# Uso de Inteligência Artificial para Predição de Sobrevivência de Pacientes com Câncer de Próstata

**Alcir Canella Filho[2], Felipe Clé Monteiro[2], Matheus do Nascimento Marques[2], Mario Olimpio de Menezes[1,2]**

[1]Sistemas de Informação
Faculdade de Computação e Informática
Universidade Presbiteriana Mackenzie
São Paulo – SP – Brasil

[2]Ciência da Computação
Faculdade de Computação e Informática
Universidade Presbiteriana Mackenzie
São Paulo – SP – Brasil

{10396389,10395521,10395894}@mackenzista.com.br, mario.menezes@mackenzie.br

***Resumo.*** *Between 1999 and 2023, the São Paulo Oncocentro Foundation registered more than 130 thousand cases of prostate cancer, the most common type being acinar cell carcinoma, with more than 72 thousand occurrences. Artificial Intelligence has been used as a methodology for predicting survival in several areas, and this study aims to apply it to data from the Fundação Oncocentro de São Paulo to predict the survival of patients with prostate cancer.*

***Resumo.*** *Entre 1999 e 2023, a Fundação Oncocentro de São Paulo registrou mais de 130 mil casos de câncer de próstata, sendo o tipo mais comum carcinoma de celulas acinosas, com mais de 72 mil ocorrências. A Inteligência Artificial tem sido utilizada como metodologia para previsão de sobrevivência em diversas áreas, e este estudo visa aplicá-la aos dados da Fundação Oncocentro de São Paulo para predição de sobrevivência de pacientes com câncer de próstata.*

***Palavras-chave:*** *Câncer, Dados, Análise, Dashboard, Predição.*

## 1. Introdução

O câncer é uma das principais causas de morbidade e mortalidade em todo o mundo[Ferlay et al. 2015], apresentando um significativo desafio para a saúde pública. De acordo com o Instituto Nacional de Câncer[INCA 2022], as projeções para o Brasil no triênio 2023-2025 indicam que, excluindo os casos de câncer de pele não melanoma, haverão cerca de 704 mil casos novos de câncer. Dentre estes, são previstos aproximadamente 71 mil casos de cancer de próstata.

Há a expectativa de que o número de pacientes com câncer aumente. A Agência Internacional de Pesquisa em Câncer (IARC - ONU) em conjunto com a The Lancet Comission[James 2024] prevê que a carga global de cancer de próstata mais que dobrará até 2040, aproximando-se de três milhões de novos casos, comparado ao número de casos estimados atualmente.

Existem bases de dados abertas com grandes volumes de dados sobre pacientes de câncer. Uma delas é a da Fundação Oncocentro de São Paulo [FOSP 2022], que armazena informações sobre pacientes com câncer no estado de São Paulo desde o ano de 1999, bem como detalhes do estadiamento clínico, faixa etária, cirurgias realizadas e sobrevivência.

É possível aplicar um modelo que estima as chances de sobrevivência de um paciente utilizando as informações disponíveis em bases de dados. Modelos de sobrevivência são amplamente utilizados para tomar decisões clínicas, utilizando diversos métodos de machine learning, para obter predições de tempo até o evento quando alguns dados estão censurados [Suresh 2022]. Nesses contextos, os modelos devem ser precisos e interpretáveis para que os utilizadores (como os médicos) possam confiar no modelo e compreender as previsões do modelo.

No artigo *Machine Learning for Predicting Survival of Colorectal Cancer Patients*[Buk Cardoso et al. 2023] são treinados três diferentes modelos para predição de sobrevivência de pessoas com câncer colorretal, utilizando Naive Bayes[Pedregosa et al. 2011], Random Forest[Ho 1995] e XGBoost[Chen and Guestrin 2016]. A análise dos resultados mostram que os modelos Random Forest e XGBoost , sendo estes dois baseados em árvores de decisão, obtiveram acurácia superior quando comparados ao modelo Naive Bayes, comumente utilizado para classificação.

