Modelagem de Sobrevida em Câncer Colorretal via Classificação Supervisionada

1st Daniel Ribeiro Trindade
Departamento de Informática
Universidade Federal do Espírito Santo

2nd Leandro Furlam Turi
Departamento de Informática
Universidade Federal do Espírito Santo

***Abstract*—O câncer colorretal (CCR) é uma das principais causas de mortalidade global, com crescente incidência no Brasil. Este estudo investiga a aplicação de algoritmos de aprendizado de máquina (MLP, Random Forest e XGBoost) para predição da sobrevida de pacientes com CCR, utilizando dados do Registro Hospitalar de Câncer do Estado de São Paulo (RHC-SP). Foram definidos quatro cenários preditivos, categorizando a sobrevida em intervalos clínicos significativos. O pré-processamento incluiu discretização de variáveis contínuas e recodificação de atributos categóricos. Os modelos foram avaliados quanto à acurácia, F1-score, precisão, recall e AUC. Os resultados indicam acurácia superior a 85% para predição de sobrevida em 1 ano e cerca de 76% para 5 anos. Em um cenário multiclasse, a acurácia caiu para 60%, com maior desempenho nas classes extremas (sobrevida menor que 1 ano e sobrevida maior que 5 anos). Estes achados destacam o potencial dos modelos para auxiliar decisões clínicas no manejo do CCR.**

***Index Terms*—câncer colorretal, aprendizado de máquina, sobrevida, predição**

I. Introdução

O câncer colorretal (CCR) representa uma das principais causas de morbidade e mortalidade globalmente, com aproximadamente 1,8 milhão de novos casos diagnosticados anualmente, correspondendo a cerca de 10% de todos os cânceres no mundo [1]. No Brasil, estimativas recentes do INCA apontam mais de 41 mil novos casos anuais, com tendência de aumento em ambos os sexos [2].

A análise de sobrevida em pacientes com câncer colorretal é utilizada no planejamento e avaliação dos serviços de saúde, assim como para identificar fatores prognósticos que possam guiar decisões terapêuticas [3]. Modelos estatísticos tradicionais, como regressões lineares ou o modelo de Cox, embora amplamente utilizados, apresentam limitações quanto à adaptação a realidades clínicas dinâmicas e potenciais reduções de acurácia com o tempo [3].

Nesse contexto, algoritmos de machine learning (ML) têm ganhado destaque. Eles são capazes de lidar com grandes volumes de dados, capturar relações complexas entre variáveis e adaptar-se rapidamente a novos cenários. O estudo de [4] demonstrou que modelos como Random Forest e XGBoost alcançam acurácias superiores a 77%, com AUCs próximas a 0,85 para a predição de sobrevida específica por câncer, superando abordagens tradicionais em alguns aspectos. Este trabalho busca expandir este estudo ao incorporar pré-processamento e categorização detalhada de variáveis para reduzir viéses.

A. O Problema de Pesquisa

A base de dados utilizada é proveniente do Registro Hospitalar de Câncer do Estado de São Paulo (RHC-SP), coordenado pela Fundação Oncocentro de São Paulo (FOSP) [5]. O conjunto de dados abrange informações sociodemográficas (idade, sexo, escolaridade), características clínicas (ano de diagnóstico, estadiamento) e dados relacionados ao tratamento (cirurgia, quimioterapia e outras modalidades).

Para os modelos propostos por [4] foram selecionados somente pacientes diagnosticados com adenocarcinoma colorretal (CID-O 3 ed.: topografias C18-C20, morfologia 8140/3) entre os anos de 2000 e 2021. Como resultado, foram obtidos 31.916 registros. Além disso, variáveis derivadas como o tempo entre consulta e diagnóstico e entre diagnóstico e início do tratamento foram calculadas para capturar aspectos temporais no manejo da doença.

