uma tendência positiva na aplicação dos métodos de AM para a predição da sobrevida para o CM feminino. A maioria dos estudos incluídos comparou diferentes métodos de AM tradicionais, analisando suas métricas de desempenho (GU et al., 2020; KALAFI et al., 2019; LI et al., 2022; BOERI et al., 2020; ZHOU et al., 2021; XIN et al., 2022; DENG et al., 2021; KLEINLEIN; RIAÑO, 2019; HAQUE et al., 2022; GANGGAYAH et al., 2019; AFSHAR et al., 2021; HUANG et al., 2022). Os métodos de AM comparados abrangem métodos como AdaBoost, ANN, DT, C5, GBC, GBDT, KNN, LightGBM, MPL, PPL-ANN, NB, RIPPER, RF, SVM, VC e XGBoost, conforme detalhado no Quadro 5. As comparações entre esses métodos basearam-se em métricas como ACC, AUC, CC, F1, *K-Statistic*, MCC, NPV, PPV, PREC, REC, ROC, SEN, SPEC e *Trade-off*, apresentado no Quadro 6.

Vários estudos aplicaram a validação cruzada K-Fold para robustez e generalização dos modelos com dados clínicos, destacando-se entre esses RF e XGBoost. O RF demonstrou desempenho superior em múltiplos estudos (ZHOU et al., 2021; DENG et al., 2021; XIN et al., 2022; HAQUE et al., 2022; GANGGAYAH et al., 2019), com alta precisão na predição de sobrevida, recorrência (HAQUE et al., 2022) e metástase em pacientes com expressão baixa de HER2 (XIN et al., 2022). Também se destacou na identificação de fatores clínicos e bioquímicos de risco para mortalidade específica por CM e na seleção de características prognósticas (ZHOU et al., 2021; GANGGAYAH et al., 2019).

O XGBoost mostrou resultados notáveis em predições precisas, conforme observado em Moncada-Torres et al. (2021), onde superou outros métodos de AM para análise de sobrevida (MONCADA-TORRES et al., 2021). Estudos como os de Jansen et al. (2020) e Gu et al. (2020) reforçam a eficácia do XGBoost e destacam sua aceitação por oncologistas, com reconhecido potencial para melhorar a precisão nas decisões clínicas (JANSEN et al., 2020; GU et al., 2020). No estudo de Li et al. (2022), o XGBoost se sobressaiu na predição de sobrevida para pacientes com metástase óssea em intervalos de 1, 3 e 5 anos, superando SVM, DT e KNN (LI et al., 2022).

Outros métodos de AM supervisionados também mostraram desempenho satisfatório. Boeri (2020) apresentou o SVM e ANN como modelos eficazes na avaliação do risco de recorrência ou mortalidade por CM (BOERI et al., 2020), enquanto Kalafi et al. (2019) destacaram o MLP, que apresentou as melhores métricas para predição de sobrevida (KALAFI et al., 2019). Kleinlein et al. (2019) relataram melhoras significativas no modelo de regressão logística (LR) com um maior conjunto de dados (KLEINLEIN; RIAÑO, 2019).

Alguns estudos exploraram técnicas adicionais para aprimorar os MAM, como imputação de dados faltantes (AFSHAR et al., 2021), adequação dos dados (SEDIGHI-MAMAN; MONDELLO, 2021), estratificação por estadiamento (HUANG et al., 2022;

HUEMAN et al., 2018) e explicabilidade dos resultados. No estudo de Huang et al. (2022), o LightGBM destacou-se na predição de sobrevida e orientação de tratamento em quimioterapia (HUANG et al., 2022). Sedighi et al. (2021) observaram que, com preparação adequada dos dados, o GLM alcançou desempenho comparável ao XGBoost e MLP, mas com menor custo computacional (SEDIGHI-MAMAN; MONDELLO, 2021).

Modelos não supervisionados, como RIPPER e SOM, também foram explorados. No estudo de Afshar et al. (2021), o RIPPER superou o C5.0 na predição de sobrevida, embora sem relevância clínica significativa nas regras extraídas (AFSHAR et al., 2021). Shukla et al. (2018) relataram melhoras na predição de sobrevida utilizando SOM e DBSCAN junto ao MLP (SHUKLA et al., 2018).

A explicabilidade é uma questão-chave para a aplicação clínica dos MAM. Jansen et al. (2020) empregaram LIME e SHAP para explicar predições de modelos, observando alta concordância, apesar de inconsistências do LIME em algumas predições (JANSEN et al., 2020). SHAP foi amplamente utilizado para detalhar a influência das variáveis nos resultados dos modelos, o que é essencial para a aceitação dos métodos de AM na prática clínica (GU et al., 2020; MONCADA-TORRES et al., 2021; JANSEN et al., 2020).

Em termos de análise de sobrevida, o estudo de Hueman et al. (2018) utilizou o modelo EACCD para criar um sistema prognóstico baseado no estadiamento antigo e atualizado, resultando em um C-Index de 0,75, considerado adequado para predições clínicas (HUEMAN et al., 2018). Comparativamente, estudos que testaram métodos de AM em paralelo com o modelo clássico de Cox Proportional Hazards (CPH) para análise de sobrevida mostraram que os métodos de AM podem ter desempenho igual ou superior ao CPH, especialmente no uso do C-Index como métrica, que integra o tempo e o status de sobrevida (LIU et al., 2020; MONCADA-TORRES et al., 2021; TENG et al., 2019).

No trabalho de Liu et al. (2020) foi desenvolvidoo modelo EXSA, que atingiu um C-Index de 0,83, comparável ao XGBoost original, e ajudou na estratificação de risco, facilitando o planejamento de tratamento (LIU et al., 2020). Moncada-Torres et al. (2021) também destacaram o XGBoost com C-Index de 0,73, reforçando que os métodos de AM podem superar o CPH (MONCADA-TORRES et al., 2021). Teng et al. (2019) utilizaram uma abordagem bayesiana para predição de sobrevida, obtendo um C-Index de 0,71 ao utilizar a proporção de linfonodos como fator prognóstico essencial para pacientes com CM, oferecendo uma ferramenta prática para médicos e pacientes (TENG et al., 2019).

Os resultados sugerem que os métodos de AM, como RF e XGBoost, apresentam um potencial significativo para a predição de sobrevida no CM feminino, oferecendo alternativas ou complementos ao modelo clássico de CPH. Esses métodos, especialmente com o uso de bibliotecas como a *"Scikit-survival"*, serão aplicados neste trabalho, visando viabilizar decisões clínicas mais precisas e personalizadas. A combinação com técnicas robustas de explicabilidade e validação fortalece ainda mais a aplicabilidade desses modelos

na prática médica (PÖLSTERL, 2020; PÖLSTERL, 2023).

Em síntese, os métodos de AM têm se destacado em relação aos métodos tradicionais de análise de sobrevida por sua habilidade em lidar com dados complexos e relações não lineares de forma eficiente e em diversos contextos. Contudo, esses avanços foram observados em populações específicas e, portanto, a generalização para outros grupos populacionais ainda demanda mais estudos.

Dessa forma, esta RS alcançou seus objetivos iniciais, contribuindo para o cumprimento de um dos objetivos específicos desta pesquisa conforme descrito na Seção 1.3 da tese, intitulada "Avanços em Modelos Computacionais para Predição da Sobrevida para o Câncer de Mama Feminino". Além de mapear os principais métodos utilizados para predição de sobrevida, a RS permitiu identificar uma biblioteca ainda pouco explorada para essa aplicação. Com base nos resultados obtidos, a biblioteca *"Scikit-survival"* (PÖLSTERL, 2023) foi selecionada para a implementação dos modelos preditivos neste trabalho.

4 MATERIAL E MÉTODOS

Neste estudo, a fim de alcançar os objetivos definidos na Seção 1.3, serão empregados métodos de AM supervisionados, os quais serão validados com dados clínicos. Os dados clínicos utilizados provêm de registros médicos autorizados, assegurando total sigilo e confidencialidade, conforme aprovado pelo Comitê de Ética da Universidade Federal de Juiz de Fora (UFJF), sob o número de parecer 5.533.296. Esses dados passarão por dois processos de pré-processamento, seguidos de análises estatísticas qualitativas e quantitativas seguidas de análise de correlação. As análises descritivas serão realizadas com o objetivo de explorar e compreender o comportamento das variáveis, preparando os dados para serem utilizados nos modelos computacionais.

Os modelos específicos para análise de sobrevida que serão aplicados incluem aqueles disponíveis na biblioteca "*Scikit-survival*"(PÖLSTERL, 2023). Entre os modelos lineares, destacam-se o CPH, CPH-EN, CPH-L, CPH-R e SSVM, enquanto, para os modelos não lineares, serão utilizados o GBS, RSD e KSSVM. O desempenho desses modelos será avaliado por meio de métricas específicas para análise de sobrevida, como o C-Index, *Brier Score* (BS) e *Integrated Brier Score* (IBS) (PÖLSTERL, 2023; PÖLSTERL, 2020). Além disso, serão aplicados métodos de seleção de atributos, incluindo técnicas lineares baseadas em CPH penalizados e não lineares, como os aplicados ao GBS, além de métodos de importância de variável, com base em extensões do RF da biblioteca "*Scikit-learn*"(PEDREGOSA et al., 2011), aplicados ao RSF e GBS (PEDREGOSA et al., 2011; PÖLSTERL, 2023).

A implementação e simulação dos modelos será realizada utilizando a linguagem *Python*, possibilitando a geração de resultados relativos às métricas de desempenho para cada modelo avaliado. A aplicação dessas metodologias é particularmente relevante, pois permite lidar de maneira eficiente com dados complexos, incluindo informações incompletas (censuradas), comuns em análises de sobrevida (PÖLSTERL, 2023; PÖLSTERL, 2020). Essa abordagem visa identificar padrões clínicos e suas relações com as respostas ao tratamento, proporcionando uma predição mais precisa da sobrevida dos pacientes com CM.

Os dados coletados orientarão não apenas a implementação dos modelos, mas também sua validação e a seleção de atributos, com o intuito de identificar o método mais representativo na predição da sobrevida das pacientes. Para facilitar a compreensão da metodologia aplicada, a explicação de cada etapa foi organizada nas seguintes seções: Pré-processamento dos Dados Clínicos, Implementação e Simulação dos Modelos Computacionais, Validação e Avaliação de Desempenho, e Seleção de Atributos.

4.1 PRE-PROCESSAMENTO DO BANCO DE DADOS CLÍNICOS

Neste estudo, os dados foram obtidos por meio de um projeto de pesquisa caracterizado como um estudo de *coorte* com base hospitalar, utilizando dados secundários de uma população de mulheres diagnosticadas com CM invasivo (CID-10 C50), conforme estudos anteriores (CINTRA, 2012; FAYER, 2014). O projeto foi submetido e aprovado pelo Comitê de Ética em Pesquisa com Seres Humanos da UFJF, sob o número de parecer 5.533.296, conforme detalhado no Anexo 6.3.

O BD Clínicos é robusto e resulta de estudos prévios que, embora não tenham empregado metodologias de modelagem computacional, fornecem dados essenciais para a validação dos modelos aplicados neste trabalho. A aprovação ética garante o alinhamento com os objetivos específicos descritos na Seção 1.3 e autoriza a utilização desses dados para a implementação dos modelos computacionais.

O pré-processamento dos dados é uma etapa fundamental ao lidar com BD Clínicos em contextos reais, dado que frequentemente há grandes volumes de dados ausentes, anormais, duplicados e inconsistentes (LIU et al., 2020). Este processo envolve a limpeza, transformação e integração dos dados, visando mitigar imprecisões e inconsistências, o que não apenas melhora a qualidade dos dados, mas também impacta positivamente o desempenho dos modelos preditivos, aumentando a precisão estatística dos estudos (KALAFI et al., 2019).

4.1.1 BANCO DE DADOS

O BD clinicos analisado neste estudo inclui registros de pacientes de centros de referência em oncologia da Zona da Mata Mineira. A população estudada é composta por mulheres diagnosticadas com CM invasivo entre janeiro de 2003 e dezembro de 2005, todas atendidas em serviços de referência em Juiz de Fora/MG. Os dados foram coletados em pesquisas anteriores conduzidas pelo Programa em Saúde Brasileira (CINTRA, 2012) e pelo Programa de Pós-Graduação em Saúde Coletiva (FAYER, 2014), que não utilizaram técnicas de modelagem computacional.

A coleta de dados do BD clínicos original (FAYER, 2014; CINTRA, 2012) foi realizada em três etapas:

- **Primeira Etapa (2009-2010):** Recrutamento de pacientes e busca ativa em arquivos de registros hospitalares de câncer, resultando em 601 casos, com acompanhamento até 31 de dezembro de 2010.
- **Segunda Etapa (2010-2011):** Avaliação das características clínicas e sociodemográficas das pacientes, excluindo aquelas que passaram por apenas um procedimento, reduzindo o número de casos para 563.

- **Terceira Etapa (Iniciada em janeiro de 2011):** Busca ativa por meio de ligações telefônicas, consulta ao CPF, contato com mastologistas e consulta ao sistema de informações de mortalidade.

Este estudo utilizou variáveis relacionadas às características clínicas, sociodemográficas, anatomopatológicas e ao uso de serviços de saúde, tratamento e acompanhamento (CINTRA, 2012; FAYER, 2014). Os dados do BD refletem o histórico clínico de pacientes diagnosticadas com CM, permitindo a aplicação e validação dos modelos em um cenário realista. Embora não sejam dados recentes, o grande diferencial deste BD é fornecer informações abrangentes da prática médica, mais próximas da realidade clínica, algo não disponível em bases de dados secundárias públicas. Para mais detalhes, consulte (CINTRA, 2012; FAYER, 2014).

4.1.2 ANÁLISE E DESCRIÇÃO DOS DADOS

Para realizar a análise e descrição dos dados clínicos utilizados neste estudo, foi empregada a biblioteca *Sweetviz*. Esta biblioteca de código aberto em *Python* que gera visualizações de alta densidade para análise exploratória dos dados numéricos e categóricos, disponibilizando medidas estatísticas e de correlação. A descrição das variáveis foi realizada por meio da das medidas de distribuição estatísticas como frequência, média, moda, valor mínimo e valor máximo, além da análise de correlação entre as variáveis pelas medidas como Coeficiente de Incerteza (CI) e Coeficiente de *Pearson* (CP). Esses descritores fornecem uma visão inicial do comportamento dos dados e de suas possíveis relações, auxiliando no entendimento da estrutura do conjunto de dados.

Entre as medidas estatísticas utilizadas, destacam-se (ASSIS et al., 2019):

1. **Desvio Padrão:** O Desvio Padrão (DB) é uma media de dispersão obtida pela raiz quadrada da variância.

2. **Frequência:** A Frequência Absoluta (FA) corresponde o número de observações amostrais, já a Frequência Relativa (FR) corresponde à proporção do número de observações em uma determinada classe em relação ao total de observações.

3. **Média:** Corresponde à distribuição obtida pelo somatório de todos os valores amostrais divididos pelo número de amostras.

4. **Moda:** Corresponde a medida de tendência central que representa o valor mais frequente em uma distribuição de dados.

5. **Valor Mínimo e Máximo:** O Valor Mínimo (Vmin) corresponde ao menor valor encontrado numa determinada variável, já o Valor Máximo (Vmax) corresponde ao maior valor encontrado.

Já as medidas de correlação determinam o grau de associação entre grandezas categóricas e numéricas, segue abaixo as medidas de correlação usadas (ASSIS et al., 2019):

1. **Coeficiente de Incerteza:** Medida de associação de variáveis qualitativas.
2. **Coeficiente de Pearson:** Medida do grau de relação linear entre duas variáveis quantitativas.

A análise exploratória foi fundamental para aprimorar o processo de pré-processamento, permitindo a identificação de valores atípicos, a detecção de correlações significativas e a realização de um segundo tratamento para os dados ausentes. Esses *insights* facilitaram decisões mais informadas na etapa de preparação dos dados, garantindo que os modelos de análise de sobrevida recebessem informações mais limpas e consistentes. A descrição detalhada dos dados também estabeleceu a base para etapas subsequentes, como a reavaliação das variáveis altamente correlacionadas após o segundo pré-processamento, além da aplicação de métodos de seleção de atributos, contribuindo para a identificação das variáveis com maior relevância para os modelos preditivos.

4.1.3 PRÉ-PROCESSAMENTO DOS DADOS

O pré-processamento dos dados será realizado em duas fases distintas, com o objetivo de corrigir inconsistências, formatar as variáveis e inferir valores ausentes, de modo que os dados possam ser utilizados como atributos de entrada para os métodos de AM supervisionados. A inferência de dados ausentes na área da saúde é uma prática comum visando melhorar a qualidade dos dados (CURIOSO et al., 2023). A primeira fase consistirá no tratamento e padronização dos dados, seguida pela imputação simples dos valores ausentes. Já a segunda fase, que ocorrerá após a etapa inicial, terá como foco a inferência de valores ausentes com base em análises descritivas e correlações entre variáveis.

4.1.3.1 Primeira Fase do Pré-Processamento

A primeira fase de pré-processamento será dedicada à limpeza e padronização dos dados, com correção de erros de digitação e remoção de acentos em *strings*. Após essa etapa, os dados ausentes serão tratados por meio de inferência sendo atribuído o valor "9"para atributos numéricos ou a palavra "ignorado"para atributos de texto. Além disso, as datas serão minuciosamente verificadas e formatadas adequadamente. Os passos específicos para esta fase estão descritos abaixo:

1. ***Strings:*** Correção de erros de digitação, remoção de acentuações, formatação para minúsculas e inserção da palavra "ignorado"nos dados ausentes;

2. **Valores Numéricos:** Formatação dos números inteiros e reais, com atribuição dos valores 9 e -1 para dados ausentes em variáveis categóricas e de contagem, respectivamente;

3. **Datas:** Verificação e formatação adequada das datas.

Após essa primeira etapa, foram selecionadas as variáveis com menos de 40% de dados faltantes. Adicionalmente, todas as *strings* foram convertidas para valores numéricos, garantindo compatibilidade com os modelos computacionais.

4.1.3.2 Segunda Fase do Pré-Processamento

Na segunda fase do pré-processamento, será realizada uma análise detalhada para inferir os valores ausentes, considerando as correlações entre variáveis e as estatísticas descritivas de cada categoria. Esse processo será conduzido em conjunto com uma análise descritiva, visando identificar variáveis correlacionadas e avaliar sua significância, especialmente quando uma variável pode ser derivada de outras.

Com base nessa análise, os valores faltantes serão inferidos utilizando variáveis correlacionadas. Para variáveis com correlação inferior a 40%, será necessário adotar um método alternativo de inferência, sempre levando em conta a relevância da variável. Nesses casos, a inferência será realizada por meio da moda, ou seja, atribuindo o valor mais frequente da variável após a análise.

Para ilustrar o processo de inferência dos dados faltantes, considere inicialmente as variáveis com baixa correlação. No caso da variável "Tabagismo", os valores ausentes foram preenchidos com a categoria “nunca fumou”, que corresponde a 62% das pacientes, possivelmente refletindo o impacto das campanhas antitabaco (MEIRELLES, 2023). De forma semelhante, na variável "Etilismo", os valores faltantes foram imputados com a categoria “nunca usou”, predominante em 64% das pacientes. Outras variáveis, como "Doença Cardiovascular"e "Doença de Hipertensão", foram preenchidas com as categorias “não”, presentes em 82% e 56% das pacientes, respectivamente. Para as variáveis relacionadas aos exames, como "Mamografia"(64%), "Citologia"(91%), "Raio X do Tórax"(86%), "Ultrassom do Abdômen"(82%) e "Recidiva Locorregional"(87%), seguiu-se o mesmo critério, adotando a categoria mais frequente para a imputação dos valores ausentes.

No caso das variáveis altamente correlacionadas, muitas das quais possuem grande relevância clínica, a inferência dos valores ausentes foi realizada por meio de um tratamento mais específico, considerando tanto a correlação estatística quanto a importância clínica no curso terapêutico da paciente.

Por exemplo, os valores ausentes na variável "Histórico de Câncer"foram preenchidos com base na variável "Histórico de CM", que indica a presença de cânceres familiares específicos. Para a variável "Menstruação", a inferência foi feita a partir do "Status

Menopausal", previamente dicotomizado no BD original (FAYER, 2014; CINTRA, 2012): mulheres com menos de 50 anos foram classificadas como pré-menopausa, enquanto aquelas com 50 anos ou mais foram classificadas como pós-menopausa. No caso da variável "Anticoncepcional", a imputação foi realizada assumindo o uso desse método quando o histórico de menstruação estava presente.

A variável "Amamentação"foi inferida a partir da variável "Gravidez", assumindo que pacientes com pelo menos uma gestação completa amamentaram após o parto. Para "Estadiamento Clínico T"a imputação foi baseada na variável "Tamanho do Tumor (X)",utilizando as seguintes faixas de classificação: T1 para tumores com $T \leq 2,0$ cm, T2 para $2,0 < T \leq 5,0$ cm e T3 para $T > 5,0$ cm, conforme preconizado na literatura (BRASIL, 2014; BARRIOS et al., 2022). A variável "Estadiamento Patológico T"foi então preenchida com base nos valores inferidos para o "Estadiamento Clínico T".

A inferência do "Tamanho do Tumor (X)"foi realizada a partir do "Estadiamento Patológico T", adotando a média do tamanho tumoral observada em cada estágio: T1 (1,5 cm), T2 (3,3 cm), T3 (6,7 cm) e T4 (6,6 cm). De forma análoga, o "Tamanho do Tumor (Y)"foi inferido com base nos dados do "Tamanho do Tumor (X)"(por conter menor quantidade de valores ausentes).

As variáveis "Estadiamento Clínico N"e "Estadiamento Patológico N"foram inferidas considerando o "Estadiamento Anatômico"e "Linfonodos Pós-cirúrgicos". O número de linfonodos comprometidos pós-cirúrgicos foi utilizado como critério para imputação, seguindo a classificação: N0 para $N0$, N1 para $0 < N \leq 3$, N2 para $3 < N \leq 9$ e N3 para $N > 10$, conforme diretrizes estabelecidas (BRASIL, 2014; BARRIOS et al., 2022). As variáveis"Linfonodos Pós-cirúrgicos"e "Linfonodos Isolados"foram preenchidas com base na média de N observada em "Estadiamento Patológico N".

As"Estadiamento Clínico M"e o "Estadiamento Patológico M"foram imputadas considerando o "Estadiamento Anatômico", sendo atribuídos valores de 0 para estágios diferentes de IV e 1 para o estágio IV, conforme descrito no Quadro 2, e tambem conforme preconizada reconizado na literatura (BRASIL, 2014; BARRIOS et al., 2022). Além disso, a variável "Metástase à Distância"foi inferida a partir da variável "Óbito por CM", uma vez que há uma forte correlação entre a presença de metástases e o desfecho óbito. Da mesma forma, a variável "Metástase Locorregional"foi preenchida com base no "Estadiamento Clínico M".

Para os receptores hormonais, a inferência da variável "Ki-67"foi realizada a partir da média desse marcador dentro de cada categoria da"Classificação Imuno-histoquímica", considerando os seguintes valores: Luminal A (1,00), Luminal B HER2- (3,14), Luminal B HER2+ (3,78), Superexpressão HER2+ (3,87), Triplo Negativo (3,77) e, nos casos não classificados, a média geral (2,77). A variável "Classificação HER2"foi inferida com base na média observada por categoria: Luminal A e Luminal B HER2- (0), Luminal B HER2+

e Superexpressão HER2+ (1). Nos casos em que a "Classificação Imuno-histoquímica"não estava disponível, a imputação foi feita utilizando a moda de cada receptor.

Para os receptores de Estrogênio e Progesterona, a imputação foi realizada com base na variável "Tratamento Sistêmico", adotando a média observada em cada categoria de tratamento: Hormonioterapia (1,00), Antraciclínico (1,00) e ausência de tratamento sistêmico (0,00). Por fim, a variável "Classificação Imuno-histoquímica"foi inferida a partir da combinação dos receptores de Estrogênio, Progesterona, HER2 e Ki-67, seguindo as diretrizes da literatura e conforme descrito no Quadro 1 (CIRQUEIRA et al., 2011).

O período de análise de sobrevida foi estabelecido em cinco anos, correspondendo a um limite de 1825 dias. A contagem do tempo de sobrevida teve início a partir do diagnóstico, definido pela data do laudo histopatológico, e foi finalizada na data do óbito por CM ou no último dia de acompanhamento, respeitando o limite de 1825 dias. Pacientes que não apresentaram o evento dentro desse intervalo foram classificadas como censuradas, com seus tempos de sobrevida encerrados na data do último acompanhamento registrado. Caso o óbito ocorra após esse período, o dado será igualmente tratado como censurado. O banco de dados inclui informações sobre a sobrevida em 5 e 10 anos, além de dados detalhados sobre o perfil imuno-histoquímico das pacientes ao longo do período máximo de acompanhamento (CINTRA, 2012; FAYER, 2014).

4.2 IMPLEMENTAÇÃO E SIMULAÇÃO DE MODELOS COMPUTACIONAIS

Os modelos foram implementados utilizando a biblioteca "*Scikit-survival*", um pacote de código aberto desenvolvido em "*Python*", especificamente voltado para análise de sobrevida. Esta biblioteca foi escolhida devido à sua especialização nesse tipo de análise, o que possibilitou a exploração de diversas de suas funcionalidades (PÖLSTERL, 2023). Por ser totalmente compatível com o "*Scikit-learn*", ela permite a utilização dos recursos dessa biblioteca amplamente adotada (PEDREGOSA et al., 2011). Dessa forma, o "*Scikit-survival*"oferece implementações de várias técnicas populares de AM adaptadas para análise de sobrevida, além de ferramentas para avaliação de desempenho, otimização e seleção de atributos. O pacote é distribuído sob a licença GPL-3, e seu código-fonte, acompanhado de instruções detalhadas, está disponível em https://github.com/sebp/scikit-survival (PÖLSTERL, 2020; PÖLSTERL, 2023).

Nesse contexto, os métodos implementados, juntamente com os hiperparâmetros utilizados, seguiram as recomendações da biblioteca, sendo apenas adaptados os métodos não lineares, como o KSSVM, e as técnicas de seleção de atributos também não lineares (PÖLSTERL, 2020; PÖLSTERL, 2023).

4.2.1 MODELOS PARA PREDIÇÃO DE SOBREVIDA

A aplicação de técnicas de AM na área da saúde possibilita o desenvolvimento de modelos computacionais promissores, conforme discutido nas Subseções2.2 e 2.2.3. No contexto da predição de sobrevida para o CM, esses métodos desempenham um papel fundamental na construção de modelos prognósticos capazes de estimar a expectativa de vida pós-diagnóstico, incorporando os avanços no tratamento e na detecção precoce da doença (BUSTAMANTE-TEIXEIRA et al., 2002; LI et al., 2021; MIN et al., 2021).

Os modelos de Regressão de Cox, tanto não penalizados quanto penalizados, descritos na Subsubseção 2.2.3.1, são amplamente utilizados na análise de sobrevida na área da oncologia devido à sua interpretabilidade. No entanto, eles apresentam limitações, como inadequação para lidar com dados de alta dimensão e dificuldade em modelar não-linearidades (LIU et al., 2020; COX, 1972; MONCADA-TORRES et al., 2021). Por outro lado, os modelos penalizados, como CPH-EN, CPH-L e CPH-R, introduzem penalidades *Ridge*, *Lasso* e *Elastic Net* nos coeficientes do CPH, visando melhorar as predições e reduzir essas limitações (SIMON et al., 2011).

Além disso, os modelos EL, como GBS e RSF, além de métodos de análise discriminante como SSVM, são ferramentas promissoras na modelagem da predição de sobrevida. O GBS, uma estrutura versátil de otimização, que utiliza árvores de decisão para aprender padrões complexos para calcular a função de sobrevida, como mostrado na Subsubseção 2.2.3.2 (LIU et al., 2020; MONCADA-TORRES et al., 2021; XIAO et al., 2022; HOTHORN et al., 2006). Já o RSF, incorpora informações de censura nas regras de divisão das árvores para estimar robustamente a função de sobrevida, como demonstrado na Subsubseção 2.2.3.3 (MONCADA-TORRES et al., 2021; XIAO et al., 2022; ISHWARAN et al., 2008). Por fim, o SSVM é projetado especificamente para dados de sobrevida, maximizando margens entre eventos e censura, conforme apresentado na Subsubseção 2.2.3.4 (MONCADA-TORRES et al., 2021; XIAO et al., 2022; PÖLSTERL et al., 2015).

4.2.2 ATRIBUTOS e HIPERPARÂMETROS DOS MODELOS

O conjunto de dados utilizado neste estudo contém informações detalhadas de pacientes diagnosticadas com CM, abrangendo aspectos relevantes da prática clínica. Após as etapas de pré-processamento, serão selecionadas as variáveis prognósticas e as variáveis de desfecho, incluindo o *status* [0-censurado ou 1-evento] e tempo de sobrevida em dias. Esses dados permitirão análises abrangentes no contexto do CM, contribuindo para a construção de modelos prognósticos mais robustos e assertivos (CINTRA et al., 2012; FAYER, 2014; CINTRA, 2012).

Para as simulações dos modelos CPH, CPH-EN, CPH-L, CPH-R, GBS, RSF, SSVM e KSSVM — métodos de AM supervisionados selecionados neste estudo — foram

seguidas as recomendações da biblioteca "*Scikit-survival*". A divisão dos dados será feita entre conjunto de treinamento (75%) e teste (25%), e os hiperparâmetros de cada modelo serão ajustados conforme as diretrizes estabelecidas pela biblioteca (PÖLSTERL, 2020; PÖLSTERL, 2023), conforme descrito a seguir:

1. **CPH (não penalizado):** O número máximo de iterações foi definido como 100, com α igual a 0, utilizando o subconjunto de recursos sem regularização. Os coeficientes foram estimados usando o método *Breslow*, considerando cada evento em um momento específico como distinto.

2. **CPH (penalizado):** As penalidades *Lasso* e *Elastic Net* foram aplicadas, com número de α no caminho da regularização igual a 100; número mínimo de α 0.01 e o parâmetro da penalidade *Lasso* igual a 1.0. Para a combinação ponderada das penalidades *Lasso* e *Ridge*, no caso do *Elastic Net* igual a 0.9. A função de sobrevida e de risco foi estimada para cada α. Para se estimar o intervalo de α para o conjunto de dados e identificar os coeficientes não nulos, foi aplicada a validação cruzada K-Fold, com $K = 5$.

3. **GBS:** O número de árvores de regressão na floresta foi igual a 100, com os tamanhos mínimos de amostra necessários para dividir um nó e permanecer em um único nó definidos como 2 e 1, respectivamente. A taxa de aprendizado foi definida igual a 1.0 para controlar a influência de cada árvore, com a aleatoriedade da amostra de *bootstrap* definida como 20. A profundidade máxima das árvores de regressão individuais foi limitada a 1 para otimizar o desempenho.

4. **RSF:** O número de árvores da floresta igual a 1.000, com tamanhos mínimos de amostra necessários para dividir um nó e permanecer em um único nó definidos como 10 e 15, respectivamente. A aleatoriedade da amostra de *bootstrap* foi configurada para 20 e utilizaremos o teste *log-rank* como critério para a construção de cada árvore.

5. **SSVM:** O número máximo de iterações definido igual a 20.000 para a penalização da função objetiva como 1 e usando o Red-black Trees (rbtree) para otimizar o equilíbrio entre a regressão e o objetivo de classificação definido igual a 0.5.

6. **KSSVM:** o número máximo de iterações como 20.000 usando a função de similaridade kernel RBF descrita na Subseção 2.2.3.4, com o parâmetro ι igual a 0,01 que representa uma configuração em que a função de similaridade é menos sensível a pequenas diferenças e considera uma similaridade mais ampla entre os pontos. Configuramos a penalização da função objetiva como 1 e usando o rbtree para otimizar o equilíbrio entre a regressão e o objetivo de classificação definido igual a 0.5.

Com exceção do modelo não linear KSSVM, que foi adaptado à biblioteca utilizada, os hiperparâmetros foram testados para identificar aqueles que garantiam a convergência do método (PÖLSTERL, 2023). Para essa tarefa, recorreu-se à formulação matemática da função de similaridade Kernel RBF, descrita na Subseção 2.2.3.4, chegando-se ao valor de $0,01$ para o parâmetro ι. Além disso, destaca-se que a implementação dos parâmetros foi mantida semelhante entre os modelos linear e não linear do SSVM, permitindo uma melhor comparação de desempenho entre eles. .

4.3 SELEÇÃO DE ATRIBUTOS

Os métodos de AM destacam-se como ferramentas inovadoras na análise de sobrevida, especialmente quando combinados com técnicas de seleção de atributos, essenciais para identificar fatores que impactam a taxa de sobrevida e aprimorar a construção de modelos preditivos. A aplicação desses métodos é amplamente aceita na área médica, como demonstram estudos de referência (XIAO et al., 2022; GANGGAYAH et al., 2019; FANIZZI et al., 2023), sendo uma abordagem valiosa para identificar variáveis clínicas relevantes por meio de métodos de seleção de características e análise da importância de variáveis (VERÍSSIMO et al., 2016; KRZYZIŃSKI et al., 2023; PÖLSTERL, 2020; SIMON et al., 2011).

Para a seleção de atributos, foram utilizados os métodos da biblioteca "*Scikit-survival*"(PÖLSTERL, 2023). Os dados foram divididos em conjuntos de treinamento (75%) e teste (25%), e a validação foi realizada com a métrica de desempenho C-Index nos dados de teste (PÖLSTERL, 2023).

4.3.1 SELEÇÃO DE CARACTERÍSTICAS

A modelagem de dados oncológicos de sobrevida apresenta desafios devido à alta dimensionalidade e à crescente disponibilidade de informações clínicas. Mesmo que os dados não possuam uma alta dimensionalidade, a aplicação de métodos de seleção de atributos é fundamental, pois aprimora a predição dos modelos ao utilizar apenas variáveis essenciais, evitando o uso de todas as variáveis disponíveis, o que pode resultar em saídas complexas e de difícil interpretação (XIAO et al., 2022).

Uma abordagem promissora nesse contexto é o uso de métodos de regularização, que penalizam estruturas complexas no espaço da solução. No caso da regressão de Cox (CPH), a penalidade *Lasso*, conforme descrita na Equação 2.17, é amplamente empregada para reduzir a dimensionalidade dos dados, promovendo uma seleção contínua de características ao forçar alguns coeficientes a zero. Já o *Elastic Net*, conforme definido na Equação 2.18, combina as penalidades *Lasso* (Equação 2.17) e *Ridge* (Equação 2.16), oferecendo maior estabilidade na seleção de características, especialmente quando há alta correlação entre variáveis (SIMON et al., 2011; PÖLSTERL, 2020; VERÍSSIMO et al., 2016).

Exemplo de aplicação de modelos de CPH penalizados pode ser encontrado no estudo de Verissimo et al. (2016), que propõe o modelo DEGREECOX. Este modelo utiliza regularizadores baseados em redes para estimar a Regressão de Cox em dados de expressão gênica, ilustrando como modelos penalizados, como *Lasso* e *Elastic Net*, melhoram a identificação de variáveis relevantes em contextos oncológicos, dada a complexidade dos dados genômicos (VERÍSSIMO et al., 2016).

A regularização visa reduzir a complexidade do modelo penalizando os coeficientes conforme critérios como tamanho ou esparsidade, conforme descrito na Subseção 2.2.3.1. A regularização *Ridge* (norma l_2), mostrada na Equação 2.16, aplica uma penalidade quadrática aos coeficientes, reduzindo seus valores proporcionalmente, mas mantendo todos no modelo, o que a torna inadequada para seleção de características. Por outro lado, a regularização *Lasso* (norma l_1), descrita na Equação 2.17, impõe uma penalidade absoluta que pode reduzir coeficientes a zero, eliminando variáveis do modelo. A regularização *Elastic Net*, definida na Equação 2.18, combina as penalidades de *Ridge* e *Lasso* (combinando normas l_1 e l_2), permitindo uma seleção de variáveis mais flexível e robusta(PÖLSTERL, 2020; PÖLSTERL, 2023; SIMON et al., 2011).