É estabelecido como objetivo deste estudo treinar modelos para predição de sobrevivência utilizando dados de câncer de próstata disponíveis no site da Fundação Oncocentro de São Paulo[FOSP 2022], tendo o método publicado por Buk Cardoso como base para o processo. Estão inclusos como objetivos os itens abaixo:

- Realizar análise exploratória e preparar dados públicos de câncer de próstata. Tal parte foi concluída neste semestre e está disponívei no *Github*. É possível visualizar endereço do repositório na seção Metodologia deste documento.

- Aplicar um modelo preditivo que integre os dados analisados para prever o impacto do câncer de próstata (carcinoma de células acinosas) na longevidade dos pacientes.

- Validar o modelo preditivo para garantir sua precisão e confiabilidade.

- Concluir a edição do artigo.

O estudo colaborará com a validação da utilização tanto do método quanto da base de dados da Fundação Oncocentro de São Paulo para predição de sobrevivência de pacientes com câncer de próstata.

## 2. Referencial Teórico

[Buk Cardoso et al. 2023] Artigo onde Buk Cardoso et al. demonstrou ser possível utilizar uma das bases de dados públicos da Fundação Oncocentro de São Paulo para aplicar modelos e prever sobrevivência de pacientes de câncer colorretal. Como o artigo teve colaboração de profissionais da área médica e está bastante detalhado, foi escolhido como base para este estudo.

[Chen and Guestrin 2016] Especificação do XGBoost, método que cria um con-

junto de modelos fracos (geralmente árvores de decisão) e os combina para formar um modelo forte. Seu principal objetivo é melhorar a precisão preditiva de modelos de aprendizado de máquina ao reduzir o erro. Este algoritimo é candidato a ser utilizado na segunda fase deste projeto.

[Ferlay et al. 2015]Artigo que contém estimativas da incidência e mortalidade de 27 tipos de câncer principais e de todos os tipos de câncer combinadas em 20 regiões do mundo.

[FOSP 2022] Base de dados pública da Fundação Oncocentro de São Paulo. Origem do conjunto de dados utilizados para este estudo.

[Ho 1995] Especificação do modelo Random Forest, um algoritmo de aprendizado de máquina que utiliza um conjunto (ou "floresta") de árvores de decisão para melhorar a precisão preditiva. Este algoritimo é candidato a ser utilizado na segunda fase deste projeto.

[INCA 2022] Publicação que apresenta estimativas no triênio 2023-2025 sobre os principais tipos de câncer no Brasil, desenvolvida com o intuito de ser utilizada em apoio a ações de prevenção e controle de câncer.

[James 2024] Publicação da The Lancet Commission em conjunto com a International Agency for Research on Cancer (ONU) que contém estimatívas sobre câncer de próstata para o ano de 2040, considerando o aumento da expectativa de vida e mudanças demográficas.

[Suresh 2022] Artigo que aborda modelos de sobrevivência aplicados em cenários de tempo até o evento, utilizando dados coletados em intervalos discretos, como anos ou meses. Tal cenário condiz com este estudo.

[Pedregosa et al. 2011] Espeficação do método Naive Bayes na biblioteca *Scikit-Learn*, sendo um conjunto de algoritmos de aprendizado de máquina supervisionados que utilizam o teorema de Bayes.

## 3. Metodologia

Na fase de *Data Understanding* foi feita a extração e revisão dos dados. Foram aplicadas as bibliotecas Pandas, Numpy , Matplotlib, Seaborn e Plotly para visualização e entendimento dos dados.

O conjunto de dados utilizado neste estudo é oriundo de uma base pública mantida pela Fundação Oncocentro de São Paulo, sendo a mesma fonte do *dataset* utilizado no artigo *Machine Learning for Predicting Survival of Colorectal Cancer Patients* [Buk Cardoso et al. 2023]. Todos os pacientes estão anonimizados, não contendo nomes ou documentos. Os dados foram tratados de forma similar à adotada no artigo, havendo necessidade de algumas mudanças pontuais. Algumas diferenças entre formato de campos de data ou atributos que não foram mencionados no artigo estavam presentes no *dataset*. Tanto a base de dados quanto a descrição dos campos estão disponíveis em: https://fosp.saude.sp.gov.br/fosp/diretoria-adjunta-de-informacao-e-epidemiologia/rhc-registro-hospitalar-de-cancer/banco-de-dados-do-rhc

O *dataset* conta com mais de 100 mil pacientes de câncer de próstata, sendo pouco mais de 72 mil com carcinoma de céculas acinosas (morfologia 85503) entre 1999 e 2023.