A abordagem proposta por [4] simplifica o problema de predição ao considerar pacientes sem registro de óbito específico por câncer como se tivessem sobrevivido até o fim do período observado ou, alternativamente, como óbitos por outras causas. Essa simplificação implica assumir que o evento de interesse (óbito por câncer colorretal) foi totalmente observado em todos os indivíduos, o que não corresponde à realidade dos dados provenientes dos registros hospitalares. Pacientes vivos no final do acompanhamento ou perdidos no seguimento representam informações parciais (censura) sobre o desfecho e, portanto, deveriam ser tratados como tal para evitar viés de sobrevivência e superestimação das taxas de óbito ou sobrevivência.

Para contornar essas limitações, este estudo categoriza o tempo até o evento (morte por câncer) em quatro faixas clinicamente significativas: menos de 1 ano, entre 1 e 3 anos, entre 3 e 5 anos e mais de 5 anos. Essa discretização reflete marcos utilizados na prática oncológica para avaliar o sucesso terapêutico e definir estratégias de acompanhamento [2]. Assim, foram definidos quatro problemas de classificação: os três primeiros com classificações binárias (prever se o paciente terá sobrevida superior a 1, 3 ou 5 anos) e um quarto, com classificação multiclasse, categorizando os pacientes nas quatro faixas de sobrevida mencionadas.

II. Tratamento e Preparação dos Dados

Para a criação das bases de dados utilizadas na execução dos modelos, partiu-se da avaliação de domínio realizada por [4],

que incluiu uma análise exploratória detalhada e procedimentos de engenharia de variáveis com forte embasamento clínico. Nesse trabalho, os autores identificaram atributos para o prognóstico da sobrevida, como estadiamento clínico, ano de diagnóstico, idade, presença de recidiva e modalidade de tratamento (cirurgia, quimioterapia, radioterapia e hormonoterapia). Além disso, realizaram transformações relevantes, como a recodificação de categorias e a criação de variáveis derivadas (tempo entre consulta e diagnóstico, tempo até o início do tratamento).

A. Seleção e Categorização de Variáveis

Partindo do conjunto original de 25 variáveis utilizadas por [4], foram aplicados critérios de relevância clínica, redundância informacional e distribuição dos dados para reduzir o número de covariáveis a 15.

- **Código da cidade de residência (`IBGE`) e Código IBGE da instituição (`IBGEATEN`):** removidas por se tratar de um identificador institucional com baixo valor preditivo.
- **Código de combinação de tratamentos (`TRATHOSP`):** excluída por ser uma variável derivada das demais modalidades de tratamento já incluídas individualmente (`CIRURGIA`, `QUIMIO`, `RADIO`, etc.).
- **Tratamento recebido no hospital = nenhum (`NENHUM`):** retirada devido à baixa frequência da categoria "`sim`" (apenas 0,3% dos casos).
- **Tratamento recebido no hospital = TMO (`TMO`):** excluída por conter apenas uma observação na categoria "`sim`".
- **Tratamento recebido no hospital = imunoterapia (`IMUNO`):** retirada pela baixa frequência da categoria "`sim`" (0,1% dos casos).
- **Tratamento fora do hospital antes da admissão = nenhum (`NENHUMANT`):** excluída devido à baixa frequência da categoria "`não`" (menos de 0,1%).
- **Diferença entre consulta e diagnóstico (`CONSDIAG`):** removida por apresentar valores negativos, inconsistentes com o fluxo clínico esperado.
- **Rede Regional de Atenção à Saúde (`RRAS`) e Departamento de Saúde Regional (`DRS`):** excluídas por serem variáveis agregadas que poderiam introduzir sobreposição com outras covariáveis contextuais.

Variáveis categóricas foram recodificadas com base no dicionário de dados oficial do RHC-SP[1] para garantir consistência semântica:

