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

Alcir Canella Filho[2], Felipe Cle´ Monteiro[2], Matheus do Nascimento Marques[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

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. In this initial phase of the study, preparation, exploratory analysis and pre-processing of data were completed to enable the application of computational predictive models, with the aim of validating their accuracy to assist clinical or public decisions.*

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. Nesta fase do estudo, foram concluídas a preparação, a análise exploratória, o pré-processamento dos dados e a aplicação de modelos preditivos computacionais, para auxiliar decisões clínicas ou públicas.*

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

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. Segundo a The Lancet Comission, se ações não forem tomadas, o aumento iminente de casos de câncer de próstata causaria 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 governamentais e clínicas. É possível aplicar modelos que estimam as chances de sobrevivência de um paciente utilizando as 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 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*[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

Segundo [Suresh 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. Os médicos, então, 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*, Buk Cardoso et al. demonstrou ser poss´ıvel utilizar uma das bases de dados públicos disponibilizada pela Fundação Oncocentro de São Paulo [FOSP 2022]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 and 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, será utilizada uma variação do algoritmo Naive Bayes representada pela fórmula abaixo, onde probabilidade dos recursos éconsiderada gaussiana[Pedregosa et al. 2011]:

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

Figura 1. Naive Bayes Gaussiano

Os dois parâmetros a seguir são estimados usando a máxima verossimilhança: σ_y e μ_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. Abaixo, um exemplo do uma implementação de árvores de decisão do artigo de [Buk Cardoso et al. 2023]

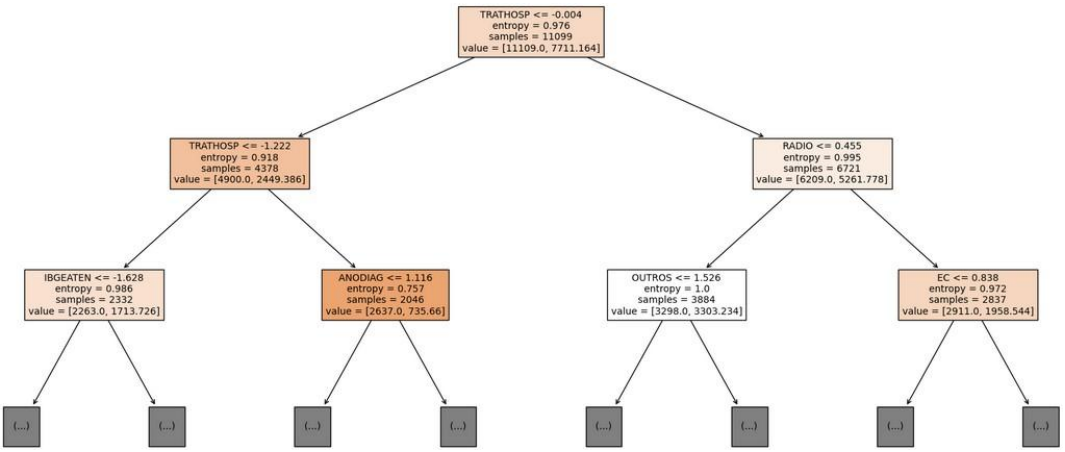


Figura 2. Random Forest

O algoritmo XGBoost é usado 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 and Guestrin 2016]. O modelo somará a previsão de múltiplas árvores juntas, como pode ser visto na figura 3

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

Figura 3. Fórmula de árvores XGBoost

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 mantenedo-

ras 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 [Buk 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 foi no ano de 2014, com aproximadamente 7 mil pacientes diagnosticados, como pode ser visto na Figura 4.

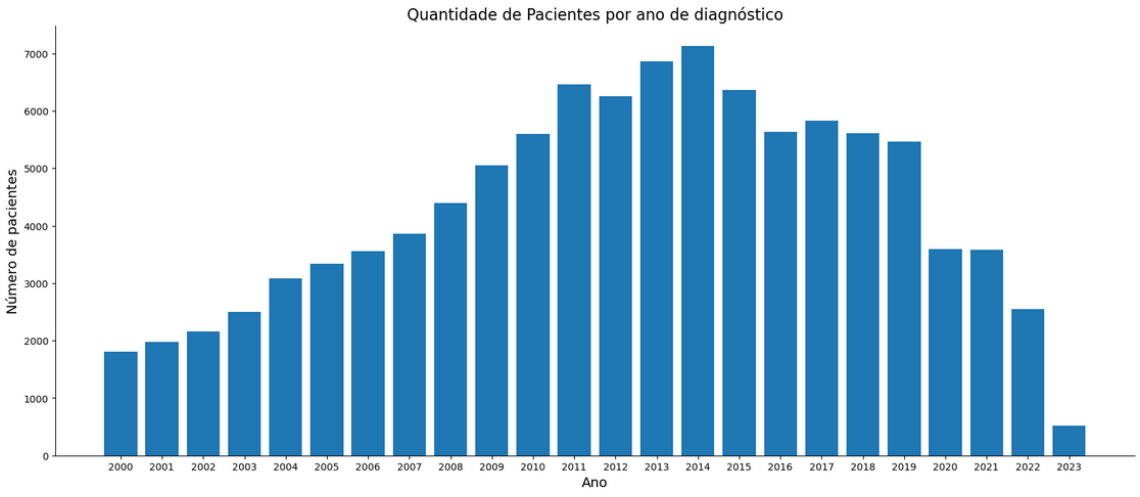


Figura 4. Quantidade de diagnosticos por ano

A maioria dos pacientes está classificada no nível 2 do atributo estadiamento clínico (ver Figura 5). Este campo determina a extensão do câncer pelo corpo do paciente, bem como é utilizado para tomada de decisão em relação a tratamentos.

