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

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

[1]Faculdade de Computação e Informática (FCI)
Universidade Presbiteriana Mackenzie São Paulo, SP – Brasil

<10396389,10395521,10395894@mackenzista.com.br>
<mario.menezes@mackenzie.br>

2025

## Resumo

*A Agência Internacional de Pesquisa em Câncer (IARC-ONU), junto a The Lancet Commission, prevê que o número de casos de câncer de próstata mais que dobrará até 2040, e que este aumento iminente implicará uma subida rápida da taxa global de mortes pela enfermidade. Tendo em vista tal cenário, este estudo foi iniciado para aplicar modelos de predição de sobrevivência utilizando dados de pacientes com câncer de próstata, focando na morfologia carcinoma de células acinosas, sendo esta a mais comum para tal tipo de câncer. Foi utilizada uma base de dados pública mantida pela Fundação Oncocentro de São Paulo, onde foram encontrados dados desde o ano 2000 até 2020 de aproximadamente 70 mil pacientes com a morfologia estudada, pertencendo a múltiplas faixas etárias e diferentes grupos de estadiamento clínico. Foram feitas a preparação, a análise exploratória e o pré-processamento dos dados, e também a aplicação de modelos preditivos computacionais, com o intuito de verificar a viabilidade dos mesmos para auxiliar decisões clínicas ou públicas.*

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

## Abstract

*The International Agency for Research on Cancer (IARC-UN), together with The Lancet Commission, predicts that the number of prostate cancer cases will more than double by 2040, and that this imminent increase will imply a rapid rise in the global rate of deaths from the disease. In view of this scenario, this study was initiated to apply survival prediction models using data from patients with prostate cancer, focusing on the morphology of acinar cell carcinoma, which is the most common for this type of cancer. A public database maintained by the Fundação Oncocentro de São Paulo was used, where data was found from the year 2000 to 2020 on approximately 70 thousand patients with the studied morphology, belonging to multiple age groups and different clinical staging groups. The data preparation, exploratory analysis and pre-processing were made, as well as the application of computational predictive models, in order to verify their viability to assist clinical or public decisions.*

**Keywords**: Cancer, Data, Analysis, Dashboard, Prediction, Machine, Learning, XGBoost.

## 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 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. Segundo a The Lancet Comission, se ações não forem tomadas, o aumento iminente de casos de câncer de próstata causará uma subida rápida da taxa global de mortes pela enfermidade.

Diante de tal cenário, faz-se necessário o estudo de métodos de análise e predição que possam auxiliar tomadas de decisão públicas e clínicas. É possível aplicar modelos que estimam as chances de sobrevivência de um paciente utilizando informações disponíveis em bases de dados. Modelos de sobrevivência são amplamente utilizados para auxiliar decisões clínicas, contando com diversos métodos de aprendizado de máquina para obter predições de tempo até o evento quando alguns dados estão censurados (SURESH K., 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.

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 2000, bem como detalhes do estadiamento clínico, faixa etária, cirurgias realizadas e sobrevivência.

No artigo *Machine Learning for Predicting Survival of Colorectal Cancer Patients*(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; 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.

Tendo o método publicado por Cardoso et al como base, foi estabelecido como objetivo deste estudo treinar modelos para predição de sobrevivência utilizando dados de câncer de próstata disponíveis no site da FOSP (2022). Foram inclusos como objetivos os itens abaixo:

- Realizar análise exploratória e preparar dados públicos de câncer de próstata.

- 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.

- Verificar a viabilidade do uso do modelo preditivo.

O estudo colaborará com a validação tanto do método quanto da base de dados da Fundação Oncocentro de São Paulo para predições de sobrevivência.

## 2 Referencial Teórico

Segundo Suresh K. (2022), modelos de sobrevivência são utilizados em algum momento inicial de um paciente, tomando por exemplo a data do diagnóstico ou do início do tratamento, para descobrir a probabilidade de sobrevivência do mesmo a determinadas janelas de tempo. Médicos utilizam os resultados de tais modelos para tomar decisões clínicas, tais como ajustar o tempo de monitoramento ou a aplicação de diferentes terapias.

No artigo *Machine Learning for Predicting Survival of Colorectal Cancer Patients*, Cardoso et al. (2023) demonstrou ser possível utilizar uma das bases de dados públicos disponibilizada pela Fundação Oncocentro de São Paulo para aplicar modelos e prever sobrevivência de pacientes com câncer colorretal. Foram utilizados os modelos Naive Bayes (PEDREGOSA et al., 2011), Random Forest (HO, 1995) e XGBoost(CHEN; GUESTRIN, 2016) para classificar a sobrevida dos pacientes em 1, 3 e 5 anos, sendo estes três modelos também objeto deste estudo.

O modelo Naive Bayes utiliza o teorema de Bayes, que pode ser implementado de diferentes maneiras e é amplamente utilizado para classificação, isto é, o modelo tenta prever o rótulo correto de um determinado dado de entrada. Neste estudo, foi utilizada uma variação do algoritmo Naive Bayes representada pela Equação 1 abaixo, onde a probabilidade dos recursos é considerada gaussiana(PEDREGOSA et al., 2011):

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \tag{1}$$

Os dois parâmetros a seguir são estimados usando a máxima verossimilhança: $\sigma_y$ e $\mu_y$

O modelo Random Forest é um algoritmo de aprendizado de máquina que utiliza um conjunto (ou "floresta") de árvores de decisão para melhorar a precisão preditiva, (HO, 1995) sendo de rápido processamento. Um conjunto de árvores é um método que utiliza múltiplos classificadores de árvore de decisão em diferentes subamostras de um conjunto de dados e emprega a média dos resultados para aumentar a precisão nas previsões e reduzir o risco de sobreajuste. Uma árvore de decisão particiona recursivamente o espaço dos atributos para agrupar amostras que possuem rótulos idênticos ou valores-alvo similares.

O algoritmo XGBoost é utilizado para problemas de aprendizagem supervisionada, onde são usados dados de treinamento (com múltiplos recursos) para prever uma variável alvo. O XGBoost é um método que cria um conjunto de árvores de classificação e regressão (CART). Uma pontuação é atribuída a cada uma das folhas da árvore, diferente das tradicionais árvores de decisão que possuem somente valores de decisão. (CHEN; GUESTRIN, 2016). O algoritmo soma a previsão de múltiplas árvores juntas, como demonstrado na Equação 2:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F} \tag{2}$$

K representa o número de árvores, $f_k$ é uma função dentro do espaço funcional F, F é o conjunto de todas as Árvores de Classificação e Regressão (CARTs) possíveis.