- **Sexo (`SEXO`):** recodificado de valores numéricos (1 = masculino, 2 = feminino) para fatores "`masc`" e "`fem`".
- **Categoria de atendimento (`CATEATEND`):** categorias `1` (Convênio) e `3` (Particular) agrupadas como "`convenio_ou_particular`", `2` como "`sus`", e `9` como "`sem_informacao`".
- **Escolaridade (`ESCOLARI2`):** categorizada em cinco níveis: "`analfabeto`", "`ens_fund_incompleto`", "`ens_fund_completo`", "`ens_medio`" e "`ens_superior`".
- **Estadiamento clínico (`EC`):** os estágios individuais (I, II, IIA, IIB, III, IIIA, IIIB, IIIC, IV, IVA, IVB, IVC) foram agrupados na variável `ECGRUP` em quatro categorias: "`I`", "`II`", "`III`" e "`IV`".
- **Idade (`IDADE`):** categorizada em três faixas etárias: "`0_a_49_anos`", "`50_a_74_anos`" e "`75_anos_mais`".
- **Tempo entre consulta e tratamento (`TRATCONS`):** categorizado em dois níveis: "`<=60_dias`" e "`>60_dias`".
- **Modalidades de tratamento (`CIRURGIA`, `QUIMIO`, `RADIO`, `HORMONIO`, `OUTROS`):** transformadas em fatores binários com níveis "`sim`" e "`nao`".

Variáveis contínuas foram discretizadas com base em pontos de corte otimizados utilizando o método *maximally selected rank statistics*, implementado pelo pacote CRAN `survminer`[2]. Esta abordagem consiste em avaliar todos os valores possíveis de uma variável como potenciais divisores da amostra em dois grupos, calculando, para cada divisão, o teste *log-rank* a fim de comparar a sobrevida entre os grupos. O ponto de corte selecionado é aquele que maximiza a estatística de teste, representando a maior diferença significativa na função de sobrevivência entre os subgrupos. Esta técnica foi aplicada às variáveis **ano de diagnóstico** e **tempo entre diagnóstico e início do tratamento**. Os seguintes pontos de corte foram identificados:

- **Ano de diagnóstico (`ANODIAG`):** "`até_2006`" e "`após_2006`".
- **Tempo entre diagnóstico e início do tratamento (`DIAGTRAT`):** "`até_81_dias`" e "`mais_de_81_dias`".

B. Criação da variável alvo Sobrevida

Para análises de sobrevida específicas pelo câncer, foram considerados 20.693 registros: os casos com falha (óbito por câncer colorretal) e os pacientes que sobreviveram mais do que o tempo de análise, garantindo os desfechos clínicos e reduzindo o viés associado a perdas de acompanhamento do paciente. A variável alvo "`sobrevida`" foi criada de acordo com 4 cenários:

- **Cenario 1**: Pacientes com sobrevida superior a 1 ano;
- **Cenario 2**: Pacientes com sobrevida superior a 3 anos;
- **Cenario 3**: Pacientes com sobrevida superior a 5 anos;
- **Cenario 4**: Sobrevida do paciente em 4 faixas:
  - sobrevida de menos de 1 ano;
  - sobrevida de entre 1 e 3 anos;
  - sobrevida de entre 3 e 5 anos;
  - sobrevida de mais de 5 anos.

Os cenários 1, 2 e 3 representam modelos onde a classificação é binária (o paciente tem sobrevida de X anos ou não), enquanto o cenário 4 é um problema de classificação

[1] https://fosp.saude.sp.gov.br/fosp/diretoria-adjunta-de-informacao-e-epidemiologia/rhc-registro-hospitalar-de-cancer/banco-de-dados-do-rhc/

[2] https://CRAN.R-project.org/package=survminer

multiclasse, onde a saída pode assumir 4 classes diferentes. Neste último cenário, deseja-se verificar se é possível classificar a sobrevida de um paciente para uma determinada faixa de anos.

III. Metodologia

Foram realizados experimentos com o objetivo de prever a sobrevida de pacientes com câncer colorretal para os 4 cenários descritos na Seção II-B, utilizando 3 algoritmos de classificação cuja modelagem é do tipo supervisionada:

- **Perceptron multicamadas (MLP)**: rede neural do tipo *feedforward* composta por múltiplas camadas de neurônios fortemente conectadas [6]. A implementação utilizada foi MLPClassifier do pacote python Scikit-learn[3] [7].
- **XGBoost**: modelo *boosting* que constrói árvores de forma sequencial, corrigindo os erros dos modelos anteriores [8]. A implementação utilizada se encontra em [9].
- **Random Forest**: *ensemble* de árvores de decisão treinadas [10].