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.

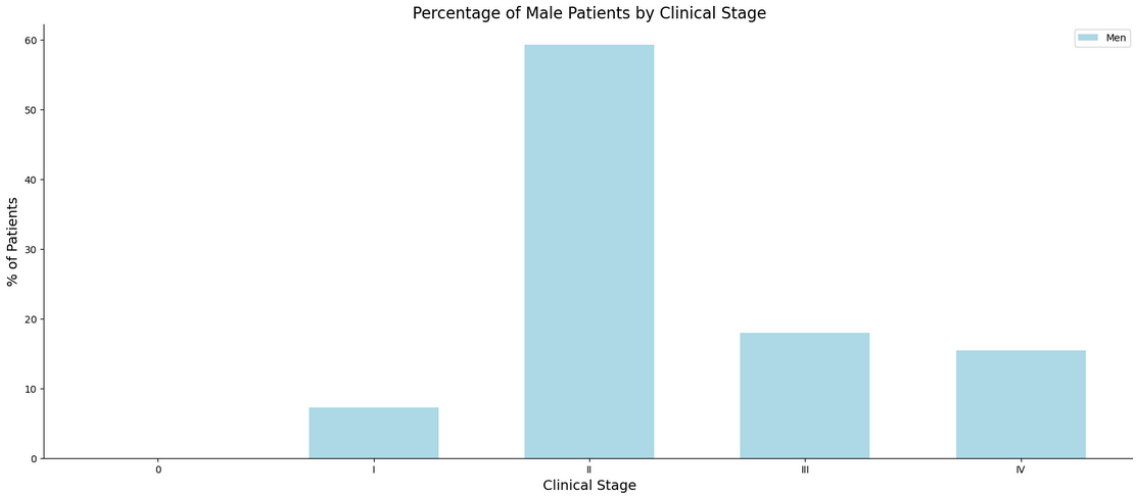


Figura 5. Porcentagem de pacientes por grupo de estadiamento cl´ınico

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’,

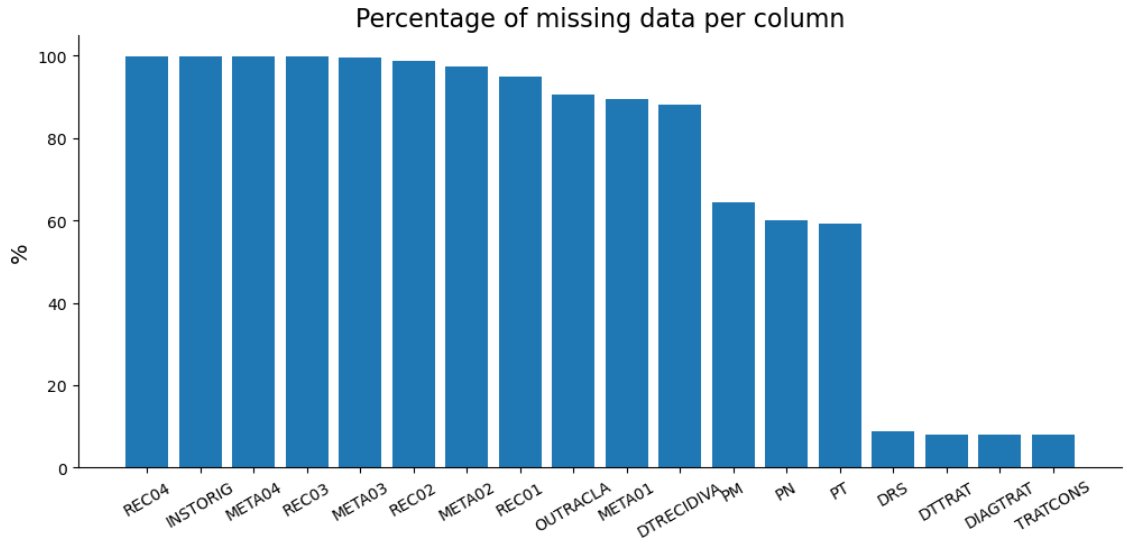


Figura 6. Colunas com maior quantidade de dados faltantes

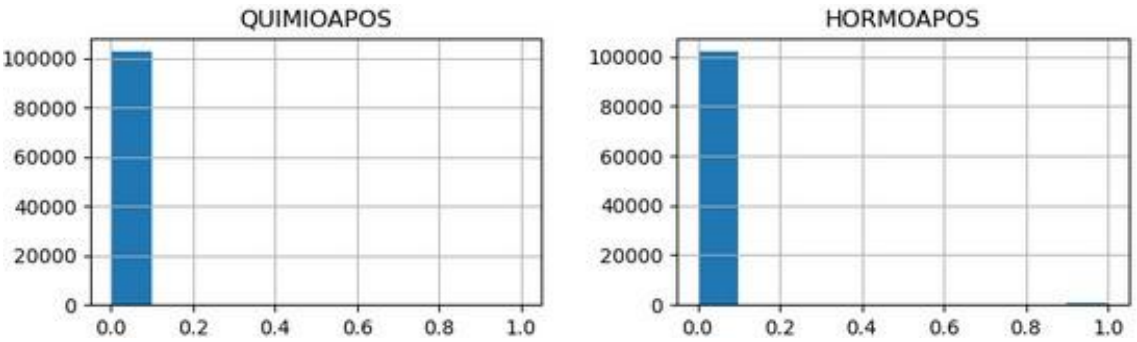


Figura 7. Exemplos de atributos com os mesmos valores para todos os pacientes

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 f ısico 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.