O conjunto de dados onde serão aplicados os modelos acima é oriundo de uma base pública mantida pela Fundação Oncocentro de São Paulo (FOSP), sendo a mesma fonte do *dataset* utilizado no artigo *Machine Learning for Predicting Survival of Colorectal Cancer Patients*. A FOSP foi fundada em 1974, sendo uma instituição vinculada à Secretaria de Saúde do Governo de São Paulo com o intuito de prover assistência à oncologia e incentivar ensino e pesquisa, e também de estimular a detecção precoce do câncer e atividades de prevenção.(FOSP, 2022)

A base de dados mantida é atualizada a cada três meses com dados gerados por instituições coordenadas pela FOSP no estado de São Paulo, sendo estas mantenedoras do Registro Hospitalar de Câncer (RHC). O objetivo da Fundação Oncocentro ao disponibilizar tais dados é auxiliar profissionais da área da saúde a gerar análises específicas.(FOSP, 2022)

## 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.

Todos os pacientes estão anonimizados, não contendo nomes ou documentos. Os dados foram tratados de forma similar à adotada no artigo de Cardoso et al. (2023), 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 2000 e 2023. Tal morfologia é o objeto deste estudo, que tem foco nos pacientes registrados de 2000 a 2020.

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

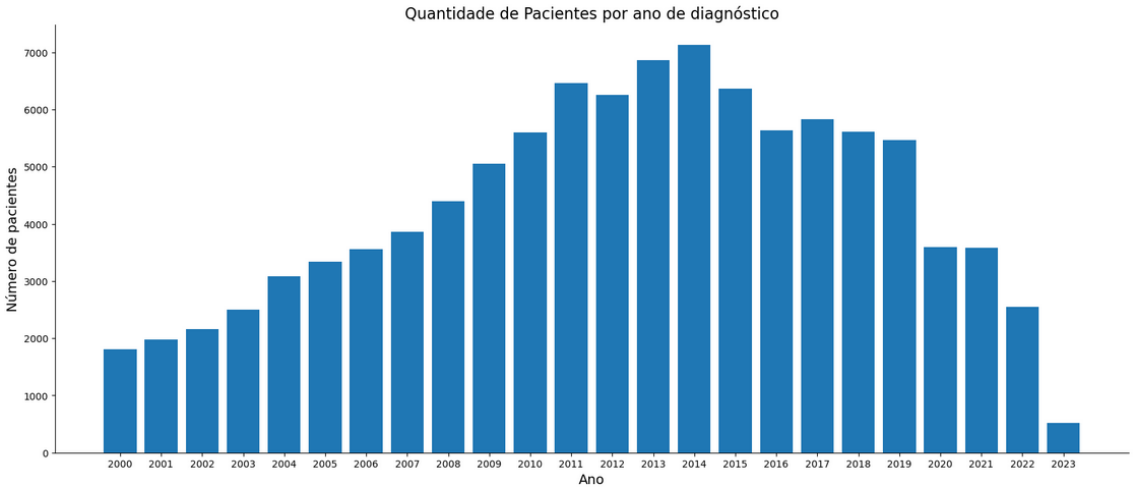


Figura 1 – Quantidade de diagnosticos por ano

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

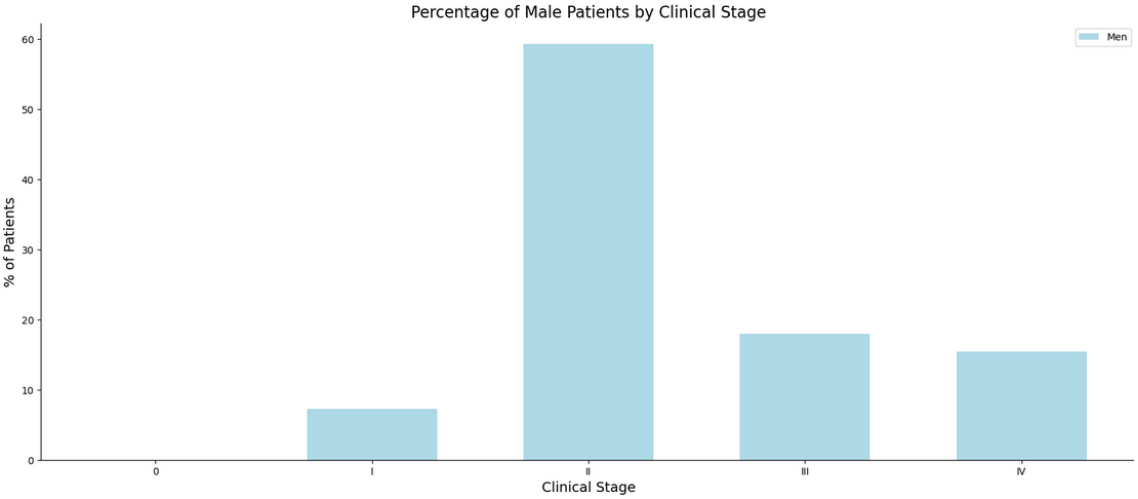


Figura 2 – Porcentagem de pacientes por grupo de estadiamento clínico

A faixa etária dos pacientes está, majoritariamente, acima dos 60 anos, composta por aproximadamente 0.8 dos pacientes. Os diagnósticos começam a partir dos 30 anos, com uma acentuação das detecções aos 40, como pode ser visto na figura 3.

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 de dados, não sendo aproveitáveis pelos algoritmos.

Foram deletadas as colunas abaixo por diversos motivos. Algumas possuem muitos dados faltantes (figura 4), 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

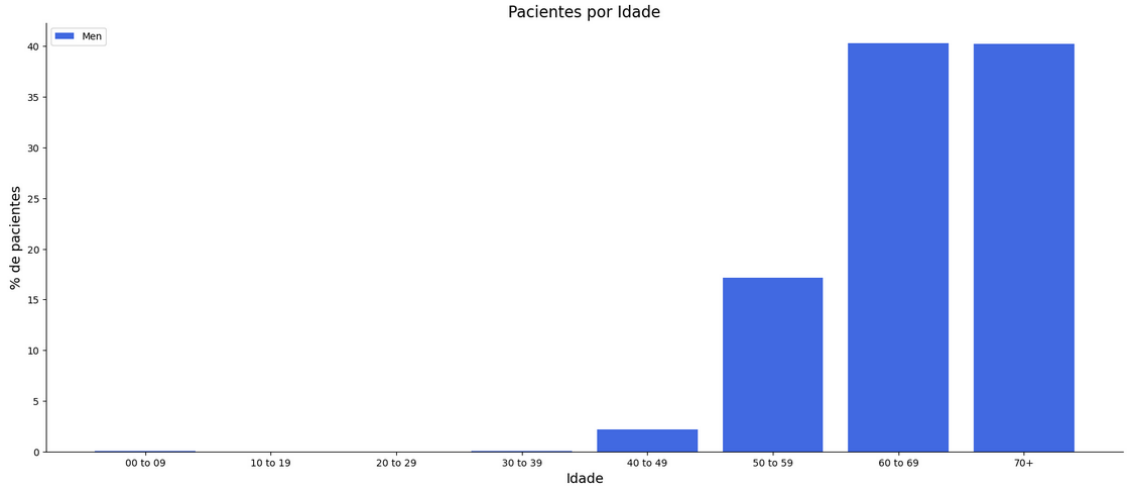


Figura 3 – Porcentagem de pacientes por faixa etária

morfologia, ou o campo ECGRUP, que possui as mesmas informações contidas no campo EC. Foram removidos os campos G, Gleason e PSA, pois tais parâmetros estão relacionados ao estadiamento clínico do paciente. Como o campo EC já está sendo utilizado, a remoção dos três campos citados aumentou a precisão do modelo por volta de 1 ponto percentual. Já os campos HABIT11, HABILIT1, e HABILIT2, por exemplo, não continham informações relacionadas ao 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', 'G', 'GLEASON' e 'PSA'.