A primeira etapa dos experimentos consistiu na geração e tratamento da base de dados. Para essa etapa, foi utilizada a linguagem R, com a seleção das variáveis de entrada, criação das variáveis-alvo (saída) e discretização das variáveis em valores contínuos, conforme descrito na Seção II.

A segunda etapa consiste no treinamento dos modelos e geração das métricas de desempenho. Nesta fase, foi utilizada a linguagem Python. Os dados foram carregados e processados com o auxílio da biblioteca Pandas [4], a fim de prepará-los para o treinamento dos modelos.

Para que seja possível o uso em algoritmos que exigem entradas numéricas, a variável alvo "`sobrevida`" foi transformada em valores numéricos (valores 0 e 1 para os cenários de classificação binária, e valores 0, 1, 2, 3 para o cenário 4 com classificação multiclasse). As variáveis de entrada que podem assumir valores com mais de 2 classes foram transformadas em variáveis binárias (codificação *one-hot*), com o auxílio da função `get_dummies` da biblioteca Pandas (note-se na Seção II que os domínios das variáveis são limitados ou foram consolidados, como no caso da idade e do tempo entre diagnóstico e início do tratamento, o que simplifica a aplicação da codificação *one-hot* e reduz a dimensionalidade do conjunto de dados)

Os dados foram divididos em conjuntos de treino e teste na proporção de 75% para treino e 25% para teste. A divisão foi estratificada, garantindo que a distribuição das classes fosse preservada em ambos os subconjuntos.

Para o modelo MLP foi realizada ainda uma etapa a mais de tratamento dos dados, com a normalização dos dados de entrada. Isso foi necessário devido à sensibilidade da arquitetura de rede neural à escala das variáveis [7].

Depois do tratamento dos dados, foi realizada uma busca pelos melhores hiperparâmetros para os modelos citados por meio da técnica de *Grid Search* com avaliação via acurácia, conforme [4], em conjunto com validação cruzada usando 3 *folds*. Para cada combinação de hiperparâmetros, foram calculadas as métricas de acurácia média e o desvio padrão nos *folds* de validação. Isso é importante para a análise de desempenho entre diferentes modelos.

O melhor modelo obtido para cada cenário e cada algoritmo foi utilizado para a obtenção das métricas de acurácia, *F1-score*, *recall* e precisão sobre o conjunto de teste. Foram geradas também as matrizes de confusão e as curvas ROC para cada modelo, possibilitando uma análise mais detalhada do desempenho preditivo.

[3]https://scikit-learn.org/

[4]https://pandas.pydata.org/

IV. Resultados

A Tabela I apresenta os valores das métricas de acurácia, *F1-score*, *recall* e precisão para os modelos treinados segundo o cenário 1 (sobrevida maior que 1 ano). Todos os modelos obtiveram acurácia próxima de 85% no conjunto de teste, o que indica uma boa capacidade em prever a sobrevida de 1 ano. Os valores obtidos nas outras métricas também são próximos entre os modelos, não havendo diferença significativa. Isso é ratificado pelos dados presentes na Tabela II, que apresenta os valores de acurácia para o melhor treinamento em cada um dos 3 *folds* na validação cruzada, além da medida *p-value* calculada através do Teste de Friedman. O *p-value* é 0,5292, indicando que não há diferença estatística entre os modelos. A notar, os modelos apresentaram valores próximos de 67% para a métrica de precisão, indicando uma tendência maior a cometer falsos positivos.

TABLE I
Desempenho dos modelos de classificação para cenário 1

Modelo	Ac	f1	Precisão	Recall
MLP	0,8505	0,7062	0,6749	0,7833
XGBoost	0,8525	0,7052	0,6722	0,7932
Random Forest	0,8531	0,7084	0,6756	0,7926

TABLE II
Acurácia por *fold* e teste de Friedman para cenário 1

Modelo	Fold 1	Fold 2	Fold 3	p-value Friedman
MLP	0,8432	0,8420	0,8452	
XGBoost	0,8439	0,8423	0,8452	0,5292
Random Forest	0,8453	0,8417	0,8428	