Para este trabalho, foram treinados e avaliados três modelos diferentes. Na exploração de dados foram criadas novas colunas que dizem respeito a sobrevivencia ou não dos pacientes em determinadas faixas de tempo, a partir do diagnóstico. São elas: vivo_ano1, vivo_ano3 e vivoano_5, o dataset de exploração de dados entra em mais detalhes de cada uma delas. Neste contexto vamos trabalhar com os modelos para prever a coluna vivo_ano5.

Antes da aplicação de cada modelo, mais etapas de pre-processamento são realizadas a partir de funções importadas do arquivo “functions.py”.Estas etapas podem incluir:

- Filtragem por intervalo de anos, no caso a coluna vivo_ano5 que escolhemos.
- Divisão Treinamento/teste
- Codificação e Normalização dos dados de treino e teste
- Balanceamento: Usa SMOTE para balancear os dados de treinamento.
- Seleção de features,etc...

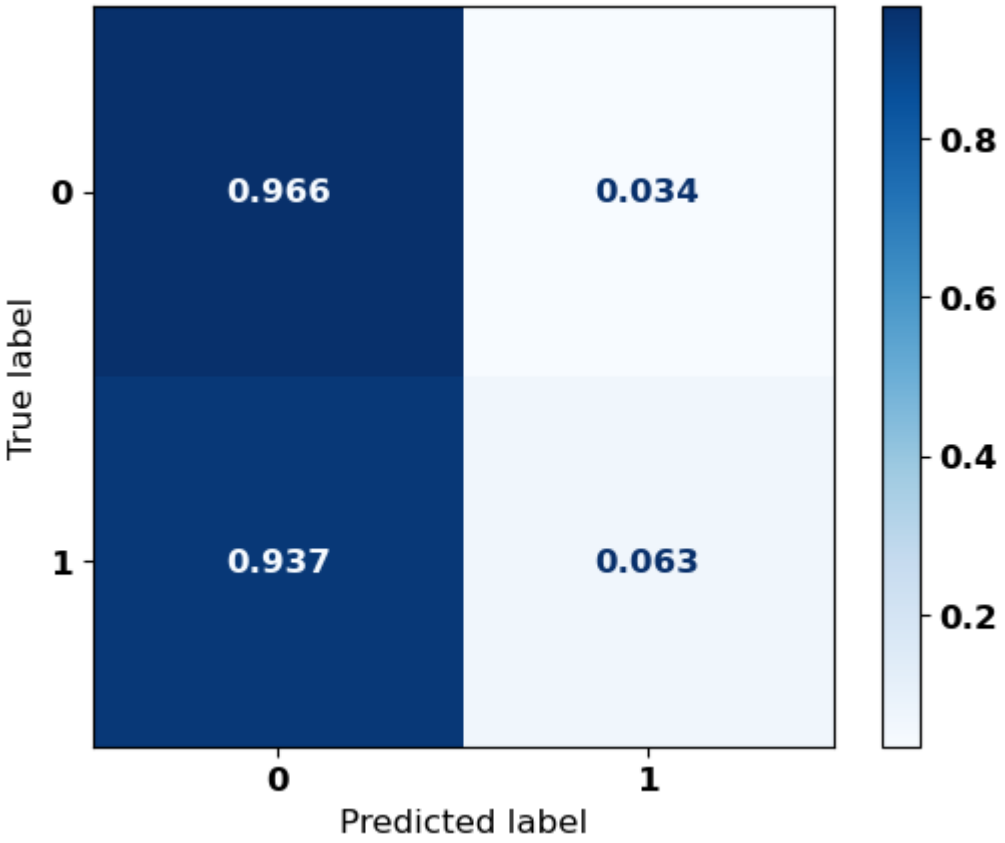
Para cada modelo, Naive Bayes, Random Forest e XGBoost realizamos o treinamento com base nos dados selecionados da coluna vivo_ano5. Para validação utilizamos matriz de confusão, curva ROC para conjunto de treino e teste e também uma análise final das features mais importantes, em alguns casos com a biblioteca e o gráfico SHAP. Também utilizamos a biblioteca Optuna do python para busca por melhores hiperâmetros nos modelos Random Forest e XGBoost.

4. Resultados

Os resultados para cada modelo estão apresentados pós-otimização de hiperparâmetros, no notebook temos o processo completo:

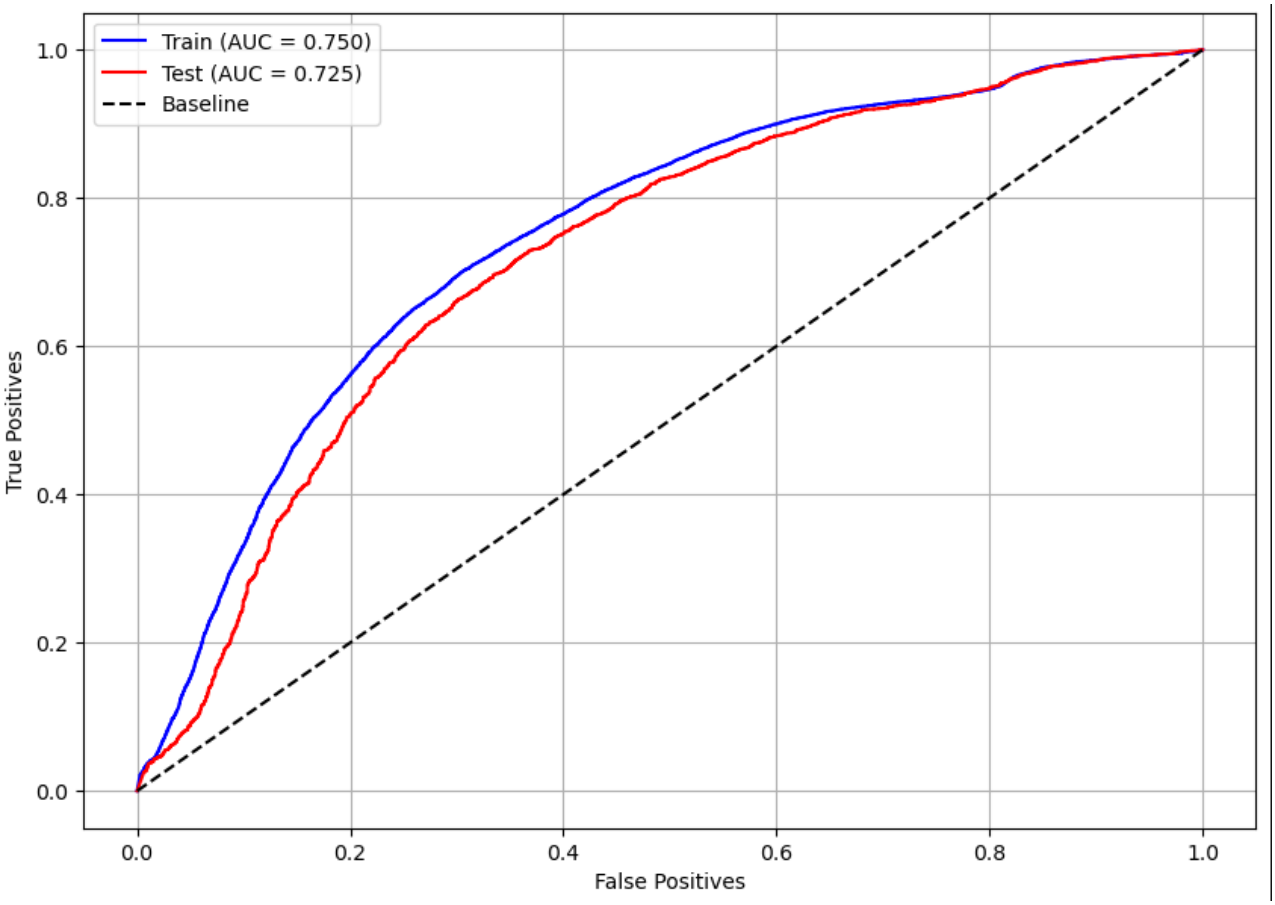
-Naive Bayes:

Matriz de confusão:



	precision	recall	f1-score	support
0	0.267	0.966	0.418	2059
1	0.841	0.063	0.118	5844
accuracy			0.298	7903
macro avg	0.554	0.515	0.268	7903
weighted avg	0.691	0.298	0.196	7903

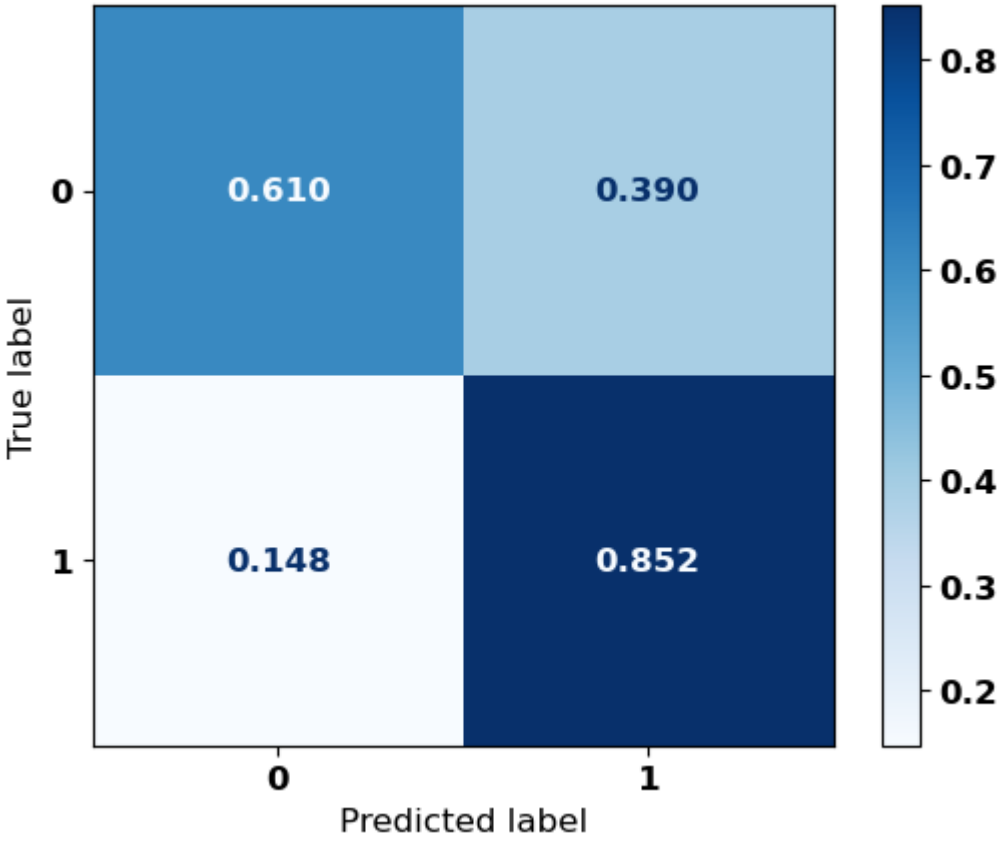
Curva ROC:



Pelas curvas não houve overfitting, mas este modelo se apresentou inadequado para previsão da classe 1, desta forma, decidimos que qualquer análise além destas não possuem tanta relevância.

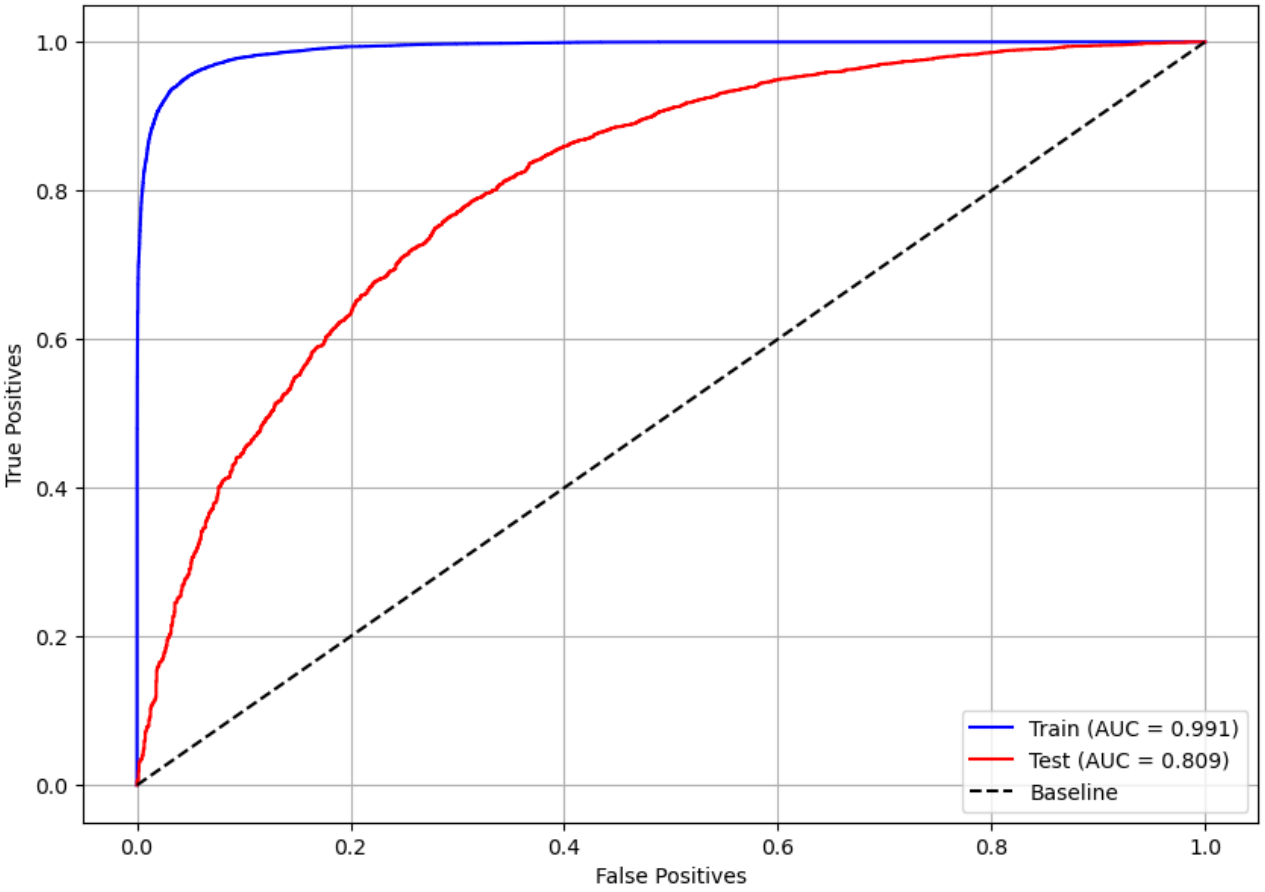
-Random Forest:

Matriz de confusão:

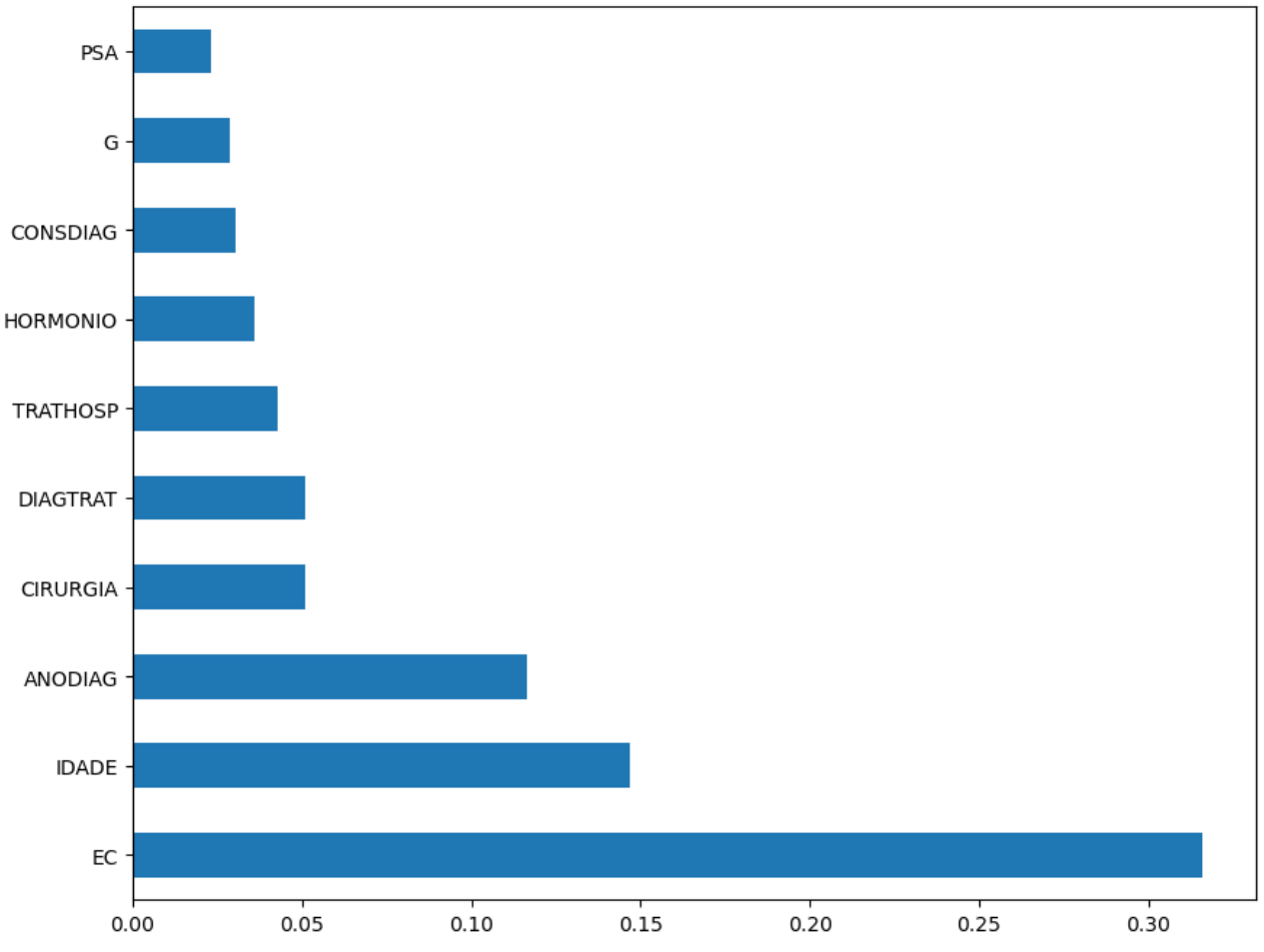


	precision	recall	f1-score	support
0	0.592	0.610	0.601	2059
1	0.861	0.852	0.857	5844
accuracy			0.789	7903
macro avg	0.727	0.731	0.729	7903
weighted avg	0.791	0.789	0.790	7903

Curva ROC:

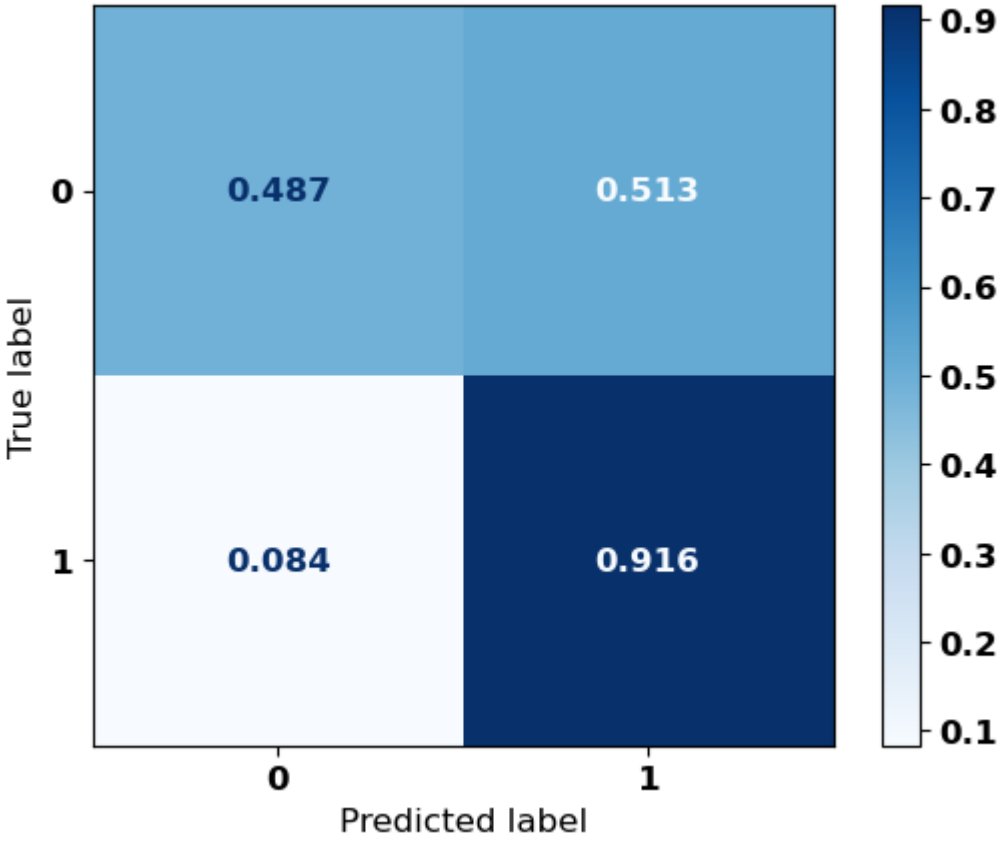


Importância das features:



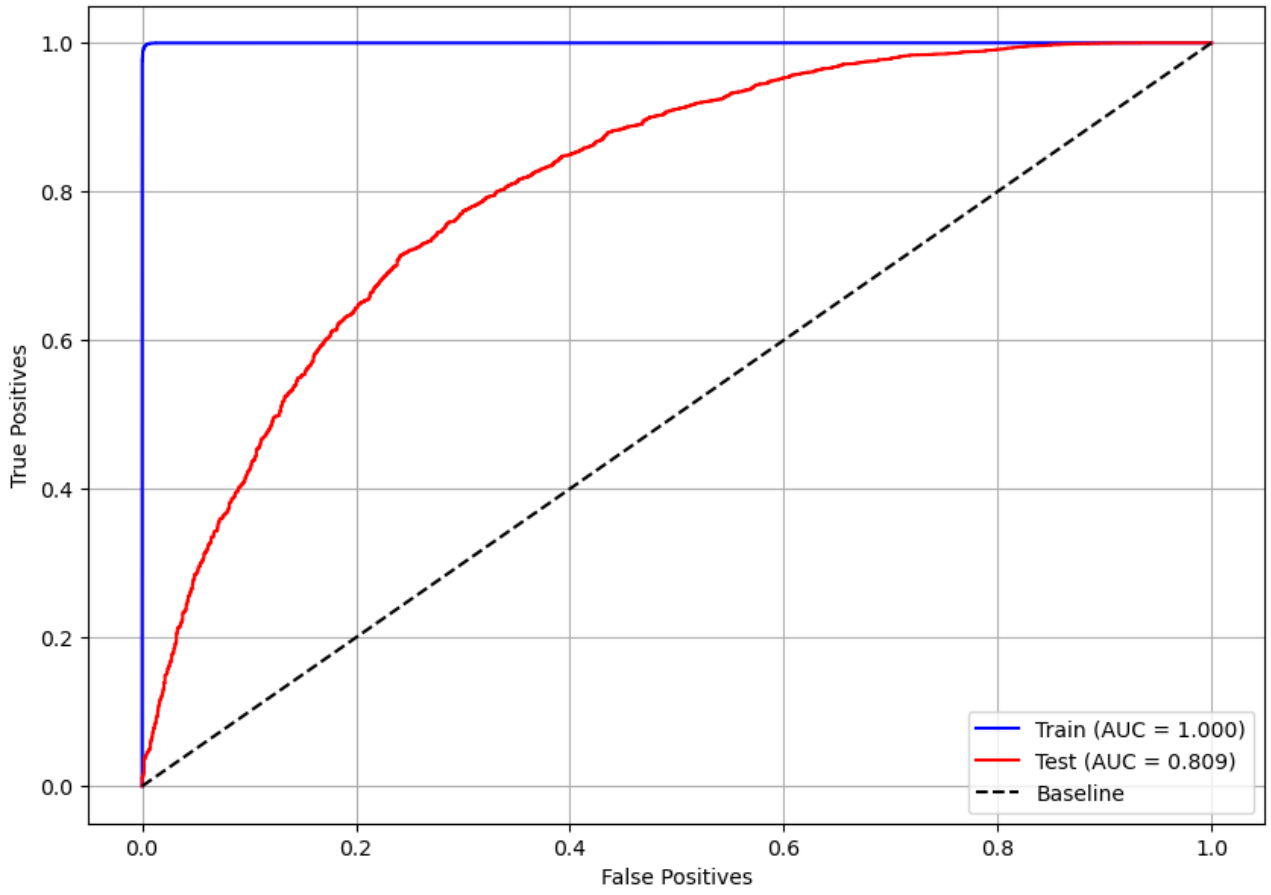
-XGBoost:

Matriz de confusão:

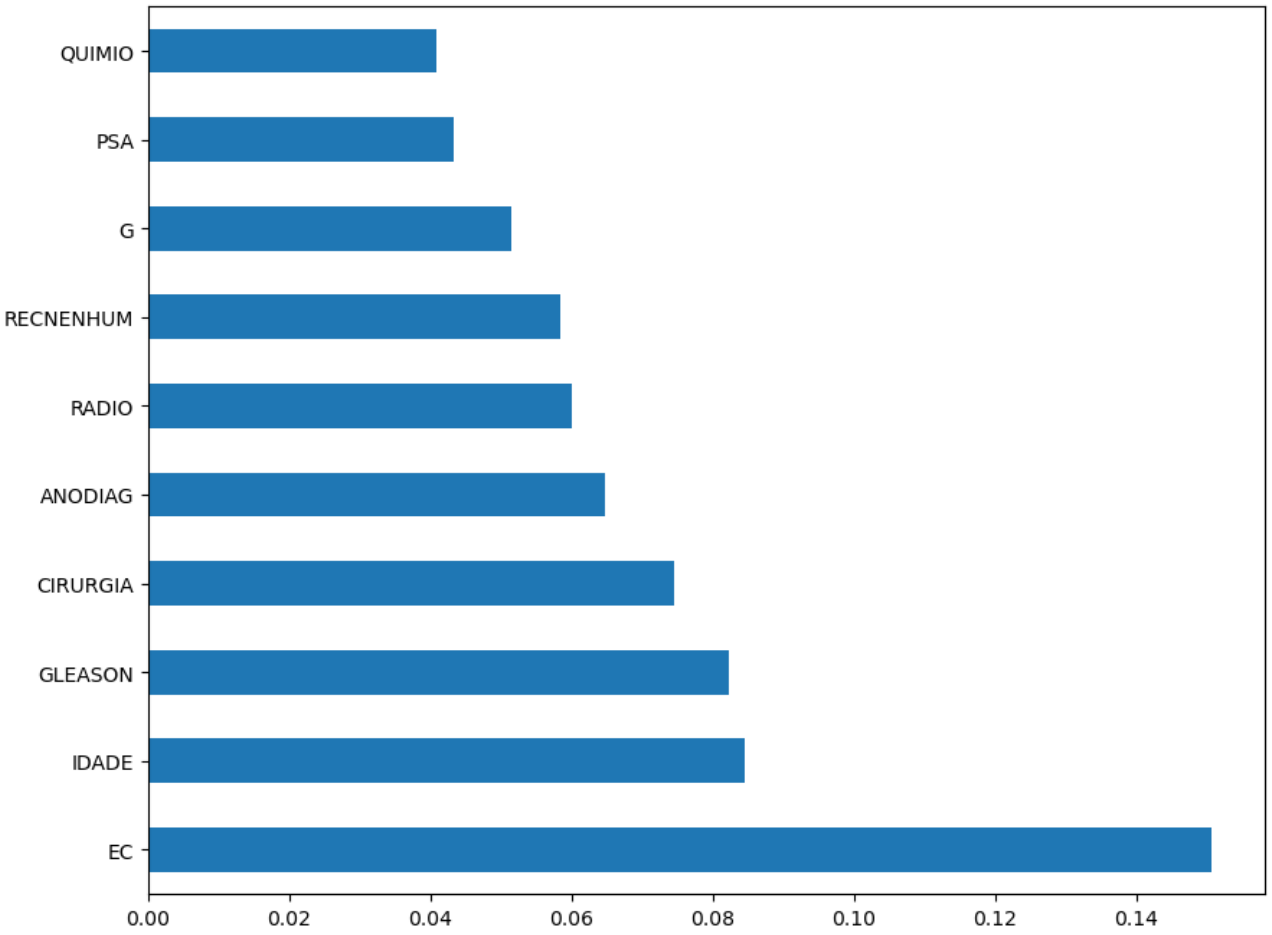


	precision	recall	f1-score	support
0	0.672	0.487	0.565	2059
1	0.835	0.916	0.874	5844
accuracy			0.804	7903
macro avg	0.754	0.701	0.719	7903
weighted avg	0.793	0.804	0.793	7903

Curva ROC:


True Positives
False Positives
0.0
0.2
0.4
0.6
0.8
1.0
Train (AUC = 1.000)
Test (AUC = 0.809)
Baseline

Importância das features:



5. Conclusão

Podemos concluir que o modelo Naïve Bayes se mostrou inadequado para este estudo, quando analisamos os resultados obtidos pelos outros modelos tivemos valores aceitáveis nas métricas da matriz de confusão. Com base no proposto inicialmente os valores de Accuracy atingiram o esperado, entretanto com o desenvolvimento do projeto foi também possível notar que eles podem e devem ser melhorados. Como este trabalho será trabalhado ainda mais no futuro, é importante revisar a exploração e preparação dos dados a fim de tentar melhorar estas métricas.

Podemos observar que o melhor modelo que foi obtido foi o Random Forest, isto pois apesar de ser possível obter uma accuracy maior com o XGBoost o Random Forest apresenta um equilibrio maior para prever os valores de classe 1, tornando este modelo mais adequado.

Como próximos passos também vamos aplicar os mesmos modelo realizados aqui para todas as outras colunas “vivo” criadas.

6. Links de acesso:

Todas as modificações feitas no dataset bem como o dicionário de dados, o note-book Jupyter contendo a análise exploratória, a preparação dos dados e aplicação e validação dos modelos estão disponíveis no seguinte repositório do GitHub:

https://github.com/acanellafilho/ia

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

Youtube com explicação:

https://www.youtube.com/watch?v=97VI90zBh0E

7. Referências Bibliográficas

Buk Cardoso, L., Cunha Parro, V., Verzinhasse Peres, S., Curado, M. P., Fer- nandes, G. A., Wünsch Filho, V., and Natasha Toporcov, T. (2023). Machine learning for predicting survival of colorectal cancer patients. Dispon´ıvel 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. Dispon´ıvel em: https://xgboost.readthedocs.io/en/stable/tutorials/model.html. Acesso em: 20 out.2024.

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. Dispon´ıvel em: https://onlinelibrary.wiley.com/doi/abs/10.1002/ijc.29210. Acesso em: 05 mai.2024.

FOSP (2022). Banco de dados do rhc. Dispon´ıvel 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. Dispon´ıvel 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* : 27*out.*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