Além disso, a seleção não linear de características utilizando GBS, com a aplicação dos mínimos quadrados nos componentes dos alunos de base, permite substituir o modelo linear por um modelo não linear, adicionando a função aditiva do GBS. Ao recuprar o valor dos coeficientes não nulos reduzindo o número de variáveis e tornando-se fácil de interpretação, mesmo após muitas iterações (PÖLSTERL, 2023).

Os hiperparâmetros de cada método de seleção de características foram definidos com base nas recomendações da biblioteca "*Scikit-survival*", sendo estabelecidos da seguinte forma:

1. **CPH-L:** Na seleção linear usando a regularização *Lasso* aplicadaa extensão da penalidade *Elastic Net*, com número de α no caminho da regularização igual a 100, número mínimo de α igual a 0.01 e o parâmetro de regularização da penalidade *Lasso* igual a 1.0. A função de sobrevida e de risco será estimada para cada α, selecionando um conjunto reduzido de características para valores elevados de α.

2. **CPH-EN:** Na seleção linear aplicada a regularização *Elastic Net* utilizando a extensão da penalidade *Elastic Net*, com número de α no caminho da regularização igual a 100, número mínimo de α igual a 0.01 e o parâmetro da regularização para a combinação ponderada das penalidades *Lasso* e *Ridge*, igual a 0.9. A função de sobrevida e de risco será estimada para cada α, oferecendo maior estabilidade através da combinação das penalidades e permitindo a seleção de todos os recursos em grupos altamente correlacionados.

3. **GBS:** Na seleção não linear usando os mínimos quadrados nos componentes dos

alunos de base do GBS, por meio da otimização da função de perda do CPH. A contribuição da taxa de aprendizado de cada aluno de base igual a 1 variando o número de estimadores até 300, após várias interações resulta na seleção das variáveis com coeficiantes difentes de 0.

Para identificar o conjunto ótimo de parâmetros α e suas generalizações, será utilizada uma funcionalidade da biblioteca "*Scikit-survival*"(PÖLSTERL, 2023). Esta ferramenta permite ajustar a intensidade da penalidade α aplicada nas regularizações *Lasso* e *Elastic Net*, visando otimizar a previsão do tempo de sobrevida. A seleção de parâmetros será realizada por validação cruzada com $k = 5$, conforme descrito em Polsterl (2020) (PÖLSTERL, 2020; PÖLSTERL, 2023). A validação cruzada divide os dados em k subconjuntos, treina o modelo em $k - 1$ subconjuntos e avalia o desempenho no subconjunto restante, repetindo o processo k vezes e determinando a eficácia do modelo pela média das predições (PÖLSTERL, 2023).

Na análise, foram utilizados o estimador "*GridSearch CV*"e o normalizador "*StandardScaler*"da biblioteca "*Scikit-learn*"(PEDREGOSA et al., 2011), compatíveis com a "*Scikit-survival*"(PÖLSTERL, 2020; PÖLSTERL, 2023). O "*GridSearch CV*"define os valores de α a serem testados nos modelos CPH-L e CPH-EN, enquanto o "*StandardScaler*"ajusta as variáveis para uma escala comum, facilitando a comparação entre os coeficientes e a seleção de atributos relevantes (PÖLSTERL, 2023; PEDREGOSA et al., 2011).

4.3.2 SELEÇÃO DE IMPORTÂNCIA DE VARIÁVEL

No contexto dos métodos de AM supervisionados aplicados à análise de sobrevida, a seleção de importância das variáveis tem como objetivo classificar o impacto das variáveis prognósticas na predição do modelo, visando criar modelos interpretáveis e identificar os indicadores mais relevantes por meio de técnicas de permutação. A métrica utilizada para avaliação é o C-Index, que compara o desempenho do modelo antes e após a permutação das colunas de características no conjunto de validação (dados de teste) (PÖLSTERL, 2020; PÖLSTERL, 2023; FANIZZI et al., 2023; KRZYZIŃSKI et al., 2023).

Neste trabalho, foram implementados os métodos de permutação de importância de variáveis da biblioteca "*Scikit-survival*"(PÖLSTERL, 2023). Esses métodos avaliam o impacto de cada variável na predição do modelo, baseando-se na implementação do RF da biblioteca "*Scikit-learn*"(PEDREGOSA et al., 2011), adaptada para os modelos GBS e RSF, conforme as recomendações da "*Scikit-survival*"(PÖLSTERL, 2020; PÖLSTERL, 2023).

A avaliação do desempenho com o C-Index (Eq.4.7), descrita na Subseção4.4.1, nos dados de teste após a permutação das variáveis nos dados de treinamento, é essencial para identificar a contribuição específica de cada variável para o modelo (PÖLSTERL,

2020; PÖLSTERL, 2023; PEDREGOSA et al., 2011). A análise foi realizada utilizando o método de permutação de importância para GBS e RSF, com 15 permutações, e a função *permutation importance* definida por Breiman (2001) e implementada no "*Scikit-learn*"(BREIMAN, 2001; PÖLSTERL, 2023; PEDREGOSA et al., 2011).

Portanto, a metodologia proposta integra métodos lineares de seleção de características, como *Lasso* e *Elastic Net*, e não lineares, como GBS, além de avaliar a importância das variáveis por permutação nos modelos GBS e RSF. Essa combinação proporciona uma abordagem abrangente para identificar variáveis relevantes, aprimorando a interpretabilidade e o desempenho dos modelos preditivos de sobrevida, além de destacar as variáveis mais importantes para a inclusão nos modelos.

4.4 VALIDAÇÃO E AVALIAÇÃO DE DESEMPENHO

De acordo com a revisão realizada por Min et al. (2021) sobre os avanços em modelos prognósticos para o CM, a avaliação dos métodos prognósticos deve considerar quatro dimensões principais (MIN et al., 2021):

1. **Discriminação:** Mede a capacidade do modelo de distinguir entre pacientes com diferentes desfechos clínicos, utilizando métricas como o C-Index.
2. **Calibração:** Avalia a concordância entre as previsões do modelo e os resultados observados, podendo utilizar testes de adequação, como o de *Hosmer-Lemeshow*.
3. **Desempenho Geral:** Integra as dimensões de discriminação e calibração, recorrendo a métricas como BS e IBS.
4. **Utilidade Clínica:** Examina a capacidade do modelo de fornecer orientações úteis para decisões médicas, utilizando métricas como taxa de precisão (sensibilidade/especificidade), análise de curvas de sobrevida de *Kaplan-Meier* e a análise de curvas de decisão.

Este estudo focou no uso da biblioteca *Scikit-survival* para avaliar o desempenho dos modelos de AM supervisionado aplicados à predição de sobrevida, considerando as particularidades dos dados de análise de sobrevida, como a censura (PÖLSTERL, 2023). As métricas implementadas foram desenvolvidas especificamente para esse tipo de análise e incluem o C-Index, que mede a capacidade discriminativa do modelo, além do BS e IBS, que avaliam a calibração levando em conta a dependência temporal. A aplicação dessas métricas permitirá analisar a eficácia do modelo na predição dos tempos de sobrevida, considerando tanto a censura dos dados quanto às particularidades do contexto (PÖLSTERL, 2020; PÖLSTERL, 2023; GRAF et al., 1999; UNO et al., 2011).

As métricas de desempenho implementadas na biblioteca *Scikit-survival* podem ser descritas com base nos dados de sobrevida $(X_1, Y_1, \delta_1), \ldots, (X_n, Y_n, \delta_n)$, onde n é o número de pacientes, X_i é o vetor de covariáveis do paciente i, Y_i é o último tempo observado do paciente i e δ_i é a variável indicadora, sendo que $\delta_i = 1$ indica a ocorrência do evento de interesse e $\delta_i = 0$ indica que o evento foi censurado (PÖLSTERL, 2023). A seguir, são apresentadas as descrições matemáticas de cada uma das métricas aplicadas (C-Index, BS e IBS).

4.4.1 *CONCORDANCE INDEX*

O C-index quantifica a capacidade de discriminação do modelo medindo a concordância entre os pares de eventos, comparando com os casos discordantes. Os pares descartados incluem aqueles em que o menor tempo de evento (morte) é o mesmo, a menos que o evento ocorra em ambos, e aqueles em que nenhum dos eventos ocorre. Assim, o método avalia a habilidade dos modelos em classificar corretamente os tempos de morte dos pacientes (MONCADA-TORRES et al., 2021; PÖLSTERL, 2020; UNO et al., 2011).

Esta métrica proposta inicialmente por Harrell et al. (1996), definida como a proporção de todos os pares comparáveis em que as previsões e os resultados são concordantes e discordantes (JR et al., 1996; PENCINA; D'AGOSTINO, 2004; UNO et al., 2011). Matematicamente, podemos expressar esta métrica introduzindo a seguinte notação:

$$\pi_C = P(\widehat{Y}_i < \widehat{Y}_j \; e \; Y_i < Y_j) + P(\widehat{Y}_i > \widehat{Y}_p \; e \; Y_i > Y_j), \tag{4.1}$$

$$\pi_D = P(\widehat{Y}_i < \widehat{Y}_j \; e \; Y_i > Y_j) + P(\widehat{Y}_i > \widehat{Y}_j \; e \; Y_i < Y_j), \tag{4.2}$$

onde π_C é a probabilidade de concordância e π_D é a probabilidade de discordância; $\widehat{Y}_i$ e $\widehat{Y}_j$ são as probabilidades previstas de sobrevida, considerando os pares i e j, com $i \neq j$; e Y_i e Y_j são os pares de tempos observados (JR et al., 1996; PENCINA; D'AGOSTINO, 2004).

Primeiramente, vamos mostrar o C-index C_H proposto por Harrel (1996), que considera todas as observações, expresso por:

$$C_H = \frac{\sum_{i \neq j}^{n} \delta_i (Y_i < Y_j) I(\widehat{\beta} X_i > \widehat{\beta} X_j)}{\sum_{i \neq j}^{n} \delta_i (Y_i < Y_j)}, \tag{4.3}$$

onde I corresponde à função indicadora de concordância e $\widehat{\beta}X$ é o estimador da função de verossimilhança parcial (2.15) para o vetor de covariáveis X, considerando os pares i e j, com $i \neq j$ (JR et al., 1996; UNO et al., 2011).

Agora, introduzindo a técnica proposta por Cheng et al. (1995) para a estimação generalizada nas análises de sobrevida, temos:

$$\widehat{C}_H = \frac{\sum_i^n \sum_j^n \delta_i \widehat{G}(Y_i)^2 I(Y_i < Y_j) I(\widehat{\beta} X_i > \widehat{\beta} X_j)}{\sum_i^n \sum_j^n \delta_i \widehat{G}(Y_i)^2 I(Y_i < Y_j)}, \tag{4.4}$$

onde $\widehat{G}(Y)$ corresponde ao estimador de *Kaplan-Meier* (Eq. 2.21) (CHENG et al., 1995; UNO et al., 2011).

Já o C-Index C_P proposto por Pencina e D'Agostino (2004), que considera apenas as observações até um ponto de tempo τ pré-definido, é expresso matematicamente por:

$$C_P = \frac{\sum_{i \neq j}^n \delta_i (Y_i < Y_j, Y_i < \tau) I(\widehat{\beta} X_i > \widehat{\beta} X_j)}{\sum_{i \neq j} \delta_i (Y_i < Y_j, Y_i < \tau)}, \tag{4.5}$$

com τ limitado $P(\widehat{\beta} X_i > \widehat{\beta} X_j \mid Y_i < Y_j, Y_i \leq \delta_i \wedge \delta_j, Y_i < \tau)$ (PENCINA; D'AGOSTINO, 2004; UNO et al., 2011).

Agora, introduzindo a técnica proposta por Cheng et al. (1995) (CHENG et al., 1995; UNO et al., 2011), temos:

$$\widehat{C}_P = \frac{\sum_i^n \sum_j^n \delta_i \widehat{G}(Y_i)^2 (Y_i < Y_j, Y_i < \tau) I(\widehat{\beta} X_i > \widehat{\beta} X_j)}{\sum_i^n \sum_j^n \delta_i \widehat{G}(Y_i)^2 (Y_i < Y_j, Y_i < \tau)}. \tag{4.6}$$

Podemos estabelecer o intervalo de confiança entre as métricas de C-index propostas, tomando $\epsilon = C_H - C_P$ e $\widehat{\epsilon} = \widehat{C}_H - \widehat{C}_P$. Note que essas duas métricas estão correlacionadas, resultando em:

$$E_\epsilon = n^{\frac{1}{2}} (\widehat{\epsilon} - \epsilon), \tag{4.7}$$

onde E_ϵ corresponde ao estimador de distribuição assintótica dos intervalos de confiança ϵ e $\widehat{\epsilon}$. Neste trabalho, o C-index aplicado baseia-se na proposta de Uno et al. (2011), que utiliza um estimador de probabilidade inversa que não depende da distribuição dos tempos de censura nos dados do teste, e é disponibilizado pela biblioteca "*Scikit-survival*"(UNO et al., 2011; PÖLSTERL, 2023).

O C-Index é uma métrica que quantifica a concordância entre os escores de risco previstos pelo modelo e os tempos de eventos observados nos dados de sobrevivência, medindo a capacidade do modelo em prever e classificar os tempos de morte dos pacientes. O intervalo de desempenho do método assume valores de $0,0$ a $1,0$, onde $0,5$ indica desempenho médio, sem capacidade discriminativa, e $1,0$ indica desempenho ótimo, referente a um modelo capaz de separar com precisão pacientes com diferentes desfechos (MONCADA-TORRES et al., 2021; XIAO et al., 2022; PINHEIRO et al., 2022; UNO et al., 2011).

4.4.2 *BRIER SCORE*

O BS corresponde a uma medida similar ao erro quadrático médio, avaliando a calibração do modelo, refletindo o quão bem as previsões de sobrevida correspondem aos eventos reais observados nos momentos t selecionados do teste (PÖLSTERL, 2020; PÖLSTERL, 2023; GRAF et al., 1999). Por ser uma métrica dependente do tempo, o BS é calculado da seguinte forma:

$$BS(t) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq t \wedge \delta_i = 1)\frac{(0 - \hat{\pi}(t \mid X_i))^2}{\hat{G}(Y_i)} + I(Y - i \geq t)\frac{(1 - \hat{\pi}(t \mid Xi))^2}{\hat{G}(t)}, \quad (4.8)$$

onde $\hat{\pi}(t \mid X_i)$ representa a probabilidade prevista de um indivíduo i permanecer livre de eventos até o tempo t para o vetor de covariáveis X_i, e $1 - \hat{G}(t)$ é a probabilidade inversa de censura, calculada usando o estimador *Kaplan-Meier* (Eq. 2.21) (GRAF et al., 1999; PÖLSTERL, 2023).

O BS é uma métrica que compara as probabilidades previstas pelo modelo com o *status* real dos pacientes em momentos específicos do teste. Desta forma, avalia a qualidade de predição do método através da probabilidade do paciente permanecer livre do evento (morte), onde valores mais baixos sugerem melhores resultados, indicando uma melhor calibração do modelo e representando de forma satisfatória a previsão individual do método para cada paciente (MONCADA-TORRES et al., 2021; XIAO et al., 2022; GRAF et al., 1999).

4.4.3 *INTEGRATED BRIER SCORE*

O IBS é uma extensão do BS (4.8), avaliando a qualidade de predição dos modelos ao longo de todo o tempo de estudo, ou seje nos dados do treino e teste, fornecendo um único do desempenho preditivo geral e da calibração ao longo do tempo (KRZYZIŃSKI et al., 2023; PÖLSTERL, 2020; PINHEIRO et al., 2022; GRAF et al., 1999). O IBS é definido como a integração do BS ao longo do intervalo $[t_1; t_{\max}]$, expresso como:

$$IBS = \int_{t_1}^{t_{\max}} BS(t)dW(t), \quad (4.9)$$

onde $W(t) = t_{\max} - t$ é a função de ponderação, e a integral é estimada através da regra dos trapézios (PÖLSTERL, 2023; GRAF et al., 1999).

Em resumo, tanto o BS quanto o IBS são métricas dependente do tempo e importantes para avaliar a precisão e calibração dos modelos de predição de sobrevida, com o BS avaliando momentos específicos e o IBS fornecendo uma visão geral do desempenho ao longo do tempo (PÖLSTERL, 2023).

5 RESULTADOS E DISCUSSÃO

Os resultados apresentados e discutidos conforme detalhado no Capítulo 4, incluem uma análise detalhada dos dados clínicos após o pré-processamento, destacando as variáveis selecionadas ao longo das etapas de tratamento dos dados, análise de correlação e seleção de atributos. Além disso, é avaliado o desempenho de cada modelo em relação aos diferentes conjuntos de atributos selecionados pelos métodos aplicados. Essa análise busca aprimorar a compreensão dos fatores prognósticos ao longo do curso clínico-terapêutico e sua influência na sobrevida de pacientes com CM femino, avaliando a precisão dos modelos preditivos propostos. Os resultados estão organizados em três seções principais: análise descritiva dos dados, seleção de atributos e desempenho dos modelos computacionais.

5.1 ANÁLISE DESCRITIVA DOS ATRIBUTOS PROGNÓSTICOS PARA O CÂNCER DE MAMA

A análise descritiva dos dados é um passo fundamental para melhorar a compreensão das características da amostra utilizada neste estudo. O conjunto de dados clínicos foi extraído de bases de estudos anteriores (CINTRA, 2012; FAYER, 2014), conforme detalhado na Seção 4.1, onde são apresentados os processos de construção, tratamento e limpeza da base de dados. Após o pré-processamento, a amostra final compreendeu informações de 558 pacientes.

Os estudos de sobrevida desempenham um papel crucial na pesquisa oncológica, contribuindo para o aprimoramento das estratégias clínico-terapêuticas, o aumento do tempo de sobrevivência e a redução do risco de recidiva em pacientes com câncer (TIZI; BERRADO, 2023; BUSTAMANTE-TEIXEIRA et al., 2002). Neste estudo, a análise de sobrevida foi conduzida em um período de cinco anos para avaliar a taxa de sobrevivência das pacientes dentro desse intervalo (LI et al., 2021; CINTRA, 2012).

O tempo de observação foi contabilizado a partir da data do laudo histopatológico até a ocorrência do evento adverso ou censura, abrangendo um total de 1825 dias, conforme definido no BD original (CINTRA, 2012; FAYER, 2014). O evento foi definido como o óbito por CM, caracterizando o término do tempo de observação, enquanto a censura representou os casos em que o evento não ocorreu até o período pré-determinado para a análise de sobrevida (CARVALHO et al., 2011).

Durante esse período, foram registrados 113 óbitos por CM, correspondendo a 20,25% das pacientes, enquanto 79,75% foram censuradas. Essa censura, classificada como censura à direita, é comum nesse tipo de análise, pois indica que, embora o tempo até o óbito seja desconhecido, a paciente permaneceu viva até o limite de tempo estabelecido para o estudo. Esse procedimento reduz a superestimação do risco e melhora a acurácia das estimativas (CARVALHO et al., 2011).

Vale destacar que a biblioteca ("*Scikit-survival*") utilizada para as simulações, além de ser específica para a análise de sobrevida, incorpora esse tipo de censura, o que a torna particularmente adequada para este estudo. Essa característica foi um dos fatores determinantes para sua escolha, garantindo que os modelos computacionais considerassem corretamente a estrutura dos dados censurados e proporcionando uma avaliação mais precisa dos desfechos clínicos (PÖLSTERL, 2023).

5.1.1 DESCRIÇÃO DOS DADOS CLÍNICO

O conjunto de dados compreende informações detalhadas sobre pacientes diagnosticadas com CM invasivo (CID-10 C50), abrangendo uma ampla gama de variáveis prognósticas. Entre essas variáveis, incluem-se idade ao diagnóstico, hábitos como tabagismo e consumo de álcool, histórico familiar de CM e outros tipos de câncer, uso de anticoncepcionais, antecedentes de gravidez e amamentação, além de condições de saúde como hipertensão e doenças cardiovasculares. Também estão contemplados os resultados de exames (mamografia, raio-X, ultrassonografia, citologia e imuno-histoquímica), características histopatológicas do tumor, estadiamento, subclassificação Luminal, esquema terapêutico, registros clínicos ao longo do tratamento, presença de metástases (locorregional e à distância), *status* menopausal e desfechos clínicos.

Inicialmente, o conjunto de dados clínicos incluía 563 pacientes e aproximadamente 220 variáveis, abrangendo aspectos sociodemográficos, clínicos, anatomopatológicos, uso de serviços de saúde, tratamento e acompanhamento de pacientes diagnosticadas com CM entre 2003 e 2005. A coleta de dados foi realizada entre 2009 e 2010, com buscas ativas em 2011, conforme descrito na Subseção 4.1.1.

Os dados foram obtidos manualmente a partir de prontuários hospitalares de centros de referência em oncologia na Zona da Mata Mineira. Devido ao alto índice de valores ausentes, o pré-processamento incluiu a padronização de *strings*, valores numéricos e datas, conforme descrito na Subseção 4.1.3. Após a padronização e inferência de dados ausentes, foram selecionadas variáveis com pelo menos 60% de dados completos, resultando em um conjunto reduzido de 69 variáveis prognósticas e 2 variáveis preditivas—correspondendo a aproximadamente 32% do total inicial. Essa redução foi realizada para minimizar ruídos e enviesamentos, garantindo maior estabilidade e melhor desempenho dos modelos preditivos de sobrevida (LIU et al., 2020).

As pacientes que não possuíam informações essenciais, como data do laudo histopatológico, data de seguimento ou data da cirurgia, foram excluídas do estudo, pois esses dados são fundamentais para estimar o tempo de sobrevida e o intervalo entre o diagnóstico e a intervenção cirúrgica. O laudo histopatológico marca o início da análise de sobrevida, enquanto a data da cirurgia é necessária para calcular o tempo entre diagnóstico e intervenção cirúrgica. Como resultado, duas pacientes foram excluídas devido à ausência

dessas datas, e outras três foram removidas por falta de informações sobre o estadiamento anatômico (descrito no Quadro 2), uma variável fundamental para a inferência e tratamento de outras variáveis. Após esse refinamento, o conjunto final compreendeu 558 amostras e 71 variáveis.

A distribuição etária das pacientes revela que a maioria (71,68%) tem entre 40 e 69 anos, com idade média de 58 anos. Essa faixa etária é relevante, pois influencia tanto o prognóstico quanto as estratégias de tratamento. Entre os tratamentos mais comuns estão cirurgia (curativa em 84,95% dos casos e diagnóstica em 14,52%), terapia sistêmica (quimioterapia em 59,14% e hormonioterapia em 28,14%) e radioterapia (73,83%). A combinação desses tratamentos varia conforme o estadiamento do câncer e outros fatores clínicos, como a expressão de marcadores imuno-histoquímicos. Durante o período de estudo, 4,84% das pacientes apresentaram recidiva, enquanto 27,06% desenvolveram metástases locorregionais ou à distância, destacando a importância do monitoramento contínuo e da adaptação dos planos terapêuticos.

Tabela 1 – Resumo da descrição das tipos de dados do Banco de Dados Clínicos.

Dados	**Descrição**	**FA**	**FR**
Categóricos	*Booleanos*	43	61,56%
	Não *Booleanos*	20	28,17%
Numéricos	Inteiros	4	5,63%
	Reais	2	2,82%
Preditores	Evento	1	1,41%
	Tempo	1	1,41%
Total		**1**	**100,00%**

Fonte: Elaborada pela autora (2024).

Essas informações proporcionam uma visão abrangente da população estudada, permitindo uma análise detalhada das variáveis no contexto clínico. Para uma melhor organização, os dados foram categorizados em três grupos: categóricos (*booleanos* e não *booleanos*), numéricos (inteiros e reais) e preditores (evento e tempo), com um resumo apresentado na Tabela 1. Nessa tabela, observa-se o conjunto final de 71 atributos selecionados após o pré-processamento, sendo 69 variáveis prognósticas e 2 variáveis preditoras. Dentre elas, $88,73\%$ correspondem a variáveis categóricas, $8,45\%$ a variáveis numéricas e $2,82\%$ a variáveis preditoras. A descrição completa de cada tipo de variável está detalhada nos Quadros 7, 8, 9, 10, 11, 12 e 13, enquanto a análise descritiva dos dados encontra-se nas Tabelas 2, 3, 4, 5, 6 e 7.

5.1.1.1 Variáveis Categóricas *Booleanas*

As variáveis categóricas *booleanas* incluem informações completas de todas as pacientes selecionadas no estudo após o processo de pré-processamento de dados, que

também abrangeu a inferência de valores ausentes, conforme descrito na Subseção 4.1.3. A descrição detalhada dessas variáveis encontra-se nos Quadros 7, 8 e 9, e uma análise estatística descritiva aprofundada pode ser consultada na Tabelas 2 e 3.

5.1.1.2 Variáveis Categóricas não *Booleanas*

As variáveis categóricas não *booleanas* passaram pelo mesmo pré-processamento, incluindo a inferência de valores faltantes. As variáveis do tipo *string* foram convertidas em valores numéricos (reais) para facilitar a aplicação dos métodos utilizados neste estudo. As descrições dessas variáveis estão organizadas nos Quadros 10 e 11, e a análise estatística descritiva pode ser conferida na Tabela 4.

5.1.1.3 Variáveis Numéricas

As variáveis numéricas, apresentadas no Quadro 12, compreendem dados em formatos inteiros e reais. Os valores ausentes ou não coletados receberam o valor de -1 durante o primeiro pré-processamento; posteriormente, foram inferidos no segundo pré-processamento, como descrito na Subseção 4.1.3. Uma análise detalhada dessas variáveis está documentada nas Tabelas 5 e 6 .

5.1.1.4 Variáveis Preditivas

As variáveis preditivas, descritas no Quadro 13, indicam a ocorrência ou não do evento adverso (óbito por CM), incluindo informações de censura e o tempo de sobrevida em dias de cada paciente, com um limite de 1825 dias. Essas variáveis são essenciais para a avaliação do desempenho dos modelos de predição de sobrevida. A análise estatística detalhada dessas variáveis está disponível na Tabela 7.

Essas 69 variáveis prognósticas, compõem o BD clínico que servirá como base para as análises de correlação e seleção de atributos subsequentes. Esse banco será então utilizado na criação do **BD 1**, destinado à aplicação dos métodos de seleção de atributos para os modelos de predição.

5.1.2 ANÁLISE DESCRITIVA DOS DADOS CLÍNICO

A análise descritiva é uma etapa estatística que resume as principais características de um conjunto de dados, proporcionando uma compreensão inicial das tendências e variações presentes nas informações. Neste estudo, foram calculadas as medidas de média, DP, moda, Vmin e Vmax para cada variável do conjunto de dados clínicos. A média fornece uma visão do valor central de cada variável, o DP indica a dispersão dos dados, a moda representa o valor mais frequente, e Vmin e Vmax evidenciam os valores extremos de cada atributo (ASSIS et al., 2019).

Quadro 7 – Descrição dos dados categóricos *booleanos* do Banco de Dados Clínicos.

Atributo	**Descrição**
Amamentação	Se realizou amamentação após o nascimento (1- Sim; 0- Não)
Anticoncenpcionais	Se fez uso de contraceptivos orais (1- Sim; 0- Não)
Atraso Quimioterapia	Se ocorreu atraso ou interrupção da quimioterapia (1- Sim; 0- Não)
Atraso Radioterapia	Se ocorreu atraso ou interrupção da radioterapia (1- Sim; 0- Não)
Cintilografia Óssea	Resultados do exame de cintilografia óssea (1- Sim; 0- Não)
Citologia	Resultados do exame de citologia (1- Presente; 0- Ausente)
Classificação Cirurgia	Classificação do tipo de cirurgia realizada (1- Curativa; 0- Diagnóstica)
Classificação Estrogênio	Classificação do resultado do receptor de estrogênio (1- Positivo; 0- Negativo)
Classificação HER2	Classificação do resultado do receptor HER2 (1- Positivo; 0- Negativo)
Classificação Progesterona	Classificação do resultado do receptor de progesterona (1- Positivo; 0- Negativo)
Componente Intraductal	Comprometimento do componente intraductal (1- Sim; 0- Não)
Diagnóstico Clínico	Diagnóstico clínico primeira suspeita (1- Presente; 0- Ausente)
Doença Cardiovascular à Distância	Diagnóstico de doença cardiovascular (1- Sim; 0- Não)
Doença de Hipertensão	Diagnóstico de doença de hitensão arterial (1- Sim; 0- Não)
Estadiamento Clínico M	Estadiamento clínico da presença de metástases (0- 0; 1- 1)
Estadiamento Patológico M	Estadiamento patológico da presença de metastase (0- 0; 1- 1)
Esvaziamento Axilar	Se realizou esvaziamento axilar na cirurgia (1- Sim; 0- Não) 4- Nunca usou)
Extensao do Tumor	Extensão do tumor para pele da mama, mamilo, musculatura peitoral (1- Sim; 0- Não)
Gravidez	Se já teve alguma gestação completa (1- Sim; 0- Não)

Fonte: Elaborada pela autora (2024).

Nesta análise descritiva, as variáveis foram agrupadas em dados categóricos, numéricos e preditivos, conforme detalhado na subseção anterior. A Figura 13 apresenta a

Quadro 8 – Descrição dos dados categóricos *booleanos* do Banco de Dados Clínicos.

Atributo	**Descrição**
Histórico de Câncer	Histórico Familiar de Câncer de seus parentes de primeiro e segundo graus (1- Presente; 0- Ausente)
Histórico de CM	Histórico Familiar de CM de seus parentes de primeiro e segundo graus (1- Presente; 0- Ausente)
Hormonioterapia	Paciente submetido à hormonioterapia (1- Sim; 0- Não)
Infiltrado Inflamatório	Identificação de infiltrado inflamatório no resultado histopatológico (1- Sim; 0- Não)
Invasao Perineural	Identificação de invasão perineural no resultado histopatológico (1- Sim; 0- Não)
Invasão Vascular	Identificação de invasão vascular no resultado histopatológico (1- Sim; 0- Não)
Linfonodo Sentinela	Identificação de linfonodos sentinelas (1- Sim; 0- Não)
Mamografia	Informações dos resultados da mamografia (1- Presente; 0- Ausente)
Marcadores Tumorais	Informações dos marcadores tumorais (1- Sim; 0- Não)
Menstruação	Se ainda tem ciclo menstrual (1- Presente; 0- Ausente)
Metástase à Distância	Diagnóstico de metástase à distância (1- Sim; 0- Não)
Metástase Locorregional	Diagnóstico de metástase locorregional
Multicentricidade	Identificação de multicentricidade no resultado histopatológico (1- Sim; 0- Não)
Multifocalidade	Identificação de multifocalidade no resultado histopatológico (1- Sim; 0- Não)

Fonte: Elaborada pela autora (2024).

distribuição percentual das variáveis por tipo de dados. A descrição estatística das variáveis categóricas, numéricas e preditivas utilizadas no banco de dados clínicos está organizada nas Tabelas 2, 3, 4, 5, 6 e 7, permitindo uma visão abrangente das características clínicas e prognósticas das pacientes diagnosticadas com CM feminino consideradas neste estudo.

Quadro 9 – Descrição dos dados categóricos *booleanos* do Banco de Dados Clínicos.

Atributo	**Descrição**
Perfil Imuno-histoquímico	Identificação dos marcadores imuno-histoquímico (1- Presentes; 0- Ausentes)
Quimioterapia	Paciente submetido à quimioterapia (1- Sim; 0- Não)
Radioterapia	Paciente fez tratamento radioterápico (1- Sim; 0- Não)
Raio X do Tórax	Resultados do exame de raio X do tórax (1- Sim; 0- Não)
Recidiva Locorregional	Paciente apresentou recidiva locorregional durante os tratamentos realizados (1- Sim; 0- Não)
Reposição Hormonal	Se fez reposição hormonal (1- Sim; 0- Não) (1- Sim; 0- Não)
Status da Menopausa	*Status* da paciente na pré ou pós-menopausa (0- Pós-menopausa; 1- Pré-menopausa)
Toxicidade	Toxicidade durante o tratamento quimioterápico (1- Sim; 0- Não)
Tratamento Complementar	Paciente fez tratamento complementar (1- Sim; 0- Não)
Ultrasson do Abdomem	Resultados do exame de ultrasson do abdomen (1- Sim; 0- Não)

Fonte: Elaborada pela autora (2024).

5.1.2.1 Análise Descritiva dos Dados Categóricos

Os dados categóricos correspondem a observações que podem ser classificadas em categorias, podendo ser *booleanas* (com dois valores possíveis) ou não *booleanas* (com múltiplos valores possíveis). As variáveis categóricas *booleanas*, apresentadas nas Tabelas 2 e 3, e descritas nos Quadros 7, 8 e 9, incluem informações em formato binário que indicam presença ou ausência de uma característica. Por exemplo, a variável "Hormonioterapia" possui uma média de $0,28$, indicando que menos de um terço das pacientes receberam este tratamento, enquanto "Marcadores Tumorais" apresenta média de $0,94$, sugerindo que a maioria das pacientes apresentaram esta característica clínica.

Já os dados categóricos não *booleanas*, descritos nos Quadros 10 e 11 e analisados na Tabela 4, exibem uma gama mais ampla de valores (de 0 a 7 em alguns casos), indicando múltiplas categorias para cada variável. Por exemplo, a variável "Estadiamento Clínico T" possui uma média de $2,83$ e DP de $1,06$, refletindo variação considerável entre os casos. A amplitude de valores para variáveis como "Estadiamento Anatômico" (Vmin de $0,0$ e Vmax de $7,0$) evidencia a diversidade de estágios clínicos entre as pacientes.

Quadro 10 – Descrição dos dados categóricos não *booleanos* do Banco de Dados Clínicos.

Atributo	**Descrição**
Classificação Doença	Classificação do estágio da doença (1- Inicial; 2- Intermediária; 3- Avançada)
Classificação Histológica	Classificação do tipo histologico do tumor (0- *In situ* 1- Lobular Invasivo; 2- Ductal Invasivo; 3- Outros Tipos)
Classificação Imuno-histoquímica	Classificação do perfil imuno-histoquímico (1- Luminal A; 2- Luminal B HER2-; 3- Luminal B HER2+; 4- Superexpressão HER2+; 5- Triplo Negativo)
Estadiamento Anatômico	Classificação do estadiamento anatômico - Sistema TNM (0- 0; 1- I; 2- IIa; 3- IIb; 4- IIIa; 5- IIIb; 6- IIIc; 7- IV)
Estadiamento Clínico T	Estadiamento clínico do tamanho do tumor (0- 0; 1- is; 2- 1; 3- 2; 4- 3; 5- 4)
Estadiamento Clínico N	Estadiamento clínico do número de linfonodos comprometidos (0- 0; 1- 1; 2- 2; 3- 3)
Estadiamento Patológico T	Estadiamento patológico do tamanho do tumor (0- 0; 1- is; 2- 1; 3- 2; 4- 3; 5- 4)
Estadiamento Patológico N	Estadiamento patológico do número de linfonodos comprometidos (0- 0; 1- 1; 2- 2; 3- 3)
Etilista	Se faz ou fez uso de álcool (1- Etilista Social; 2- Atual; 3- Passado;
Grau Histopatológico	Classificação do grau histopatológico do tumor (1- Bem Diferenciado; 2- Moderadamente Diferenciado; 3- Pouco Diferenciado)

Fonte: Elaborada pela autora (2024).

5.1.2.2 Análise Descritiva dos Dados Numéricos

Os dados numéricos, que podem assumir valores inteiros ou reais, abrangem variáveis quantitativas utilizadas para entender diferentes aspectos clínicos das pacientes (ASSIS et al., 2019). A análise dessas variáveis está apresentada nas Tabelas 5 e 6, e as variáveis inteiras e reais estão listadas no Quadro 12.