A Tabela III apresenta os valores das métricas de acurácia, *F1-score*, *recall* e precisão para os modelos treinados segundo o cenário 2 (sobrevida maior que 3 anos). Todos os modelos obtiveram acurácia próxima de 77% no conjunto de teste, o que indica uma boa capacidade em prever a sobrevida de 3 anos, mas menos eficaz do que para sobrevida de 1 ano. Os valores obtidos nas outras métricas também são próximos de 77%, não havendo diferença significativa entre os modelos. Todos os modelos alcançaram *F1-scores* acima de 0,75, indicando boa precisão e sensibilidade combinadas. A Tabela IV apresenta o *p-value* de 0,7165, indicando que não há diferença estatística entre os modelos.

TABLE III
DESEMPENHO DOS MODELOS DE CLASSIFICAÇÃO PARA CENÁRIO 2

Modelo	Ac	f1	Precisão	Recall
MLP	0,7711	0,7616	0,7580	0,7731
XGBoost	0,7696	0,7609	0,7576	0,7702
Random Forest	0,7669	0,7572	0,7537	0,7688

TABLE IV
ACURÁCIA POR *fold* E TESTE DE FRIEDMAN PARA CENÁRIO 2

Modelo	Fold 1	Fold 2	Fold 3	p-value Friedman
MLP	0,7803	0,7697	0,7591	
XGBoost	0,7811	0,7695	0,7642	0,7165
Random Forest	0,7783	0,7708	0,7639	

A Tabela V apresenta os valores das métricas de acurácia, *F1-score*, *recall* e precisão para os modelos treinados segundo o cenário 3 (sobrevida maior que 5 anos). Os resultados são próximos aos obtidos no cenário 2, com valores próximos de 76% de acurácia. Em relação a outras métricas, também não há diferenças significativas. A Tabela VI apresenta o *p-value* de 0,2636, indicando que não há diferença estatística entre os modelos.

TABLE V
DESEMPENHO DOS MODELOS DE CLASSIFICAÇÃO PARA CENÁRIO 3

Modelo	Ac	f1	Precisão	Recall
MLP	0,7652	0,7598	0,7623	0,7582
XGBoost	0,7634	0,7581	0,7608	0,7565
Random Forest	0,7640	0,7588	0,7618	0,7572

TABLE VI
ACURÁCIA POR *fold* E TESTE DE FRIEDMAN PARA CENÁRIO 3

Modelo	Fold 1	Fold 2	Fold 3	p-value Friedman
MLP	0,7618	0,7742	0,7622	
XGBoost	0,7660	0,7745	0,7637	0,2636
Random Forest	0,7633	0,7695	0,7668	

A Figura 1 apresenta as curvas ROC para os cenários 1, 2 e 3. As curvas são bastante semelhantes entre si, com valores de AUC elevados ($\sim$0.83–0.84), o que indica uma boa capacidade de classificação dos modelos.

A Tabela VII apresenta as métricas para os modelos treinados segundo o cenário 4 (faixas de sobrevida). Todos os modelos apresentaram desempenho semelhante, com acurácia em torno de 60%. Dentre os modelos, o Random Forest obteve uma leve vantagem, apresentando os maiores valores de *F1-score* (0,4461) e *recall* (0,5291), o que indica melhor capacidade de identificar corretamente todas as faixas de sobrevida. A Tabela VIII apresenta o *p-value* de 0,5292, indicando que não há diferença estatística entre os modelos.

A Figura 2 apresenta também as matrizes de confusão para os modelos. A partir dela é possível identificar que os modelos apresentaram melhor desempenho ao classificar pacientes com sobrevida superior a 5 anos ("`>5anos`"), com uma grande concentração de acertos na última linha da matriz. O mesmo desempenho superior é visto de forma menos pronunciada para a faixa "`<1ano`". Isso indica que essas faixas de sobrevidas são mais facilmente distinguíveis pelas variáveis do conjunto de dados. Por outro lado, os modelos se confundem mais com as faixas "`1-3anos`" e "`3-5anos`". As curvas ROC para o cenário 4 (Figura 3) corroboram: classes "`<1ano`" e "`>5anos`" apresentam as maiores AUCs em todos os modelos, com valores de aproximadamente 0,84 a 0,85, indicando uma maior capacidade dos modelos em distinguir essas classes. Já faixas intermediárias ("`1-3anos`" e "`3-5anos`") obtiveram valores de AUC inferiores, entre 0,71 e 0,74, indicando dificuldade dos modelos em distinguir corretamente essas faixas.