Tal morfologia é o objeto deste estudo, que tem foco nos pacientes registrados de 2000 a 2020.

O pico de diagnosticos foi no ano de 2014, com aproximadamente 7 mil pacientes diagnosticados, como pode ser visto na Figura 1.

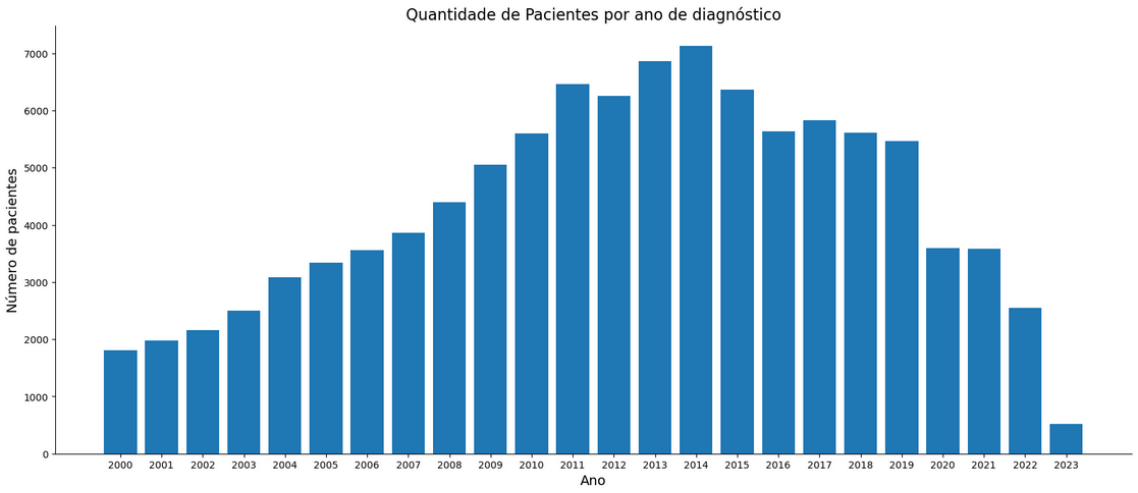


**Figura 1. Quantidade de diagnosticos por ano**

A maioria dos pacientes está classificada no nível 2 do atributo estadiamento clínico (ver Figura 2). Este campo determina a extensão do câncer pelo corpo do paciente, bem como é utilizado para tomada de decisão em relação a tratamentos.