Na análise descritiva dos dados inteiros, como por exemplo, o atributo “Dias até a Cirurgia” possui uma média de $29,92$ dias e um DP de $89,57$, sugerindo grande variação entre as pacientes no tempo de espera para o procedimento cirúrgico (com valores variando de 0 a 1253 dias). A variável “Idade”, com média de $58,15$ anos e DP de $13,69$, reflete uma ampla faixa etária entre as pacientes (de 26 a 91 anos). Nas variáveis “Linfonodos Isolados” e “Linfonodos Pós-cirúrgicos” mostra amplitude de valores (Vmin de 0 e Vmax de 56 e 37, respectivamente), indicando variação considerável nessas características clínicas.

Quadro 11 – Descrição dos dados categóricos não *booleanos* do Banco de Dados Clínicos.

Atributo	**Descrição**
Marcador Estrogênio	Informação dos resultados do receptor de estrogênio em + (1,2,3,4,5) (0- -; 1- +; 2- ++; 3- +++; 4- ++++; 5- +++++)
Marcador HER2	Informação dos resultados do receptor HER2 em + (1,2,3,4,5) (0- -; 1- +; 2- ++; 3- +++; 4- ++++; 5- +++++)
Marcador Ki67	Informação dos resultados do receptor Ki67 em + (1,2,3,4,5) (0- -; 1- +; 2- ++; 3- +++; 4- ++++; 5- +++++)
Marcador P53	Informação dos resultados do receptor P53 em + (1,2,3,4,5) (0- -; 1- +; 2- ++; 3- +++; 4- ++++; 5- +++++)
Marcador Progesterona	Informação dos resultados do receptor de progesterona em + (1,2,3,4,5) (0- -; 1- +; 2- ++; 3- +++; 4- ++++; 5- +++++)
Marcadores Imuno-histoquímicos	Informações dos resultados dos marcadores do perfil imuno-histoquimico (0- Luminal A e B HER2-; 1- Luminal B HER2+; 2- Superexpressão HER2+; 3- Triplo Negativo; 4- Desconhecido)
Tabagista	Se faz ou fez uso de cigarro (1- Fumante; 2- Ex-fumante; 3- Nunca fumou)
Tipo de tratamento	Tipos de tratamentos realizados pelo paciente(1- Neoadjuvante; 2- Adjuvante; 3- Neoadjuvante e Adjuvante; 4- Paliativa)
Tratamento Quimioterápico	Paciente realizou ou não os seguintes esquemas de tratamento quimioterápico (0- Antraciclico e Taxano; 1- Antraciclico; 2- CMF; 3- Não realizou)
Tratamento Sistêmico	Paciente realizou ou não tratamento sistêmico quimioterápico/hormonioterápico (1- Hormonioterapia, 2- Antraciclico; 3-Antraciclico e Taxano; 4- CMF; 5- Não fez tratamento sistêmico)

Fonte: Elaborada pela autora (2024).

Para os dados numéricos reais, as dimensões tumorais (eixo X e eixo Y) exibem médias de $3,22$ cm e $2,70$ cm, respectivamente, com DP de $2,79$ cm e $2,45$ cm, indicando

Quadro 12 – Descrição dos dados numéricos do Banco de Dados Clínicos.

Dados	**Atributo**	**Descrição**
Inteiro	Dias Cirurgia	Dias para a realização da cirurgia a contar da data do Laudo Histopatológico
	Idade	Idade do paciente na primeira consulta
	Linfonodos Isolados	Número de linfonodos comprometidos isolados
	Linfonodos Pós-cirúrgico	Número de linfonodos comprometidos pós-cirúrgico
Real	Tamanho Tumor (X)	Tamanho do tumor em cm (eixo x)
	Tamanho Tumor (Y)	Tamanho do tumor em cm (eixo y)

Fonte: Elaborada pela autora (2024).

Quadro 13 – Descrição dos dados preditivos do Banco de Dados Clínicos.

Atributo	**Descrição**
Óbito CM	Pacientes com óbito por CM no período analisado (113)
Dias Sobrevida	Dias de sobrevida após o diagnóstico limitado a 5 anos (1825)

Fonte: Elaborada pela autora (2024).

grande variabilidade na extensão tumoral, que varia entre 0 e 25 cm em ambas as direções.

5.1.2.3 Análise Descritiva dos Dados Preditivos

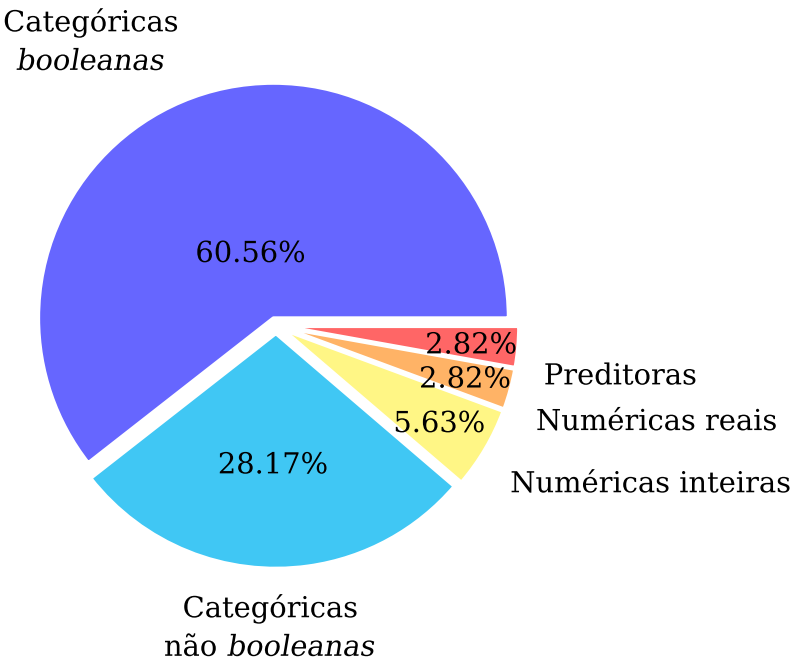
As variáveis preditivas, incluindo o *status* do evento (0 para censurado e 1 para óbito por CM) e o tempo de sobrevida em dias, são essenciais para os modelos computacionais de predição da sobrevida(PÖLSTERL, 2023). Esses dados estão detalhados no Quadro 13, com a análise descritiva apresentada na Tabela 7.

A variável *booleanas* "Óbito por CM" apresenta média de $0,20$, com DP de $0,40$, indicando que aproximadamente 20% das pacientes no conjunto de dados sofreram o evento adverso durante o período de estudo e que aproximadamente 80% das pacientes foram censuradas. Já a variável "Dias de Sobrevida" mostra uma média de $1558,26$ dias e DP de $504,88$, indicando que a maioria das pacientes sobreviveram ao menos até o final do período de análise (1825 dias).

5.1.2.4 ANÁLISE DE CORRELAÇÃO DOS DADOS CLÍNICOS

A análise de correlação foi empregada para investigar as relações entre pares de variáveis clínicas. A correlação, expressa por um coeficiente adimensional que varia de -1 a 1, reflete a intensidade da associação entre as variáveis: valores próximos de zero indicam

Figura 13 – Grafico do tipos de dados do Banco de Dados Clínicos.



Fonte: Elaborada pela autora (2024).

ausência de correlação, enquanto valores absolutos próximos de 1 denotam uma forte correlação (ASSIS et al., 2019). Durante o processo de pré-processamento, foram identificadas variáveis redundantes que, apesar de representarem informações semelhantes, foram coletadas ou expressas de maneiras diferentes. Por exemplo, as variáveis "Classificação Estrogênio"(positivo ou negativo) e "Marcador Estrogênio"(positivo em cruzes ou negativo) representam essencialmente a mesma informação de forma distinta. Da mesma forma, a variável "Estadiamento Anatômico"abrange dados dos estadiamentos clínico e patológico (T, N, M). Esse procedimento teve como objetivo minimizar os ruídos decorrentes da alta correlação entre variáveis semelhantes e evitar problemas de singularidade na matriz de entrada (ASSIS et al., 2019; PÖLSTERL, 2023). A exclusão dessas variáveis permitiu a convergência dos métodos aplicados em cada simulação, sem a necessidade de recorrer a técnicas adicionais, priorizando as variáveis com maior relevância clínica, comumente utilizadas na prática médica (BRASIL, 2014).

Com base nesse raciocínio, as variáveis altamente correlacionadas foram excluídas como parte do aprimoramento do conjunto de dados. A seleção das variáveis a serem removidas seguiu critérios baseados nas medidas de CI para as variáveis qualitativas e CP para as variáveis quantitativas, além de considerar a significância clínica das variáveis correlacionadas. Essas variáveis foram agrupadas em blocos, conforme mostrado nas Tabelas 8 e 9. A exclusão dessas variáveis redundantes e de alta correlação facilitou a

Tabela 2 – Análise Descritiva dos dados categóricos *booleanos*.

Atributo	**Média**	**DP**	**Moda**	**Vmin**	**Vmax**
Amamentação	0,78	± 0,42	1,00	0,00	1,00
Anticoncepcional	0,65	± 0,48	1,00	0,00	1,00
Atraso Quimioterapia	0,13	± 0,33	0,00	0,00	1,00
Atraso Radioterapia	0,03	± 0,18	0,00	0,00	1,00
Cintilografia Óssea	0,80	± 0,40	1,00	0,00	1,00
Citologia	0,08	± 0,28	0,00	0,00	1,00
Classificação Cirurgia	0,85	± 0,35	1,00	0,00	1,00
Classificação Estrogênio	0,66	± 0,47	1,00	0,00	1,00
Classificação HER2	0,19	± 0,39	0,00	0,00	1,00
Classificação Progesterona	0,63	±0,48	1,00	0,00	1,00
Componente Intraductal	0,39	± 0,49	0,00	0,00	1,00
Diagnóstico Clínico	0,71	± 0,45	1,00	0,00	1,00
Doenca Cardiovascular	0,12	± 0,32	0,00	0,00	1,00
Doença de Hipertenção	0,40	± 0,49	0,00	0,00	1,00
Estadiamento Clínico M	0,06	± 0,25	0,00	0,00	1,00
Estadiamento Patológico M	0,07	± 0,25	0,00	0,00	1,00
Esvaziamento Axilar	0,73	± 0,44	1,00	0,00	1,00
Extensão do Tumor	0,12	± 0,33	0,00	0,00	1,00
Gravidez	0,86	± 0,35	1,00	0,00	1,00

Fonte: Elaborada pela autora (2024).

execução das simulações dos modelos de predição, garantindo a estabilidade do processo e a qualidade das análises subsequentes.

Nas Tabelas 8e9, são apresentadas as variáveis que apresentaram alta correlação ($> 0{,}65$) após o pré-processamento e a análise de correlação descrita na Subseção 4.1.2 do Capítulo de Material e Métodos. Nessas Tabelas, as variáveis altamente correlacionadas são identificadas por valores próximos de 1, com as respectivas medidas de correlação associada, tanto numérica quanto categórica, fornecidas pelo relatório da biblioteca *Sweetviz*, a qual foi utilizada tanto para a análise descritiva quanto para a análise de correlação dos dados.

As variáveis altamente correlacionadas, que representam $44,93\%$ do total das variáveis do banco de dados clínico, foram agrupadas em blocos conforme suas características clínicas comuns. A análise de correlação levou em consideração a relevância clínica de cada variável em seu respectivo grupo, além da representatividade de cada uma, priorizando aquelas com o maior número de valores coletados antes do segundo pré-processamento dos dados, como exemplificado pelas variáveis "Tamanho Tumor (X)"e "Tamanho Tumor (Y)". Nas tabelas, as variáveis excluídas são destacadas com um asterisco (*) (19 variáveis), enquanto as mantidas são destacadas em grafite (12 variáveis).

Com a exclusão dessas variáveis altamente correlacionadas, o conjunto de dados foi reduzido em $27,54\%$, passando de 69 para 50 variáveis. Esse novo conjunto, denominado

Tabela 3 – Análise Descritiva dos dados categóricos *booleanos*.

Atributo	**Média**	**DP**	**Moda**	**Vmin**	**Vmax**
Histórico Cancer	0,57	± 0,50	1,00	0,00	1,00
Histórico CM	0,26	± 0,44	0,00	0,00	1,00
Hormonioterapia	0,28	± 0,45	0,00	0,00	1,00
Infiltrado	0,41	± 0,49	0,00	0,00	1,00
Invasão Perineural	0,05	± 0,21	0,00	0,00	1,00
Invasão Vascular	0,11	± 0,31	0,00	0,00	1,00
Linfonodo	0,14	± 0,35	0,00	0,00	1,00
Mamografia	0,65	± 0,48	1,00	0,00	1,00
Marcadores Tumorais	0,94	± 0,23	1,00	0,00	1,00
Menstruação	0,31	± 0,46	0,00	0,00	1,00
Metástase à Distância	0,24	± 0,43	0,00	0,00	1,00
Metástase Locorregional	0,07	± 0,25	0,00	0,00	1,00
Multicentralidade	0,05	± 0,23	0,00	0,00	1,00
Multifocalidade	0,09	± 0,28	0,00	0,00	1,00
Perfil Imuno-histoquímico	0,81	± 0,39	1,00	0,00	1,00
Quimioterapia	0,59	± 0,49	1,00	0,00	1,00
Radioterapia	0,82	± 0,38	1,00	0,00	1,00
Raio X do Tórax	0,89	± 0,31	1,00	0,00	1,00
Recidiva Locorregional	0,05	± 0,21	0,00	0,00	1,00
Resposição Hormonal	0,11	± 0,31	0,00	0,00	1,00
Status Menopausal	0,33	± 0,47	,00	0,00	1,00
Toxicidade Quimioterapia	0,37	± 0,48	0,00	0,00	1,00
Tratamento Complementar	0,96	± 0,19	1,00	0,00	1,00
Ultrasson do Abdomem	0,86	± 0,35	1,00	0,00	1,00

Fonte: Elaborada pela autora (2024).

BD 1, será utilizado nas simulações dos métodos de aprendizado de máquina e na aplicação dos métodos de seleção de atributos. A comparação de desempenho do **BD 1**, com diferentes configurações de subconjuntos de variáveis selecionadas pelos métodos de seleção, proporcionará uma avaliação mais detalhada e abrangente dos modelos de predição de sobrevida para o CM.

5.2 RESULTADOS DA SELEÇÃO DOS ATRIBUTOS PROGNÓSTICOS PARA MODELOS DE SOBREVIDA

Neste estudo, a aplicação de métodos de seleção de atributos foi essencial para aprimorar a análise e predição de sobrevida no CM feminino, além de permitir a validação com dados clínicos utilizados na prática médica. A seleção das variáveis foi realizada por métodos lineares e não lineares, incluindo técnicas baseadas na escolha de características e avaliação da importância das variáveis, com a divisão dos conjuntos de treinamento (75%) e teste (25%). A seleção foi avaliada pela métrica de desempenho C-Index (dados de teste),

Tabela 4 – Análise Descritiva dos dados categóricos não*booleanos*.

Atributo	**Média**	**DP**	**Moda**	**Vmin**	**Vmax**
Classificação Doença	2,00	± 0,80	2,00	1,00	3,00
Classificação Histológica	1,79	± 0,69	2,00	0,00	3,00
Classificação Imuno-histoquímica	2,68	± 1,48	2,00	1,00	5,00
Estadiamento Anatômico	2,80	± 1,99	1,00	0,00	7,00
Estadiamento Clínico N	0,76	± 1,00	0,00	0,00	3,00
Estadiamento Clínico T	2,83	± 1,06	3,00	0,00	5,00
Estadiamento Patológico N	0,75	± 1,00	0,00	0,00	3,00
Estdiamento Patológico T	2,83	± 1,06	3,00	0,00	5,00
Etilista	3,42	± 1,15	4,00	1,00	4,00
Grau Histológico	1,93	± 0,62	2,00	1,00	3,00
Marcador Estrogênio	2,39	± 2,05	0,00	0,00	5,00
Marcador HER2	0,72	± 1,41	0,00	0,00	4,00
Marcador Ki67	2,61	± 1,16	3,00	0,00	4,00
Marcador p53	0,99	± 1,57	0,00	0,00	5,00
Marcador Progesterona	2,09	± 1,92	0,00	0,00	5,00
Marcadores Imuno-histoquímicos	1,42	± 1,66	0,00	0,00	4,00
Tabagista	2,62	± 0,68	3,00	1,00	3,00
Tipo de Tratamento	2,15	± 0,55	2,00	1,00	4,00
Tratamento Quimioterápico	1,79	± 1,09	3,00	0,00	3,00
Tratamento Sistêmico	2,36	± 1,31	2,00	1,00	5,00

Fonte: Elaborada pela autora (2024).

Tabela 5 – Análise Descritiva dos dados numéricos inteiros.

Atributo	**Média**	**DP**	**Moda**	**Vmin**	**Vmax**
Dias Cirurgia	29,92	± 89,57	6,00	0,00	1253,00
Idade	58,15	± 13,69	62,00	26,00	91,00
Linfonoso Isolado	12,91	± 9,72	0,00	0,00	56,00
Linfonodo Pós-cirúrgico	2,54	± 4,78	0,00	0,00	37,00

Fonte: Elaborada pela autora (2024).

Tabela 6 – Análise Descritiva dos dados numéricos reais.

Atributo	**Média**	**DP**	**Moda**	**Vmin**	**Vmax**
Tamanho Tumor (X)	3,22	± 2,79	1,50	0,00	25,00
Tamanho Tumor (Y)	2,70	± 2,45	2,00	0,00	25,00

Fonte: Elaborada pela autora (2024).

conforme detalhado na Seção 4.3 (PÖLSTERL, 2023).

A seleção de características desempenha um papel importante na simplificação da modelagem, ao reduzir a dimensionalidade dos dados e facilitar a visualização. Essa

Tabela 7 – Análise Descritiva dos dados preditivos.

Atributo	**Média**	**DP**	**Moda**	**Vmin**	**Vmax**
Dias Sobrevida	1558,26	± 504,88	1825,00	9,00	1825,00
Óbito CM	0,20	± 0,40	0,00	0,00	1,00

Fonte: Elaborada pela autora (2024).

abordagem contribui para o desenvolvimento de modelos preditivos mais eficientes, evitando a complexidade de incluir indiscriminadamente todas as variáveis, ao mesmo tempo em que favorece uma interpretação mais clara dos resultados (PÖLSTERL, 2020; VERÍSSIMO et al., 2016; SIMON et al., 2011).

Além disso, os métodos de seleção baseados na permutação da importância das variáveis avaliam os fatores prognósticos que impactam a taxa de sobrevida no CM. Esses métodos permitem quantificar o efeito de cada variável na predição, facilitando a interpretação e a identificação das variáveis mais relevantes para o problema analisado (GANGGAYAH et al., 2019; VERÍSSIMO et al., 2016; KRZYZIŃSKI et al., 2023).

5.2.1 IDENTIFICAÇÃO DOS ATRIBUTOS PROGNÓSTICOS PELA SELEÇÃO DE CARACTERÍSTICA

A seleção dos subconjuntos de variáveis clínicas teve como objetivo reduzir a dimensionalidade dos dados e identificar as variáveis mais relevantes e não redundantes para a predição de risco ou do tempo até o evento (PÖLSTERL, 2023; SIMON et al., 2011). Para isso, foi aplicada uma combinação de métodos lineares, como os métodos penalizados *Ridge* (norma l_2), *Lasso* (norma l_1) e *Elastic Net* (combinação das normas l_1 e l_2), além de métodos não lineares aplicados ao GBS. Esses métodos foram implementados com base nas diretrizes estabelecidas pela biblioteca "*Scikit-survival*"(PÖLSTERL, 2020; PÖLSTERL, 2023) , conforme detalhado na Subseção 4.3.1.

O conjunto ideal de características foi determinado aplicando regularizações *Lasso* e *Elastic Net*, que selecionaram 9 variáveis com coeficientes não nulos, representando os atributos mais relevantes para a predição. Para permitir a comparação do desempenho dos métodos de AM aplicados a diferentes subconjuntos, este número de variáveis (9) foi utilizado nas demais seleções.

5.2.1.1 Regularização *Lasso* e *Elastic Net*

A penalização *Lasso*, fundamentada na norma l_1, permite a inclusão progressiva de variáveis no modelo, atribuindo coeficientes não nulos apenas às características mais relevantes. As características cujos coeficientes foram zerados foram descartadas, reduzindo o número de atributos e simplificando o modelo (PÖLSTERL, 2020; SIMON et al., 2011).

Tabela 8 – Análise de Correlação dos dados do Banco de Dados Clínicos.

Atributo	**Atributo Correlacionado**	**Valor**	**Métrica**
*Idade**	*Status* Menopausal	0,80	CP
	Menstruação	0,72	CP
Menstruação	Idade	0,72	CP
***Status* Menopausal**	Idade	0,80	CP
*Estadiamento Clínico T**	Estadiamento Patológico T	0,99	CI
	Tamanho Tumor (X)	0,70	CP
	Tamanho Tumor (Y)	0,65	CP
*Estadiamento Patológico T**	Estadiamento Clínico T	0,99	CI
	Tamanho Tumor (X)	0,70	CP
	Tamanho Tumor (Y)	0,65	CP
Tamanho Tumor (X)	Tamanho Tumor (Y)	0,95	CP
	Estdiamento Patológico T	0,70	CP
	Estadiamento Clínico T	0,70	CP
*Tamanho Tumor (Y)**	Tamanho Tumor (X)	0,95	CP
	Estdiamento Patológico T	0,65	CP
	Estadiamento Clínico T	0,65	CP
*Estadiamento Clínico N**	Estadiamento Patológico N	0,96	CI
	Linfonodo Pós-cirúrgico	0,87	CP
*Estadiamento Patológico N**	Estadiamento Clínico N	0,96	CI
	Linfonodo Pós-cirúrgico	0,87	CP
*Linfonodo Pós-cirúrgico**	Estadiamento Patológico N	0,87	CP
	Estadiamento Clínico N	0,87	CP
	Estadiamento Anatômico	0,82	CP
*Estadiamento Clínico M**	Estadiamento Anatômico	1,00	CI
	Estadiamento Patológico M	0,97	CI
	Metástase Locorregional	0,94	CI
	Tipo de Tratamento	0,91	CI
*Estadiamento Patológico M**	Estadiamento Anatômico	0,96	CI
	Tipo de Tratamento	0,95	CI
	Estadiamento Clínico M	0,95	CI
	Metástase Locorregional	0,89	CI
*Metástase Locorregional**	Estadiamento Anatômico	0,93	CI
	Estadiamento Clínico M	0,91	CI
	Estadiamento Patológico M	0,87	CI
	Tipo de Tratamento	0,84	CI
Tipo de Tratamento	Estadiamento Patológico M	0,95	CI
	Estadiamento Clínico M	0,91	CI
	Metástase Locorregional	0,84	CI

(*) Variáveis excluídas após análise de correlação.
(**Grafite**) Variáveis selecionadas após análise de correlação.

Fonte: Elaborada pela autora (2024).

A penalização *Elastic Net*, por sua vez, combina as normas l_1 e l_2, associando a abordagem de seleção do *Lasso* com a estabilidade conferida pelo *Ridge*. Essa metodologia

Tabela 9 – Análise de Correlação dos dados do Banco de Dados Clínicos.

Atributo	**Atributo Correlacionado**	**Valor**	**Métrica**
Estadiamento	Estadiamento Clínico M	1,00	CI
Anatômico	Doença	1,00	CI
	Estadiamento Patológico M	0,96	CI
	Metástase Locorregional	0,93	CI
	Linfonodo Pós-cirúrgico	0,82	CP
*Doença**	Estadiamento Anatômico	1,00	CI
Linfonoso Isolado	Esvaziamento Axilar	0,71	CP
*Esvaziamento Axilar**	Linfonoso Isolado	0,71	CP
*Hormonioterapia**	Tratamento Sistêmico	1,00	CI
*Quimioterapia**	Tratamento Quimioterápico	1,00	CI
	Tratamento Sistêmico	0,99	CI
Tratamento Sistêmico	Hormonioterapia	1,00	CI
	Tratamento Quimioterápico	0,99	CI
	Quimioterapia	0,99	CI
*Tratamento**	Quimioterapia	1,00	CI
*Quimioterápico**	Tratamento Sistêmico	0,99	CI
*Classificação Progesterona**	Marcador Progesterona	1,00	CI
Marcador Progesterona	Classificação Progesterona	1,00	CI
*Classificação Estrogênio**	Marcador Estrogênio	1,00	CI
	Classificação Imuno-histoquímica	0,77	CI
Marcador Estrogênio	Classificação Estrogênio	1,00	CI
*Classificação HER2**	Classificação Imuno-histoquímica	1,00	CI
	Marcador HER2	0,97	CI
	Marcadores Imuno-histoquímicos	0,75	CI
Marcador HER2	Classificação HER2	0,97	CI
Classificação	Classificação HER2	1,00	CI
Imuno-histoquímica	Classificação Estrogênio	0,77	CI
*Marcadores**	Perfil Imuno-histoquímico	1,00	CI
*Imuno-histoquímicos**	Classificação HER2	0,75	CI
Perfil Imuno-histoquímico	Marcadores Imuno-histoquímicos	1,00	CI

(*) Variáveis excluídas após análise de correlação.
(Grafite) Variáveis selecionadas após análise de correlação.

Fonte: Elaborada pela autora (2024).

é particularmente vantajosa para lidar com dados de alta dimensionalidade e variáveis altamente correlacionadas, embora neste estudo já tenha sido feita uma exclusão prévia dessas variáveis conforme descrito na Subseção 5.1.2.4 (PÖLSTERL, 2020; SIMON et al., 2011). As variáveis selecionadas, com coeficientes não nulos em ambos os métodos, representam 18% das variáveis do **BD 1**. Tanto *Lasso* quanto *Elastic Net* convergiram na seleção das mesmas características, com coeficientes semelhantes em ambos os métodos, como mostrado na Tabela 10. Nessa tabela, observa-se que $66,67\%$ das variáveis selecionadas (em negrito) foram confirmadas por pelo menos dois outros métodos de seleção,

constituindo o **BD 2**.

Tabela 10 – Características selecionadas pela Regularização *Lasso* e *Elastic Net*.

Atributos	**Coef. *Lasso***	**Coef. *Elastic Net***
Classificação Imuno-histoquímica	0,1039	0,1040
Dias Cirurgia	0,0012	0,0011
Estadiamento Anatômico	**0,2901**	**0,3033**
Linfonodo Isolado	0,0252	0,0249
Marcador Estrogênio	0,0803	0,0906
Marcador Progesterona	0,0174	0,0217
Metástase à Distância	**1,9234**	**1,7314**
Tamanho Tumor (X)	0,0810	0,0834
Tratamento Sistêmico	0,0228	0,0218

(**Negrito**) Variáveis selecionadas também por outros dois métodos de seleção de atributos.
(**Vermelho**) Variável com o maior coeficiente.
(**Azul**) Variável com o segundo maior coeficiente.

Fonte: Elaborada pela autora (2024).

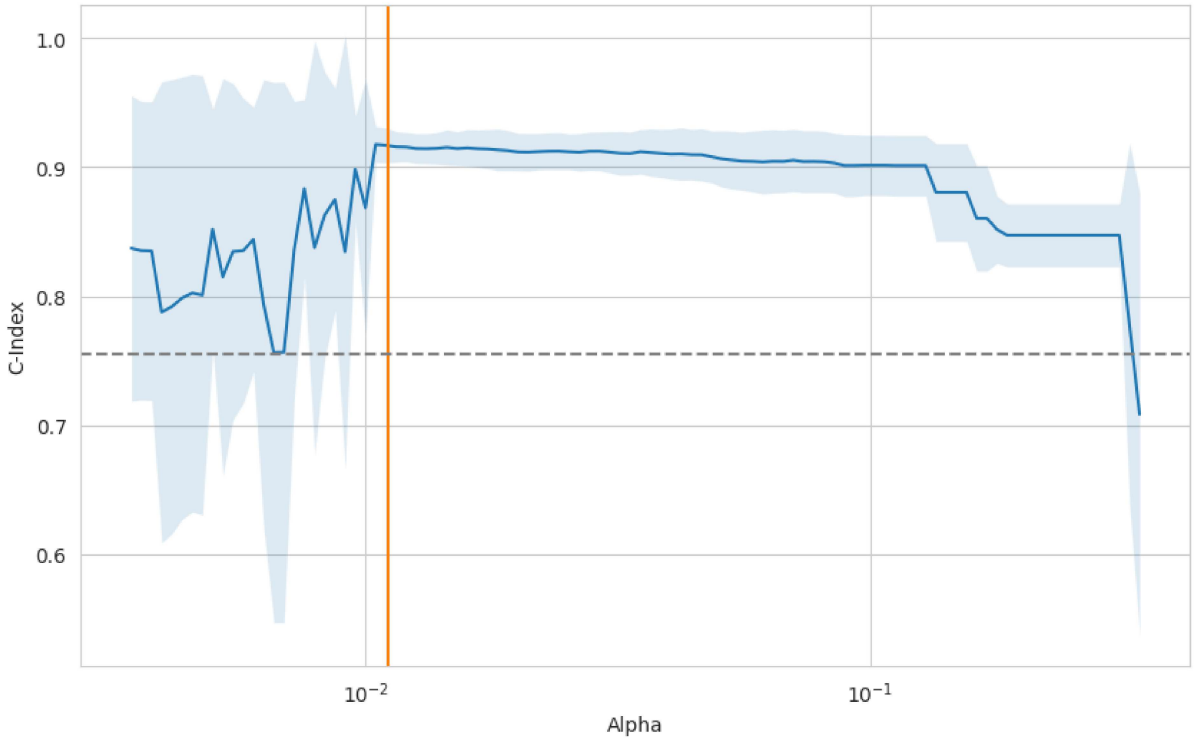
5.2.1.2 Regularização pela Validação Cruzada

A validação cruzada foi aplicada nas abordagens de regularização *Lasso* e *Elastic Net* para identificar o melhor conjunto de parâmetros α. As Figuras 14 e 15 ilustram os resultados, evidenciando o impacto direto deste parâmetro nos coeficientes β. Valores elevados de α levam à anulação dos coeficientes, o que é indicado pelo pico do C-Index, representado pela linha vertical laranja. Por outro lado, valores mais baixos de α permitem a inclusão de um maior número de variáveis, o que pode comprometer a capacidade discriminativa do modelo, como demonstrado pela linha horizontal tracejada cinza, que se aproxima do valor de 0.75 do C-Index, conforme discutido em (PÖLSTERL, 2023)

Após definir o α ideal, foram inspecionados os coeficientes não nulos e aplicada a seleção de características por validação cruzada em ambas as abordagens *Lasso* e *Elastic Net*. Este passo permitiu identificar os coeficientes mais significativos para o modelo, destes foram selecionados os 9 atributos com valores mais elevados, conforme apresentado na Tabela 11. É importante notar que tanto a Regularização *Lasso* quanto a *Elastic Net* novamente selecionaram os mesmos 9 atributos, tanto em termos de ordem quanto de proximidade dos coeficientes, assim como descrito na Subsubseção anterior.

A Tabela 11 exibe os valores dos coeficientes não nulos para cada atributo selecionado, sendo que $44,44\%$ das variáveis selecionadas também foram selecionadas por outros dois método de seleção. Estes 9 atributos selecionados pela validação cruzada formam o **BD 3**.

Figura 14 – Variação da Força de Penalidade do Método *Lasso* em Função do Parâmetro α: Indicação do pico de anulação dos coeficientes (laranja) e C-Index de referência (cinza) sem penalização.



Fonte: Elaborada pela autora (2024).

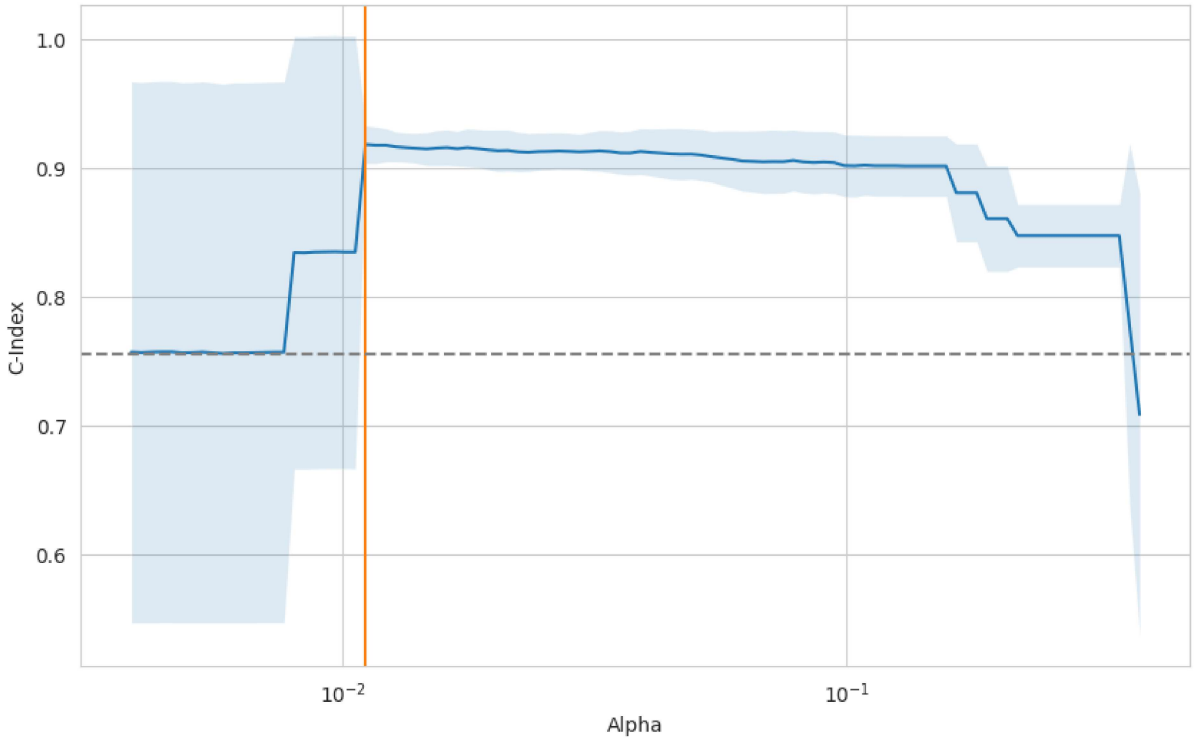
Tabela 11 – Características selecionadas pelo K-Fold.

Atributos	**Coeficiente *Lasso***	**Coeficiente *Elastic Net***
Atraso Radioterapia	0,1420	0,1444
Classificação Cirurgia	0,2370	0,2369
Estadiamento Anatômico	**0,4542**	**0,4556**
Grau Histológico	0,1532	0,1537
Metástase à Distância	**1,3596**	**1,3509**
Status Menopausal	0,1595	0,1713
Tabagista	0,2179	0,2212
Tamanho Tumor (X)	0,2217	0,2249
Tratamento Sistêmico	0,1701	0,1723

(**Negrito**) Variáveis selecionadas também por outros dois métodos de seleção de atributos.
(**Vermelho**) Variável com o maior coeficiente.
(**Azul**) Variável com o segundo maior coeficiente.

Fonte: Elaborada pela autora (2024).

Figura 14 – Variação da Força de Penalidade do Método *Elastic Net* em Função do Parâmetro α: Indicação do pico de anulação dos coeficientes (laranja) e C-Index de referência (cinza) sem penalização.



Fonte: Elaborada pela autora (2024).

5.2.1.3 Regularização pelo GBS

A aplicação do método de seleção de características não linear GBS, que utiliza uma função de perda baseada na verossimilhança parcial do modelo de Cox, resultou na identificação e seleção das variáveis com os maiores coeficientes (PÖLSTERL, 2023). Este método selecionou um conjunto de 9 variáveis, conforme apresentado na Tabela 12. Entre as variáveis selecionadas, $44,44\%$ também foram identificadas por outros métodos de seleção, e essas variáveis compõem o **BD 4**.