Os atributos a seguir foram excluídos pois demonstraram importância baixa ou nula:

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

Na Tabela 1, os atributos do dataset escolhidos para o estudo:

Seguindo a linha de 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, Todos os pacientes cuja escolaridade não foi informada foram incluídos no estudo. Esta abordagem difere da utilizada por Cardoso et al., que utilizou aprendizado de máquina para estimar a escolaridade destes pacientes.

Ao final da preparação dos dados, restaram 56874 registros de pacientes que foram utilizados para treinamento dos modelos utilizando os algoritmos Naive Bayes, Random

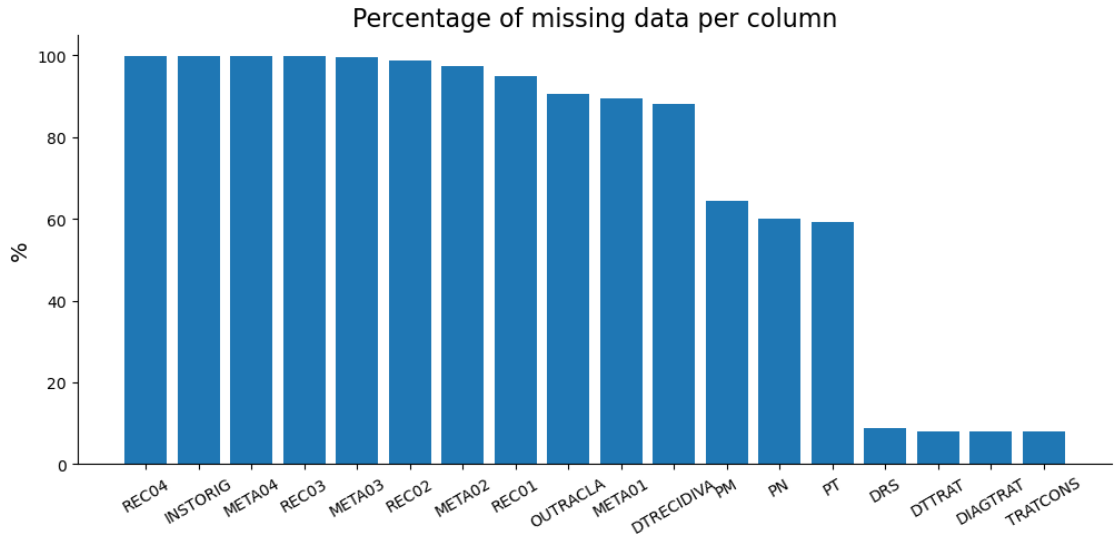


Figura 4 – Colunas com maior quantidade de dados faltantes

Forest e XGBoost. Foram feitas predições de sobrevivência para acima de 1,3 e 5 anos, bem como óbito geral e óbito por câncer.

Todas as modificações feitas no dataset bem como o dicionário de dados, o notebook Jupyter contendo a análise exploratória e a preparação dos dados, e os resultados dos algoritmos de predição 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.*

Cada modelo foi treinado com 75 por cento dos dados, sendo os 25 por cento restantes utilizados para testar o mesmo. Foi aplicado o SMOTE para balancear as classes durante o treino, que utiliza os k vizinhos mais próximos para gerar registros para a classe com menor número de pacientes. Para a fase de testes, foram utilizados apenas dados reais, sendo o número de óbitos menor do que o de sobreviventes em todos os cenários.

Tabela 1 – Descrição das Variáveis do Conjunto de Dados

| **Campo** | **Descrição** |
|---|---|
| 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 | Estadiamento clínico |
| TRATHOSP | Combinação dos tratamentos realizados no 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 entre datas de consulta e o diagnóstico |
| TRATCONS | Diferença em dias entre datas de consulta e tratamento |
| DIAGTRAT | Diferença em dias entre 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 |

# 4 Resultados e discussão

Nas matrizes de confusão dos testes de 1, 3 e 5 anos de sobrevivência, a classe 0 indica óbito e 1 indica sobrevivência. As matrizes estão mostrando valores percentuais ao invés de quantidade de pacientes.

## 4.1 Predições de Um Ano ou Mais de Sobrevivência

Para testar as predições de 1 ano de sobrevivência, foi utilizada a seguinte quantidade de pacientes: 478 óbitos (classe 0) e 12862 sobreviventes (classe 1).

### 4.1.1 Naive Bayes - 1 Ano

O algoritmo Naive Bayes, apesar de acertar a maioria dos óbitos (classe 0), demonstrou um valor de F1-Score baixíssimo para a classe 1, errando 0.9 das predições da classe

(figura 5). O comportamento foi semelhante nos testes das outras predições abordadas por este estudo, logo este algoritmo não será mais explorado.

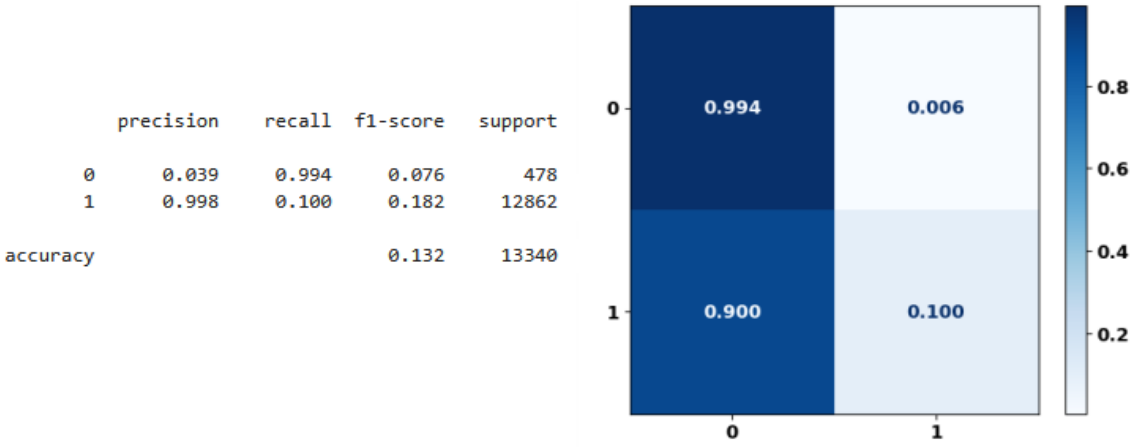


Figura 5 – Naive Bayes - 1 Ano

### 4.1.2 Random Forest e XGBoost - 1 Ano

O Random Forest mostrou um desempenho mais equilibrado em comparação ao Naive Bayes, com uma acurácia de 0.767 (figura 6). O XGBoost teve desempenho similar, figura 7.