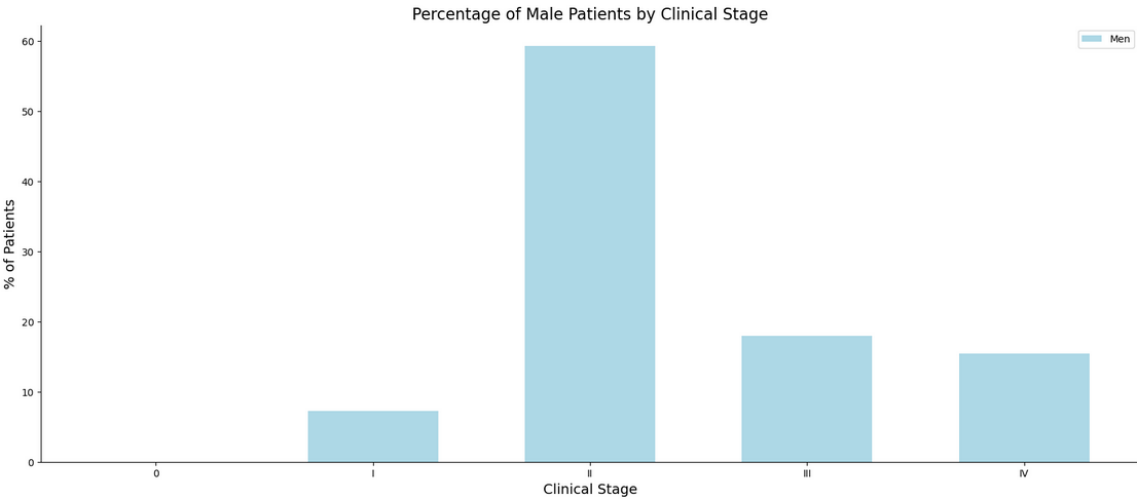


**Figura 2. Porcentagem de pacientes por grupo de estadiamento clínico**

Para a preparação dos dados, foram excluídos atributos com maior quantidade de dados faltantes no dataframe. Também foi aplicada correlação de Pearson para identificar atributos que não são relevantes. Parte dos atributos foram descartados por não possuírem variedade relevante de dados, não sendo aproveitáveis pelos modelos.

Foram deletadas as colunas abaixo por diversos motivos. Algumas possuem muitos dados faltantes, e não seria possível fazer um preenchimento preciso dos mesmos,

ou por não causarem impacto significativo (baixa ou nula variabilidade de valores). Citando como exemplo a coluna MORFO, deletada pelo fato da análise cobrir somente uma morfologia, ou o campo ECGRUP, que possui as mesmas informações contidas no campo EC. Os campos HABIT11, HABILIT1, e HABILIT2, por exemplo, não continham informações de tratamento dos pacientes:

'UFRESID', 'UFNASC', 'CIDADE', 'DESCTOPO', 'DESCMORFO','OUTRACLA','INSTORIG' 'META01', 'META02','META03', 'META04', 'REC01', 'REC02', 'REC03','REC04', 'MORFO', 'TOPO', 'TOPOGRUP', 'T', 'N', 'M', 'NAOTRAT', 'TRATAMENTO', 'TRATFAPOS', 'NENHUMAPOS', 'CIRURAPOS', 'RADIOAPOS', 'QUIMIOAPOS', 'HORMOAPOS', 'IMUNOAPOS','OUTROAPOS', 'RECLOCAL', 'RECREGIO', 'RECDIST', 'HABILIT','INSTORIG', 'CICI','CICIGRUP','CICISUBGRU','CLINICA','ECGRUP', 'TRATFANTES','FAIXAETAR','PERDASEG','HABIT11','HABILIT1','HABILIT2', 'CIDADEH',','CIRU','S','QUIMIOANT','HORMOANT','TMOANT','IMUNOANT', 'OUTROANT',