5.2.2 IDENTIFICAÇÃO DOS ATRIBUTOS PROGNÓSTICOS SELEÇÃO DE IMPORTÂNCIA DE VARIÁVEL

Os resultados apresentados derivam das simulações dos métodos de seleção de importância de variáveis, detalhados na Subseção 4.3.2. A metodologia aplicada utiliza o método de permutação de importância de variável baseado no RF da biblioteca "Scikit-learn"(PEDREGOSA et al., 2011), e adaptados para os modelos GBS e RSF, seguindo as recomendações da biblioteca "*Scikit-survival*"(PÖLSTERL, 2023).

O método de permutação de importância de variável avalia a contribuição de cada

Tabela 12 – Características selecionadas pelo método não linear do GBS.

Atributos	**Coeficiente**
Atraso Radioterapia	**1,0157**
Classificação Cirurgia	0,2728
Classificação Imuno-histoquímica	0,1350
Doença Cardiovascular	0,2704
Estadiamento Anatômico	0,1925
Extensão do Tumor	0,1100
Metástase à Distância	**3,1000**
Status Menopausal	0,1179
Tamanho Tumor (X)	0,1064

(**Negrito**) Variáveis selecionadas também por outros dois métodos de seleção de atributos.
(**Vermelho**) Variável com o maior coeficiente.
(**Azul**) Variável com o segundo maior coeficiente.

Fonte: Elaborada pela autora (2024).

atributo para a predição de sobrevida, fornecendo uma ideia clara de quais variáveis têm o impacto mais significativo (KALAFI et al., 2019; PEDREGOSA et al., 2011; PÖLSTERL, 2020). Os modelos baseados em árvores de decisão, como GBS e RSF, são frequentemente utilizados para calcular a importância das variáveis, no caso da análise de sobrevida utilizando o C-Index (dados treino) como métrica de avaliação do impacto de cada variável no conjunto de teste, priorizando aquelas que apresentam maiores médias, correspondendo ao impacto destas variáveis na capacidade preditiva do modelo (PEDREGOSA et al., 2011; PÖLSTERL, 2020).

Para assegurar uma comparação consistente entre os métodos de seleção de características e de importância de variáveis, foram selecionadas também 9 variáveis prognósticas em cada abordagem.

5.2.2.1 Permutação de Importância pelo GBS

O GBS constrói o modelo de forma iterativa, criando novas árvores de decisão para corrigir os erros residuais das árvores anteriores, melhorando progressivamente a precisão das previsões (PÖLSTERL, 2020; PÖLSTERL, 2023). Na Tabela 13 observamos as médias e DP das variáveis selecionadas pela permutação de importância. Entre as 9 variáveis de maoior média de imporância selecionadas $66,67\%$ também foram identificadas por pelo menos dois métodos, evidenciando a relevância desta abordagem. Essas variáveis compõem o conjunto denominado **BD 5**.

Tabela 13 – Variáveis selecionadas pela permutação de importância aplicadas ao GBS.

Atributos	**Média de Importância**	**DP**
Citologia	0,0026	± 0,0024
Classificação Imuno-histoquímica	0,0018	± 0,0024
Estadiamento Anatômico	0,0111	± 0,0037
Linfonodo Isolado	0,0062	± 0,0021
Mamografia	0,0030	± 0,0010
Metástase à Distância	**0,2207**	**± 0,0316**
Tamanho Tumor (X)	**0,0161**	**± 0,0050**
Tipo de Tratamento	0,0121	± 0,0018
Tratamento Sistêmico	0,0017	± 0,0019

(**Negrito**) Variáveis selecionadas também por outros dois métodos de seleção de atributos.
(**Vermelho**) Variável com o maior média de importância
(**Azul**) Variável com segunda maior média de importância.

Fonte: Elaborada pela autora (2024).

5.2.2.2 Permutação de Importância pelo RSF

O RSF combina previsões de várias árvores de decisão para estimar a função de risco de sobrevida, onde cada árvore é treinada em uma amostra *bootstrap* do conjunto de dados. A agregação das previsões individuais das árvores resulta em uma estimativa robusta e menos suscetível a *overfitting* (PÖLSTERL, 2020; PÖLSTERL, 2023). Na Tabela 14, observamos as 9 variáveis selecionadas pelo método de permutação de importância aplicado ao RSF. Dessas variáveis, apenas $33,33\%$ também foram identificadas por outros dois métodos de seleção, e compõem o **BD 6**.

Tabela 14 – Variáveis selecionadas pela permutação de importância aplicadas ao RSF.

Atributos	**Média de Importância**	**DP**
Atraso Quimioterapia	0,0004	± 0,0003
Cintilografia Óssea	0,0004	± 0,0006
Diagnóstico Clínico	0,0004	± 0,0003
Estadiamento Anatômico	**0,0131**	**± 0,0053**
Extensão do Tumor	0,0013	± 0,0021
Linfonodo Isolado	0,0011	± 0,0010
Metástase à Distância	**0,1402**	**± 0,0287**
Tipo de Tratamento	0,0112	± 0,0020
Toxicidade Quimioterapia	0,0005	± 0,0007

(**Negrito**) Variáveis selecionadas também por outros dois métodos de seleção de atributos.
(**Vermelho**) Variável com o maior média de importância
(**Azul**) Variável com segunda maior média de importância.

Fonte: Elaborada pela autora (2024).

O Quadro 14 apresenta uma comparação dos atributos selecionados por cada método de seleção de atributos. Entre eles, destaca-se que os atributos destacados de

vermelho "Estadiamento Anatômico"e "Metástase à Distância"foram selecionados por todos os métodos, destacando sua importância tanto na predição de sobrevida quanto na condução clínica-terapêutica. Atributos destacados de azul como "Tamanho do Tumor (X)", "Classificação Imuno-histoquímica", "Linfonodo Isolado"e "Tratamento Sistêmico"foram selecionados em pelo menos três métodos, reforçando sua relevância clínica.

Quadro 14 – Resumo comparativo das variáveis selecionadas por cada Banco de Dados.

Atributo	**BD 2**	**BD 3**	**BD 4**	**BD 5**	**BD 6**
Estadiamento Anatômico	X	X	X	X	X
Metástase à Distância	X	X	X	X	X
Classificação Imuno-histoquímica	X		X	X	
Linfonodo Isolado	X			X	X
Tratamento Sistêmico	X	X		X	
Tamanho Tumor (X)	X	X	X	X	
Atraso Radioterapia		X	X		
Classificação Cirurgia		X	X		
Extensão do Tumor			X		X
***Status* Menopausal**		X	X		
Tipo de Tratamento				X	X
Atraso Quimioterapia					X
Cintilografia Óssea					X
Citologia				X	
Diagnóstico Clínico				X	
Dias Cirurgia	X				
Doença Cardiovascular			X		
Grau Histológico X Mamografia				X	
Marcador Estrogênio	X				
Marcador Progesterona	X				
Tabagista		X			
Toxicidade Quimioterapia				X	

(**Vermelho**) Variáveis selecionadas por todos os métodos de seleção de atributos.
(**Azul**) Variáveis selecionadas por pelo menos três métodos de seleção de atributos.
(**Grafite**) Variáveis selecionadas por pelo menos dois métodos de seleção de atributos.

Fonte: Elaborada pela autora (2024).

A seleção de variáveis é uma prática comum na construção de modelos com múltiplas variáveis, especialmente em conjuntos de dados com estruturas complexas e interdependências entre variáveis, como na análise de sobrevida. Além dos métodos utilizados neste estudo, outras abordagens, como a análise hierárquica entre variáveis, a seleção bayesiana (TENG et al., 2019), o ganho de entropia (SHUKLA et al., 2018), e as importâncias de variáveis de RF (KALAFI et al., 2019; XIAO et al., 2022) e RSF (FANIZZI et al., 2023) são utilizados para melhorar o desempenho preditivo (FANIZZI et al., 2023; KALAFI et al., 2019; LIU et al., 2020; SHUKLA et al., 2018; TENG et al., 2019; XIAO et al., 2022).

5.3 RESULTADOS DA MODELAGEM COMPUTACIONAL PARA PREDIÇÃO DA SOBREVIDA

Os resultados apresentados foram obtidos por meio da simulação de métodos de AM supervionados para a análise de sobrevida, incluindo modelos lineares, como CPH (versões penalizada e não penalizada) e SSVM, bem como modelos não lineares, como GBS, RSF e KSSVM. Diferentes subconjuntos do **BD Clínico**, selecionados conforme a metodologia descrita na Seção 5.2, foram utilizados como entrada para os modelos implementados.

Os subconjuntos do **BD Clínico**, compostos por 69 variáveis listadas nos Quadros 7 a 12, foram preparados conforme as seguintes seleções:

BD 1 Conjunto com 50 variáveis selecionadas após pré-processamento e análise de correlação, com exclusão de variáveis (*) conforme descrição nas Tabelas 8 e 9.

BD 2 Conjunto de 9 variáveis selecionadas de forma consistente entre os métodos de regularização *Lasso* e *Elastic Net*, conforme apresentadas na Tabela 10.

BD 3 Conjunto com 9 variáveis selecionadas por validação cruzada utilizando as regularizações *Lasso* e *Elastic Net*, conforme descrito na Tabela 11.

BD 4 Seleção pelo método de característica não linear via GBS, com 9 variáveis mais relevantes, detalhadas na Tabela 12.

BD 5 Seleção por permutação de importância aplicada ao GBS, com 9 variáveis mais importantes, conforme apresentado na Tabela 13.

BD 6 Seleção por permutação de importância aplicada ao RSF, com 9 variáveis mais importantes, conforme listado na Tabela 14.

A implementação foi realizada utilizando a biblioteca **"Scikit-survival"** (PÖLSTERL, 2023), compatível com **"Scikit-Learn"** (PEDREGOSA et al., 2011), conforme descrito na Seção 4.2. Os hiperparâmetros dos modelos foram definidos na Subseção 4.2.2, e o desempenho foi avaliado com base nas métricas C-Index, BS e IBS, descritas na Seção 4.4.

Um C-Index próximo de 1 indica excelente discriminação, enquanto valores de BS e IBS próximos a 0 refletem boa calibração do modelo (PÖLSTERL, 2020; PÖLSTERL, 2023; TIZI; BERRADO, 2023). Observa-se que os modelos SSVM linear e KSSVM não possuem métricas BS e IBS, pois essas métricas são aplicáveis apenas a modelos que estimam a função de sobrevida, o que não se aplica a esses modelos (PÖLSTERL, 2020; GRAF et al., 1999). As métricas BS e IBS são indicadores da precisão das predições de modelos para análise de sobrevida, sendo que valores inferiores a $0,25$ para BS são

considerados indicativos de boa calibração e predição (XIAO et al., 2022; PÖLSTERL, 2023). Vale destacar que o C-Index é a métrica mais comumente utilizada para avaliação de desempenho, enquanto BS e IBS são frequentemente usadas em casos de empate e para avaliar a calibração dos modelos (PÖLSTERL, 2023).

Tabela 15 – Avaliação do desempenho dos métodos de AM 1ineares em diferentes subconjuntos do Banco de Dados 1.

Modelo	BD	C-Index	BS	IBS
Cox Proportional Hazards	1	0,875	0,085	**0,043**
	2	**0,934**	**0,060**	0,051
	3	0,903	0,089	0,057
	4	0,924	0,070	0,052
	5	0,931	0,067	0,050
	6	0,908	0,081	0,060
Cox Proportional Hazards Lasso	1	**0,937**	0,066	0,051
	2	**0,937**	0,066	0,051
	3	0,906	0,084	0,056
	4	0,930	0,066	0,052
	5	0,930	**0,066**	**0,050**
	6	0,909	0,078	0,060
Cox Proportional Hazards Elastic Net	1	**0,942**	0,068	0,052
	2	**0,942**	0,068	0,052
	3	0,906	0,084	0,056
	4	0,931	0,066	0,052
	5	0,931	**0,066**	**0,050**
	6	0,909	0,078	0,060
Survival Support Vector Machine	1	0,878		
	2	0,902		
	3	**0,904**		
	4	0,885		
	5	0,858		
	6	0,873		

(**Vermelho**) Melhores métricas de desempenho em cada modelo.
(**Azul**) Modelos não paramétricos.
(**Grafite**) Modelos semi-paramétricos.

Fonte: Elaborada pela autora (2024).

As Tabelas 15 e 16 apresentam um resumo detalhado do desempenho dos métodos de AM, incluindo modelos lineares, não lineares, semi-paramétricos (destacados em grafite) e não paramétricos (em azul), facilitando a comparação de cada modelo nos diferentes subconjuntos do BD Clínico. As métricas de avaliação de desempenho incluem o C-Index e o BS nos conjuntos de teste, além do IBS, que integra os resultados dos conjuntos de treino e teste. Os melhores desempenhos estão destacados em vermelho.

De forma geral, o modelo RSF se destacou em termos de desempenho discriminativo,

Tabela 16 – Avaliação do desempenho dos métodos de AM não lineares em diferentes subconjuntos do Banco de Dados 1.

Modelo	BD	C-Index	BS	IBS
Gradient Boosting Survival	1	0,932	0,056	**0,044**
	2	0,932	0,060	0,046
	3	0,927	0,056	0,053
	4	0,933	0,057	0,048
	5	0,937	0,055	0,045
	6	**0,939**	**0,052**	0,056
Random Survival Forest	1	0,934	0,068	0,055
	2	0,933	0,053	0,049
	3	0,937	0,049	0,053
	4	0,930	0,051	0,051
	5	0,938	0,046	**0,048**
	6	**0,952**	**0,041**	0,052
Kernel Survival Support Vector Machine	1	0,879		
	2	0,917		
	3	0,904		
	4	0,924		
	5	**0,936**		
	6	0,925		

(**Vermelho**) Melhores métricas de desempenho em cada modelo.
(**Azul**) Modelos não paramétricos.
(**Grafite**) Modelos semi-paramétricos.

Fonte: Elaborada pela autora (2024).

alcançando o melhor C-Index (0,952) e o melhor BS (0,041) nos dados de teste, utilizando o **BD 6**. Em relação à qualidade das predições nos dados de treino e teste, o modelo CPH obteve o melhor IBS (0,043) no **BD 1**.

Entre os modelos lineares, o CPH-EN apresentou o melhor C-Index (0,942) nos subconjuntos **BD 1** e **BD 2**, enquanto o CPH se destacou na qualidade das predições, apresentando o melhor IBS (0,043). Esse desempenho superior do CPH pode ser atribuído à redução da complexidade do modelo, com a exclusão de variáveis altamente correlacionadas e à aplicação do método de seleção de características lineares.

Entre os modelos não lineares, o RSF obteve o melhor desempenho tanto no C-Index quanto no BS. No entanto, em termos de qualidade de predição, o GBS se destacou, apresentando o melhor IBS (0,044) no **BD 1**. Nos modelos semi-paramétricos (grafite) e não paramétricos (azul), o RSF e o GBS se destacaram, respectivamente, como os melhores desempenhos, conforme ilustrado nas Tabelas 15 e 16.

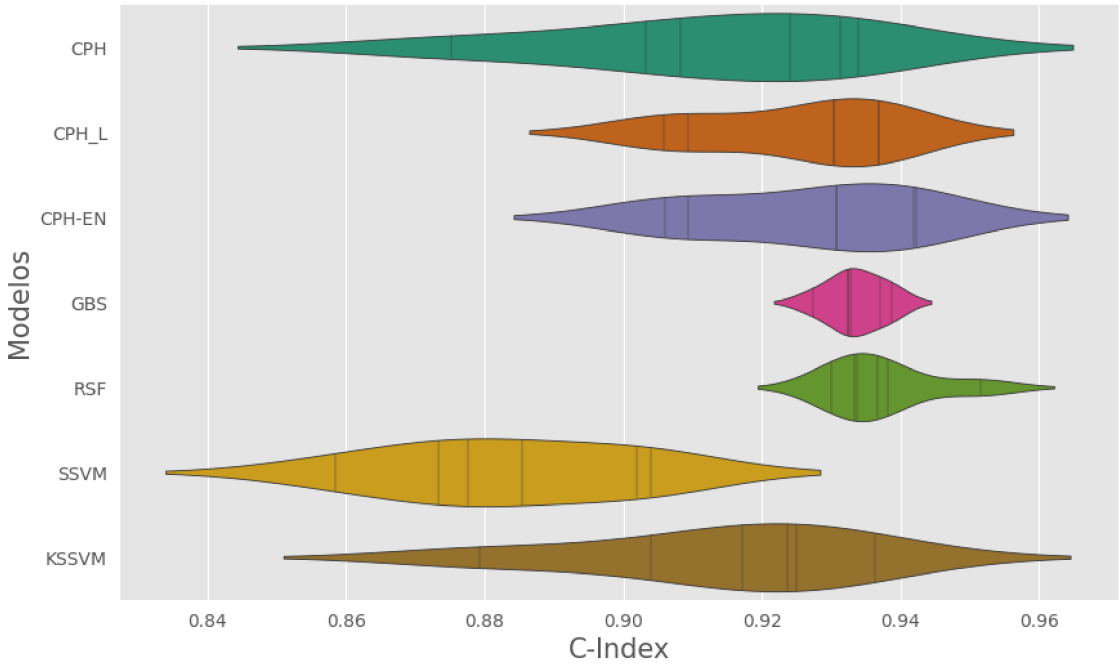
O **BD 1**, contendo 50 atributos, foi utilizado como referência para os demais subconjuntos. Os modelos aplicados ao **BD 1** apresentaram um desempenho satisfatório, com destaque para o CPH-EN, que obteve um C-Index de 0,942 e o menor BS de 0,043,

indicando boa discriminação e calibração, atribuída à redução da dimensionalidade e exclusão de variáveis correlacionadas.

Na comparação entre os subconjuntos selecionados pelos métodos de seleção de características, o CPH-EN no **BD 2** se destacou entre os métodos lineares, enquanto o RSF apresentou o melhor desempenho entre os métodos não lineares no **BD 3**. Nos subconjuntos selecionados pelos métodos de permutação de variáveis, o CPH e o CPH-EN se destacaram no **BD 5** (para modelos lineares), e o RSF obteve os melhores resultados no **BD 6** (para modelos não lineares), conforme mostrado na Tabela 16.

No que se refere à melhoria percentual do C-Index após a aplicação dos métodos de seleção de atributos, considerando o **BD 1** como referência, o CPH obteve o maior ganho de desempenho, subindo de 0,875 para 0,934 (um aumento de 6,71%) no **BD 2**, destacando-se entre os modelos lineares e semi-paramétricos. Entre os modelos não lineares, o KSSVM apresentou o maior ganho percentual, de 0,879 para 0,936 (6,50%) no **BD 5**, destacando-se também entre os modelos não paramétricos. No entanto, ao considerar os resultados médios de desempenho de todos os subconjuntos, o RSF obteve a melhor média geral (0,937), enquanto o CPH-EN obteve o melhor desempenho entre os modelos lineares (0,934).

Figura 16 – Comparação do desempenho do C-Index entre os métodos de AM: CPH (Azul-Petróleo), CPH-L (Laranja), CPH-EN (Azul), GBS (Rosa), RSF (Verde), SSVM (Amarelo) e KSSVM(marrom); considerando os dados do teste.

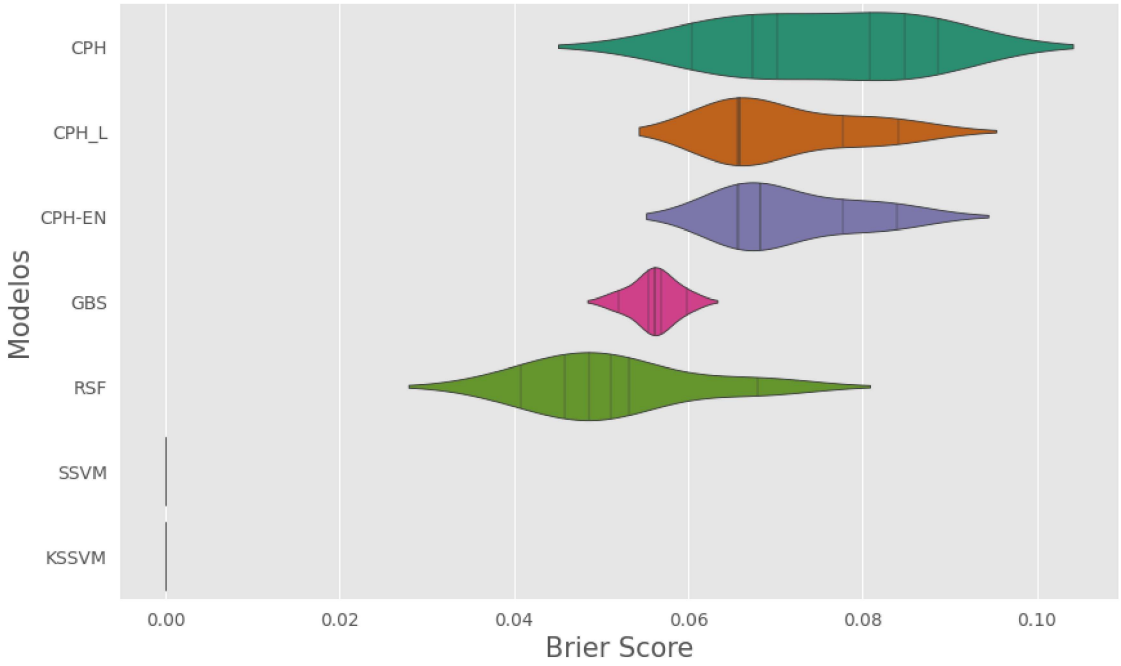


Fonte: Elaborada pela autora (2024).

A Figura 16 ilustra o desempenho discriminativo de cada modelo, medido pelo C-

Index, para todos os subconjuntos do BD Clínicos. Observa-se que os modelos não lineares, como GBS e RSF, apresentaram um comportamento mais estável e valores mais elevados entre os subconjuntos. Por outro lado, os modelos lineares exibiram maior variação nos valores de desempenho, o que pode ser atribuído às interferências mínimas de correlação entre as variáveis selecionadas pelos métodos de seleção. Em comparação, os modelos SSVM e KSSVM mostraram maior amplitude nos resultados, com desempenho inferior, sendo que o modelo não linear (KSSVM) apresentou melhores resultados em relação ao linear.

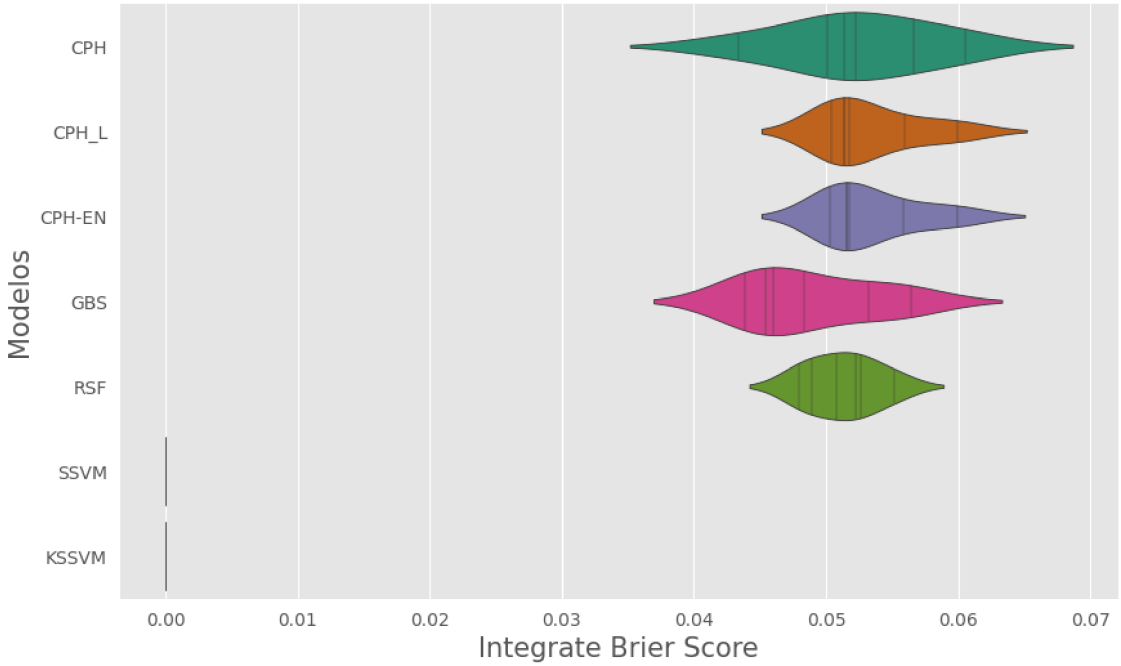
Figura 17 – Comparação do desempenho do *Brier Score* para MAM: CPH (Azul-Petróleo), CPH-L (Laranja), CPH-EN (Azul), GBS (Rosa) e RSF (Verde); considerando os dados do teste.



Fonte: Elaborada pela autora (2024).

A Figura 17 apresenta o desempenho de calibração dos modelos, medido pela métrica BS, considerando apenas os dados de teste. Os menores valores de BS foram alcançados pelos modelos GBS e RSF, indicando uma melhor calibração. Embora os modelos lineares também tenham mostrado boa calibração, estes apresentaram menor estabilidade. Por fim, a Figura 18 ilustra o desempenho do IBS para todos os tempos de análise (treino e teste), mostrando intervalos semelhantes entre os modelos, o que sugere boa calibração e robustez. Todos os métodos de AM apresentaram valores de IBS inferiores a 0,07 em todos os subconjuntos do BD Clínicos, o que pode ser atribuído aos processos de pré-processamento de dados e à redução da complexidade do modelo pela remoção das variáveis correlacionadas.

Figura 18 – Comparação do desempenho do *Integrated Brier Score* para MAM: CPH (Azul-Petróleo), CPH-L (Laranja), CPH-EN (Azul), GBS (Rosa) e RSF (Verde); considerando os dados de treino e teste.



Fonte: Elaborada pela autora (2024).

Vale ressaltar que os modelos KSSVM e SSVM não possuem métricas BS e IBS, pois essas métricas são aplicáveis apenas a modelos que estimam a função de sobrevida (GRAF et al., 1999).

Em termos de desempenho médio, o modelo RSF obteve as melhores médias de C-Index (0,937) e BS (0,051), corroborando os resultados de outros estudos (CARVALHO et al., 2024; CARVALHO et al., 2023; FANIZZI et al., 2023; LIU et al., 2020; MONCADA-TORRES et al., 2021; PINHEIRO et al., 2022; TIZI; BERRADO, 2023; XIAO et al., 2022). O CPH-EN e o GBS alcançaram as melhores médias de IBS, com 0,049. Quando comparados os modelos lineares, o CPH-EN se destacou com o melhor desempenho em termos de C-Index (0,934), BS (0,056) e IBS. Entre os modelos não lineares, o RSF e o GBS se destacaram. Por fim, entre os métodos SVM, o modelo não linear (KSSVM) obteve desempenho superior no C-Index (0,914) em relação ao modelo linear (0,883).

5.3.1 COMPARAÇÃO DO DESEMPENHO ENTRE OS MODELOS

A comparação do desempenho dos diferentes modelos aplicados aos subconjuntos do **BD Clínico** foi realizada com base nas métricas C-Index, BS e IBS. Os valores elevados de C-Index (próximos a 1) indicam uma alta concordância entre predições e observações, enquanto valores baixos de BS e IBS (próximos a 0) indicam uma boa calibração entre

as probabilidades preditas e as frequências observadas de sobrevida, conforme discutido em Polsterl (2020, 2023) (PÖLSTERL, 2020; PÖLSTERL, 2023). Nas Tabelas 17, 18, 19, 20, 21 e 22, os modelos semi-paramétricos estão destacados em grafite, os modelos não paramétricos em azul, e os melhores desempenhos de cada métrica de avaliação são indicados em vermelho.

O desempenho dos modelos foi inicialmente avaliado no **BD 1**, que contém 50 variáveis prognósticas, conforme apresentado na Tabela 17. Os modelos CPH-EN (0,942) e RSF (0,934), representando abordagens lineares e não lineares, respectivamente, apresentaram desempenho superior em discriminação (C-Index) quando comparados aos modelos de regressão de Cox (CPH e CPH-L), aos modelos EL (GBS) e aos modelos de classificação (KSSVM e SSVM). Entre os modelos não paramétricos (azul), o GBS obteve o melhor desempenho em todas as métricas. Quanto à calibração (BS e IBS), os modelos CPH-L (0,066) e GBS (0,044) destacaram-se, com os melhores desempenhos entre os lineares e não lineares, respectivamente.

Tabela 17 – Comparação do desempenho dos métodos de AM no Banco de Dados 1.

Linearidade	**Modelo**	**C-Index**	**BS**	**IBS**
Linear	***Cox Proportional Hazards***	0,875	0,085	0,043
	CPH Lasso	0,937	**0,066**	**0,051**
	CPH Elastic Net	**0,942**	0,068	0,052
	Survival SVM	0,878		
Não Linear	***Gradient Boosting Survival***	0,932	**0,056**	**0,044**
	Random Survival Forest	**0,934**	0,068	0,055
	Kernel Survival SVM	0,879		
Média		**0,911**	**0,069**	**0,049**

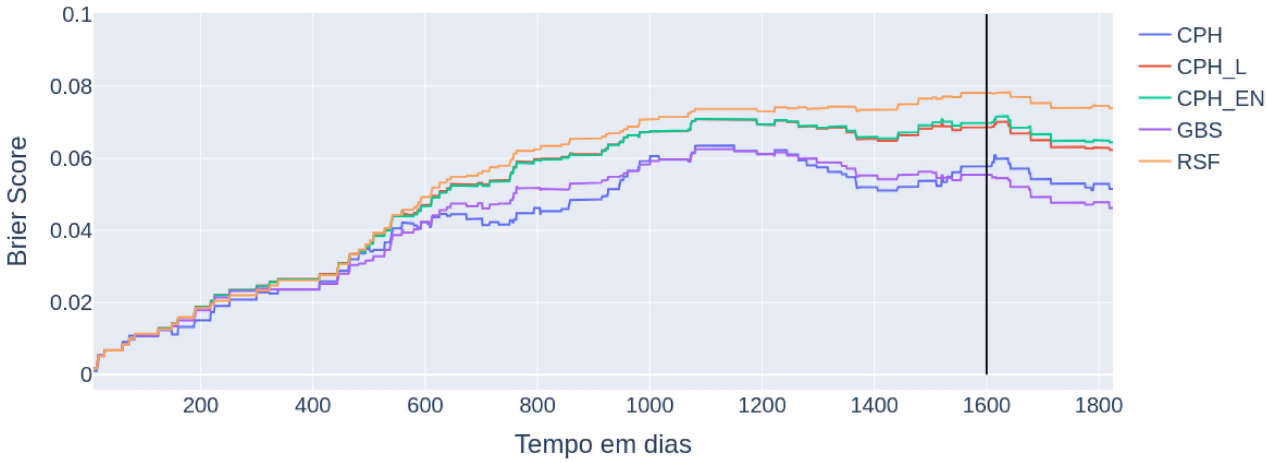
(**Vermelho**) Melhores métricas de desempenho entre os modelos lineares e não linares.
(**Azul**) Modelos não paramétricos.
(**Grafite**) Modelos semi-paramétricos.

Fonte: Elaborada pela autora (2024).

As Figuras 19 a 24 ilustram a variação do BS ao longo do tempo para os diferentes métodos de AM aplicados aos subconjuntos do **BD Clínico**, considerando os dados de treino e teste. A linha preta destaca o intervalo de tempo entre 1600 e 1825 dias, com o maior percentual de censura (75, 08%), entre pacientes censurados (97, 14%) e não censurados (2, 86%). Nesse intervalo, observa-se um declínio das linhas de BS, possivelmente associado ao aumento da quantidade de observações.

A Figura 19 mostra a dinâmica do BS ao longo do tempo para o **BD 1**, onde o modelo não linear GBS (roxo) se destaca com os menores valores de BS, seguido pelo modelo CPH (azul), destacando a eficácia de ambos ao longo do período analisado. Esses resultados corroboram os dados da Tabela 17, na qual o GBS exibe o menor BS, e o CPH apresenta o melhor IBS, sugerindo maior precisão na predição do risco ao longo do tempo.

Figura 19 – Comparação do desempenho do *Brier Score* entre os métodos de AM no BD 1: CPH (azul), CPH-L (vermelho), CPH-EN (verde), GBS (roxo) e RSF (laranja) em dados do treino e teste ao longo do tempo. A linha vertical preta delimita o intervalo de tempo que se concentram 75% dos dados.



Fonte: Elaborada pela autora (2024).

Tabela 18 – Comparação do desempenho dos métodos de AM no Banco de Dados 2.

Linearidade	Modelo	C-Index	BS	IBS
Linear	***Cox Proportional Hazards***	0,934	**0,060**	**0,051**
	CPH Lasso	0,937	0,066	0,051
	CPH Elastic Net	**0,942**	0,068	0,052
	Survival SVM	0,902		
Não Linear	***Gradient Boosting Survival***	0,932	0,060	**0,046**
	Random Survival Forest	**0,933**	**0,053**	0,049
	Kernel Survival SVM	0,917		
Média		**0,928**	**0,062**	**0,050**

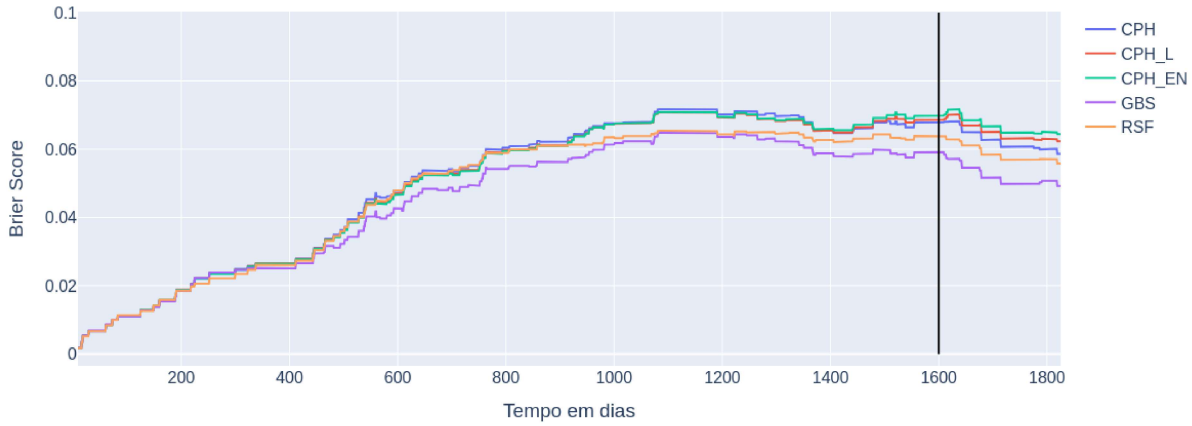
(**Vermelho**) Melhores métricas de desempenho entre os modelos lineares e não linares.
(**Azul**) Modelos não paramétricos.
(**Grafite**) Modelos semi-paramétricos.

Fonte: Elaborada pela autora (2024).

Para o **BD 2**, composto por 9 variáveis selecionadas por regularização via *Lasso* e *Elastic Net*, o modelo CPH-EN $(0,942)$ e o RSF $(0,933)$ novamente se destacaram em termos de C-Index. O modelo GBS $(0,932)$, entre os não paramétricos, apresentou o melhor desempenho nas métricas de calibração (BS e IBS). Em contraste, os modelos KSSVM $(0,917)$ e SSVM $(0,907)$ apresentaram os piores resultados em discriminação. A Figura 20 destaca a dependência temporal dos modelos nos dados de treino e teste para o **BD 2**, mostrando que os modelos não lineares, GBS (roxo) e RSF (laranja), se mostraram

mais robustos, enquanto os modelos de Cox, tanto penalizado quanto não penalizado, exibiram desempenho inferior, corroborando os resultados da Tabela 18.