O Random Forest demonstrou 0.934 AUC no treinamento e 0.849 no teste, o que sugere baixo *overfitting*. Já a AUC do XGBoost ficou em 0.986 no treino e 0.841 no teste, o que evidencia um sobreajuste maior. Isto não impactou os resultados, pois o XGBoost lida melhor com *overfitting* do que o Random Forest.

As importâncias dos principais atributos utilizados destacam a participação majoritária do Estadiamento Clínico e DIAGTRAT (diferença em dias entra as datas de tratamento e diagnóstico), demonstrado nas figuras 8 e 9.

Não houve melhora nos teste quandos utilizada a otimização de hiperparâmetros.

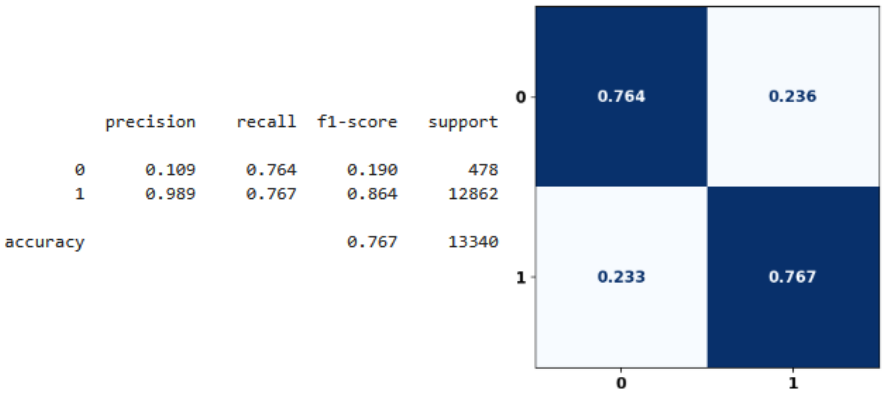


Figura 6 – Random Forest - 1 Ano

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.109 | 0.768 | 0.191 | 478 |
| 1 | 0.989 | 0.767 | 0.864 | 12862 |
| accuracy | | | 0.767 | 13340 |

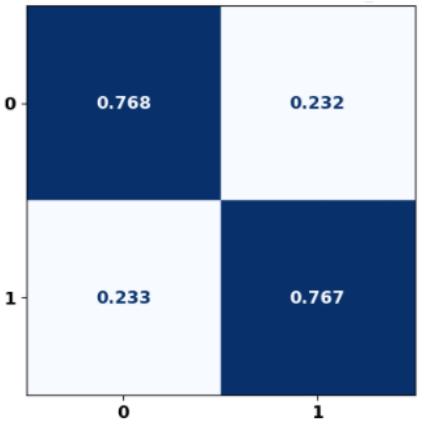


Figura 7 – XGBoost - 1 Ano

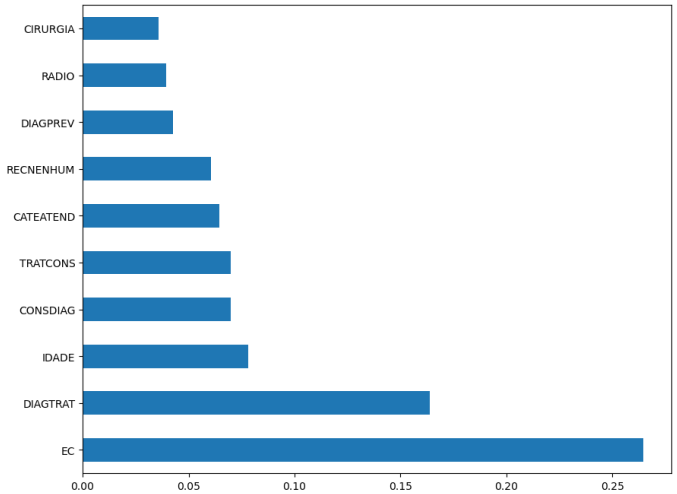


Figura 8 – Importância dos atributos - Random Forest 1 Ano

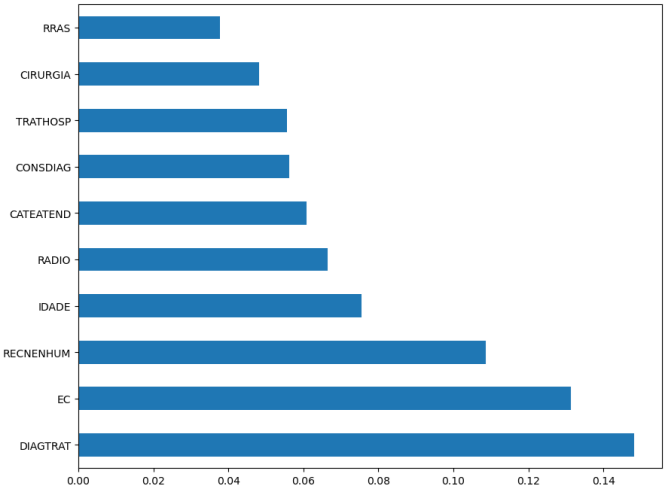


Figura 9 – Importância dos atributos - XGBoost - 1 Ano

## 4.2 Predições de Três Anos ou Mais de Sobrevivência

Para as predições de sobrevivência de 3 anos, há 1606 óbitos (classe 0) e 9916 sobreviventes (classe 1).

A acurácia do Random Forest é um pouco menor do que a demonstrada pelo mesmo algoritmo na predição de um ano. O modelo ficou equilibrado em 0.74 nas classes 0 e 1, abaixo dos 0.76 registrados pelo mesmo algoritmo anteriormente (figura 10). O algoritmo XGBoost teve um desempenho superior, atingindo acurácia um pouco maior que 0.75 (figura 11).

Ambos algoritmos demonstraram overfitting menor em relação às predições de 1 ano, sendo AUC de 0.889 no treinamento e 0.814 durante o teste do Random Forest, 0.959 e 0.823 para treinamento e teste, respectivamente, do XGBoost.

Nos cenários de sobrevivência de três anos, o atributo EC (estadiamento clínico) demonstrou ser mais determinante do que o atributo DIAGTRAT (diferença em dias entre o diagnóstico e início do tratamento), e o atributo IDADE também teve uma relevância maior (figuras 12 e 13).

A otimização de hiperparâmetros não mostrou resultados satisfatórios neste cenário.