TABLE VII
DESEMPENHO DOS MODELOS DE CLASSIFICAÇÃO PARA CENÁRIO 4

Modelo	Ac	f1	Precisão	Recall
MLP	0,6071	0,4445	0,4621	0,5162
XGBoost	0,6051	0,4417	0,4590	0,4962
Random Forest	0,6077	0,4461	0,4613	0,5291

TABLE VIII
ACURÁCIA POR *fold* E TESTE DE FRIEDMAN PARA CENÁRIO 4

Modelo	Fold 1	Fold 2	Fold 3	p-value Friedman
MLP	0,6019	0,6058	0,6013	
XGBoost	0,6050	0,6040	0,6100	0,5292
Random Forest	0,6050	0,6035	0,6052	

Ainda, os resultados[5] demonstram que não houve *overfitting*, uma vez que as métricas de desempenho no conjunto de treino e teste foram muito próximas em todos os modelos avaliados. Além disso, a validação cruzada apresentou baixa variabilidade entre *folds*, e nenhuma diferença significativa foi observada entre as métricas dos conjuntos, reforçando a boa capacidade de generalização do modelo.

V. CONCLUSÃO

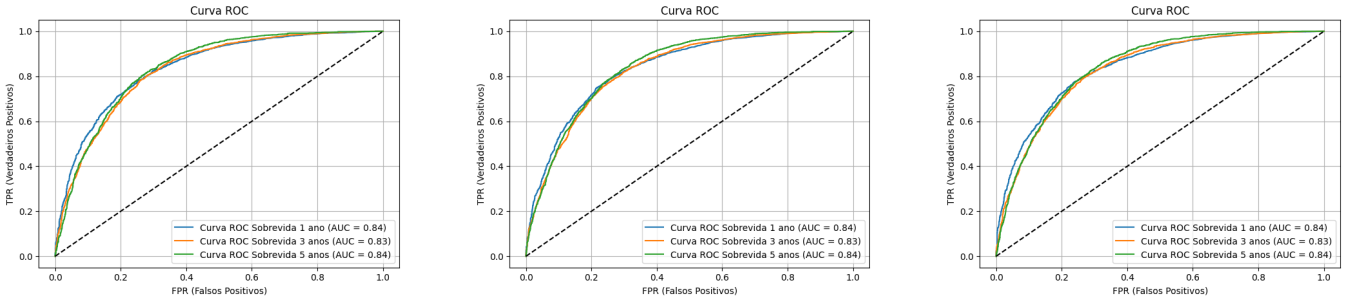
Os modelos de aprendizado de máquina avaliados demonstraram capacidade para prever a sobrevida de pacientes com câncer colorretal, especialmente nos cenários de classificação binária. A abordagem proposta, com discretização de tempo de sobrevida em intervalos clínicos, permitiu uma análise alinhada à prática oncológica.

Embora os resultados para o cenário multiclasse tenham apresentado menor acurácia, os modelos demonstraram maior sensibilidade para identificar pacientes com prognóstico extremo. Estes achados reforçam o potencial de técnicas baseadas em IA como suporte à decisão clínica, podendo contribuir para o planejamento terapêutico e alocação de recursos em saúde. Futuras pesquisas devem considerar o uso de dados contínuos e estratégias para tratar censura nos dados, visando aprimorar a precisão preditiva.