Figura 20 – Comparação do desempenho do *Brier Score* entre os métodos de AM no BD 2: CPH (azul), CPH-L (vermelho), CPH-EN (verde), GBS (roxo) e RSF (laranja) em dados do treino e teste ao longo do tempo. A linha vertical preta delimita o intervalo de tempo que se concentram 75% dos dados.



Fonte: Elaborada pela autora (2024).

A Tabela 19 apresenta os resultados para o **BD 3**, com 9 variáveis selecionadas por validação cruzada. Neste conjunto, o RSF obteve o melhor desempenho tanto em discriminação (C-Index = 0,937) quanto em calibração (BS = 0,049 e IBS = 0,053). No entanto, o **BD 3** apresentou a pior média de desempenho (0, 912) entre os subconjuntos selecionados por métodos de seleção de atributos. A Figura 21 ilustra a variação do BS ao longo do tempo, onde o GBS (roxo) apresentou o melhor desempenho, seguido pelo RSF (laranja), com todos os modelos exibindo um comportamento semelhante, como evidenciado na Tabela 19.

No **BD 4**, com 9 variáveis selecionadas pelo método não linear de seleção de características via GBS (Tabela 12), os modelos CPH-EN (0, 931) e GBS (0, 933) se destacaram em termos de C-Index. Em relação ao BS, os modelos CPH (0, 052) e RSF (0, 052) apresentaram os melhores resultados. A Figura 22 mostra a dependência temporal desses modelos no **BD 4**, onde o GBS (roxo) manteve seu bom desempenho, enquanto os outros modelos exibiram comportamentos semelhantes, com valores próximos aos do GBS.

No **BD 5**, com 9 variáveis selecionadas pelo método de permutação de importância de variáveis aplicado ao GBS (Tabela 12), os modelos CPH-EN (0, 931) e RSF (0, 938) novamente lideraram em desempenho, especialmente em C-Index e BS, entre os lineares e não lineares, respectivamente. Os métodos não lineares mostraram desempenho superior em discriminação em comparação aos lineares. A Figura 23 ilustra o comportamento do

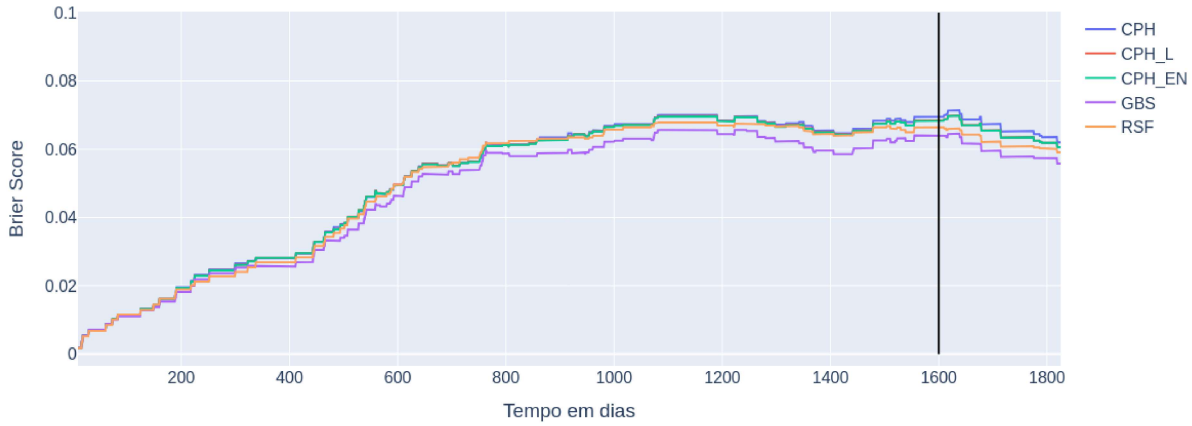
Tabela 19 – Comparação do desempenho dos métodos de AM no Banco de Dados 3.

Linearidade	Modelo	C-Index	BS	IBS
Linear	***Cox Proportional Hazards***	0,903	0,089	0,057
	CPH Lasso	**0,906**	**0,084**	**0,056**
	CPH Elastic Net	**0,906**	**0,084**	**0,056**
	Survival SVM	0,904		
Não Linear	***Gradient Boosting Survival***	0,927	0,056	**0,053**
	Random Survival Forest	**0,937**	**0,049**	**0,053**
	Kernel Survival SVM	0,904		
Média		**0,912**	**0,072**	**0,055**

(**Vermelho**) Melhores métricas de desempenho entre os modelos lineares e não linares.
(**Azul**) Modelos não paramétricos.
(**Grafite**) Modelos semi-paramétricos.

Fonte: Elaborada pela autora (2024).

Figura 21 – Comparação do desempenho do *Brier Score* entre os métodos de AM no BD 3: CPH (azul), CPH-L (vermelho), CPH-EN (verde), GBS (roxo) e RSF (laranja) em dados do treino e teste ao longo do tempo.



Fonte: Elaborada pela autora (2024).

BS ao longo do tempo para o **BD 5**, com o GBS (roxo) se destacando, seguido pelo RSF (laranja), como indicado na Tabela 21.

Por fim, para o **BD 6**, construído com 9 variáveis selecionadas pelo método de permutação de importância aplicado ao RSF (Tabela 14), o RSF obteve o melhor desempenho em C-Index $(0,952)$, seguido pelo GBS $(0,939)$ e KSSVM $(0,925)$ entre os não lineares. Os modelos não lineares também superaram os lineares nas métricas BS e IBS, como mostrado na Tabela 22 e ilustrado na Figura 24, que apresenta o comportamento do BS ao longo do tempo para o **BD 6**. Nesse subconjunto, o RSF (laranja) obteve o melhor

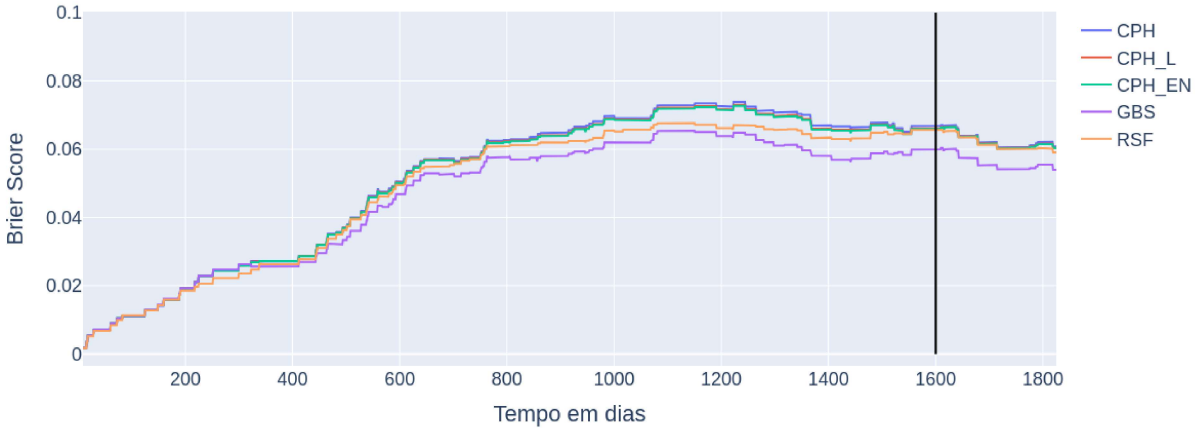
Tabela 20 – Comparação do desempenho dos métodos de AM no Banco de Dados 4.

Linearidade	Modelo	C-Index	BS	IBS
Linear	***Cox Proportional Hazards***	0,924	**0,069**	**0,052**
	CPH Lasso	0,930	0,069	0,053
	CPH Elastic Net	**0,931**	0,072	0,053
	Survival SVM	0,885		
Não Linear	***Gradient Boosting Survival***	**0,933**	0,062	**0,048**
	Random Survival Forest	0,930	**0,052**	0,050
	Kernel Survival SVM	0,924		
Média		**0,922**	**0,065**	**0,051**

(**Vermelho**) Melhores métricas de desempenho entre os modelos lineares e não linares.
(**Azul**) Modelos não paramétricos.
(**Grafite**) Modelos semi-paramétricos.

Fonte: Elaborada pela autora (2024).

Figura 22 – Comparação do desempenho do *Brier Score* entre os métodos de AM no BD 4: CPH (azul), CPH-L (vermelho), CPH-EN (verde), GBS (roxo) e RSF (laranja) em dados do treino e teste ao longo do tempo.



Fonte: Elaborada pela autora (2024).

desempenho, seguido pelo GBS (roxo), enquanto os modelos de Cox exibiram desempenho semelhante e sobreposto.

Os resultados deste estudo evidenciam o desempenho superior do modelo RSF, baseado em árvores, na predição de sobrevida de pacientes com CM feminino. O RSF se destacou principalmente por sua excelente capacidade de discriminação e calibração, apresentando resultados consistentes em diversos subconjuntos de dados testados. Dentre os modelos não lineares, tanto o RSF quanto o GBS demonstraram um desempenho sólido, enquanto, entre os lineares, o CPH-EN se destacou com a melhor *performance*. Esses

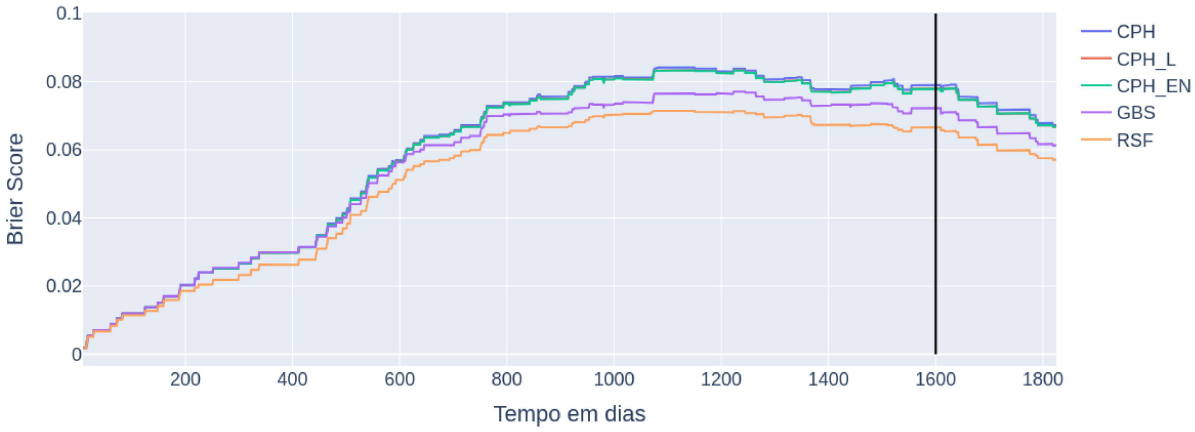
Tabela 21 – Comparação do desempenho dos métodos de AM no Banco de Dados 5.

Linearidade	Modelo	C-Index	BS	IBS
Linear	***Cox Proportional Hazards***	0,931	0,069	0,051
	CPH Lasso	0,930	0,066	0,051
	CPH Elastic Net	**0,931**	**0,066**	**0,051**
	Survival SVM	0,858		
Não Linear	***Gradient Boosting Survival***	0,937	0,056	**0,046**
	Random Survival Forest	**0,938**	**0,050**	0,050
	Kernel Survival SVM	0,936		
Média		**0,923**	**0,062**	**0,050**

(**Vermelho**) Melhores métricas de desempenho entre os modelos lineares e não linares.
(**Azul**) Modelos não paramétricos.
(**Grafite**) Modelos semi-paramétricos.

Fonte: Elaborada pela autora (2024).

Figura 23 – Comparação do desempenho do *Brier Score* entre os métodos de AM no BD 6: CPH (azul), CPH-L (vermelho), CPH-EN (verde), GBS (roxo) e RSF (laranja) em dados do treino e teste ao longo do tempo. A linha vertical preta delimita o intervalo de tempo que se concentram 75% dos dados.



Fonte: Elaborada pela autora (2024).

achados reforçam a eficácia dos métodos não lineares, particularmente na captura de padrões complexos dos dados, essenciais para análises mais robustas em contextos clínicos. Em termos de avaliação média de desempenho, o **BD 2** obteve o melhor resultado global entre os subconjuntos selecionados pelos métodos de seleção de atributos, destacando-se como o subconjunto com o melhor desempenho neste estudo.

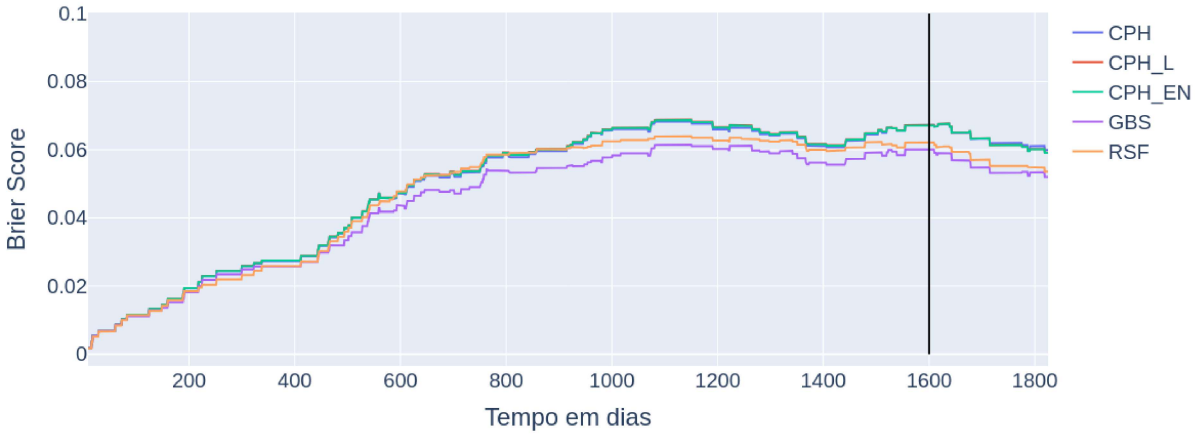
Tabela 22 – Comparação do desempenho dos métodos de AM no Banco de Dados 6.

Linearidade	Modelo	C-Index	BS	IBS
Linear	***Cox Proportional Hazards***	0,908	0,081	0,060
	CPH Lasso	**0,909**	**0,078**	**0,060**
	CPH Elastic Net	**0,909**	**0,078**	**0,060**
	Survival SVM	0,873		
Não Linear	***Gradient Boosting Survival***	0,939	0,052	0,056
	Random Survival Forest	**0,952**	**0,041**	**0,052**
	Kernel Survival SVM	0,925		
Média		**0,917**	**0,066**	**0,058**

(**Vermelho**) Melhores métricas de desempenho entre os modelos lineares e não linares.
(**Azul**) Modelos não paramétricos.
(**Grafite**) Modelos semi-paramétricos.

Fonte: Elaborada pela autora (2024).

Figura 24 – Comparação do desempenho do *Brier Score* entre os métodos de AM no BD 5: CPH (azul), CPH-L (vermelho), CPH-EN (verde), GBS (roxo) e RSF (laranja) em dados do treino e teste ao longo do tempo. A linha vertical preta delimita o intervalo de tempo que se concentram 75% dos dados.



Fonte: Elaborada pela autora (2024).

5.3.2 DISCUSSÃO DOS RESULTADOS

Os resultados deste estudo destacam o papel dos modelos computacionais na predição da sobrevida de pacientes com CM feminino, demonstrando a eficácia das abordagens utilizadas para aprimorar a precisão preditiva. A análise revelou o impacto positivo de um pré-processamento rigoroso dos dados, com um aumento médio de 15,88% no desempenho discriminativo dos modelos após a aplicação de estratégias de pré-processamento. O primeiro processo de pré-processamento, utilizando imputação simples nos dados ausentes

e resultando em 69 variáveis prognósticas, levou a um C-Index médio de 0,784. Após a aplicação do segundo pré-processamento, que envolveu a análise de relevância das variáveis clínicas seguido da remoção das altamente correlacionadas, o conjunto foi reduzido para 50 variáveis, com um C-Index médio elevado para 0,911 (**BD 1**). Esses resultados evidenciam como o tratamento cuidadoso dos dados contribui para o desenvolvimento de modelos mais precisos e adaptados ao contexto clínico.

Quanto à seleção de atributos, os métodos aplicados resultaram em melhorias significativas na robustez preditiva e ajudaram a identificar variáveis com grande valor prognóstico para a prática clínica. No **BD 2**, a aplicação dos métodos de regularização *Lasso* e *Elastic Net* levou a um aumento de 1,89% no C-Index e uma redução de 82% no número de variáveis em comparação ao **BD 1**, obtendo o melhor desempenho médio de C-Index, de 0,928. Nos métodos de permutação, o aplicado ao GBS no **BD 5** gerou uma melhoria de 1,33% no C-Index, resultando no segundo melhor desempenho médio (0,923), reforçando a eficácia dessas técnicas na precisão da predição de risco, conforme detalhado na Subseção 5.3.1.

Ao comparar o desempenho dos modelos (Tabelas 15 e 16), o modelo RSF se destacou pela sua superioridade tanto em discriminação (C-Index) quanto em calibração (BS e IBS). Esses resultados são consistentes com estudos prévios sobre CM feminino (LIU et al., 2020; MONCADA-TORRES et al., 2021; PINHEIRO et al., 2022; FANIZZI et al., 2023; TIZI; BERRADO, 2023; XIAO et al., 2022), apesar de utilizarem bases de dados e atributos diferentes. O RSF obteve um C-Index de 0,952 no **BD 6** com 9 atributos, superando estudos como os de Carvalho et al. (2023 e 2024), que reportaram valores de 0,802 e 0,917 (com 9 e 70 atributos, respectivamente) (CARVALHO et al., 2023; CARVALHO et al., 2024), Fanizzi et al. (2023) com 0,65 (18 atributos) (FANIZZI et al., 2023), e Liu et al. (2020) com 0,814 (24 atributos) (LIU et al., 2020). A performance superior do RSF também se reflete nos seus resultados consistentes em todos os subconjuntos do **BD 1**, com C-Index variando de $0,930$ a $0,952$, e valores de BS abaixo de $0,25$ (variando de $0,041$ a $0,053$), indicando excelente calibração na predição ao longo do tempo (XIAO et al., 2022).

O GBS obteve o segundo melhor desempenho discriminativo entre os modelos não lineares, alcançando um C-Index de $0,939$ no **BD 6** com 9 atributos, superando os valores encontrados em estudos anteriores, como o de Carvalho et al. (2024), com $0,914$ (CARVALHO et al., 2024), e e Liu et al. (2020), com $0,823$ (LIU et al., 2020). O GBS também se destacou pela calibração, apresentando os menores valores do IBS, com $0,044$ no **BD 1**, o que indica um desempenho superior ao longo do tempo. Esses resultados são consistentes com estudos que utilizaram métodos baseados em GB, como o XGBoost. Por exemplo, Moncada-Torres et al. (2021) relataram um C-Index de $0,73$, inferior aos resultados obtidos neste estudo (MONCADA-TORRES et al., 2021), e Liu et al. (2020) observaram um C-Index de $0,834$, o que também indica que o GBS aqui utilizado fornece

uma calibração e discriminação aprimoradas (LIU et al., 2020).

O modelo linear SSVM apresentou desempenho intermediário, com um C-Index de 0,904 no **BD 3**, superior aos resultados de Carvalho et al. (2024) com 0,746 (CARVALHO et al., 2024), Moncada-Torres et al. (2021), com 0,63 (MONCADA-TORRES et al., 2021), e Xiao et al. (2022), com 0,812 (XIAO et al., 2022). No entanto, ajustes adicionais podem ser necessários no SSVM devido à sua natureza de otimização por hiperplano. A versão não linear, KSSVM, obteve um C-Index de 0,936 no **BD 5**, mostrando uma melhora no desempenho.

Os modelos clássicos de regressão de Cox (CPH) e suas variantes penalizadas (CPH-L e CPH-EN) apresentaram desempenho satisfatório, próximo ao dos modelos baseados em árvores. O CPH obteve um C-Index de 0,934 no **BD 2**, superando outros estudos como o de Carvalho et al. (2024), com 0,743 (CARVALHO et al., 2024), e Fanizzi et al. (2023), com 0,55 (FANIZZI et al., 2023). As variantes penalizadas (CPH-L e CPH-EN) atingiram C-Index de 0,937 e 0,942, respectivamente, superando os resultados de Xiao et al. (2022), com 0,816 (XIAO et al., 2022). Embora o CPH continue a ser amplamente utilizado, ele exige suposições adicionais, como a proporcionalidade dos riscos e a linearidade dos efeitos (KRZYZIŃSKI et al., 2023; TIZI; BERRADO, 2023; FANIZZI et al., 2023).

Tabela 23 – Comparação do desempenho do C-Index dos métodos de AM com os da Literatura.

Estudos	**Atributos**	**CPH**	**GBS**	**RSF**	**SSVM**
Presente Estudo	9	**0,934**	**0,939**	**0,952**	**0,904**
Carvalho *et al.* 2024	70	0,743	0,914	0,917	0,746
Fanizzi *et al.* 2023	18	0,55		0,65	
Liu *et al.* 2020	10	0,759	0,823	0,814	
Moncada-Torres *et al.* 2021	9	0,63		0,63	0,63
Xiao *et al.* 2022	21	0,814		0,827	0,812

(**Vermelho**) Melhores desempenhos do C-Index dos métodos de AM entre os estudos.

Fonte: Elaborada pela autora (2024).

A Tabela 23 compara os resultados do presente estudo com os de pesquisas recentes, evidenciando que, apesar das diferenças na origem dos BD, os resultados obtidos se destacam em relação aos outros estudos que utilizaram dados secundários de diferentes populações (China, Itália e Holanda) (FANIZZI et al., 2023; LIU et al., 2020; MONCADA-TORRES et al., 2021; XIAO et al., 2022). O desempenho deste estudo, destacado em vermelho na tabela, pode ser associado à qualidade das informações clínicas do BD, ao pré-processamento dos dados e à análise de correlação. Os modelos de regressão de Cox, amplamente utilizados na análise de sobrevida por sua interpretabilidade, requerem redução da dimensionalidade, flexibilização de suposições restritivas e tratamento de relações não lineares (FANIZZI et al., 2023; TIZI; BERRADO, 2023; XIAO et al., 2022).

Em contraste, modelos baseados em árvores apresentam maior capacidade preditiva e adaptabilidade, mesmo com grandes volumes de dados (FANIZZI et al., 2023; LIU et al., 2020; MONCADA-TORRES et al., 2021; XIAO et al., 2022). Nesse contexto, a integração dessas abordagens surge como estratégia promissora para aprimorar a predição da sobrevida e otimizar a tomada de decisões clínicas. Essa sinergia possibilita abordagens mais precisas e individualizadas para pacientes com CM, impulsionando avanços na condução clínica (CARVALHO et al., 2023; CARVALHO et al., 2024; LAI et al., 2019).

5.3.3 IMPÁCTOS CLÍNICOS E PRÁTICOS DOS RESULTADOS

A aplicação de técnicas de AM na predição da sobrevida para pacientes com CM feminino, validados com dados clínicos usuais da prática médica, não apenas melhora significativamente a precisão preditiva dos modelos, mas também tem o potencial de impactar positivamente a gestão clínica e o planejamento terapêutico (CARVALHO et al., 2023; FANIZZI et al., 2023; MONCADA-TORRES et al., 2021). Os resultados deste estudo demonstram que a identificação de variáveis prognósticas relevantes pode guiar decisões clínicas mais informadas, contribuindo para melhores desfechos para as pacientes e oferecendo um melhor enfrentamento da doença no contexto da saúde pública (CINTRA et al., 2012; KRZYZIŃSKI et al., 2023; LIU et al., 2020; XIAO et al., 2022).

Além de avançar na predição de sobrevida para CM, este estudo preenche lacunas importantes na interseção entre oncologia e modelagem computacional, propondo uma abordagem inovadora. A melhoria na predição de sobrevida e aa identificação de fatores prognósticos críticos podem influenciar diretamente políticas públicas de saúde e estratégias clínicas, promovendo um cuidado mais eficiente e individualizado para as pacientes com CM (CARVALHO et al., 2023).

Na Seção 5.2, foram destacados os métodos de seleção de atributos e modelagem preditiva aplicados à oncologia, ressaltando seu impacto na prática clínica. A análise revelou a identificação de 22 variáveis prognósticas importantes, como documentada no Quadro 14, das quais seis se destacam pela sua relevância no prognóstico e na condução terapêutica. Estas variáveis são discutidas a seguir:

1. **Estadiamento Anatômico:** O estadiamento da doença é um dos principais determinantes para a escolha do tratamento. A maioria das pacientes diagnosticadas com CM estavam nos estágios II e III, representando 60% dos casos. A correta classificação do estágio da doença ao diagnóstico, levando em consideração sua extensão locorregional e à distância, é fundamental para guiar decisões clínicas (BARRIOS et al., 2022; BRASIL, 2014).

2. **Metástase à Distância:** A presença de metástase é um indicador da gravidade do CM, também utilizado para definição do Estadiamento Anatômico. Aproxima-

damente 24,37% das pacientes apresentavam forma mais agressiva da doença, a metástases à distância, condição que agrava o prognóstico. Esta variável é particularmente importante, pois pacientes em estágios iniciais podem ter metástases não detectadas, especialmente em casos de cânceres triplo-negativos ou HER2+ (BARRIOS et al., 2022).

3. **Classificação Imuno-histoquímicos:** A avaliação do perfil imuno-histoquímico, como descrito no Quadro 1, permite classificar os tumores em subtipos com prognósticos distintos. Entre as pacientes analisadas, 24,55% eram Luminais A, 34,41% Luminal B HER2-, 11,29% Luminal B HER2+, 7,77% com superexpressão do HER2+ e 22,04% com tumores triplo-negativos. Esta classificação é determinante para a escolha do tratamento, uma vez que os tumores triplo-negativos e com superexpressão de HER2+ apresentam prognósticos mais desfavoráveis (CIRQUEIRA et al., 2011; CINTRA, 2012).

4. **Linfonodo Isolado:** O comprometimento dos linfonodos é um marcador importante para a extensão da doença e a escolha do tratamento. A média de linfonodos isolados encontrados na população de estudo foi de 12,91%. A presença de linfonodos comprometidos reflete a agressividade do câncer e a eficácia das terapias administradas (BRASIL, 2014; CINTRA, 2012).

5. **Tratametno Sistêmico:** O tratamento sistêmico, que inclui quimioterapia ou hormonoterapia, é definido com base no estadiamento e no perfil imuno-histoquímico. Entre as pacientes, 28,14% receberam hormonioterapia, 40,68% receberam antraciclínico, 10,57% foram tratadas com uma combinação de antraciclico e taxano, 8,04% com CMF, e 12,54% não utilizaram tratamento medicamentoso. A escolha do tratamento tem impacto direto sobre o prognóstico das pacientes (BRASIL, 2014; CINTRA, 2012).

6. **Tamanho do Tumor (X):** O tamanho do tumor primário é outro marcador importante do estadiamento e está associado à agressividade da doença. No estudo, a média do tamanho dos tumores no eixo X foi de 3,21 cm. Os tumores maiores estão frequentemente associados a um pior prognóstico e maior risco de disseminação para outros órgãos (BARRIOS et al., 2022; CINTRA, 2012).

As variáveis prognósticas destacadas acima, juntamente com uma extensa lista de características clínicas que descrevem o estado dos pacientes e sua doença, contribuem para melhorar o curso clínico terapêutico (BARRIOS et al., 2022; BRASIL, 2014; TIZI; BERRADO, 2023). As variáveis clínicas selecionadas pelos métodos de seleção de atributos sugerem a necessidade de um acompanhamento rigoroso e classificação precisa do tumor para direcionar corretamente a escolha terapêutica.

É importante reconhecer que conjuntos de dados médicos não representam apenas a doença em si, mas também como esta doença está sendo tratada, ou seja, representam uma amostragem do acesso do paciente ao sistema de saúde, mostrando a necessidade de maior utilização dessas informações disponíveis nos serviços de saúde para a produção de conhecimento. Isso possibilita que modelos prognósticos treinados em diferentes populações e períodos de tempo minimizem falhas ao capturar o risco real para o paciente em questão (CINTRA et al., 2012; TIZI; BERRADO, 2023).

6 CONCLUSÃO

Este estudo demonstrou que os métodos de AM, especialmente o RSF, apresentam desempenho superior na predição de sobrevida de pacientes com CM feminino. Os resultados ressaltaram a importância do pré-processamento dos dados e da análise de correlação, que vão além da simples significância estatística das variáveis, contribuindo diretamente para a melhoria da precisão preditiva. A aplicação de estratégias rigorosas de pré-processamento resultou em um aumento médio de 15,88% no desempenho discriminativo dos modelos, evidenciando o impacto positivo dessas abordagens na modelagem. Além disso, a seleção criteriosa de atributos foi essencial para identificar as variáveis mais relevantes, destacando-se "Estadiamento Anatômico", "Metástase à Distância", "Classificação Imuno-histoquímica", "Linfonodo Isolado", "Tratamento Sistêmico" e "Tamanho do Tumor (X)". Esses achados reforçam a necessidade de um tratamento adequado dos dados e de uma escolha bem fundamentada das variáveis, visando maximizar a acurácia preditiva e a aplicabilidade clínica dos modelos.

6.1 CONTRIBUIÇÕES DA MODELAGEM COMPUTACIONAL NA PREDIÇÃO DE SOBREVIDA PARA O CÂNCER DE MAMA

A integração de técnicas de modelagem computacional na medicina tem transformado a teoria em prática. Embora esses métodos ainda estejam em desenvolvimento, há evidências crescentes de seu impacto na avaliação diagnóstica, terapêutica e prognóstica (PAIXÃO et al., 2022). Em particular, a aplicação da IA, especialmente dos métodos de AM, tem demonstrado sua importância prática ao fornecer *insights* valiosos sobre o prognóstico do CM, permitindo um cuidado mais personalizado e adaptado às necessidades individuais das pacientes. Além disso, a incorporação dessas tecnologias nos registros médicos eletrônicos viabiliza o processamento eficiente de grandes volumes de dados. O futuro da IA e das técnicas de AM na pesquisa médica e na saúde é promissor, com potencial significativo de expansão e impacto (RUBINGER et al., 2022).

Neste estudo, a análise da eficácia dos métodos de AM na predição da sobrevida em mulheres com CM destacou a importância do pré-processamento dos dados, da análise de correlação e da seleção criteriosa de atributos. O impacto dessa etapa foi evidenciado pelo aumento significativo no desempenho discriminativo dos modelos, que passou de 0,784 no primeiro pré-processamento para 0,911 após a segunda etapa de refinamento dos dados e análise de correlação. A combinação dessas abordagens com modelos de AM resultou em ganhos significativos na acurácia preditiva, como detalhado na Subseção 5.3.1.

A identificação das variáveis prognósticas, realizada por meio de métodos de seleção de atributos e análise de permutação de importância, permitiu detectar fatores críticos para decisões clínicas mais informadas, conforme descrito na Seção 5.2 e na Subsubseção 5.3.3.

Variáveis como "Estadiamento Anatômico"e "Metástase à Distância"foram selecionadas por todos os métodos de seleção de atributos, destacando-se como essenciais para a predição de sobrevida nos métodos de AM. Esses achados podem orientar decisões clínicas mais precisas, auxiliando na definição de terapias alinhadas ao perfil individual das pacientes (BARRIOS et al., 2022).

Os resultados obtidos evidenciam o avanço da aplicação de modelos computacionais na predição de sobrevida de pacientes com CM feminino, com destaque para o RSF, que apresentou o melhor desempenho entre os modelos testados, seguido pelos modelos GBS e CPH-EN, ambos superando os resultados descritos na literatura, conforme discutido na Subseção 5.3.2. Os modelos clássicos, como a Regressão de Cox, também se beneficiaram do pré-processamento e da redução de dimensionalidade, resultando em melhorias na acurácia. Esses resultados corroboram estudos anteriores (CARVALHO et al., 2023; LIU et al., 2020; MONCADA-TORRES et al., 2021; XIAO et al., 2022; PINHEIRO et al., 2022; FANIZZI et al., 2023) e reforçam o papel dos métodos de AM na prática clínica.

Um dos principais desafios — e também uma das maiores contribuições deste estudo — foi lidar com BD clínicos cujos dados, coletados entre 2003 e 2005 e acompanhados até 2011 (CINTRA, 2012; FAYER, 2014), embora antigos, preservam a riqueza das informações típicas da prática clínica. Durante o desenvolvimento do trabalho, o pré-processamento dos dados e a seleção criteriosa das variáveis, baseada na análise de correlação, foram aprendizados fundamentais. Além disso, o uso de métodos de seleção de atributos permitiu identificar as variáveis mais relevantes nas rotinas clínicas, fornecendo *insights* valiosos sobre os fatores que influenciam o prognóstico e o tratamento de pacientes com CM.

As implicações clínicas dos achados são significativas, permitindo uma visão mais detalhada das variáveis prognósticas, tanto para melhorar a predição dos métodos de AM quanto para o planejamento de tratamentos mais ajustados às necessidades individuais. A integração desses modelos aos registros médicos eletrônicos pode aprimorar a gestão de casos e otimizar o curso clínico-terapêutico, melhorando a tomada de decisões clínicas e a estratégia de saúde pública (CARVALHO et al., 2023; CARVALHO et al., 2024).

Apesar dos avanços, o estudo enfrentou outro desafio, como a aquisição de dados de diferentes fontes, o tratamento de dados faltantes, a seleção de variáveis clínicas relevantes e a interpretação das variáveis que refletem a prática clínica. Esses fatores impactaram a análise da evolução do curso clínico, essencial para a interpretação dos resultados e sua aplicação na medicina. Além disso, o método SSVM não linear (KSSVM) exigiu ajustes técnicos, principalmente na aplicação da função Kernel RBF, que não estava definida pela biblioteca, necessitando de um aprofundamento na fundamentação matemática (PÖLSTERL, 2023).

Embora os resultados obtidos sejam promissores, o estudo apresenta algumas limitações que precisam ser consideradas. A principal limitação é a quantidade de dados

ausentes no BD clínico, que restringiu o uso de variáveis disponíveis. Além disso, embora o BD utilizado contenha informações valiosas sobre o curso clínico-terapêutico, ele está desatualizado, o que pode limitar a representatividade dos dados frente aos avanços mais recentes no tratamento do CM. Outro ponto crítico é a não validação dos modelos em outros BD clínicos, o que impede a verificação da robustez e da capacidade de generalização dos resultados. As análises, fundamentadas em dados históricos, também podem ser influenciadas pela evolução dos tratamentos e protocolos clínicos, impactando os resultados futuros, especialmente com a implementação de novas inovações terapêuticas.

Um exemplo destas evoluções do curso clínico terapêutico é o Estadiamento Clínico-Prognóstico, que além do sistema TNM e do Estadiamento Anatômico, considera também o Grau Histológico do tumor e os resultados dos receptores (RE, RP e HER2) na classificação, conforme definido por Barrios (2022) (BARRIOS et al., 2022). Assim, a validação dos modelos em diferentes conjuntos de dados atuais e diferentes populações é fundamental para confirmar sua robustez, capacidade de generalização e aplicabilidade clínica.

Em conclusão, este estudo destacou a variabilidade no desempenho dos AM na análise de sobrevida, com resultados que dependem das características dos dados e dos critérios de avaliação adotados. No contexto clínico do CM feminino, o modelo supervisionado RSF se destacou pelo seu desempenho superior. Contudo, a escolha do modelo ideal deve levar em conta não apenas a acurácia, mas também fatores cruciais como interpretabilidade, eficiência e validade clínica. Assim, pesquisas futuras devem se concentrar na melhoria contínua do pré-processamento de dados, na otimização dos parâmetros dos modelos e na seleção cuidadosa de atributos, a fim de aprimorar seu desempenho e garantir maior aplicabilidade clínica. A evolução dessas abordagens pode fortalecer ainda mais a integração dos modelos de AM na prática clínica, oferecendo suporte decisivo para um cuidado mais personalizado e eficaz.