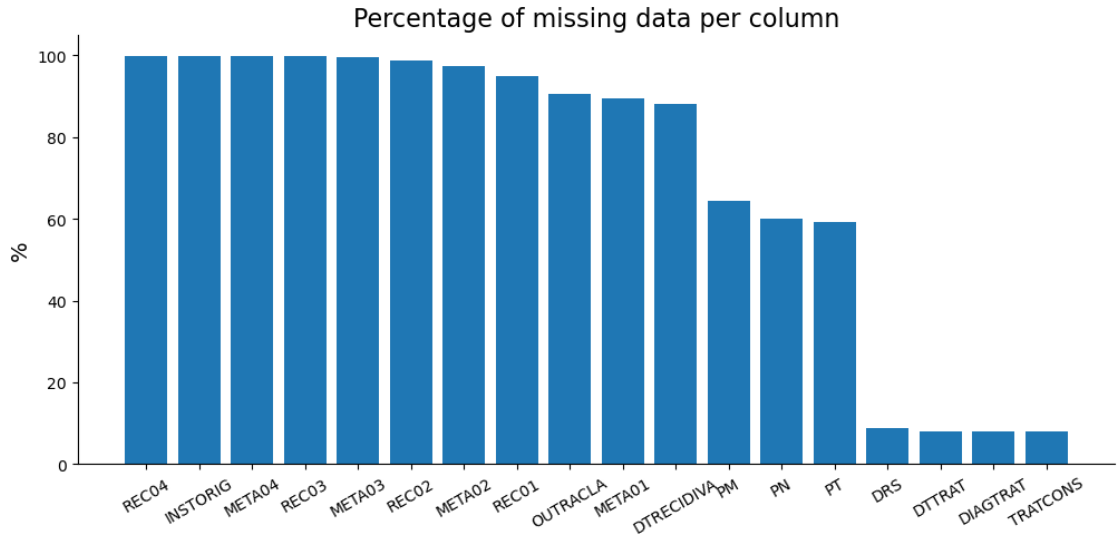


**Figura 3. Colunas com maior quantidade de dados faltantes**

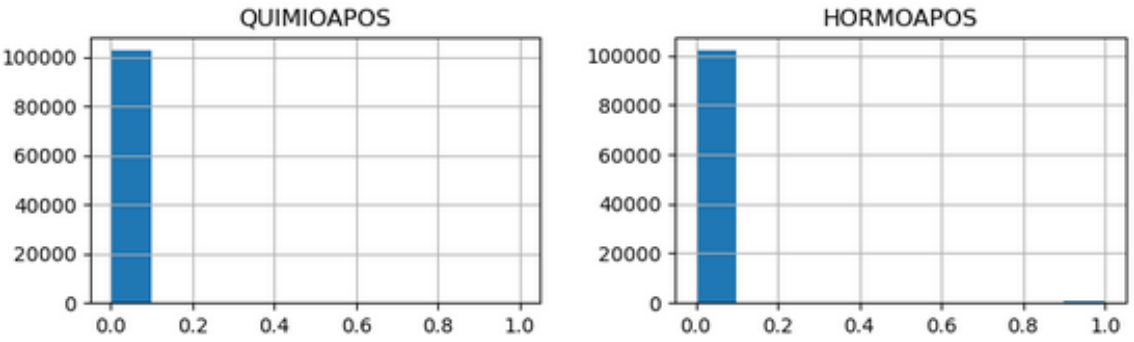


**Figura 4. Exemplos de atributos com os mesmos valores para todos os pacientes**

Os atributos abaixo foram excluídos após aplicação da correlação de Pearson, tendo mostrado importância baixa ou nula:

'SEXO','LOCALTNM','IDMITOTIC','LATERALI','ERRO', 'TMOAPOS', 'CIRURANT','RADIOANT'

Abaixo, os atributos do dataset escolhidos para o estudo:

| Campo | Descrição |
|---|---|
| ESCOLARI | Código para escolaridade do paciente |
| IDADE | Idade do paciente |
| IBGE | Código da cidade de residência do paciente |
| CATEATEND | Categoria de atendimento ao diagnóstico |
| DIAGPREV | Diagnóstico e tratamento anterior |
| BASEDIAG | Código da base do diagnóstico |
| EC | Estadio clinico |
| G | Classificação TNM – G (Grau) |
| PSA | Classificação TNM - PSA |
| GLEASON | Classificação TNM - Gleason |
| TRATHOSP | combinação dos tratamentos realizadosno hospital |
| NENHUM | Tratamento recebido no hospital = nenhum |
| CIRURGIA | Tratamento recebido no hospital = cirurgia |
| RADIO | Tratamento recebido no hospital = radioterapia |
| QUIMIO | Tratamento recebido no hospital = quimioterapia |
| HORMONIO | Tratamento recebido no hospital = hormonioterapia |
| TMO | Tratamento recebido no hospital = tmo |
| IMUNO | Tratamento recebido no hospital = imunoterapia |
| OUTROS | recebido no hospital = outros |
| NENHUMANT | Nenhum tratamento recebido fora do hospital e antes da admissão |
| ULTINFO | Última informação sobre o paciente |
| CONSDIAG | Diferença em dias datas de consulta o diagnóstico |
| TRATCONS | Diferença em dias datas de consulta e tratamento |
| DIAGTRAT | Diferença em dias datas de tratamento e diagnóstico |
| ANODIAG | Ano de diagnóstico |
| DRS | Departamento Regional de Saúde |
| RRAS | Rede Regional de Atenção à Saúde |
| RECNENHUM | Sem recidiva |
| IBGEATEN | Código IBGE da instituição |
| ULTICONS | Diferença de dias entre última informação e consulta |
| ULTIDIAG | Diferença de dias entre última informação e diagnóstico |
| ULTITRAT | Diferença de dias entre última informação e tratamento |

Foram excluídas linhas de pacientes cuja morfologia é diferente de carcinoma de células acinosas, sendo esta o objeto do estudo.