REFERÊNCIAS

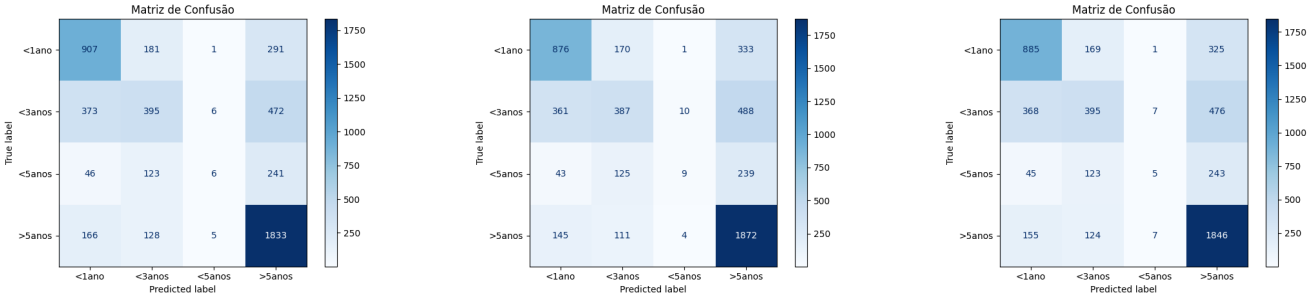
[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, May 2021.

[5]https://github.com/intel-comp-saude-ufes/2025-1-P1-colon-rectal-cancer-survival



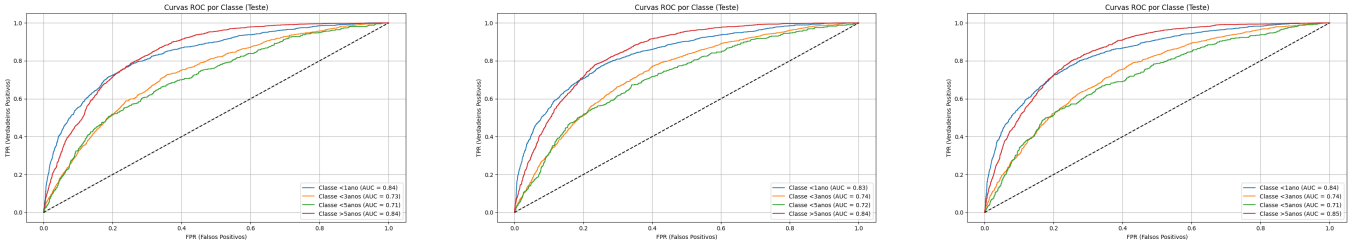
(a) MLPClassifier (b) Random Forest (c) XGBoost

Fig. 1. Curvas ROC para cenários 1, 2 e 3



(a) MLPClassifier (b) Random Forest (c) XGBoost

Fig. 2. Matrizes de confusão para o cenário 4



(a) MLPClassifier (b) Random Forest (c) XGBoost

Fig. 3. Curvas ROC para o cenário 4

[2] Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA), "Estimativa 2020: Incidência de câncer no brasil," https://www.inca.gov.br/sites/ufu.sti.inca.local/files/media/document/estimativa-2020-incidencia-de-cancer-no-brasil.pdf, 2019, acessado em: 11 jul. 2025.

[3] M. A. Freitas and E. A. Colosimo, *Confiabilidade: Análise de tempo de falha e testes de vida acelerados*. Belo Horizonte: UFMG, Escola de Engenharia: Fundação Christiano Ottoni, 1997.

[4] L. B. Cardoso, V. C. Parro, S. V. Peres, M. P. Curado, G. A. Fernandes, V. W. Filho, and T. N. Toporcov, "Machine learning for predicting survival of colorectal cancer patients," *Scientific Reports*, vol. 13, no. 1, p. 8874, Jun 2023. [Online]. Available: https://doi.org/10.1038/s41598-023-35649-9

[5] Fundação Oncocentro de São Paulo (FOSP), "Registro hospitalar de câncer (rhc) – banco de dados," https://fosp.saude.sp.gov.br/fosp/diretoria-adjunta-de-informacao-e-epidemiologia/rhc-registro-hospitalar-de-cancer/banco-de-dados-do-rhc/, 2025, acessado em: 11 jul. 2025.

[6] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[8] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[9] T. Chen and Contributors, "Xgboost github repository," https://github.com/dmlc/xgboost, 2016, a scalable, portable gradient boosting library.

[10] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.