6.2 AVANÇOS DOS OBJETIVOS ESPECÍFICOS

Os objetivos específicos delineados para este trabalho, descritos na Subseção 1.3.2, foram cumpridos conforme segue:

1. **Revisão Sistemática:** O primeiro objetivo, que envolveu a condução de uma RS, dos estudos publicados de 2018 a 2022 sobre a aplicação de métodos de AM na predição de sobrevida no CM, foi alcançado conforme descrito no Capítulo 3. De maneira surpreendente, dentre os 19 estudos incluídos, apenas dois (MONCADA-TORRES et al., 2021; LIU et al., 2020) utilizaram a biblioteca específica para análise de sobrevida, Scikit-survival, a mesma empregada nesta pesquisa, demonstrando assim a relevância deste trabalho.

2. **BD Clínicos:** O segundo objetivo, envolveu a submissão e aprovação do projeto

(parecer Anexo 6.3) para obtenção dos daddos para a construção do BD Clínicos, a realização dos pré-processamento dos dados e análise descritiva e de correlação dos dados, conforme descrito na Subseção 4.1.3, do Capítulo 4, Material e Métodos. A descrição dos dados encontra-se na Seção 5.1 do Capítulo 5, Resultados e Discussão.

3. **Simulação e Avaliação dos MAM:** O terceiro objetivo, que envolveu a simulação dos métodos de AM supervisionados e a avaliação de seu desempenho aplicado aos dados clínicos, foi alcançado e está descrito na Seção 5.3 do Capítulo 5, Resultados e Discussão. A metodologia de implementação e validação está descrita nas Subseções 4.2 e 4.4, do Capítulo 4, Material e Métodos.

4. **Métodos de Seleção de Atributos:** O quarto objetivo, que envolveu a aplicação de métodos de seleção de atributos para melhorar o desempenho e identificar variáveis clínicas relevantes, foi realizado conforme descrito na Seção 5.2 e na Subseção 5.3.3 do Capítulo 5, Resultados e Discussão. A metodologia de seleção de características e permutação da importância das variáveis está detalhada na Seção 4.3 no Capítulo 4, Material e Métodos.

5. **Comparação dos Resultados com a Literatura:** O quinto objetivo, envolve defender a aplicabilidade dos resultados em termos de desempenho e de utilidade na prática clínica, conforme apresentado e discutido na Subseção 5.3.2.

Durante o doutorado, foram submetidos dois artigos e um capítulo de livro com os resultados da pesquisa. O artigo "Aplicação do Random Survival Forest na Análise da Sobrevida para Câncer de Mama", publicado no Journal of Health Informatics, está disponível na íntegra no Anexo 6.3 (CARVALHO et al., 2023). O artigo "Comparative Analysis of Machine Learning Models for Breast Cancer Patients' Survival Prediction", publicado no Journal Intelligent Systems Design and Applications: Machine Learning Solutions, também está disponível na íntegra no Anexo 6.3 (CARVALHO et al., 2024). O capítulo de livro está em processo de publicação, intitulado "IA na Saúde: Avanços e Desafios", no livro INOVAÇÃO NA ÁREA DA SAÚDE E OS DESAFIOS DA INDÚSTRIA 4.0. Parte dos objetivos específicos foi desenvolvido durante o doutorado sanduíche na Faculdade de Engenharia da Universidade do Porto, Portugal, iniciado em abril e finalizado em setembro deste ano, contribuindo positivamente para o andamento e conclusão deste trabalho.

6.3 PERSPECTIVA PARA PESQUISAS FUTURAS

A pesquisa desenvolvida tem caráter inovador ao aplicar métodos de AM para predição de sobrevida no CM feminino, utilizando dados clínicos da Zona da Mata Mineira. Pesquisas futuras devem focar na ampliação da validação dos modelos com dados clínicos

de diferentes populações e na integração de novas técnicas de pré-processamento, seleção de atributos e otimização de parâmetros. Especificamente, a incorporação de métodos como o algoritmo k-Nearest Neighbors (k-NN) aliado à análise da importância clínica das variáveis pode contribuir para uma seleção mais criteriosa dos atributos mais relevantes para a modelagem, melhorando a precisão e interpretabilidade dos modelos desenvolvidos (CURIOSO et al., 2023). A expansão desta pesquisa, unindo oncologia e modelagem computacional para identificar fatores prognósticos específicos ao CM feminino, é essencial. Além disso, a aplicação de técnicas avançadas de otimização e métodos mais robustos de validação poderá garantir maior confiabilidade e replicabilidade dos resultados obtidos.

Investimentos em pesquisa e no desenvolvimento de modelos computacionais podem não apenas melhorar a predição de sobrevida, mas também influenciar aspectos emocionais e culturais das pacientes. A integração desses modelos na prática clínica pode revolucionar o tratamento do CM, proporcionando melhores prognósticos e qualidade de vida para as pacientes (LI et al., 2021; MIN et al., 2021). Futuros estudos devem se concentrar na promoção da colaboração interdisciplinar entre oncologistas e cientistas de dados, visando ao desenvolvimento de modelos computacionais interpretáveis e validados com dados clínicos, capazes de ser aplicados na prática clínica para o planejamento do tratamento e predição da sobrevida no CM feminino. A sinergia entre o conhecimento especializado e a análise avançada de dados, como sugerido por Fanizzi et al. (2023), pode levar a avanços significativos neste campo (FANIZZI et al., 2023).

Este estudo contribuiu significativamente para o campo da oncologia computacional, demonstrando que os métodos de AM podem oferecer predições precisas e clinicamente relevantes para a sobrevida de pacientes com CM feminino. A implementação efetiva dessas ferramentas na prática clínica tem o potencial de melhorar o tratamento do CM, oferecendo esperança e melhores prognósticos para milhões de pacientes em todo o mundo. A aplicação ampla e efetiva das técnicas de AM na prática clínica exigirá esforços conjuntos de pesquisadores, profissionais de saúde e gestores de políticas públicas, com o objetivo de melhorar os desfechos clínicos e a qualidade de vida das pacientes.

REFERÊNCIAS

AFSHAR, Hadi Lotfnezhad; JABBARI, Nasrollah; KHALKHALI, Hamid Reza; ESNAASHARI, Omid. Prediction of breast cancer survival by machine learning methods: An application of multiple imputation. **Iranian Journal of Public Health**, Tehran University of Medical Sciences, v. 50, n. 3, p. 598, 2021.

AIVALIOTIS, Georgios; PALCZEWSKI, Jan; ATKINSON, Rebecca; CADE, Janet E; MORRIS, Michelle A. A comparison of time to event analysis methods, using weight status and breast cancer as a case study. **Scientific reports**, Nature Publishing Group, v. 11, n. 1, p. 1–9, 2021.

ALLEMANI, Claudia; MATSUDA, Tomohiro; CARLO, Veronica Di; HAREWOOD, Rhea; MATZ, Melissa; NIKŠIĆ, Maja; BONAVENTURE, Audrey; VALKOV, Mikhail; JOHNSON, Christopher J; ESTÈVE, Jacques et al. Global surveillance of trends in cancer survival 2000–14 (concord-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. **The Lancet**, Elsevier, v. 391, n. 10125, p. 1023–1075, 2018.

AMIN, Mahul B; EDGE, Stephen B; GREENE, Frederick L; BYRD, David R; BROOKLAND, Robert K; WASHINGTON, Mary Kay; GERSHENWALD, Jeffrey E; COMPTON, Carolyn C; HESS, Kenneth R; SULLIVAN, Daniel C et al. **AJCC cancer staging manual**. [S.l.]: Springer, 2017. v. 1024.

ARNOLD, Melina; MORGAN, Eileen; RUMGAY, Harriet; MAFRA, Allini; SINGH, Deependra; LAVERSANNE, Mathieu; VIGNAT, Jerome; GRALOW, Julie R; CARDOSO, Fatima; SIESLING, Sabine et al. Current and future burden of breast cancer: Global statistics for 2020 and 2040. **The Breast**, Elsevier, v. 66, p. 15–23, 2022.

ASSIS, Janilson Pinheiro de; SOUSA, Roberto Pequeno de; DIAS, Carlos Tadeu Dos Santos. **Glossário de estatística**. [S.l.]: Edufersa, 2019.

BARRIOS, Carlos Henrique Escosteguy; AMORIM, Gilberto Luiz da Silva; TAVARES, Maíra; CRUZ, Marcelo; GONÇALVES, Marina Sahade; BEDIN, Sabrina Richter; REINERT, Tomás. Mama: Estadiamento. **Diretrizes de Tratamentos Oncológicos Recomendados pela Sociedade Brasileira de Oncologia Clínica**, SBOC, 2022.

BELLE, Vanya Van; PELCKMANS, Kristiaan; SUYKENS, Johan AK; HUFFEL, Sabine Van. Learning transformation models for ranking and survival analysis. **Journal of machine learning research**, v. 12, n. 3, 2011.

BOERI, Carlo; CHIAPPA, Corrado; GALLI, Federica; BERARDINIS, Valentina De; BARDELLI, Laura; CARCANO, Giulio; ROVERA, Francesca. Machine learning techniques in breast cancer prognosis prediction: A primary evaluation. **Cancer medicine**, Wiley Online Library, v. 9, n. 9, p. 3234–3243, 2020.

BRASIL. **MINISTÉRIO DA SAÚDE. Protocolos Clínicos e Diretrizes Terapêuticas em Oncologia**. Secretaria de Atenção à Saúde. Brasília - DF: Ministério da Saúde, 2014.

BRASIL. **Ministério da Saúde. DATASUS - TabNEt**. 2021. Acesso em: 2023-09-20. Disponível em: https://datasus.saude.gov.br/informacoes-de-saude-tabnet/.

BRASIL. **Ministério da Saúde. DATASUS Tecnologia da Informação a Serviço do SUS. MORTALIDADE - BRASIL**. 2021. Acesso em: 2023-07-19. Disponível em: http://tabnet.datasus.gov.br/cgi/tabcgi.exe?sim/cnv/obt10uf.def.

BRASIL. **Ministério da Saúde. DATASUS Tecnologia da Informação a Serviço do SUS. MORTALIDADE - BRASIL**. 2022. Acesso em: 2024-07-01. Disponível em: http://tabnet.datasus.gov.br/cgi/tabcgi.exe?sim/cnv/obt10uf.def.

BRASIL; CIêNCIA TECNOLOGIA, Inovação e Insumos Estratégicos em Saúde. Departamento de Gestão e Incorporação de Tecnologias em Saúde Ministério da Saúde. Secretaria de. **Diretrizes metodológicas: elaboração de revisão sistemática e metanálise de ensaios clínicos randomizados**. Brasília Internet, 2012. ISSN 978-65-5993-021-0. Disponível em: https://rebrats.saude.gov.br/images/Documentos/2021/20210622_Diretriz_Revisao_Sistematica_2021.pdf.

BRASIL, Ministério da Saúde. Estimativa 2023 : incidência de câncer no brasil / instituto nacional de câncer. **Rio de Janeiro: INCA**, 2022.

BRAY, F.; FERLAY, J.; SOERJOMATARAM, I.; SIEGEL, R. L.; TORRE, L. A.; JEMAL, A. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. **CA: a cancer journal for clinicians**, v. 68, n. 6, p. 394–424, 2018.

BRAY, Freddie; LAVERSANNE, Mathieu; SUNG, Hyuna; FERLAY, Jacques; SIEGEL, Rebecca L; SOERJOMATARAM, Isabelle; JEMAL, Ahmedin. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. **CA: a cancer journal for clinicians**, Wiley Online Library, v. 74, n. 3, p. 229–263, 2024.

BREIMAN, Leo. Random forests. **Machine learning**, Springer, v. 45, p. 5–32, 2001.

BRITTON, Nicholas F; BRITTON, NF. **Essential mathematical biology**. [S.l.]: Springer, 2003. v. 453.

BURGES, Christopher JC. A tutorial on support vector machines for pattern recognition. **Data mining and knowledge discovery**, Springer, v. 2, n. 2, p. 121–167, 1998.

BUSTAMANTE-TEIXEIRA, Maria Teresa; FAERSTEIN, Eduardo; LATORRE, Maria do Rosário. Técnicas de análise de sobrevida. **Cadernos de Saúde Pública**, SciELO Public Health, v. 18, p. 579–594, 2002.

CARVALHO, André; FACELI, K; LORENA, A; GAMA, J. Inteligência artificial–uma abordagem de aprendizado de máquina. **Rio de Janeiro: LTC**, v. 2, p. 45, 2011.

CARVALHO, Daniela Schimitz de. **Modelagem matemática do crescimento tumoral mamário**. 2016. Dissertação (Dissertação, Programa de Pós-Graduação em Modelagem Computacional da Faculdade de Engenharia) — Universidade Federal de Juiz de Fora, MG, 2016.

CARVALHO, Daniela Schimitz de; GUERRA, Maximiliano Ribeiro; BARRA, Luis Paulo da Silva; QUEIROZ, Rafael Alves Bonfim de. Modelagem computacional do crescimento tumoral mamÁrio. **Anais do Seminário Científico do UNIFACIG**, n. 3, 2018.

CARVALHO, Daniela Schimitz de; GUERRA, Maximiliano Ribeiro; BARRA, Luis Paulo da Silva; QUEIROZ, Rafael Alves Bonfim de. Aspectos gerais epidemiológicos da mortalidade por câncer de mama feminino no brasil e no mundo. **Anais do Simpósio de Enfermagem**, v. 1, n. 1, 2019.

CARVALHO, Daniela Schimitz de; NOGUEIRA, Thallys da Silva; GOLIATT, Priscila Vanessa Zabala Caprile. Aplicação do random survival forest na análise da sobrevida para câncer da mama. **Journal of Health Informatics**, v. 15, n. Especial, 2023.

CARVALHO, Daniela Schimitz de; CAPRILES, Priscila; GOLIATT, Leonardo. Comparative analysis of machine learning models for breast cancer patients' survival prediction. **Intelligent Systems Design and Applications: Machine Learning Solutions, Volume 7**, Springer Nature, v. 7, p. 172, 2024.

CARVALHO, Marilia Sá; ANDREOZZI, Valeska Lima; CODEÇO, Claudia Torres; CAMPOS, Dayse Pereira; BARBOSA, Maria Tereza Serrano; SHIMAKURA, Silvia Emiko. **Análise de sobrevivência: teoria e aplicações em saúde**. [S.l.]: SciELO-Editora FIOCRUZ, 2011.

CHENG, SC; WEI, Lee J; YING, Zhiliang. Analysis of transformation models with censored data. **Biometrika**, Oxford University Press, v. 82, n. 4, p. 835–845, 1995.

CHU, Wei; KEERTHI, S Sathiya. Support vector ordinal regression. **Neural computation**, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , v. 19, n. 3, p. 792–815, 2007.

CINTRA, JRD. **Sobrevida e fatores associados em pacientes com câncer de mama, com diagnóstico entre 2003 e 2005 no município de Juiz de Fora-Minas Gerais**. 2012. Tese (Tese (Doutorado) Programa de Pós-Graduação em Saúde Brasileira da Faculdade de Mediciana) — Universidade Federal de Juiz de Fora, 2012.

CINTRA, Jane Rocha Duarte; TEIXEIRA, Maria Teresa Bustamante; DINIZ, Roberta Wolp; JUNIOR, Homero Gonçalves; FLORENTINO, Thiago Marinho; FREITAS, Guilherme Fialho De; OLIVEIRA, Luiz Raphael Mota; NEVES, Mariana Teodoro Dos Reis; PEREIRA, Talita; GUERRA, Maximiliano Ribeiro. Perfil imuno-histoquímico e variáveis clinicopatológicas no câncer de mama. **Revista da Associação Médica Brasileira**, SciELO Brasil, v. 58, p. 178–187, 2012.

CIRQUEIRA, Magno Belém; MOREIRA, Marise Amaral Rebouças; SOARES, Leonardo Ribeiro; FREITAS-JÚNIOR, Ruffo. Subtipos moleculares do câncer de mama. **Femina**, 2011.

CORTES, Corinna; VAPNIK, Vladimir. Support-vector networks. **Machine learning**, Springer, v. 20, p. 273–297, 1995.

COX, David R. Regression models and life-tables. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 34, n. 2, p. 187–202, 1972.

CURIOSO, Isabel; SANTOS, Ricardo; RIBEIRO, Bruno; CARREIRO, André; COELHO, Pedro; FRAGATA, José; GAMBOA, Hugo. Addressing the curse of missing data in clinical contexts: A novel approach to correlation-based imputation. **Journal of King**

Saud University-Computer and Information Sciences, Elsevier, v. 35, n. 6, p. 101562, 2023.

DAVENPORT, Thomas; KALAKOTA, Ravi. The potential for artificial intelligence in healthcare. **Future healthcare journal**, Royal College of Physicians, v. 6, n. 2, p. 94, 2019.

DENG, Fei; HUANG, Jibing; YUAN, Xiaoling; CHENG, Chao; ZHANG, Lanjing. Performance and efficiency of machine learning algorithms for analyzing rectangular biomedical data. **Laboratory Investigation**, Elsevier, v. 101, n. 4, p. 430–441, 2021.

D'EREDITA, G; GIARDINA, C; MARTELLOTTA, M; NATALE, T; FERRARESE, F. Prognostic factors in breast cancer: the predictive value of the nottingham prognostic index in patients with a long-term follow-up that were treated in a single institution. **European Journal of Cancer**, Elsevier, v. 37, n. 5, p. 591–596, 2001.

DICUONZO, Grazia; DONOFRIO, Francesca; FUSCO, Antonio; SHINI, Matilda. Healthcare system: Moving forward with artificial intelligence. **Technovation**, Elsevier, v. 120, p. 102510, 2023.

DUVENAUD, David. **The kernel cookbook: Advice on covariance functions**. [S.l.: s.n.], 2014.

EDELSTEIN-KESHET, Leah. **Mathematical models in biology**. [S.l.]: SIAM, 2005.

ENGLAND, Public Health. **Predict Breast Cancer**. 2022. Acesso em: 2023-08-09. Disponível em: https://breast.predict.nhs.uk/about/technical/publications.

FANIZZI, Annarita; POMARICO, Domenico; RIZZO, Alessandro; BOVE, Samantha; COMES, Maria Colomba; DIDONNA, Vittorio; GIOTTA, Francesco; FORGIA, Daniele La; LATORRE, Agnese; PASTENA, Maria Irene et al. Machine learning survival models trained on clinical data to identify high risk patients with hormone responsive her2 negative breast cancer. **Scientific Reports**, Nature Publishing Group UK London, v. 13, n. 1, p. 8575, 2023.

FAYER, Vivian Assis. **Sobrevida de 10 anos e fatores prognósticos em coorte hospitalar de pacientes com câncer de mama assistidas em Juiz de Fora, Minas Gerais, Brasil**. 2014. Dissertação (Dissertação, Programa de Pós-Graduação em Saúde Coletiva da Faculdade de Medicina) — Universidade Federal de Juiz de Fora, MG, 2014.

FERLAY, Jacques; COLOMBET, M; SOERJOMATARAM, Isabelle; DYBA, T; RANDI, Giorgia; BETTIO, Manola; GAVIN, Anna; VISSER, Otto; BRAY, Freddie. Cancer incidence and mortality patterns in europe: Estimates for 40 countries and 25 major cancers in 2018. **European journal of cancer**, Elsevier, v. 103, p. 356–387, 2018.

FERLAY, Jacques; COLOMBET, Murielle; SOERJOMATARAM, Isabelle; MATHERS, Colin; PARKIN, Donald M; PIÑEROS, Marlon; ZNAOR, Ariana; BRAY, Freddie. Estimating the global cancer incidence and mortality in 2018: Globocan sources and methods. **International journal of cancer**, Wiley Online Library, v. 144, n. 8, p. 1941–1953, 2019.

FERLAY, Jacques; COLOMBET, Murielle; SOERJOMATARAM, Isabelle; PARKIN, Donald M; PIÑEROS, Marion; ZNAOR, Ariana; BRAY, Freddie. Cancer statistics for the year 2020: An overview. **International Journal of Cancer**, Wiley Online Library, 2021.

FRIEDMAN, Jerome H. Greedy function approximation: a gradient boosting machine. **Annals of statistics**, JSTOR, p. 1189–1232, 2001.

FRIEDMAN, Jerome H. Stochastic gradient boosting. **Computational statistics & data analysis**, Elsevier, v. 38, n. 4, p. 367–378, 2002.

GANGGAYAH, Mogana Darshini; TAIB, Nur Aishah; HAR, Yip Cheng; LIO, Pietro; DHILLON, Sarinder Kaur. Predicting factors for survival of breast cancer patients using machine learning techniques. **BMC medical informatics and decision making**, Springer, v. 19, p. 1–17, 2019.

GOMES, Mariana Aquaroni Farão; FRAGA, André Veloce; GOMES, Hélio Aquaroni Farão. Câncer de mama masculino: um assunto que deve ser abordado. **CuidArte, Enferm**, p. 253–258, 2022.

GRAF, Erika; SCHMOOR, Claudia; SAUERBREI, Willi; SCHUMACHER, Martin. Assessment and comparison of prognostic classification schemes for survival data. **Statistics in medicine**, Wiley Online Library, v. 18, n. 17-18, p. 2529–2545, 1999.

GU, Dongxiao; SU, Kaixiang; ZHAO, Huimin. A case-based ensemble learning system for explainable breast cancer recurrence prediction. **Artificial Intelligence in Medicine**, Elsevier, v. 107, p. 101858, 2020.

HAQUE, Mohammad Nazmul; TAZIN, Tahia; KHAN, Mohammad Monirujjaman; FAISAL, Shahla; IBRAHEEM, Sobhee Md; ALGETHAMI, Haneen; ALMALKI, Faris A. Predicting characteristics associated with breast cancer survival using multiple machine learning approaches. **Computational and Mathematical Methods in Medicine**, Hindawi, v. 2022, 2022.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome; HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. Boosting and additive trees. **The elements of statistical learning: data mining, inference, and prediction**, Springer, p. 337–387, 2009.

HESS, Viviane. Adjuvante chemotherapie–entscheidungshilfe aus dem internet: Adjuvant! online. **Therapeutische Umschau**, Verlag Hans Huber, v. 65, n. 4, p. 201–205, 2008.

HOTHORN, Torsten; BÜHLMANN, Peter; DUDOIT, Sandrine; MOLINARO, Annette; LAAN, Mark J Van Der. Survival ensembles. **Biostatistics**, Oxford University Press, v. 7, n. 3, p. 355–373, 2006.

HOWARD, Grant R; JOHNSON, Kaitlyn E; AYALA, Areli Rodriguez; YANKEELOV, Thomas E; BROCK, Amy. A multi-state model of chemoresistance to characterize phenotypic dynamics in breast cancer. **Scientific reports**, Nature Publishing Group, v. 8, n. 1, p. 1–11, 2018.

HUANG, Kaiyan; ZHANG, Jie; YU, Yushuai; LIN, Yuxiang; SONG, Chuangui. The impact of chemotherapy and survival prediction by machine learning in early elderly triple negative breast cancer (etnbc): a population based study from the seer database. **BMC geriatrics**, Springer, v. 22, n. 1, p. 268, 2022.

HUEMAN, Mathew T; WANG, Huan; YANG, Charles Q; SHENG, Li; HENSON, Donald E; SCHWARTZ, Arnold M; CHEN, Dechang. Creating prognostic systems for cancer patients: A demonstration using breast cancer. **Cancer medicine**, Wiley Online Library, v. 7, n. 8, p. 3611–3621, 2018.

IARC, International Agency for Research on Cancer. **Cancer Incidence in Five Continents.** 2021. Acesso em: 2023-08-14. Disponível em: https://ci5.iarc.fr/Default.aspx.

IARC, International Agency for Research on Cancer. **Cancer Today**. 2023. Acesso em: 2023-08-14. Disponível em: https://gco.iarc.fr/today/home.

IARC, International Agency for Research on Cancer. **Cancer Today**. 2024. Acesso em: 2024-06-06. Disponível em: https://gco.iarc.fr/today.

INCA. **Instituto Nacional do Câncer. Registros de Câncer de Base Populacional**. 2022. Acesso em: 2023-09-20. Disponível em: https://www.gov.br/inca/pt-br/assuntos/cancer/numeros/registros/base-populacional.

INCA. **Instituto Nacional do Câncer. Registros Hospitalares de Câncer (RHC)**. 2022. Acesso em: 2023-09-20. Disponível em: https://www.gov.br/inca/pt-br/assuntos/cancer/numeros/registros/rhc.

INCA. **Instituto Nacional do Câncer. Estatísticas de câncer**. 2023. Acesso em: 2023-07-21. Disponível em: https://www.inca.gov.br/numeros-de-cancer.

INCA, Ministério da Saúde. Instituto Nacional de Câncer. **Atlas on-line de mortalidade.** 2022. Acesso em: 2024-07-01. Disponível em: https://mortalidade.inca.gov.br/MortalidadeWeb/pages/Modelo01/consultar.xhtml#panelResultado.

ISHWARAN, H; KOGALUR, UB. **Random survival forests for R. R News. 7 (2): 25–31**. 2007.

ISHWARAN, H.; KOGALUR, U.B. **Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC)**. [S.l.], 2023. R package version 3.2.3. Disponível em: https://cran.r-project.org/package=randomForestSRC.

ISHWARAN, Hemant; KOGALUR, Udaya B; BLACKSTONE, Eugene H; LAUER, Michael S. **Random survival forests**. 2008.

JANSEN, Tom; GELEIJNSE, Gijs; MAAREN, Marissa Van; HENDRIKS, Mathijs P; TEIJE, Annette Ten; MONCADA-TORRES, Arturo. Machine learning explainability in breast cancer survival. In: **Digital Personalized Health and Medicine**. [S.l.]: IOS Press, 2020. p. 307–311.

JR, Frank E Harrell; LEE, Kerry L; MARK, Daniel B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. **Statistics in medicine**, Wiley Online Library, v. 15, n. 4, p. 361–387, 1996.

KALAFI, EY; NOR, NAM; TAIB, NA; GANGGAYAH, MD; TOWN, C; DHILLON, SK. Machine learning and deep learning approaches in breast cancer survival prediction using clinical data. **Folia biologica**, Charles University in Prague, First Faculty of Medicine, v. 65, n. 5/6, p. 212–220, 2019.

KAPLAN, Edward L; MEIER, Paul. Nonparametric estimation from incomplete observations. **Journal of the American statistical association**, Taylor & Francis, v. 53, n. 282, p. 457–481, 1958.

KITSIOS, Fotis; KAMARIOTOU, Maria; SYNGELAKIS, Aristomenis I; TALIAS, Michael A. Recent advances of artificial intelligence in healthcare: a systematic literature review. **Applied Sciences**, MDPI, v. 13, n. 13, p. 7479, 2023.

KLEINLEIN, Ricardo; RIAÑO, David. Persistence of data-driven knowledge to predict breast cancer survival. **International journal of medical informatics**, Elsevier, v. 129, p. 303–311, 2019.

KRZYZIŃSKI, Mateusz; SPYTEK, Mikołaj; BANIECKI, Hubert; BIECEK, Przemysław. Survshap (t): Time-dependent explanations of machine learning survival models. **Knowledge-Based Systems**, Elsevier, v. 262, p. 110234, 2023.

KöHN-LUQUE, Alvaro; LAI, Xiaoran; FRIGESSI, Arnoldo. Towards personalized computer simulations of breast cancer treatment. **arXiv preprint arXiv:2007.15934**, 2020.

LAI, Xiaoran; GEIER, Oliver M; FLEISCHER, Thomas; GARRED, Oystein; BORGEN, Elin; FUNKE, Simon W; KUMAR, Surendra; ROGNES, Marie E; SEIERSTAD, Therese; BORRESEN-DALE, Anne-Lise et al. Toward personalized computer simulation of breast cancer treatment: A multiscale pharmacokinetic and pharmacodynamic model informed by multitype patient data. **Cancer research**, AACR, v. 79, n. 16, p. 4293–4304, 2019.

LI, Chaofan; LIU, Mengjie; LI, Jia; WANG, Weiwei; FENG, Cong; CAI, Yifan; WU, Fei; ZHAO, Xixi; DU, Chong; ZHANG, Yinbin et al. Machine learning predicts the prognosis of breast cancer patients with initial bone metastases. **Frontiers in Public Health**, Frontiers Media SA, v. 10, 2022.

LI, Jiaxin; ZHOU, Zijun; DONG, Jianyu; FU, Ying; LI, Yuan; LUAN, Ze; PENG, Xin. Predicting breast cancer 5-year survival using machine learning: A systematic review. **PloS one**, Public Library of Science San Francisco, CA USA, v. 16, n. 4, p. e0250370, 2021.

LIU, Pei; FU, Bo; YANG, Simon X; DENG, Ling; ZHONG, Xiaorong; ZHENG, Hong. Optimizing survival analysis of xgboost for ties to predict disease progression of breast cancer. **IEEE Transactions on Biomedical Engineering**, IEEE, v. 68, n. 1, p. 148–160, 2020.

LUDERMIR, Teresa Bernarda. Inteligência artificial e aprendizado de máquina: estado atual e tendências. **Estudos Avançados**, SciELO Brasil, v. 35, p. 85–94, 2021.

MAABREH, Roqia Saleem Awad; ALAZZAM, Malik Bader; ALGHAMDI, Ahmed S et al. Machine learning algorithms for prediction of survival curves in breast cancer patients. **Applied Bionics and Biomechanics**, Hindawi, v. 2021, 2021.

MCKENNA, Matthew T; WEIS, Jared A; BROCK, Amy; QUARANTA, Vito; YANKEELOV, Thomas E. Precision medicine with imprecise therapy: Computational modeling for chemotherapy in breast cancer. **Translational oncology**, v. 11, n. 1, p. 732–742, 2018.

MEIRELLES, Ricardo Henrique Sampaio. **Os avanços do controle do tabagismo no Brasil**. [S.l.]: SciELO Brasil, 2023. e33SP100 p.

MIN, Ningning; WEI, Yufan; ZHENG, Yiqiong; LI, Xiru. Advancement of prognostic models in breast cancer: a narrative review. **Gland Surgery**, AME Publications, v. 10, n. 9, p. 2815, 2021.

MITCHELL, Tom M; LEARNING, Machine. The mcgraw-hill companies. **Inc., New York**, 1997.

MONCADA-TORRES, Arturo; MAAREN, Marissa C van; HENDRIKS, Mathijs P; SIESLING, Sabine; GELEIJNSE, Gijs. Explainable machine learning can outperform cox regression predictions and provide insights in breast cancer survival. **Scientific Reports**, Nature Publishing Group, v. 11, n. 1, p. 1–13, 2021.

NAVE, OPhir. Adding features from the mathematical model of breast cancer to predict the tumour size. **International Journal of Computer Mathematics: Computer Systems Theory**, Taylor & Francis, v. 5, n. 3, p. 159–174, 2020.

NICOLÒ, Chiara; PÉRIER, Cynthia; PRAGUE, Melanie; BELLERA, Carine; MACGROGAN, Gaëtan; SAUT, Olivier; BENZEKRY, Sébastien. Machine learning and mechanistic modeling for prediction of metastatic relapse in early-stage breast cancer. **JCO clinical cancer informatics**, American Society of Clinical Oncology, v. 4, p. 259–274, 2020.

OKAGBUE, Hilary I; ADAMU, Patience I; OGUNTUNDE, Pelumi E; OBASI, Emmanuela CM; ODETUNMIBI, Oluwole A. Machine learning prediction of breast cancer survival using age, sex, length of stay, mode of diagnosis and location of cancer. **Health and Technology**, Springer, p. 1–7, 2021.

OPAS, Organização Pan-America de Saúde. **Cancer**. 2020. Acesso em: 2023-04-14. Disponível em: https://www.paho.org/pt/topicos/cancer.

OUZZANI, Mourad; HAMMADY, Hossam; FEDOROWICZ, Zbys; ELMAGARMID, Ahmed. Rayyan—a web and mobile app for systematic reviews. **Systematic Reviews**, v. 5, n. 1, p. 210, 2016. ISSN 2046-4053. Disponível em: http://dx.doi.org/10.1186/s13643-016-0384-4.

PAGE, Matthew J; MCKENZIE, Joanne E; BOSSUYT, Patrick M; BOUTRON, Isabelle; HOFFMANN, Tammy C; MULROW, Cynthia D; SHAMSEER, Larissa; TETZLAFF, Jennifer M; AKL, Elie A; BRENNAN, Sue E et al. A declaração prisma 2020: diretriz atualizada para relatar revisões sistemáticas. **Revista panamericana de salud publica**, SciELO Public Health, v. 46, p. e112, 2023.

PAIXÃO, Gabriela Miana de Mattos; SANTOS, Bruno Campos; ARAUJO, Rodrigo Martins de; RIBEIRO, Manoel Horta; MORAES, Jermana Lopes de; RIBEIRO, Antonio L. Machine learning na medicina: Revisão e aplicabilidade. **Arquivos Brasileiros de Cardiologia**, SciELO Brasil, v. 118, p. 95–102, 2022.

PANCH, Trishan; SZOLOVITS, Peter; ATUN, Rifat. Artificial intelligence, machine learning and health systems. **Journal of global health**, International Society for Global Health, v. 8, n. 2, 2018.

PAPA, Joao Paulo; FALCAO, Alexandre Xavier. Optimum-path forest: a novel and powerful framework for supervised graph-based pattern recognition techniques. **Institute of Computing University of Campinas**, v. 4148, 2010.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PENCINA, Michael J; D'AGOSTINO, Ralph B. Overall c as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. **Statistics in medicine**, Wiley Online Library, v. 23, n. 13, p. 2109–2123, 2004.

PINHEIRO, Talita Santos; YAHATA, Erika; SANTOS, Pablo Deoclecia dos; OLIVEIRA, Fellipe Soares de; TAKAHATA, André Kazuo; SUYAMA, Ricardo; TANAKA, Harki; OLIVEIRA, Tiago Ribeiro; ROMANI, Ana Paula; SIMOES, Priscyla Waleska. Machine learning e análise multivariada aplicados à sobrevida do câncer mama. **Journal of Health Informatics**, v. 14, 2022.

PÖLSTERL, Sebastian. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. **The Journal of Machine Learning Research**, JMLRORG, v. 21, n. 1, p. 8747–8752, 2020.

PÖLSTERL, Sebastian. **Scikit-Survival**. 2023. Acesso em: 2024-04-14. Disponível em: https://scikit-survival.readthedocs.io/en/stable/index.htmlr.

PÖLSTERL, Sebastian; NAVAB, Nassir; KATOUZIAN, Amin. Fast training of support vector machines for survival analysis. In: SPRINGER. **Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part II 15**. [S.l.], 2015. p. 243–259.

PÖLSTERL, Sebastian; NAVAB, Nassir; KATOUZIAN, Amin. An efficient training algorithm for kernel survival support vector machines. **arXiv preprint arXiv:1611.07054**, 2016.

PREZIOSI, Luigi. **Cancer modelling and simulation**. [S.l.]: CRC Press, 2003.

PRISMA. **PRISMA TRANSPARENT REPORTING of SYSTEMATIC REVIEWS and META-ANALYSES**. 2021. Acesso em: 2023-01-12. Disponível em: http://www.prisma-statement.org/.

RIBEIRO, Caroline Madalena; CORREA, Flávia de Miranda; MIGOWSKI, Arn. Efeitos de curto prazo da pandemia de covid-19 na realização de procedimentos de rastreamento, investigação diagnóstica e tratamento do câncer no brasil: estudo descritivo, 2019-2020. **Epidemiologia e Serviços de Saúde**, SciELO Brasil, v. 31, p. e2021405, 2022.

RIDGEWAY, Greg. The state of boosting. **Computing science and statistics**, Citeseer, p. 172–181, 1999.