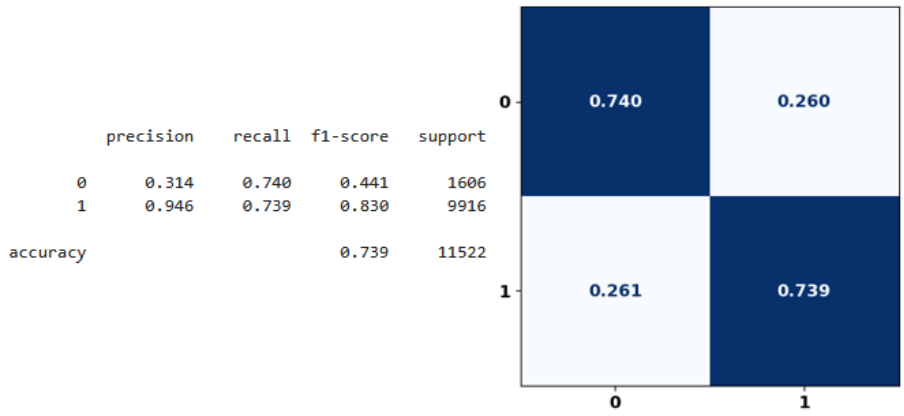


Figura 10 – Random Forest - 3 Anos

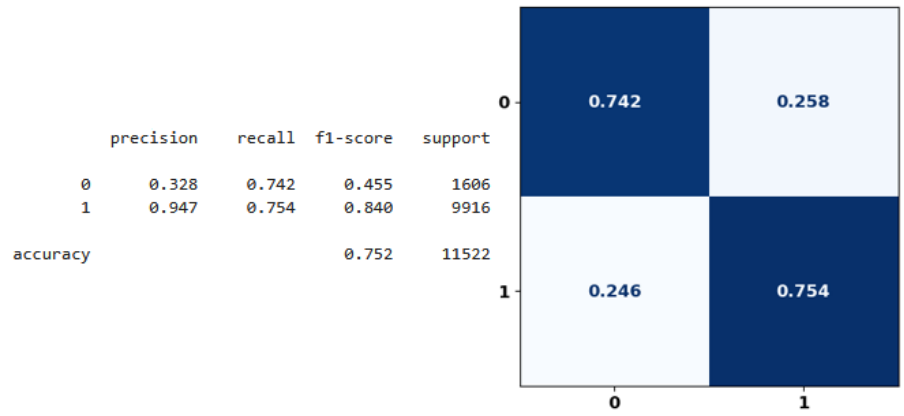


Figura 11 – XGBoost - 3 Anos

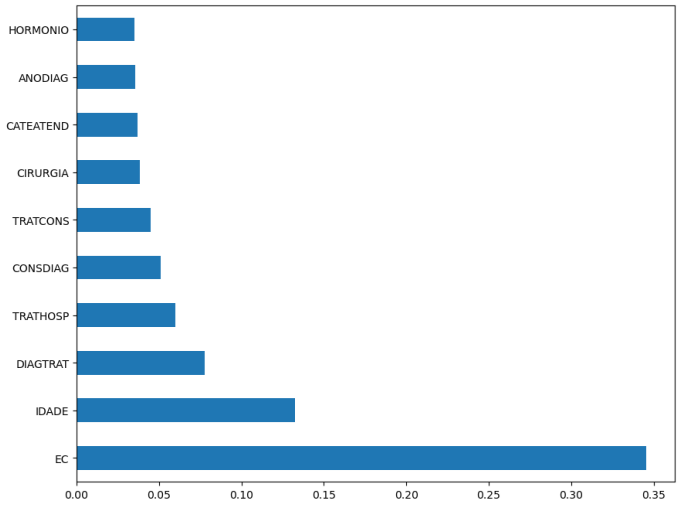


Figura 12 – Importâncias Random Forest - 3 Anos

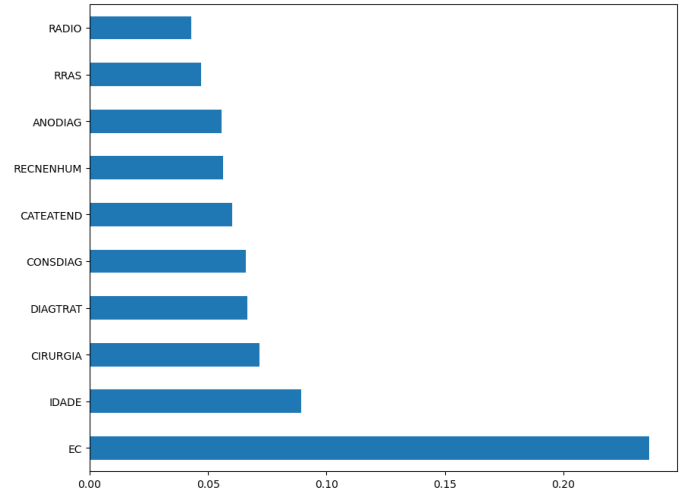


Figura 13 – Importâncias XGBoost - 3 Anos

## 4.3 Predições de Cinco Anos ou Mais de Sobrevivência

Para os testes de sobrevivência de cinco anos, foram utilizados 2492 registros de óbito e 7341 de sobrevivência.

O Random Forest no cenário de cinco anos de sobrevivência foi o único algoritmo que demonstrou leve melhora quando utilizada a otimização de hiperparâmetros (figura 14), aumentando em quase 0.15 as previsões da classe 1 e atingindo uma acurácia de 0.744. O XGBoost novamente é o algoritmo com melhor desempenho, com uma acurácia de 0,746 (figura 15).

É demonstrado um overfitting mais acentuado neste cenário para o Random Forest, com a AUC do treino 0.980 e teste em 0.816, enquanto o do XGBoost diminuiu em relação à predição de três anos (0.930 e 0.822).

Os parâmetros mais importantes após a otimização foram EC, IDADE e ANODIAG, respectivamente (figuras 16 e 17).