Seguindo a linha de Buk Cardoso et al. também foram excluídos pacientes cujo valor do campo ECGRUP fosse igual a $x$ ou $y$. Estes valores indicam, respectivamente, casos em que o tumor primário, linfonodos regionais ou metástases não possam ser avaliados pelo exame físico ou exames complementares, ou estadiamento feito durante ou

após o tratamento. Logo, pacientes nestas categorias não colaborariam para uma análise de sobrevivência precisa. Por fim, foram excluídos todos os pacientes cuja escolaridade não teria sido informada. Tal abordagem difere da utilizada por Buk Cardoso et al., que preferiu utilizar aprendizado de máquina para estimar a escolaridade de pacientes que não a haviam informado.

Ao final da preparação dos dados, restaram 44094 registros de pacientes que serão utilizados para treinamento dos modelos.

Todas as modificações feitas no dataset bem como o dicionário de dados, o notebook Jupyter contendo a análise exploratória, a preparação dos dados e os gráficos gerados estão visíveis no seguinte repositório do GitHub:

https://github.com/acanellafilho/tccsobrevivencia

*Nota: Foi necessário comprimir o dataset, pois o mesmo ultrapassava o limite de 25mb imposto pelo github.*

Durante a próxima fase do projeto, o grupo verificará a viabilidade do uso de modelos como *Naive Bayes*, *Random Forest* e *XGBoost*, sendo estes utilizados no artigo de Buk Cardoso et al. Cada modelo será testado com 25 por cento dos dados, sendo os 75 por cento restantes utilizados para treinar o mesmo.

## 4. Cronograma

Para a fase do TCC 1 foram cumpridos todos os requisitos, que eram a finalização do pré-processamento e a análise exploratória.

**Tabela 1. Cronograma para o Desenvolvimento do TCC**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Levantamento Bibliográfico | ✓ | ✓ | | | | | | | | |
| Coleta de Dados | | ✓ | ✓ | | | | | | | |
| Tratamento de Dados | | | ✓ | ✓ | ✓ | ✓ | | | | |
| Pré-processamento de Dados | | | ✓ | ✓ | ✓ | ✓ | | | | |
| Análise Exploratória | | | | ✓ | ✓ | ✓ | | | | |
| Treinamento do Modelo Preditivo | | | | | X | X | X | | | |
| Validação do Modelo | | | | | | X | X | X | | |
| Analise dos Resultados | | | | | | | X | X | X | |
| Redação do Artigo Científico | | | | | | | | X | X | |
| Revisão e Ajustes | | | | | | | | | X | X |

## Referências Bibliográficas


Buk Cardoso, L., Cunha Parro, V., Verzinhasse Peres, S., Curado, M. P., Fernandes, G. A., Wünsch Filho, V., and Natasha Toporcov, T. (2023). Machine learning for predicting survival of colorectal cancer patients. Disponível em: http://dx.doi.org/10.1038/s41598-023-35649-9. Acesso em:5 mai.2024.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D. M., Forman, D., and Bray, F. (2015). Cancer incidence and mortality worldwide: Sources, methods and major patterns in globocan 2012. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1002/ijc.29210. Acesso em: 05 mai.2024.

FOSP (2022). Banco de dados do rhc. Disponível em:https://fosp.saude.sp.gov.br/fosp/diretoria-adjunta-de-informacao-e-epidemiologia/rhc-registro-hospitalar-de-cancer/banco-de-dados-do-rhc. Acesso em:5 mai.2024.

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE.

INCA (2022). Estimativa 2023: Incidência de câncer no brasil. Disponível em:https://www.inca.gov.br/publicacoes/livros/estimativa-2023-incidencia-de-cancer-no-brasil. Acesso em: 16 out.2024.

James, N. D. e. a. (2024). The lancet commission. Disponível em: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(24)00651-2/fulltext". Acesso em: 23 out.2024.

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., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. "https://scikit-learn.org/1.5/modules/naive$_b ayes.html$". $Acessoem$ : $27out.2024$.

Suresh, K., S. C. . G. D. (2022). Survival prediction models: an introduction to discrete-time modeling. https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-022-01679-6citeas. Acesso em: 27 out.2024.