ROSEN, Ryan D; SAPRA, Amit. Tnm classification. In: **StatPearls [Internet]**. [S.l.]: StatPearls Publishing, 2023.

RUBINGER, Luc; GAZENDAM, Aaron; EKHTIARI, Seper; BHANDARI, Mohit. Machine learning and artificial intelligence in research and healthcare. **Injury**, Elsevier, 2022.

SANTOS, Marceli de Oliveira; LIMA, Fernanda Cristina da Silva de; MARTINS, Luís Felipe Leite; OLIVEIRA, Julio Fernando Pinto; ALMEIDA, Liz Maria de; CANCELA, Marianna de Camargo. Estimativa de incidência de câncer no brasil, 2023-2025. **Revista Brasileira de Cancerologia**, v. 69, n. 1, 2023.

SAÚDE, Organização Mundial da. **CID-10: Classificação Estatística Internacional de Doenças com disquete Vol. 1**. [S.l.]: Edusp, 1994.

SAúDE, Ministério da Saúde/Secretaria de Atenção à. Portaria conjunta nº 5, de 18 de abril de 2019, aprova as diretrizes diagnósticas e terapêuticas do carcinoma de mama. **Diário Oficial [da] República Federativa do Brasil**, Brasília, DF, 2019. ISSN 1677-7042. Disponível em: http://pesquisa.in.gov.br/imprensa/jsp/visualiza/index.jsp?data=29/04/2019&jornal=515&pagina=44.

SCHAAR, Mihaela Van der; ALAA, Ahmed M; FLOTO, Andres; GIMSON, Alexander; SCHOLTES, Stefan; WOOD, Angela; MCKINNEY, Eoin; JARRETT, Daniel; LIO, Pietro; ERCOLE, Ari. How artificial intelligence and machine learning can help healthcare systems respond to covid-19. **Machine Learning**, Springer, v. 110, p. 1–14, 2021.

SEDIGHI-MAMAN, Zahra; MONDELLO, Alexa. A two-stage modeling approach for breast cancer survivability prediction. **International Journal of Medical Informatics**, Elsevier, v. 149, p. 104438, 2021.

SHUKLA, Nagesh; HAGENBUCHNER, Markus; WIN, Khin Than; YANG, Jack. Breast cancer data analysis for survivability studies and prediction. **Computer methods and programs in biomedicine**, Elsevier, v. 155, p. 199–208, 2018.

SIEGEL, Rebecca L; MILLER, Kimberly D; WAGLE, Nikita Sandeep; JEMAL, Ahmedin. Cancer statistics, 2023. **Ca Cancer J Clin**, v. 73, n. 1, p. 17–48, 2023.

SIMON, Noah; FRIEDMAN, Jerome; HASTIE, Trevor; TIBSHIRANI, Rob. Regularization paths for cox's proportional hazards model via coordinate descent. **Journal of statistical software**, NIH Public Access, v. 39, n. 5, p. 1, 2011.

SMOLA, Alex J; SCHÖLKOPF, Bernhard. A tutorial on support vector regression. **Statistics and computing**, Springer, v. 14, p. 199–222, 2004.

SOBIN, Leslie H; GOSPODAROWICZ, Mary K; WITTEKIND, Christian. **TNM classification of malignant tumours**. [S.l.]: John Wiley & Sons, 2011.

SOUZA, Tatiana Rocha de. **A influência da Noética na proliferação de células do adenocarcinoma de mama sob quimioterapia**. 2018. Dissertação (Tese, Programa de Pós-Graduação em Matemática Aplicada) — Instituto de Matemática, Estatística e Computação da Universidade Estadual de Campinas, SP, 2018.

SUNG, Hyuna; FERLAY, Jacques; SIEGEL, Rebecca L; LAVERSANNE, Mathieu; SOERJOMATARAM, Isabelle; JEMAL, Ahmedin; BRAY, Freddie. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. **CA: a cancer journal for clinicians**, Wiley Online Library, v. 71, n. 3, p. 209–249, 2021.

TAHMASSEBI, Amirhessam; WENGERT, Georg J; HELBICH, Thomas H; BAGO-HORVATH, Zsuzsanna; ALAEI, Sousan; BARTSCH, Rupert; DUBSKY, Peter; BALTZER, Pascal; CLAUSER, Paola; KAPETAS, Panagiotis et al. Impact of machine learning with multiparametric magnetic resonance imaging of the breast for early prediction of response to neoadjuvant chemotherapy and survival outcomes in breast cancer patients. **Investigative radiology**, NIH Public Access, v. 54, n. 2, p. 110, 2019.

TAPAK, Leili; SHIRMOHAMMADI-KHORRAM, Nasrin; AMINI, Payam; ALAFCHI, Behnaz; HAMIDI, Omid; POOROLAJAL, Jalal. Prediction of survival and metastasis in breast cancer patients using machine learning classifiers. **Clinical Epidemiology and Global Health**, Elsevier, v. 7, n. 3, p. 293–299, 2019.

TENG, Jing; ABDYGAMETOVA, Assem; DU, Jing; MA, Bian; ZHOU, Rong; SHYR, Yu; YE, Fei. Bayesian inference of lymph node ratio estimation and survival prognosis for breast cancer patients. **IEEE journal of biomedical and health informatics**, IEEE, v. 24, n. 2, p. 354–364, 2019.

TIZI, Wafaa; BERRADO, Abdelaziz. Machine learning for survival analysis in cancer research: A comparative study. **Scientific African**, Elsevier, v. 21, p. e01880, 2023.

TORRE, Lindsey A; ISLAMI, Farhad; SIEGEL, Rebecca L; WARD, Elizabeth M; JEMAL, Ahmedin. Global cancer in women: burden and trends. **Cancer epidemiology, biomarkers & prevention**, AACR, v. 26, n. 4, p. 444–457, 2017.

UNO, Hajime; CAI, Tianxi; PENCINA, Michael J; D'AGOSTINO, Ralph B; WEI, Lee-Jen. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. **Statistics in medicine**, Wiley Online Library, v. 30, n. 10, p. 1105–1117, 2011.

VELTEN, Kai. **Mathematical modeling and simulation: introduction for scientists and engineers**. [S.l.]: John Wiley & Sons, 2009.

VERÍSSIMO, André; OLIVEIRA, Arlindo Limede; SAGOT, Marie-France; VINGA, Susana. Degreecox–a network-based regularization method for survival analysis. **BMC bioinformatics**, Springer, v. 17, n. 16, p. 109–121, 2016.

VINAYAK, Rashmi Korlakai; GILAD-BACHRACH, Ran. Dart: Dropouts meet multiple additive regression trees. In: PMLR. **Artificial Intelligence and Statistics**. [S.l.], 2015. p. 489–497.

WEINBERG, Robert A; WEINBERG, Robert A. **The biology of cancer**. [S.l.]: WW Norton & Company, 2006.

WHO, World Health Organisation. **Cancer**. 2022. Acesso em: 2023-01-12. Disponível em: https://www.who.int/news-room/fact-sheets/detail/cancer.

WHO, World Health Organisation. **Breast Cancer**. 2024. Acesso em: 2024-06-06. Disponível em: https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

WODARZ, Dominik; KOMAROVA, Natalia. **Computacional Biology of cancer Lecture notes and mathematical modeling**. [S.l.]: World Scientific, 2005.

XIAO, Jialong; MO, Miao; WANG, Zezhou; ZHOU, Changming; SHEN, Jie; YUAN, Jing; HE, Yulian; ZHENG, Ying et al. The application and comparison of machine learning models for the prediction of breast cancer prognosis: Retrospective cohort study. **JMIR medical informatics**, JMIR Publications Inc., Toronto, Canada, v. 10, n. 2, p. e33440, 2022.

XIN, Ling; WU, Qian; ZHAN, Chongming; QIN, Hongyan; XIANG, Hongyu; XU, Ling; YE, Jingming; DUAN, Xuening; LIU, Yinhua; (CSBRS), Chinese Society of Breast Surgery et al. Multicenter study of the clinicopathological features and recurrence risk prediction model of early-stage breast cancer with low-positive human epidermal growth factor receptor 2 expression in china (chinese society of breast surgery 021). **Chinese Medical Journal**, Chinese Medical Journals Publishing House Co., Ltd. 42 Dongsi Xidajie ..., v. 135, n. 06, p. 697–706, 2022.

ZHOU, Cheng-Mao; XUE, Qiong; WANG, Ying; TONG, Jianhuaa; JI, Muhuo; YANG, Jian-Jun. Machine learning to predict the cancer-specific mortality of patients with primary non-metastatic invasive breast cancer. **Surgery Today**, Springer, v. 51, p. 756–763, 2021.

ANEXO A – Parecer Nº 5.533.296 Aprovado pelo Comitê de Ètica da UFJF

UFJF - UNIVERSIDADE FEDERAL DE JUIZ DE FORA - MG

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: Estudo comparativo de modelos computacionais na previsão da sobrevida de pacientes submetidas ao tratamento de câncer de mama

Pesquisador: Daniela Schimitz de Carvalho

Área Temática:

Versão: 3

CAAE: 58758822.6.0000.5147

Instituição Proponente: Instituto de Ciências Exatas

Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 5.533.296

Apresentação do Projeto:

As informações elencadas nos campos “Apresentação do Projeto”, “Objetivo da Pesquisa” e “Avaliação dos Riscos e Benefícios” foram retiradas do arquivo Informações Básicas da Pesquisa."Resumo:O câncer de mama femino é um dos principais problemas no âmbito da saúde pública mundial, sendo o mais incidente entre os cânceres, mais prevalente entre as mulheres, com elevadas taxas de mortalidade, quando associados ao diagnóstico e tratamento tardio, e estágio mais avançado da doença, que podem levar a forma mais agressiva, a metástase sistêmica que corresponde a primeira causa de óbito por câncer na população

feminina. Sabendo-se da carência de trabalhos multidisciplinares integrando a expertise das áreas da saúde e ciência de dados, que vem promovendo inovações e resultando no esclarecimento e direcionamento de tomada de decisão para vários problemas clínicos. Sendo assim, esta pesquisa tem por objetivo aplicar métodos de modelos computacionais a um conjunto de dados clínicos robustos provenientes de pesquisas já realizadas que não utilizaram esta metodologia, no intuito de avaliar o melhor modelo que represente a predição da sobrevida do câncer de mama".

Objetivo da Pesquisa:

Comunicação de conclusão da pesquisa."Objetivo Primário:Esta pesquisa tem por objetivo primário aplicar métodos de modelos computacionais a um conjunto de dados clínicos robusto de pacientes diagnosticadas com câncer de mama proveniente de pesquisas já realizadas que não utilizaram

Endereço: JOSE LOURENCO KELMER S/N
Bairro: SAO PEDRO **CEP:** 36.036-900
UF: MG **Município:** JUIZ DE FORA
Telefone: (32)2102-3788 **E-mail:** cep.propp@ufjf.br

esta metodologia, com a finalidade de avaliar o modelo que melhor representa a predição da sobrevida do câncer de mama".

"Objetivo Secundário:Para se alcançar os objetivos primários, os objetivos secundários são descritos abaixo: - tratamento dos dados para garantir a eficácia dos modelos aplicados a este banco de dados clínicos; - testar os modelos com a base de dados clínicos tratados;

- avaliar os resultados obtidos a fim de validar os modelos computacionais; e - gerar produção científica".

Avaliação dos Riscos e Benefícios:

Não foi registrada a ocorrência de algum evento adverso ou intercorrência durante a implementação do estudo ou em decorrência do mesmo.

"Riscos: O potencial de danos desta pesquisa é mínimo, pois trata-se de estudo coorte, de um banco de dados clínico de trabalhos já concluídos, do Programa em Saúde Brasileira (CINTRA, 2012) e do Programa de Pós Graduação em Saúde Coletiva (FAYER, 2014), aprovadas pelo CEP da UFJF sob os pareceres de nº 293/3006 e nº 151/219 respectivamente. No qual, não serão feitas intervenções diretas sobre as pacientes, e os seus dados pessoais e sensíveis, que possibilitam a sua identificação individualizada, não serão selecionados. Exclusivamente os dados clínicos (destacados no ANEXO 1) constituirão o banco de dados da pesquisa e os mesmos serão mantidos sob sigilo e confidencialidade; apenas informações agregadas serão tornadas públicas pelas publicações da tese e dos artigos submetidos aos periódicos científicos, obedecendo os requisitos da Lei nº 13.709 e Lei nº 13.853, Lei Geral de Proteção de Dados Pessoais (BRASIL, 1988). O presente será composto pelas informações clínicas provenientes dos registros médicos que foram devidamente autorizados pela direção dos serviços de saúde no estudo e serão totalmente confidenciais para todos os casos, sendo usados estritamente para fins científicos e submetido ao Comitê de Ética em Pesquisa Humana da Universidade Federal de Juiz de Fora".

"Benefícios: Apresentação de modelos computacionais mais eficientes para predição de sobrevida do câncer de mama invasivo contribuindo na orientação da tomada de decisão clínico-terapêutica e para o planejamento de estratégias de tratamento anti-câncer".

Comentários e Considerações sobre a Pesquisa:

O projeto está bem estruturado, delineado e fundamentado, sustenta os objetivos do estudo em sua metodologia de forma clara e objetiva, e se apresenta em consonância com os princípios éticos norteadores da ética na pesquisa científica envolvendo seres humanos elencados na resolução 466/12 do CNS e com a Norma Operacional Nº 001/2013 CNS.

Considerações sobre os Termos de apresentação obrigatória:

O protocolo de pesquisa está em configuração adequada, apresenta FOLHA DE ROSTO devidamente preenchida,com o título em português, identifica o patrocinador pela pesquisa, estando de acordo com as atribuições definidas na Norma Operacional CNS 001 de 2013 item 3.3 letra a; e 3.4.1 item 16. Apresenta o TERMO DE DISPENSA DO TCLE de acordo com a Resolução CNS 466 de 2012, item: IV.8. O Pesquisador apresenta titulação e experiência compatível com o projeto de pesquisa, estando de acordo com as atribuições definidas no Manual Operacional para CPEs. Apresenta DECLARAÇÃO de infraestrutura e de concordância com a realização da pesquisa de acordo com as atribuições definidas na Norma Operacional CNS 001 de 2013 item 3.3 letra h.

Conclusões ou Pendências e Lista de Inadequações:

Diante do exposto, o projeto está aprovado, pois está de acordo com os princípios éticos norteadores da ética em pesquisa estabelecido na Res. 466/12 CNS e com a Norma Operacional Nº 001/2013 CNS. Data prevista para o término da pesquisa: Dezembro de 2025.

Considerações Finais a critério do CEP:

Diante do exposto, o Comitê de Ética em Pesquisa CEP/UFJF, de acordo com as atribuições definidas na Res. CNS 466/12 e com a Norma Operacional Nº001/2013 CNS, manifesta-se pela APROVAÇÃO do protocolo de pesquisa proposto. Vale lembrar ao pesquisador responsável pelo projeto, o compromisso de envio ao CEP de relatórios parciais e/ou total de sua pesquisa informando o andamento da mesma, comunicando também eventos adversos e eventuais modificações no protocolo.

Este parecer foi elaborado baseado nos documentos abaixo relacionados:

Tipo Documento	Arquivo	Postagem	Autor	Situação
Informações Básicas do Projeto	PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_1947074.pdf	18/07/2022 11:54:01		Aceito
Projeto Detalhado / Brochura Investigador	Projeto_Daniela.pdf	18/07/2022 11:53:24	Daniela Schimitz de Carvalho	Aceito
TCLE / Termos de Assentimento / Justificativa de Ausência	DispensaTCLE1.pdf	06/06/2022 18:48:20	Daniela Schimitz de Carvalho	Aceito

Endereço: JOSE LOURENCO KELMER S/N
Bairro: SAO PEDRO **CEP:** 36.036-900
UF: MG **Município:** JUIZ DE FORA
Telefone: (32)2102-3788 **E-mail:** cep.propp@ufjf.br

Declaração de Instituição e Infraestrutura	DeclaracaoInfraestrutura.pdf	06/06/2022 18:46:04	Daniela Schimitz de Carvalho	Aceito
Outros	Anexo_1.pdf	16/05/2022 13:21:41	Daniela Schimitz de Carvalho	Aceito
Parecer Anterior	ProjCaMama_1.pdf	16/05/2022 13:19:40	Daniela Schimitz de Carvalho	Aceito
Parecer Anterior	Parecer_29330061.pdf	16/05/2022 13:18:32	Daniela Schimitz de Carvalho	Aceito
Folha de Rosto	folhaderosto_assinada.pdf	16/05/2022 09:00:22	Daniela Schimitz de Carvalho	Aceito

Situação do Parecer:
Aprovado

Necessita Apreciação da CONEP:
Não

JUIZ DE FORA, 18 de Julho de 2022

Assinado por:
Jubel Barreto
(Coordenador(a))

ANEXO B – Artigo I: Aplicação do *Random Survival Forest* na análise da sobrevida para o câncer de mama

CBIS'22
XIX Congresso Brasileiro de Informática em Saúde
29/11 a 02/12 de 2022 - Campinas/SP - Brasil

Aplicação do *Random Survival Forest* na análise da sobrevida para câncer da mama

Application of Random Survival Forest in breast cancer survival analysis

Aplicación del *Random Survival Forest* en el análisis de la supervivencia del cáncer de mama

Daniela Schimitz de Carvalho[1],Thallys da Silva Nogueira[1], Priscila Vanessa Zabala Capriles Goliatt [1]

1 Programa de Pós-Graduação em Modelagem Computacional, Universidade Federal de Juiz de Fora – UFJF, Juiz de Fora (MG), Brasil.

Autor correspondente: Daniela Schimitz de Carvalho
E-mail: daniela.schimitz@estudante.ufjf.br

Resumo

Este trabalho tem por objetivo aplicar um método de aprendizado de máquina supervisionado a um conjunto de dados clínicos da Zona da Mata Mineira, para se avaliar o desempenho da precisão da predição de sobrevida para câncer de mama. O banco de dados utilizado passou por pré-processamento fornecendo as variáveis a serem empregadas no *Random Survival Forest*. Os resultados apresentam as métricas de desempenho satisfatória para métodos de predição da sobrevida. Sendo concluído, que os métodos de aprendizagem de máquina são promissores na assistência e orientação na prática clínica.

Descritores: Câncer de Mama; Aprendizado de Máquina; Análise de Sobrevida

Abstract

This paper aims to apply a supervised machine learning method to a clinical dataset from Zona da Mata Mineira, to evaluate the performance of survival prediction accuracy for breast cancer. The database utilized went through pre-processing providing the variables used in the Random Survival Forest. The results show satisfactory

performance metrics for survival prediction methods. Concluding that, the machine learning methods are promising assisting and guiding clinical practice.
Keywords: Breast Neoplasms; Machine Learning; Survival Analysis

Resumen

Este trabajo tiene como objetivo aplicar un método de aprendizaje automático supervisado a un conjunto de datos clínicos de la Zona da Mata Mineira, para evaluar el rendimiento de la precisión de la predicción de la supervivencia para el cáncer de mama. La base de datos utilizada pasó por un preprocesamiento que proporcionó las variables que se emplearían en el *Random Survival Forest*. Los resultados presentan métricas de rendimiento satisfactorias para los métodos de predicción de la supervivencia. Concluyendo que los métodos de aprendizaje automático son prometedores en la asistencia y orientación en la práctica clínica.
Descriptores: Neoplasias de mama; Machine Learning; Análisis de Supervivencia

Introdução

O câncer de mama (CM) é um dos principais problemas no âmbito da saúde pública mundial, sendo o mais diagnosticado entre todos os cânceres em ambos os sexos.[1,2] Segundo, estimativas de 2020 da Organização Mundial de Saúde (OMS) esta patologia é mais prevalente entre as mulheres, com 2,26 milhões de novos casos e 685 mil óbitos.[2] Já no Brasil, 66 mil novos casos e 18 mil óbitos, segundo estimativas do Instituto Nacional do Câncer (INCA) de 2020.[3]

No que tange a saúde da mulher, o CM continua com elevadas taxas de incidência, prevalência e mortalidade. Estas altas taxas são decorrentes dos fatores de risco, fatores prognósticos e diagnósticos tardios, que podem evoluir para a forma mais agressiva da doença, a metástase sistêmica e consequentemente a morte.[1,4,6]

Entre os fatores de risco destaca-se os relacionados à vida reprodutiva da mulher: menopausa tardia, menarca precoce, não ter filhos, idade avançada no nascimento do primeiro filho e menos filhos, uso de anticoncepcionais orais e reposição hormonal.[5,7,8] E os relacionados à contemporaneidade como: estilo de vida,

envelhecimento, sedentarismo, obesidade, tabagismo e etilismo. Por fim, enfatiza-se a relevância dos fatores protetores como amamentação e atividade física.[4,7]

Já entre os fatores prognósticos, vale ressaltar os que interferem na sobrevida. O termo sobrevida indica o tempo específico em que pacientes sobreviveram a uma determinada patologia após o diagnóstico.[5,10] No caso do CM alguns fatores como: idade ao diagnóstico, tamanho do tumor, comprometimento linfonodal, estadiamento, tipo e grau histológico, marcadores moleculares e imuno-histoquímicos; influenciam a sobrevida. Na prática clínica, também se avalia os status dos receptores de estrogênio, progesterona e fator de crescimento humano epidérmico receptor-2 (HER2) para a subclassificação do CM e indicações clínico-terapêuticas específicas.[5,8,9]

Por fim, o diagnóstico precoce está diretamente relacionado ao bom prognóstico e consequentemente ao início do tratamento, evitando-se a evolução do CM para estágios mais avançados e assim melhorar a sobrevida e diminuir o sofrimento do paciente.[4,11,12] Este diagnóstico é efetuado pelo rastreamento mamográfico e exame físico (presença de nódulo), logo após, são avaliados a história pregressa, os fatores de risco e prognósticos, os resultados dos marcadores tumorais e estadiamento para assim se estabelecer as estratégias de tratamento.[5,8,13]

Diante disto, o conhecimento destes fatores são de extrema importância para o planejamento terapêutico e avaliação do curso clínico-terapêutico.[7,11,13] Atualmente, este processo é realizado de forma holística, subsequente a ponderação dos fatores significativos envolvidos no processo evolutivo desta doença.[7,8,14]

Neste sentido, os modelos computacionais fornecem informações preditivas importantes nas tomadas de decisões clínico-terapêuticas, além de preencherem lacunas da prática médica guiadas atualmente por variáveis clínicas observáveis.[9,11,14] Visto que, vive-se o desdobramento do emprego da inteligência computacional na prestação de entendimento e elucidação de problemas nas diversas áreas.[10,15]

Nesta perspectiva, os métodos de aprendizado de máquina (MAM) vêm ganhando espaço e aplicabilidade na área oncológica, fornecendo informações significativas, como, por exemplo, para o diagnóstico, manejo e prognóstico dos pacientes.[11,12,16] Consequentemente, com o avanço da ciência de dados e

inovações das tecnologias surgem novos modelos prognósticos para CM, como ferramentas iminentes de alto potencial e aplicabilidade na prática médica.[9,12,16,17]

Como, por exemplo, o MAM supervisionado *Random Survival Forest* (RSF)[18], amplamente aplicado na predição de sobrevida.[9,12,16,17] O RSF se fundamenta nos princípios do método original, *Random Forest* (RF)[19], em que se cultiva árvores usando dados de *bootstrap*, os nós se dividem usando uma seleção aleatória de recursos, e geram uma predição constituída pela média dos preditores de cada árvore.[17,18,20]

Recentemente, vários métodos de predição prognóstica foram propostos, porém, apresentam como limitação a não validação com dados locais, resultando em uma capacidade preditiva abaixo do ideal.[9,12,16,18] Deste modo, esta pesquisa tem por objetivo aplicar o RSF a um banco de dados clínicos de pacientes diagnosticadas com CM, tratadas e acompanhadas em centros de referência oncológica da Zona da Mata Mineira, a fim de avaliar o desempenho da predição de sobrevida para o CM.

Material e Métodos

Neste trabalho, aplicou-se um modelo computacional a um banco de dados clínico de pacientes diagnosticadas com CM. O MAM supervisionado selecionado RSF, implementado na linguagem Python por meio da biblioteca Scikit-survival[21]. Esta biblioteca, permite a realização de análises de sobrevida através da correlação entre as covariáveis e o tempo de evento, e também métricas de desempenho específicas.[21,22,23]

O tratamento e formatação do banco de dados foram realizados nas etapas de pré-processamento e análise descritiva dos dados. Após estas etapas, os dados resultantes foram utilizados na implementação do método para a predição da sobrevida para CM, e validação do RSF pelas métricas de desempenho.[22,24,25]

Banco de dados

Os dados são provenientes de uma coorte de base hospitalar, coletados manualmente nos prontuários dos registros hospitalares dos centros de referência de oncologia da região da Zona da Mata Mineira. O banco de dados refere-se a uma

população de mulheres diagnosticadas com CM entre janeiro de 2003 e janeiro de 2005, submetidas à terapêutica local e/ou sistêmica. Vale enfatizar que estes dados são oriundos de pesquisas já concluídas[5], que não utilizaram desta metodologia. Após a submissão e aprovação pelo Comitê de Ética na pesquisa com seres humanos da Universidade Federal de Juiz de Fora, CAAE:58758822.6.0000.51.47.

A coleta dos dados constituída pelo: recrutamento (2009 até 2010) nos arquivos de registro de câncer de base hospitalar; seguida da avaliação das características clínicas, sociodemográficas e exclusão das pacientes que realizaram apenas um único procedimento nas instituições; e buscas (2011) por ligações telefônicas, consultas de CPF válidos e com mastologistas, e no sistema de informações de mortalidade.[5]

Pré-processamento dos dados

Nos dados brutos do banco original, em consequência da forma de coleta (manual) e do número de dados faltantes, foram realizados dois pré-processamentos. O primeiro pré-processamento, através de tratamento dos dados faltantes e padronização da formatação dos dados, como segue discriminado abaixo:

1. **Strings**: corrigido os erros de digitação; retirando os acentos e caracteres; formatando em minúsculas; inserindo a palavra 'ignorado' nos dados ausentes;
2. **Valores numéricos:** formatando os números inteiros e contínuos; para os dados faltantes de variáveis categorizadas e de contagem atribuindo, respectivamente, os valores 9 e -1;
3. **Datas**: verificando os valores ausentes da data de seguimento e laudo histopatológico (critério de exclusão), e formatadas as datas (ano, mês e dia).

O segundo pré-processamento, estabelecendo um novo tratamento, resultando na construção de 3 bancos de dados:

a) **Banco I:** constituído pelos dados do primeiro pré-processamento, com imputação dos termos: 'ignorado', 9 e -1 para os dados faltantes;
b) **Banco II**: os dados não coletados foram convertidos para a média ou moda das variáveis após análise das mesmas; e
c) **Banco III**: as linhas com os dados: 'ignorado', 9 e -1 foram excluídas, aplicando apenas ao modelo os valores coletados das variáveis.

MAM para predição de sobrevida

Os MAM para análise de sobrevida tem por objetivo estabelecer a relação entre as variáveis e o momento de ocorrência do evento, e assim, aprender com estes dados para fornecer a previsão do evento de interesse. A relação entre o tempo de sobrevivência e a ocorrência do evento se dá pelas funções de sobrevida (retorna a probabilidade de tempo de sobrevida) e de risco (retorna a probabilidade de ocorrência do evento). Onde, a previsão da sobrevida é decorrente da avaliação da medida de correlação entre o risco previsto e observado no teste.[21,22,23]

O RSF fornece a previsão do paciente sobreviver após um determinado tempo de seguimento, através das funções de sobrevida Kaplan-Meier e de risco Nelson-Aalen. No qual, o cultivo das árvores incorporado às informações de censura é realizado pela regra de divisão aleatória long-rank, que visa dividir os nós com sobrevidas diferentes, e assim maximizar a diferença de sobrevida entre nós.[9,16,18,20]

Portanto, o RSF tem por objetivo prever a sobrevida de cada paciente determinando o quão bem o modelo generaliza a predição do tempo de sobrevida. Para se alcançar esta finalidade, o método não aplica apenas dois conjuntos de valores de entrada e saída; mas sim três conjuntos: um constituído pelas variáveis independentes (prognósticas), os outros formados por um vetor com o indicador do evento (censurado ou não censurado) e dos tempos de sobrevida (dias).[21,22]

Atributos e hiperparâmetros do RSF

Na construção do RSF, as variáveis prognósticas selecionadas foram: idade, tamanho do tumor, status menopausal, receptores estrogênio e progesterona, hormonioterapia, linfonodos positivos e grau do tumor, como proposto pela biblioteca; e inclusão da variável HER2 justificada pela construção do banco de dados.[5,21,22] Já as variáveis de valor: status (0-censurado ou 1-evento) e tempo de sobrevida (dias).[21,22]

Como o banco de dados original tem por finalidade analisar as taxas de sobrevida de 5 anos e os principais fatores associados ao perfil imuno-histoquímico.[5] E também, como na prática clínica os resultados dos receptores estrogênio, progesterona e HER2 resultam na subclassificação do CM e direcionam a condução clínico-terapêutica.[5,8,9] Assim, optou-se pela inserção da variável prognóstica HER2.

Já os hiperparâmetros utilizados no RSF, foram, respectivamente, o número de árvores na floresta de 1000, o número mínimo de amostras necessárias para dividir um nó e em um único nó, respectivamente, 10 e 15. Os dados foram divididos 75% para treinamento e 25% para testes, onde os critérios de divisão aplicados para a construção de cada árvore se baseiam no teste de log-rank.[21]

Métrica de Desempenho

As medidas de desempenho aplicadas aos MAM para análise de sobrevida tem a funcionalidade de avaliar o quão bem o método prever os diferentes tempos de sobrevida. Porém, como estes tempos estão sujeitos à censura, não se utilizam métricas usuais como: erro quadrático médio ou correlação.[21,22,23] Logo, para se avaliar o RSF foram selecionadas as métricas: índice de concordância (C-index)[24] e *Brier Score*[25], mais robustas e específicas para captar o comportamento analisado.[12,21]

O C-index quantifica a capacidade em discriminar os escores de risco dos diferentes tempos previstos no método pelos observados nos dados (teste), por intermédio de uma medida de concordância entre os pares concordantes pelos não descartados. Os pares descartados tem como característica: o menor tempo do evento (morte); pares com o mesmo tempo, a menos que o evento ocorra em um ou em ambos; e por fim, a não ocorrência do evento em ambos os elementos do par.[9,21,24]

Em resumo, o C-index quantifica o poder do método em predizer e classificar os tempos de morte dos pacientes. O intervalo do desempenho do método, assume valores de 0,0 até 1,0, com a seguinte interpretação: 0,5 para desempenho médio, sem discriminação preditiva; e 1,0 para desempenho perfeito, referente a um modelo capaz de separar os pacientes com diferentes desfechos. [9,17,23]

O *Brier Score* corresponde a uma medida similar ao erro quadrático médio, através de testes entre a precisão das probabilidades previstas nas funções de sobrevida, com o status observado nos dados para os momentos (T) selecionados do teste, ou seja, a calibração do modelo. Desta forma, avalia a qualidade de predição do método através da probabilidade do paciente permanecer livre do evento (morte), em

que valores mais baixos sugerem melhores resultados, representando de forma satisfatória a previsão individual do método para cada paciente.[21,25]

Resultados e Discussão

Os resultados apresentados são decorrentes da aplicação do RSF a um banco de dados clínico composto por uma população de 563 mulheres diagnosticadas com CM, limitando-se o tempo do seguimento por 5 anos (1825 dias), em consequência do banco original ter como proposta a análise de 5 anos, e assim, identificar a capacidade de sobrevivência neste período. A contagem de tempo de sobrevida de 5 anos se inicia com a data do laudo histopatológico (diagnóstico) e se finaliza com a data do evento adverso (óbito por CM) ou da censura (último dia de acompanhamento, ou limitado ao tempo proposto de análise). Portanto, após o pré-processamento apenas uma paciente foi excluída por não possuir a data do laudo histopatológico, restando 562 pacientes.[5,10]

Análise descritiva dos dados

O banco de dados original coletado de forma manual (prontuários), justificando assim os tratamentos e descrições detalhadas dos dados.[5] Deste modo, o Quadro I especifica as características das variáveis utilizadas pelo RSF.[5,18]

Quadro 1 – Descrição das variáveis clínicas aplicadas ao RSF [5]

Variáveis clínicas do banco de dados	Características das variáveis
Idade	Idade da paciente na primeira consulta oncológica (suspeita de diagnóstico)
Tamanho do tumor	Medida do tamanho do tumor (em cm)
Linfonodos comprometidos	Número de linfonodos comprometidos (após a cirurgia)
Receptor de Estrogênio	Resultado do exame imuno-histoquímico (em cruzes)
Receptor de Progesterona	Resultado do exame imuno-histoquímico (em cruzes)
Receptor Her2	Resultado do exame imuno-histoquímico (em cruzes)

Tempo de sobrevida 5 anos	O tempo de sobrevivência dos pacientes em dias (limitado a 1825 dias)
Status menopausal	Estado menopausal (pós-menopausa-1 ou pré-menopausa-0)
Hormonioterapia	Administrada a hormonioterapia (sim-1 ou não-0)
Grau histopatológico	Grau histopatológico tumoral (bem diferenciado-1 ou moderadamente diferenciado-2 ou pouco diferenciado-3)
Seguimento	Status do seguimento, vivas ou óbito por CM (censurado-0 ou não censurado-1)

A Tabela 1 mostra a análise descritiva dos dados de cada banco (I, II e III), já definidas anteriormente no Quadro 1; sumarizando e explorando o comportamento de cada variável.

Tabela 1 – Análise descritiva das variáveis utilizadas no RSF.

Variáveis	**Banco I**	**Banco II**	**Banco III**
Idade * (mínimo - máximo)	58.14 ± 13.67 (26 - 91)	58.14 ± 13.67 (26 - 91)	56.57 ± 12.90 (26 - 91)
Tamanho do tumor * (mínimo - máximo)	2.99 ± 2.91 (-1.0 - 25.0)	3.22 ± 2.75 (0.0 - 25.0)	3.20 ± 2.51 (0.4 - 15.0)
Linfonodos comprometidos* (mínimo - máximo)	2.39 ± 4.74 (-1 - 37)	2.53 ± 4.69 (0 - 37)	2.90 ± 4.95 (0 - 37)
Receptor Estrogênio * (mínimo - máximo)	2.89 ± 2.52 (0 - 9)	2.35 ± 2.06 (0 - 5)	2.38 ± 2.02 (0 - 5)
Receptor Progesterona * (mínimo - máximo)	2.59 ± 2.47 (0 - 9)	2.05 ± 1.93 (0 - 5)	2.16 ± 1.89 (0 - 5)
Receptor Her2 * (mínimo - máximo)	1.82 ± 3.02 (0 - 9)	0.83 ± 1.38 (0 - 5)	0.76 ± 1.41 (0 - 5)
Tempo de sobrevida* (mínimo - máximo)	1557.72 ± 505.60 (9 -1825)	1557.72 ± 505.60 (9 -1825)	1621.85 ± 429.74 (149 -1825)
Status menopausal **			
Pós-menopausa	375	375	215
Pré-menopausa	187	187	123
Hormonioterapia **			
Uso	157	157	86
Não uso	405	405	252
Grau histopatológico **			
1 (bem diferenciado)	127	127	110
2 (moderadamente)	184	346	160

3 (pouco diferenciado)	89	89	68
Seguimento **			
Censurado (vivas)	433	433	269
Não censurado (óbito)	129	129	69
Total **			
(população de pacientes)	562	562	338

* Média e desvio padrão
** Frequência

As análises descritivas das variáveis são apresentadas na Tabela 1, contendo nas variáveis numéricas: as médias, desvios padrões, valores mínimos e máximos; e nas categóricas as frequências numéricas de cada categoria. Desta forma, nota-se como as diferenças formas de pré-processamento dos dados em cada banco (I, II e III) influenciaram respectivamente sua descrição. Para exemplificar, observa-se: no Banco I e II o mesmo valor amostral e de frequência próximos, mas valores de média, desvio padrão, mínimos e máximos bem diferentes (exceto na idade e tempo de sobrevida 5 anos que foram coletados todos os valores); já entre os bancos II e III com valores mínimos e máximos iguais, exceto para variável tempo de sobrevida 5 anos, mas com valores amostrais, de frequências, média e desvio padrão diferentes.