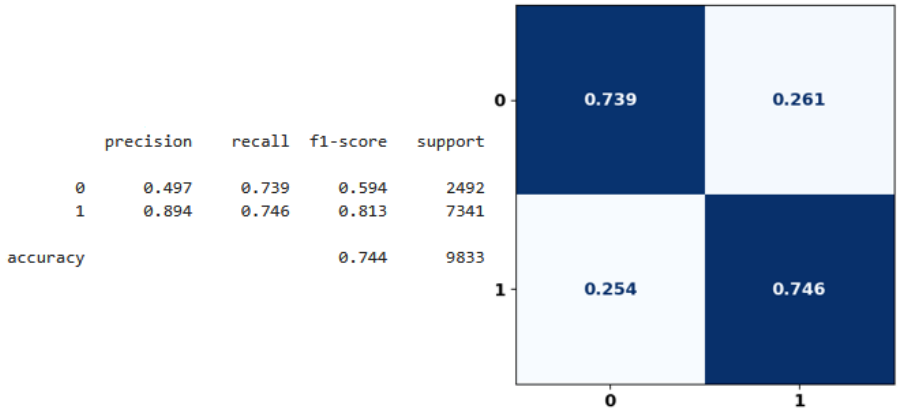


Figura 14 – Random Forest - 5 Anos

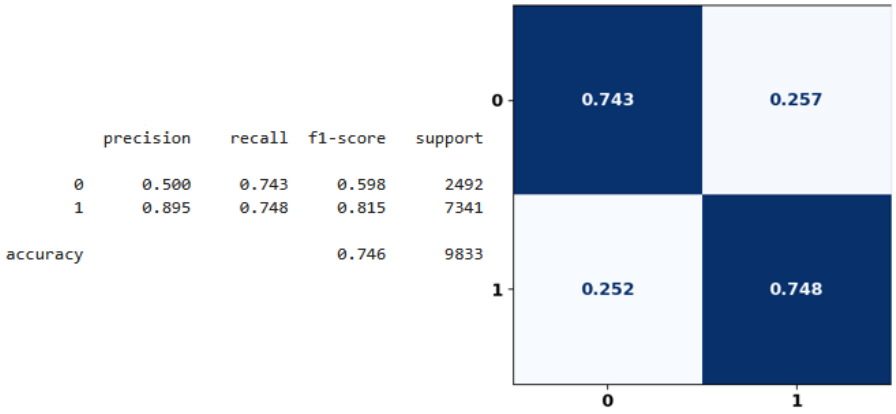


Figura 15 – XGBoost - 5 Anos

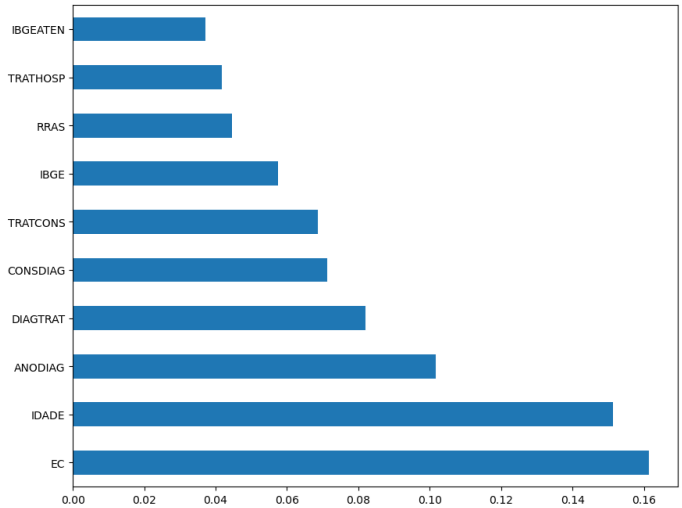


Figura 16 – Importâncias Random Forest - 5 Anos

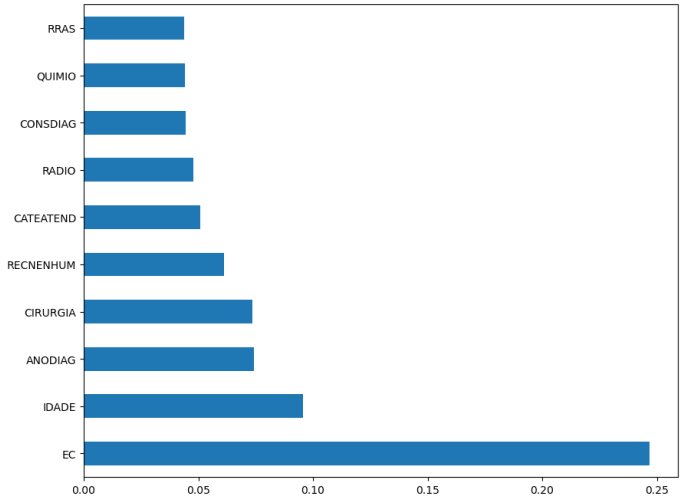


Figura 17 – Importancias XGBoost - 5 Anos

## 4.4 Óbito Geral

Nas previsões de óbito geral, 0 indica sobrevivência e 1 indica óbito. Foram utilizados 9639 registros de sobrevivência e 4580 registros de óbito para testes.

O Random forest demonstrou acurácia de 0.760 (figura 18), enquanto o XGBoost chegou a 0.765 (figura 19). O Random Forest teve o ano do diagnóstico como atributo mais importante (figura 20), enquanto para o XGBoost o atributo estadiamento clínico teve maior importância(figura 21).

A área sob a curva do Random Forest foi de 0.878 para treinamento e 0.839 para teste. Já o XGBoost alcançou 0.917 em treinamento e 0.847 em teste, sendo a predição que menos demonstrou overfitting para o algoritmo.

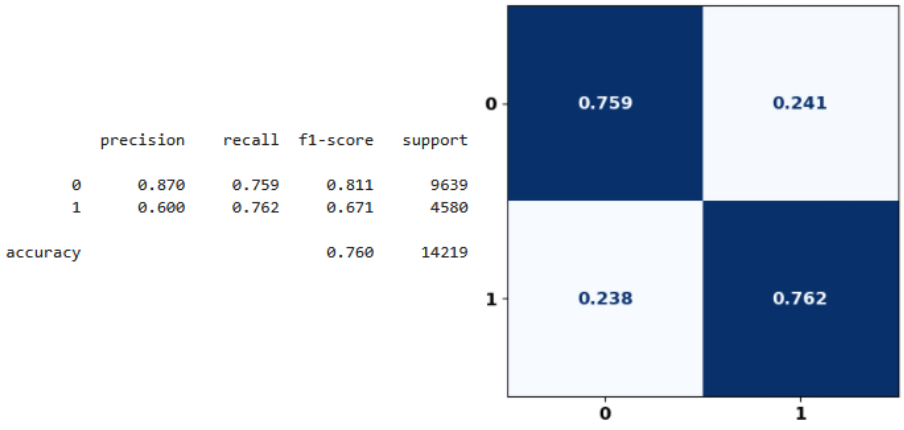


Figura 18 – Random Forest - Óbito Geral

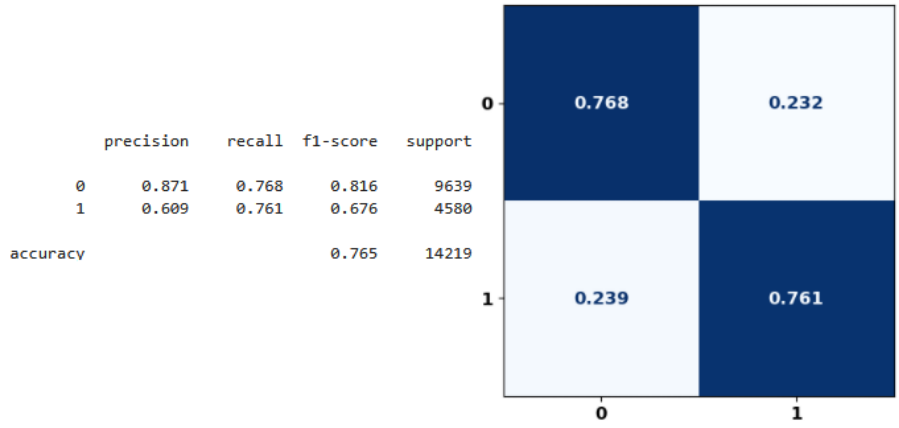


Figura 19 – XGBoost - Óbito Geral

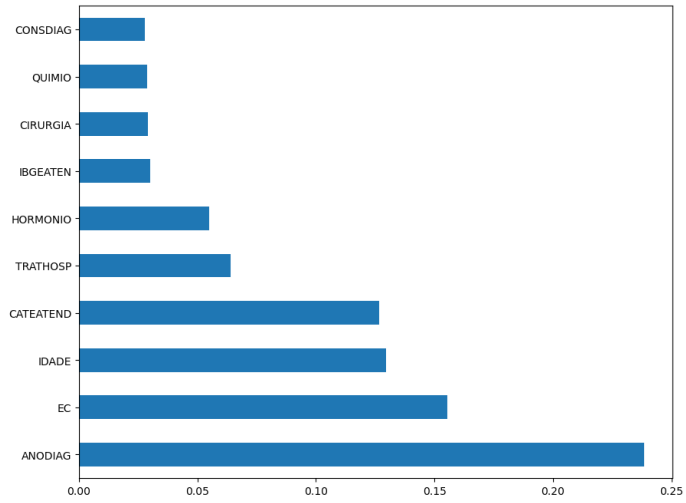


Figura 20 – Random Forest - Importâncias - Óbito Geral

## 4.5 Óbito por Câncer

Nas previsões de óbito por câncer, 0 indica sobrevivência e 1 indica óbito. Foram utilizados 12301 registros de sobrevivência e 1918 registros de óbito.

O algoritmo Random Forest alcançou uma acurácia de 0.813 (figura 22), e o XGboost obteve uma acurácia de 0.815 (figura 23), desempenho consideravelmente superior do algoritmo quando comparado aos cenários anteriores.

A área sob a curva de treinamento e teste do Random Forest foi de 0.928 e 0.888, respectivamente. O XGBoost demonstrou 0.974 de AUC em treinamento e 0.887 em teste.

No Random Forest, os atributos de estadiamento clínico, tratamentos realizados no hospital e hormonioterapia tiveram a maior importância, como pode ser visto na figura 24, enquanto no XGBoost o fato de haver ou não recidiva foi mais importante do que os três atributos citados anteriormente (figura 25) .

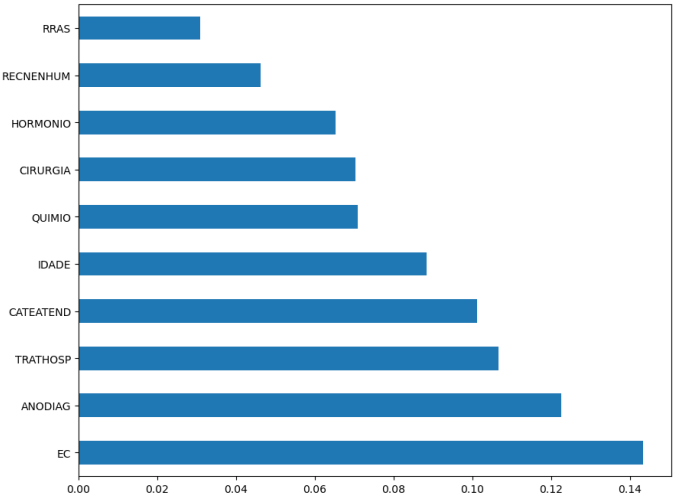


Figura 21 – XGBoost - Importâncias - Óbito Geral

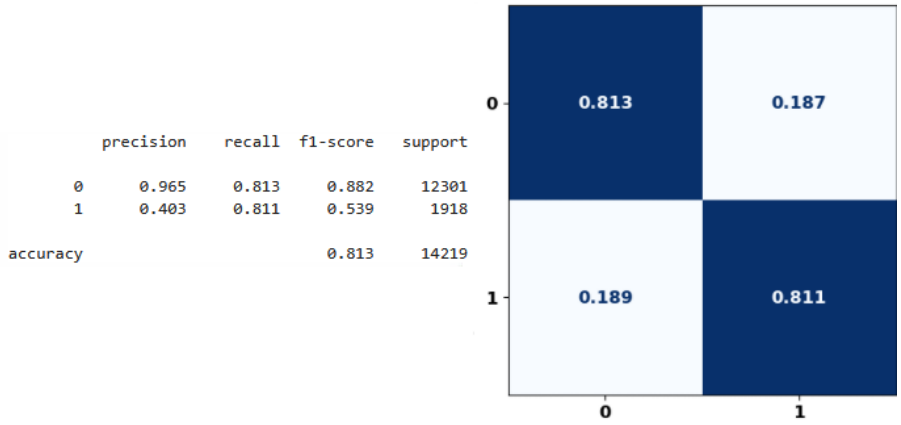


Figura 22 – Random Forest - Óbito por Câncer

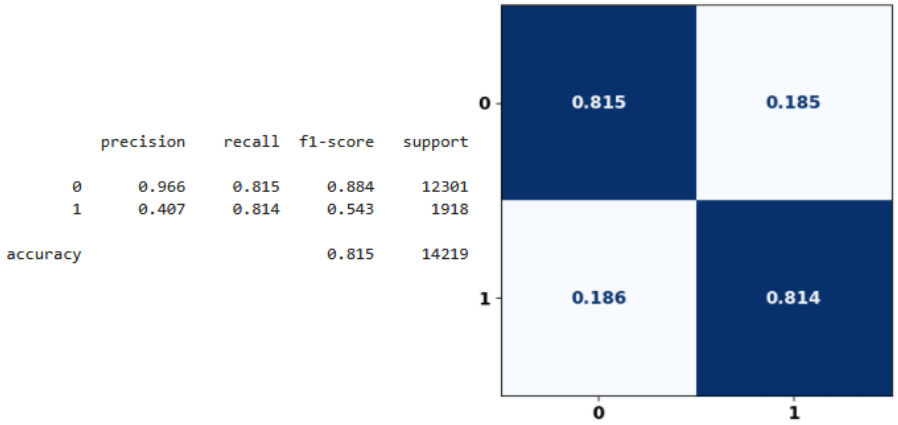


Figura 23 – XGBoost - Óbito por Câncer

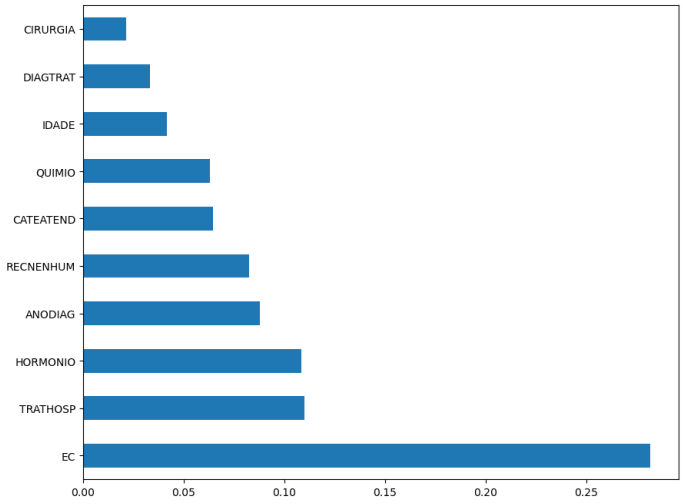


Figura 24 – Random Forest - Importâncias - Óbito por Câncer

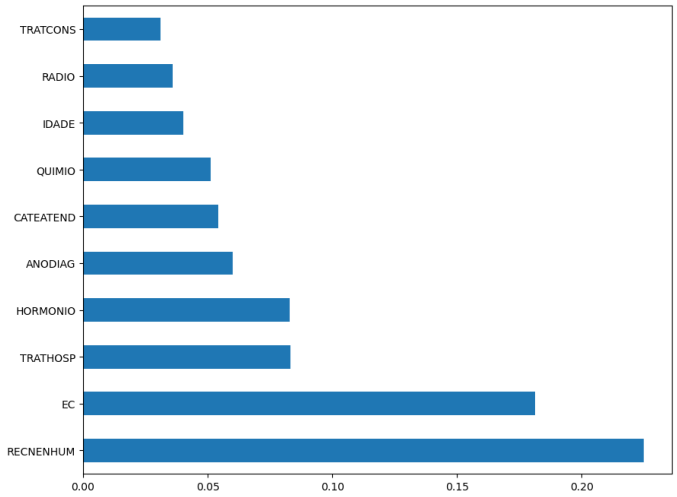


Figura 25 – XGBoost - Importâncias - Óbito por Câncer

## 4.6 Discussão

O Naive Bayes, apesar de ser um algoritmo eficiente para classificação de textos e problemas simples, mostrou limitações quando aplicado aos cenários de predição de sobrevivência do modelo. Este cenário foi praticamente o mesmo visto no artigo de Cardoso et al. (2023), e por este motivo o algoritmo não foi mais explorado durante o estudo.

Em contrapartida, o XGBoost obteve os resultados mais precisos do estudo em todos os cenários, tendo mostrado que a técnica de *boosting* criando múltiplas árvores, cada uma corrigindo os erros da anterior, é mais adequada para o modelo do que o Random Forest.

Foi observado que, à medida que o tempo de sobrevivência previsto aumenta, as variáveis preditivas se tornaram mais incertas, dado que a acurácia diminuiu quando o cenário mudou de 1 para 3 e 5 anos. Outra área que demonstrou mudança com o aumento do tempo de sobrevivência foi a otimização de hiperparâmetros, que passou a ter efeito

positivo na análise de 5 anos ou mais de sobrevivência quando aplicada ao Random Forest.

O modelo atingiu acurácia de 0.765 para óbito geral e 0.815 para óbito por câncer. Este último representa um salto de performance, sendo superior a algoritmos mais tradicionais no âmbito da predição de sobrevivência.

Para fins de comparação, o algoritmo *c-statistic* utilizado no artigo *PREDICT: model for prediction of survival in localized prostate cancer*(LG MONNINKHOF EM, 2016), atingiu acurácia de 0.78 na predição de sobrevivência específica de câncer, e 0.68 para sobrevivência geral. Já no artigo *DREAM challenge*, promovido pela The Lancet Commission(GUINNEY JUSTINABDALLAH, 2022), o melhor algoritmo explorado calcula sobrevivência global utilizando ePCR (regressão de Cox), demonstrando acurácia de 0.791. Outro artigo comparável é o *Survival analysis of localized prostate cancer with deep learning* (DAI X., 2022), que atingiu acurácia de mortalidade de câncer de próstata de 0.85, 0.80 e 0.76 para 2, 5 e 10 anos, respectivamente, utilizando RDSM.

## 5 Conclusão

Os resultados do modelo estão próximos dos demonstrados no artigo de Cardoso et al. (2023), tendo as predições de óbito por câncer atingido acurácia superior às obtidas no artigo de referência, o que evidencia uma validação empírica dos procedimentos adotados.