Análise do desempenho do RSF

O desempenho do RSF apresentado por meio da análise das métricas C-index e *Brier Score*, usualmente aplicadas a métodos de análise de sobrevida, como demonstrados na Tabela 2.[12,21,22] Onde, destaca-se a realização de dois testes: teste 1 sem a inclusão da variável HER2 e teste 2 com a inclusão da variável HER2; aplicados em cada banco I, II e III, caracterizados e analisados descritivamente, respectivamente no Quadro I e Tabela I.

Tabela 2– Representação do desempenho do RSF.

Medidas de desempenho	**Banco I [a]**	**Banco II [b]**	**Banco III [c]**
C-index (sem HER2)	0,784295	0,790064	0,746695
C-index (com HER2)	0,792308	**0,801603**	0,739736

Brier Score (sem HER2)	0,125033	0,120964	0,156970
Brier Score (com HER2)	0,123146	0,118892	0,156214

[a] População de 562 pacientes diagnosticadas com CM.
[b] População de 562 pacientes diagnosticadas com CM.
[c] População de 338 pacientes diagnosticadas com CM.

De maneira geral, as medidas de desempenho do C-index visam verificar a precisão da previsão da sobrevida.[13] Os resultados obtidos nos bancos I, II e III são superiores aos alcançados de 0,68 pelo método proposto RSF por Ishwaran [17]. Agora, ao se comparar o valor de desempenho do C-index entre os testes, ou seja, a inclusão ou não da variável HER2 e entre os bancos I, II e III, observa-se:

- Nos bancos I e II, a inclusão da variável HER2 melhora-se sutilmente a eficácia de predição do RSF, com um aumento da precisão (C_{dist} = C_{indexH} - C_{index}): no banco I, em *0,792308 – 0,784295 = 0,008013*; e no banco II, em *0,801603 – 0,790064 = 0,011539*, esta diferença pode ser atribuída ao pré-processamento.
- No banco III se inverte esta classificação, onde a inclusão obteve uma pequena diminuição da precisão, em *0,739736 - 0,746695 = -0,006959*, não melhorando o desempenho do RSF, que pode ter sido influenciada pela redução da quantidade amostral.

Já as medidas de desempenho do *Brier Score* apontam para uma boa calibração e predição do RSF, com métrica inferior a 0,25 em todos os bancos e testes.[12,25] Os resultados desta métrica no banco III são um pouco maiores que dos outros bancos, e oposto entre os testes (com ou sem HER2), mas ainda na faixa de efetividade do desempenho do RSF que pode ter sido influenciada pelo número amostral reduzido.

As medidas de desempenho do C-index obtidas em todos os testes e bancos deste trabalho, são superiores quando comparados aos resultados de outros trabalhos da literatura que também aplicaram o RSF a um banco de dados de mulheres diagnosticadas com CM.[9,16,17] Porém, no estudo de Aivaliotis et al.[16] o valor de C-index foi de aproximadamente 0,57, entretanto o objetivo não era associar a sobrevida, mas sim a incidência do CM ao Índice de Massa Corporal, comparando o clássico modelo *Cox Proportional Hazards* (CPH) ao RSF.

O estudo de Moncada-Torres et al.[9], usou dados de pacientes com CM não metastático para comparar o desempenho do RSF a outros métodos como CPH, *Survival Support Vector Machines* (SSVM) e *Extreme Gradient Boosting* (XGB), demonstrando que os métodos RSF e SSVM apresentam um C-index de 0,63 enquanto o XGB 0,73, valores estes menores que os obtidos neste trabalho (Tabela 2).

O trabalho de Pinheiro et al.[17] também verifica os resultados dos modelos CPH e RSF aplicados a um conjunto de dados de CM, resultando em um C-index de 0,68, menor que os alcançados neste trabalho (Tabela 2). Vale destacar, que os dados das pacientes do presente trabalho constitui-se: todos os tipos histológicos de CM e também as metastáticas.

Já os resultados do trabalho Xiao et al.[12] também comparando-se os modelos CPH, SSVM e RSF, verificando-se que o RSF superou outros modelos em capacidade discriminativa com um C-index de 0,85 e um Brier Score de 0,045, resultados estes melhores que demonstrados na Tabela 2. Vale enfatizar que o estudo[12], constituído por uma população de mais 22 mil pacientes e 11 atributos de dados prognósticos aplicados ao RSF, reflete a importância de bancos robustos de alta qualidade para serem aplicados aos MAM.[9,10,13] Representando uma das limitações encontradas na aplicação destes métodos, cujo objetivo é obter uma previsão satisfatória do tempo de sobrevida para CM.[10,11,12,13]

Conclusão

Os MAM tornaram-se uma metodologia inovadora para previsão da sobrevida, mesmo que existam ainda melhorias e potencial para se agregar modelos computacionais adicionais. Os resultados provenientes deste trabalho, mostraram que o RSF possibilita análises de sobrevida promissoras para o CM, principalmente quando validadas por um banco de dados clínico robusto.

Afinal, nota-se que as pesquisas multidisciplinares, com construção de bancos de dados robustos e de alta qualidade, baseadas em descobertas anteriores e orientadas pelos especialistas da área médica, podem resultar em métodos eficazes para serem aplicados na prática clínica.

Agradecimentos

Agradecimentos à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro por bolsa concedida; ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora pelo aprendizado; ao professor Dr. Maximiliano Ribeiro Guerra e à mastologista Dra. Jane Rocha Duarte Cintra pela colaboração e incentivo disponibilizando o banco de dados clínicos; e aos Hospitais 9 de Julho e Instituto Oncológico apoio nesta pesquisa.

Referências

1 Ferlay J et al. Cancer statistics for the year 2020: An overview. International Journal of Cancer. 2021;149(4),p.778-89.

2 World Health Organization. Cancer [Internet]; c2022 [cited 2022 Set 12]. Available from: https://www.who.int/news-room/fact-sheets/detail/cancer

3 Instituto Nacional de Câncer. Estatísticas de câncer [Internet]; c2022 [cited 2022 Set 10]. Available from: https://www.gov.br/inca/pt-br/assuntos/cancer/numeros/

4 Carvalho DS, Guerra MR, Barra LP, Queiroz RA. Aspectos gerais epidemiológicos da mortalidade por câncer de mama feminino no brasil e no mundo. Anais Simpósio de Enfermagem [Internet]. 2019 [cited 2022 Out 27];3:[about 1 p.]. Available from: http://pensaracademico.facig.edu.br/index.php/simposioenfermagem/article/view/1116

5 Cintra JR. Sobrevida e fatores associados em pacientes com câncer de mama, com diagnóstico entre 2003 e 2005 no município de Juiz de Fora – MG. [dissertation]. Juiz de Fora (JF): Universidade Federal de Juiz de Fora, 2012.

6 Torre LA, Islami F, Siegel RL, Ward EM, Jemal A. Global Cancer in Women: Burden and Trends. Cancer Epidemiol Biomarkers Prev. 2017;26(4):444-457.

7 Carvalho DS, Guerra MR, Barra LP, Queiroz RA. Modelagem computacional do crescimento tumoral mamário. Anais Seminário Científico UNIFACIG [Internet]. 2017 [cited 2022 Out 27];3:[about 1 p.]. Available from: http://pensaracademico.facig.edu.br/index.php/semiariocientifico/article/view/438

8 Ministério da Saúde (BR). Secretaria de Atenção à Saúde. Protocolos clínicos e diretrizes terapêuticas em Oncologia/Ministério da Saúde, Secretaria de Atenção à Saúde – Brasília : Ministério da Saúde, 2014.

9 Moncada-Torres A et al. Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. Scientific Reports. 2021;11(1):p.1-13.

10 Li J et al. Predicting breast cancer 5-year survival using machine learning: a systematic review. PloS one [Internet]. 2021 [cited 2022 Out 27];16(4):[about 1 p.]. Available from: https://doi.org/10.1371/journal.pone.0250370

11 Tapak L et al. Prediction of survival and metastasis in breast cancer patients using machine learning classifiers. Clinical Epidemiology and Global Health. 2019;7(3):p.293-9.

12 Xiao J, Mo M, Wang Z, Zhou C, Shen J, Yuan J, He Y, Zheng Y. The Application and Comparison of Machine Learning Models for the Prediction of Breast Cancer Prognosis: Retrospective Cohort Study. JMIR medical informatics [Internet]. 2022[cited 2022 Out 27]; 10(2):[about 1 p.]. Available from: https://medinform.jmir.org/2022/2/e33440

13 Hueman MT et al. Creating prognostic systems for cancer patients: A demonstration using breast cancer. Cancer medicine. 2018;7(8):p.3611-21.

14 Lai X et al. Toward Personalized Computer Simulation of Breast Cancer Treatment: A Multiscale Pharmacokinetic and Pharmacodynamic Model Informed by Multitype Patient Data. Cancer research. 2019;79(16):p.4293-304.

15 Nave O. Adding features from the mathematical model of breast cancer to predict the tumour size. International Journal of Computer Mathematics: Computer Systems Theory. 2020;5(3):p.159-174.

16 Aivaliotis G et al. A comparison of time to event analysis methods, using weight status and breast cancer as a case study. Scientific reports. 2021;11(1):p. 1-9.

17 Pinheiro TS et al. Machine Learning e Análise Multivariada aplicados à Sobrevida do Câncer Mama. J Health Inform [Internet]. 2022[cited 2022 Out 27];(14).[about 1 p.]. Available from: https://jhi.sbis.org.br/index.php/jhi-sbis/article/view/971

18 Ishwaran H et al. Random survival forests. The annals of applied statistics. 2008;2(3):p.841-60.

19 Breiman L. Random forests. Machine Learning, Springer Science and Business Media LLC. 2001;45(1):p.5–32.

20 Ishwaran H, Lu M. Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. Statistics in medicine. 2019;38(4):p.558-82.

21 Understanding Predictions in Survival Analysis. [Internet];[cited 2022 Set 12]. Available from: https://scikit-survival.readthedocs.io/en/stable/user_guide/

22 Pölsterl S. scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. J. Mach. Learn. Res. 2020;21(212):p.1-6.

23 Fast Unified Random Forests with random. [Internet];[cited 2022 Set 12]. Available from: https://www.randomforestsrc.org/articles/survival.html

24 Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On The C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statistics In medicine. 2011;30(10):1105-17.

25 Graf E, Schmoor C, Sauerbrei W, Schumacher M.Assessment and comparison of prognostic classification schemes for survival data. Statistics in medicine. 1999;18(17-18):2529-45.

ANEXO C – Artigo II: *Comparative Analysis of Machine Learning Models for Breast Cancer Patients' Survival Prediction*

Comparative Analysis of Machine Learning Models for Breast Cancer Patients' Survival Prediction

Daniela Schimitz de Carvalho[1], Priscila Capriles[2], and Leonardo Goliatt[3(✉)]

[1] Federal University of Juiz de Fora, Juiz de Fora, Brazil
daniela.schimitz@estudante.ufjf.br
[2] Department of Computer Science, Federal University of Juiz de Fora, Juiz de Fora, Brazil
capriles@ice.ufjf.br
[3] Department of Applied and Computational Mechanics, Federal University of Juiz de Fora, Juiz de Fora, Brazil
leonardo.goliatt@ufjf.edu.br

Abstract. Breast cancer (BC) is one of the most frequently diagnosed neoplasms worldwide and remains the leading cause of mortality among women. The recent application of machine learning methods has successfully predicted BC survival. In this study, we conducted a comparative analysis of specific survival analysis models applied to clinical data. Ensemble models, Gradient Boosting Survival (GBS), and Random Survival Forest (RSF) outperformed traditional approaches. The results of this study reinforce the promising potential of machine learning methods in analyzing the survival of breast cancer patients, with GBS and RSF models standing out as highly effective approaches to enhance the prediction of these patients' survival. This advancement is relevant in oncology, making substantial contributions to clinical decision-making and therapeutic planning.

1 Introduction

Breast cancer (BC) ranks among the most prevalent neoplasms globally, surpassing lung cancer in incidence. It is essential to underscore that this disease predominantly affects women, standing as the primary cause of cancer-related mortality in the female population, regardless of a country's economic and social context. According to the 2020 estimates by the World Health Organization (WHO), there were approximately 2.26 million new cases of breast cancer recorded, resulting in 685,000 deaths worldwide [1,2]. In Brazil, BC statistics align with this global trend, with the National Cancer Institute (INCA) estimating around 66,000 new cases in 2020, leading to 18,000 deaths attributed to the disease in the same year [3].

BC is a diverse and invasive form of cancer and manifests in highly varied ways. The complexity and heterogeneity of this disease, coupled with the need

A. Abraham et al. (Eds.): ISDA 2023, LNNS 1052, pp. 172–181, 2024.
https://doi.org/10.1007/978-3-031-64776-5_17

for personalized prognostics, present a significant challenge in women's public health. Controlling BC demands a comprehensive approach encompassing prevention, screening, early diagnosis, and appropriate treatment [4,5,8].

Mammographic screening and physical examination primarily achieve early diagnosis. At the same time, therapeutic strategies are established based on assessing risk factors, prognosis, staging, and examination results (location, histopathological, immunohistochemical, molecular, and genetic). Early detection plays a central role in increasing survival rates and is closely related to prognostic factors that assist in predicting the risk of recurrence and metastatic evolution, influenced by various aspects such as age at diagnosis, staging, family history, and reproductive factors, among others [5–7].

The application of Machine Learning Methods (MLMs) in oncology has proven to be essential for the development of computational prognostic models aimed at predicting the survival of breast cancer patients. The accuracy in this prediction plays a vital role in personalized clinical decision-making and optimizing financial resources, contributing to improvements in both patient survival and treatment effectiveness. However, to attain these objectives, a multidisciplinary approach is required, involving experts in Data Science and Oncology, along with the creation of robust data repositories for model validation and continuous improvement [4,5,9].

In this context, developing and validating computational models that utilize widely recognized clinical measures in medical practice are imperative. These models aim to enhance the diagnosis, treatment, monitoring, and survival prediction for breast cancer patients. These advancements result in a significant reduction in patient suffering and improvements in both life expectancy and quality of life [4,5]. Moreover, addressing this cancer in women benefits public health and recognizes and emphasizes the pivotal role of women as active participants in society, both from a social and economic perspective. It also acknowledges the significant contribution of women as family caregivers and addresses the complex issue of gender inequality [8].

Survival analysis is a vital tool in the medical field for investigating disease progression and estimating the likelihood of survival at specific milestones. This analysis takes a personalized approach, considering various factors to produce individualized estimates of the survival function and associated risks. "Survival", which refers to the period after diagnosis during which a patient remains alive, is a commonly used metric for assessing survivability. Survival analysis deals with censored data, which encompasses incomplete information about individuals' survival times, representing a portion of the studied population [5,10,11].

In this light, the primary objective of this study is to apply and compare the classic Cox Proportional Hazards model (CPH) [12] with the performance of specific MLMs for survival analysis [13] on a robust clinical dataset of patients diagnosed with breast cancer and monitored in reference centers in the Zona da Mata Mineira. The research aims to contribute substantially to advancing breast cancer treatment and prognosis, aligning with global efforts to combat this devastating disease and promoting improved outcomes for patients and society.

2 Machine Learning Methods Applied to Survival Analysis

In the context of cancer, these algorithms leverage clinical data to predict prognosis by acquiring insights from patterns derived from historical data. They achieve this by creating and refining statistical models using previously observed real-world data. Supervised MLMs, which learn the connection between inputs and correct outputs, are particularly valuable for addressing classification and regression challenges. Prominent examples of such methods encompass Gradient Boosting Survival (GBS), Random Survival Forest (RSF), and Survival Support Vector Machine (SSVM), all of which have found successful applications in the field of survival analysis [9,14,15].

Recently, supervised MLMs have gained prominence in breast cancer prediction, surpassing conventional statistical approaches such as the CPH model. The primary advantage of MLMs lies in their ability to handle complex data, including incomplete information, known as censored data, which poses a significant challenge in survival analysis. These methods aim to uncover hidden connections between clinical patterns and treatment responses, enabling highly personalized prognoses. Enhancing the reliability of these methods is paramount in oncological clinical practice, as it aids in decision-making and provides more accurate treatment guidance for breast cancer patients. Recent research works corroborate and strengthen this trend [4,11,16–18,25].

2.1 Cox Proportional Hazards

Due to its interpretability, the CPH model is extensively used in oncology survival analysis. This model calculates the likelihood of a specific event occurring at a given time based on predictor variable values. In survival analysis, the CPH evaluates how covariates influence event hazard rates, such as mortality, facilitating the discovery of prognostic factors. However, the CPH has certain limitations, including its unsuitability for high-dimensional data, restrictive assumptions, and challenges in modeling nonlinearities and intricate variable interactions. Therefore, it is essential to bear these constraints when applying the CPH to more intricate survival analyses [11,12,16].

2.2 Gradient Boosting Survival

The GBS does not denote a specific model but is a highly adaptable framework for optimizing various loss functions. It takes the form of a learning method structured as an optimization problem, wherein a loss function is systematically minimized by progressively incorporating weak learners, such as decision trees. Each iteration generates a fresh decision tree to grasp the residuals left by the previous model. This iterative tree-adding process continues until further significant enhancements are unattainable, culminating in the final prediction derived from the weighted combination of each tree's forecasts [11,16,18,19].

2.3 Random Survival Forest

The RSF is an approach derived from the original Random Forest method, which employs bootstrapping and random feature selection to develop decision trees. RSF stands out by incorporating censoring information into the tree-splitting rules introducing randomness during construction. The log-rank test splits the survival trees to maximize the survival difference between nodes. RSF estimates the probability of patient survival after a specified time, using estimators such as Kaplan-Meier and Nelson-Aalen for survival and cumulative hazard functions in terminal nodes. The final result is obtained by averaging the predictions from each tree, providing a robust estimate of the survival function [16,18,20].

2.4 Survival Support Vector Machine

The SSVM is an extension of the Support Vector Machines (SVM) method designed to handle survival data. The primary purpose of SSVM is to find a hyperplane in the feature space that maximizes the margins between event and censoring classes. SSVM is a supervised Machine Learning technique that applies to classification and regression tasks, using iterations to find a hyperplane that minimizes error and maximizes the margin between classes. The unique feature of SSVM is its incorporation of survival data through ranking based on the number of instances. This capability facilitates precise estimation of survival probabilities and the prediction of event risks for patients [16,18,21].

3 Material and Methods

In order to exemplify the application of supervised MLMs, we compared the performance of the classical CPH model to the GBS, RSF, and SSVM models for the survival analysis of female BC patients. We utilized a robust clinical database that included patients diagnosed between January 2003 and January 2005, treated and monitored by oncology centers in the Zona da Mata Mineira, with active follow-ups until 2011 [6,7].

3.1 Database and Data Preprocessing

The clinical database employed in this study contains information from hospital records of oncology reference centers in the Zona da Mata Mineira region. It encompasses a population of women diagnosed with invasive breast cancer who received local and systemic treatments. It is important to note that these data were initially gathered from prior research conducted by the Brazilian Health Program [6] and the Postgraduate Program in Collective Health [7], which employed statistical methodologies Furthermore, the research project underwent submission and received approval from the Human Research Ethics Committee of the Federal University of Juiz de Fora, with the approval number 5.533.296.

Data collection from the original database occurred in three distinct stages. In the first stage, conducted between 2009 and 2010, patients were recruited by

searching cancer hospital records, resulting in 601 cases and a follow-up deadline of December 31, 2010. In the second stage, we evaluated the clinical and sociodemographic characteristics of the patients, and we excluded those who had undergone only one procedure in the institutions, resulting in 563 cases. The third stage started in January 2011 and involved conducting phone calls, validating through the Brazilian individual taxpayer registry number (CPF), contacting mastologists, and consulting the mortality information system [6,7,22].

In the original database's raw data, we initiated preprocessing procedures. Initially, we addressed missing data and standardized formatting by rectifying typing errors and eliminating accents from text strings. We also performed specific value imputation for missing numerical data. Dates underwent meticulous verification and were appropriately formatted. Our analysis covered five years, with a maximum limit of 1826 days. The 5-year survival time calculation began at the time of diagnosis and concluded either at breast cancer-related death or on the last day of follow-up, restricted to 1826 days. If a patient passed away after this defined period, her data was recorded as "censored", as the database contained survival data for 5 and 10 years, including comprehensive details about the immunohistochemical profile of patients monitored during this extended period [6,7]. As a result, after preprocessing and specific value imputation, the resulting database consists of numerical data with a value of 9 for missing data, numerical count data with a value of -1, and strings containing the term "unknown", which subsequently converted them into numerical values.

3.2 Methods, Attributes, and Hyperparameters

The CPH, GBS, RSF, and SSVM models were implemented in Python using the Scikit-survival library [13]. This library enables survival analysis by correlating covariates with event times and provides performance metrics specific to this type of analysis [15]. In summary, the attributes corresponding to prognostic variables were age, staging, family history, reproductive and other disease history, hormonal and molecular receptor status, immunohistochemical profile, tumor type and grade, and treatments administered, among others. The value variables included status (0-censored or 1-event) and survival time (days) [6,7,22].

To apply the selected models in our study, we followed the standard practice of splitting the data into training (75%) and testing (25%) sets, as the Scikit-survival library recommended. Additionally, we configured the hyperparameters of each model according to the guidelines established by Polsterl et al. (2020), which are summarized below:

1. CPH: We set the maximum number of iterations to 100, with alpha equal to 0, using the feature subset without regularization. Coefficients were estimated using the Breslow method, considering each event at a specific time as distinct.
2. GBS: We defined the number of regression trees in the forest as 100, with minimum sample sizes required to split a node and to stay in a single node as 2 and 1, respectively. The learning rate was set at 1.0 to control the influence of each tree, with bootstrapping sample randomness set at 0. The maximum depth of individual regression trees was limited to 1 to optimize performance.

3. RSF: We employed 1000 trees in the forest, with minimum sample sizes required for splitting a node and staying in a single node set at 10 and 15, respectively. We configured bootstrapping sample randomness at 20 and utilized the log-rank test as the criterion for constructing each tree.
4. SVM: We set the maximum number of iterations to 100 for optimization using the Newton-Raphson method. We configured the weight to penalize the loss of the objective function as 1, and we set the regression parameter at 1 for classification.

3.3 Performance Metrics

The performance assessment of survival analysis methods involved the consideration of three metrics: Harrell's C-index for assessing discriminative capacity, and the Brier Score (BS) and Integrated Brier Score (IBS) for evaluating the calibration ability of the models [15,23,24].

The C-index is a metric that quantifies the agreement between the risk scores predicted by the model and the observed event times in survival data, measuring the model's ability to predict and classify patients' time of death. It ranges from 0.0 to 1.0, where 0.5 indicates average performance with no discriminative capacity, and 1.0 represents optimal performance, with the model accurately distinguishing patients with different outcomes [16–18,24].

The BS metric compares the model's predicted probabilities with the actual patient status at specific time points during the test. Lower BS values signify improved calibration, indicating more precise predictions of a patient's probability of remaining event-free. This metric underscores the model's ability to predict individual outcomes for each patient [16,18,23]. It is important to note that the BS only applies to models capable of estimating a survival function and cannot be used with methods like SSVM [15,23].

Lastly, the IBS, like the BS, is a time-dependent measure that provides a single value to better assess the prediction quality of the models by comparing all time points. Lower values suggest a better outcome obtained through survival curves that the model produces compared to actual data, facilitating the comparison of calibration among the models [10,15,17,23].

4 Results and Discussion

In this study, we compared four MAMs for survival analysis using clinical breast cancer data. Our dataset comprised 558 women who had received a BC diagnosis, and the analysis focused on a 5-year follow-up period to evaluate the survival capacity of these patients during this timeframe. The time count commenced from the date of the histopathological report. He ended either at the occurrence of an adverse event or the censoring date, representing the last day of the predefined follow-up, totaling 1826 days. Over this period, we observed 113 BC-related deaths. In the study, we treated patients who did not experience the event during this period as censored. We terminated their survival data on the last day of the previously defined follow-up.

The dataset covered information about patients diagnosed with BC, encompassing various prognostic variables such as age at diagnosis, habits related to smoking and alcohol consumption, family history of cancer, contraceptive use, pregnancy, and breastfeeding, as well as health conditions like hypertension and cardiovascular disease. Additionally, clinical details include examination results (mammography, X-ray, ultrasound, cytology, and immunohistochemistry), tumor characteristics, treatments received, presence of metastases, menopausal status, and clinical outcomes. The dataset utilized 70 prognostic variables and two value variables representing events and days of survival, allowing comprehensive analyses in the context of BC.

In this study, we utilized four distinct models: CPH, GBS, RSF, and SSVM, implementing them with the scikit-survival library [13], an extension of scikit-learn tailored for survival analysis. Our evaluation of these models encompassed three key metrics: the C-index, the BS, and the IBS [13,15], with a summary of results provided in Table 1. Upon analyzing the data in this table, it became evident that ensemble-based models, namely GBS and RSF, consistently outperformed their regression and classification-based counterparts, such as CPH and SSVM, such as CPH and SSVM, across all the metrics employed. However, it is noteworthy that the SSVM model exhibited intermediate performance, as elucidated in Sect. 3.3. These findings align with previous studies [10,11,16–18,25], which also compared machine learning models tailored for survival analysis in various contexts. These studies demonstrated that ensemble methods like GBS and RSF tend to surpass traditional models like CPH in performance metrics designed explicitly for survival data [11,16,25].

Table 1. Performance of Machine Learning Models for Survival Analysis

Modelo	C-Index[a]	Brier Score[b]	Integrated BS[b]
Cox Proportional Hazards	0.743212	0.155655	0.056657
Gradient Boosting Survival	0.914180	0.065183	0.043731
Random Survival Forest	0.917219	0.084619	0.054676
Survival Support Vector Machine	0.746564	————-	————-

[a] Values close to 1 indicate excellent model discrimination capability.
[b] Values close to 0 highlight outstanding model calibration.

Figure 1 provides a comprehensive view of the evolving BS over time, showcasing predictions made by the CPH, GBS, and RSF models. This visual representation allows a direct and intuitive comparison of the models' performance at each time interval [10,25]. The GBS model consistently achieved the lowest BS and Integrated Brier Score (IBS) values, underscoring its precision in long-term risk prediction, as highlighted by the data in Fig. 1. In stark contrast, the RSF model recorded the highest Harrell C-index, as detailed in Table 1, indicating a robust agreement with the observed data. It is worth noting that the RSF's performance surpassed the reported results of Carvalho et al. (2022), Liu et al.

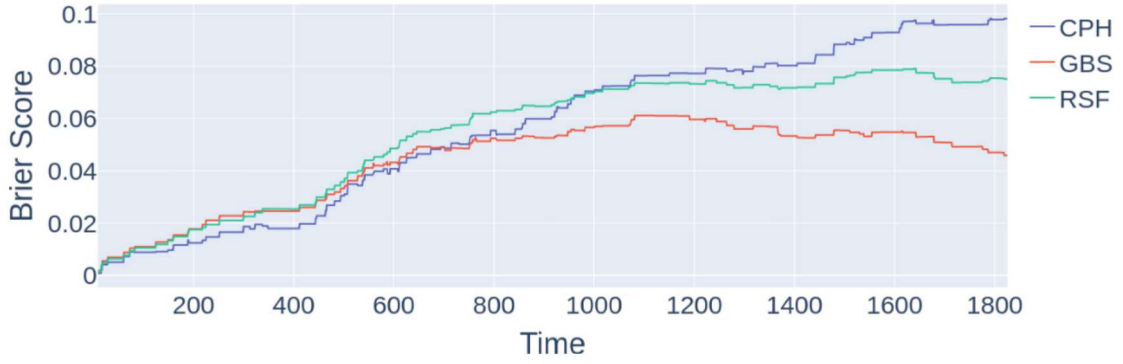


Fig. 1. Temporal Variations of Brier Score for Three Models: CPH (Blue), GBS (Red), and RSF (Green), Considering Both Training and Testing Data.

(2020), and Xiao et al. (2020), yielding respective C-indices of 0.802, 0.814, and 0.827 [4,11,18].

Meanwhile, the SSVM model, while displaying intermediate performance, outperformed the findings of Moncada et al. (2021), with a C-index of 0.63. However, it fell short of the work by Xiao et al. (2022), with a C-index of 0.812 [16,18]. This variance can be attributed to the necessity of parameter adjustments in the SSVM, as it seeks to find an optimal hyperplane for separating data into two groups with distinct risk profiles [13].

Lastly, the classical CPH model exhibited the weakest performance, even though it surpassed the IBS value (0.152) found in Krzyzinski et al.'s (2023) work [10]. This model, widely used in survival analysis, relies on assumptions such as proportional hazards and linear effects, making it sensitive to tied events where multiple events coincide [10,18,25].

5 Conclusion

In conclusion, this study underscores the varied performance of machine learning models for survival analysis, a performance contingent upon data characteristics and evaluation criteria. In our clinical context of breast cancer, ensemble models like GBS and RSF stood out. Nevertheless, choosing the best model must encompass other critical considerations, such as interpretability, efficiency, and validity. Future endeavors should enhance data preprocessing, parameter optimization, and careful feature selection to bolster each model's performance.

Furthermore, fostering interdisciplinary collaboration between oncologists and data scientists is essential to develop more robust and interpretable BC survival analysis models. As Fanizzi et al. (2023) aptly noted, "the synergy of domain expertise and advanced data analytics can lead to significant advancements in the field" [25].

Acknowledgement. Coordination for the Improvement of Higher Education Personnel (CAPES) for the financial support through a granted scholarship to the Computational Modeling Graduate Program at the Federal University of Juiz de Fora,

of Dr. Maximiliano Ribeiro Guerra and Dr. Jane Rocha Duarte Cintra for providing clinical data, and to the 9 de Julho Hospital and Oncological Institute for their support in the research.

References

1. Ferlay, J., Colombet, M., Soerjomataram, I., Parkin, D.M., Piñeros, M., Znaor, A., et al.: Cancer statistics for the year 2020: an overview. Int. J. Cancer (2021). https://doi.org/10.1002/ijc.33588
2. World Health Organisation, in Cancer. (Available via WHO, 2022). https://www.who.int/news-room/fact-sheets/detail/cancer. Cited 12 Jan 2023
3. Instituto Nacional do Câncer, in Estatísticas de câncer. (Available via Ministério da Saúde, 2021). https://www.gov.br/inca/pt-br/assuntos/cancer/numeros/. Cited 27 Jul 2023
4. Carvalho, D.S., Nogueira, T.S., Goliatt, P.V.C.Z.: Aplicação do Random Survival Forest na análise da sobrevida para câncer da mama. J. Health Inf. (2023). https://doi.org/10.59681/2175-4411.v15.iEspecial.2023.1113
5. Li, J., Zhou, Z., Dong, J., Fu, Y., Li, Y., Luan, Z., et al.: Predicting breast cancer 5-year survival using machine learning: a systematic review. J. PloS one (2021). https://doi.org/10.1371/journal.pone.0250370
6. Cintra, J.R.D.: Sobrevida e fatores associados em pacientes com câncer de mama, com diagnóstico entre 2003 e 2005 no município de Juiz de Fora Minas Gerais. Programa de Pós-Graduação em Saúde Brasileira da Faculdade de Mediciana da Universidade Federal de Juiz de Fora, Juiz de Fora (2012)
7. Fayer, V.A(2012) Sobrevida de 10 anos e fatores prognósticos em coorte hospitalar de pacientes com câncer de mama assistidas em Juiz de Fora, Minas Gerais, Brasil. Programa de Pós-Graduação em Saúde Coletiva da Universidade Federal de Juiz de Fora, Juiz de Fora
8. Torre, L.A., Islami, F., Siegel, R.L., Ward, E.M., Jemal, A.: A Global cancer in women: burden and trendsglobal cancer in women: burden and trends. Can. Epi. Bio. Prev. (2017). https://doi.org/10.1158/1055-9965.EPI-16-0858
9. Min, N., Wei, Y., Zheng, Y., Li, X.: Advancement of prognostic models in breast cancer: a narrative review. J. Gland. Surg. (2021). https://doi.org/10.21037/gs-21-441
10. Krzyziński, M., Spytek, M., Baniecki, H., Biecek, P.: SurvSHAP (t): Time-dependent explanations of machine learning survival models. Knowl. Inf. Syst. (2023). https://doi.org/10.1016/j.knosys.2022.110234
11. Liu, P., Fu, B., Yang, S.X., Deng, L., Zhong, X., Zheng, H.: Optimizing survival analysis of XGBoost for ties to predict disease progression of breast cancer. IEEE Trans. Biomed. Eng. (2020). https://doi.org/10.1109/TBME.2020.2993278
12. Cox, D. R.: scikit-survival: regression models and life-tables. J. R. Stat. Soc. Series B Stat. Methodol. **34**, 187–202 (1972)
13. scikit-survival 0.21.0, in scikit-survival, (Available via Pölsterl, S., et al., 2015–2023). https://scikit-survival.readthedocs.io/en/stable/index.html. Cited 26 Set 2023
14. Panch, T., Szolovits, P., Atun, R.: Artificial intelligence, machine learning and health systems. J. Global Health (2018). https://doi.org/10.7189/jogh.08.020303
15. Pölsterl, S.: scikit-survival: a Library for Time-to-Event Analysis Built on Top of scikit-learn. J. Mach. Learn. Res. **21**, 8747–8752 (2020)

16. Moncada-Torres, A., van Maaren, M.C., Hendriks, M.P., Siesling, S., Geleijnse, G.: Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. J. Sci. Rep. (2021). https://doi.org/10.1038/s41598-021-86327-7
17. Pinheiro, T., et al.: Machine Learning e Análise Multivariada aplicados à Sobrevida do Câncer Mama. J. Health Inf. **14** (2022)
18. Xiao, J., Mo, M., Wang, Z., Zhou, C., Shen, J., Yuan, J., et al.: The application and comparison of machine learning models for the prediction of breast cancer prognosis: retrospective cohort study. JMIR Med. Inf. (2022). https://doi.org/10.2196/33440
19. Hothorn, Bühlmann, P., Dudoit, S., Molinaro, A., Van Der Laan, M.J.T.: Survival ensembles. J. Biost. (2006). https://doi.org/10.1093/biostatistics/kxj011
20. Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S.: Random survival forests. Ann. Appl. Stat. (2008). https://doi.org/10.1214/08-AOAS169
21. Pölsterl, S., Navab, N., Katouzian, A.: Fast training of support vector machines for survival analysis. In: Appice, A., Rodrigues, P.P., Costa, V.S., Gama, C.S.J., Jorge, A. (eds.) Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7–11, 2015, Proceedings, Part II 15, pp. 243–259. Springer, Portugal (2015). https://doi.org/10.1007/978-3-319-23525-7_15
22. Cintra J.R.D., et al.: Perfil imuno-histoqímico e variáveis clinicopatológicas no câncer de mama. Rev. Assoc. Med. (2012). https://doi.org/10.1590/S0104-42302012000200013
23. Granf, E., Schmoor, C., Sauerbrei, W., Schumancher, M.: Clinical implications of dysregulated cytokine production. Stat. Med. **18**, 2529–2545 (1999)
24. Uno, H., Cai, T., Pencina, M.J., D'Agostino, R.B., Wei, L.J.: On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat. Med. (2011). https://doi.org/10.1002/sim.4154
25. Fanizzi, A., Pomarico, D., Rizzo, A., Bove, S., Comes, M.C., Didonna, V., et al.: Machine learning survival models trained on clinical data to identify high risk patients with hormone responsive HER2 negative breast cancer. Sci. Rep. (2023). https://doi.org/10.1038/s41598-023-35344-9