O modelo produziu resultados satisfatórios quando utilizados os algoritmos XGBoost e Random Forest. Destaca-se o fato de o modelo estudado utilizar uma base de dados pública, com menos atributos clínicos, demonstrando resultados comparáveis aos de pesquisas financiadas por grandes laboratórios, que utilizam bases de dados privadas e possuem maior variabilidade de atributos. Isto demonstra que modelos preditivos eficazes para a sobrevivência de câncer de próstata podem ser desenvolvidos utilizando bases de dados públicas, servindo como alternativa às pesquisas que enfrentam limitações no acesso a dados privados.

O fato do modelo utilizar dados públicos abre possibilidades de políticas públicas mais eficientes. Futuros estudos podem tirar proveito do modelo, como predições utilizando outras morfologias de câncer ou a possível aplicação do modelo em outras regiões do país, coletando mais dados e possibilitando treinamento mais extenso dos algoritmos, a fim de obter maior precisão.

## Referências


CARDOSO, L. B. et al. *Machine learning for predicting survival of colorectal cancer patients*. Springer Science and Business Media LLC, 2023. Disponível em: http://dx.doi.org/10.1038/s41598-023-35649-9. Acesso em:5 mai.2024. Disponível em: <http://dx.doi.org/10.1038/s41598-023-35649-9>.

CHEN, T.; GUESTRIN, C. *XGBoost: A Scalable Tree Boosting System*. New York, NY, USA: ACM, 2016. 785–794 p. Disponível em: https://xgboost.readthedocs.io/en/stable/tutorials/model.html. (KDD '16). Acesso em: 20 out.2024. Disponível em: <http://doi.acm.org/10.1145/2939672.2939785>.

DAI X., P. J. Y. S. e. a. *Survival analysis of localized prostate cancer with deep learning*. 2022. Acesso em: 11 mai.2025. Disponível em: <https://www.nature.com/articles/s41598-022-22118-y>.

FERLAY, J. et al. *Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012*. 2015. E359-E386 p. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1002/ijc.29210. Acesso em: 05 mai.2024. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/ijc.29210>.

FOSP. *Banco de Dados do RHC*. 2022. Acesso em:5 mai.2024. 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>.

GUINNEY JUSTINABDALLAH, K. e. a. *Prediction of overall survival for patients with metastatic castration-resistant prostate cancer: development of a prognostic model through a crowdsourced challenge with open clinical trial data*. 2022. Acesso em: 11 abr.2025. Disponível em: <https://www.thelancet.com/journals/lanonc/article/PIIS1470-2045(16)30560-5/abstract>.

HO, T. K. Random decision forests. In: IEEE. *Proceedings of 3rd international conference on document analysis and recognition*. [S.l.], 1995. v. 1, p. 278–282.

INCA. *Estimativa 2023: Incidência de Câncer no Brasil*. Ministério da Saúde, 2022. Disponível em:https://www.inca.gov.br/publicacoes/livros/estimativa-2023-incidencia-de-cancer-no-brasil. Acesso em: 16 out.2024. Disponível em: <https://www.inca.gov.br/publicacoes/livros/estimativa-2023-incidencia-de-cancer-no-brasil>.

JAMES, N. D. e. a. *The Lancet Commission*. Elsevier Ltd, 2024. Disponível em: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(24)00651-2/fulltext". Acesso em: 23 out.2024. Disponível em: <https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(24)00651-2/fulltext>.

LG MONNINKHOF EM, v. O. I. v. d. P. H. d. M. G. v. V. M. K. *PREDICT: model for prediction of survival in localized prostate cancer*. 2016. Acesso em: 27 mar. 2025. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/26420595/>.

PEDREGOSA, F. et al. *Scikit-learn: Machine Learning in Python*. 2011. 2825–2830 p. "https://scikit-learn.org/1.5/modules/naive$_b$*ayes.html*".*Acessoem* : 27*out*.2024.

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