Survival Prediction for Oral Cancer Patients: A Machine Learning Approach

Murilo Cruz Lopes[1], Marília de Matos Amorim[1], Valéria Souza Freitas[1], Rodrigo Tripodi Calumby[1]

Universidade Estadual de Feira de Santana, Brazil
mclopes@ecomp.uefs.br, amorim.mah@hotmail.com, vfreitas@uefs.br, rtcalumby@uefs.br

Abstract. There is a high incidence of oral cancer in Brazil, with 150,000 new cases estimated for 2020-2022. In most cases, it is diagnosed at an advanced stage and are related to many risk factors. The Registro Hospitalar de Câncer (RHC), managed by Instituto Nacional de Câncer (INCA), is a nation-wide database that integrates cancer registers from several hospitals in Brazil. RHC is mostly an administrative database but also include clinical, socioeconomic and hospitalization data for each patient with a cancer diagnostic in the country. For these patients, prognostication is always a difficult task a demand multi-dimensional analysis. Therefore, exploiting large-scale data and machine intelligence approaches emerge as promising tool for computer-aided decision support on death risk estimation. Given the importance of this context, some works have reported high prognostication effectiveness, however with extremely limited data collections, relying on weak validation protocols or simple robustness analysis. Hence, this work describes a detailed workflow and experimental analysis for oral cancer patient survival prediction considering careful data curation and strict validation procedures. By exploiting multiple machine learning algorithms and optimization techniques the proposed approach allowed promising survival prediction effectiveness with F1 and AuC-ROC over 0.78 and 0.80, respectively. Moreover, a detailed analysis have shown that the minimization of different types of prediction errors were achieved by different models, which highlights the importance of the rigour in this kind of validation.

CCS Concepts: • **Computing methodologies** → **Machine learning algorithms**.

Keywords: RHC, health, oral cancer, machine learning

1. INTRODUÇÃO

O Instituto Nacional de Câncer (INCA), órgão auxiliar do Ministério da Saúde, define câncer como: "um conjunto de mais de 100 doenças que têm em comum o crescimento desordenado de células, que invadem tecidos e órgãos" [INCA 2019a]. Estas células tendem a tornarem-se muito agressivas à medida que passam a dividir-se rapidamente, resultando na formação de tumores. Os tipos de câncer correspondem aos diversos tipos de células no corpo, por exemplo, quando começam em tecidos epiteliais, como pele ou mucosas, são chamados de carcinomas. Por vez, os sarcomas são aqueles com origem em tecidos conjuntivos, como osso, músculo ou cartilagem. A velocidade de multiplicação das células e a capacidade de invadir tecidos e órgãos vizinhos são outras características do câncer, caracterizando o processo chamado de metástase.

Especificamente, o câncer de boca é um tumor maligno que afeta lábios, estruturas da boca, como gengivas, bochechas, céu da boca, língua e a região abaixo da língua. Na maioria dos casos, o câncer de boca é diagnosticado em estado avançado. Os riscos para o desenvolvimento de um câncer de boca envolvem o uso de tabaco, consumo regular de bebidas alcoólicas, exposição ao sol sem proteção, entre outros [INCA 2021]. Assim, o câncer de boca é considerado um problema de saúde pública e sua importância é ratificada por altas taxas de incidência e mortalidade em todo o mundo. Segundo os dados da Agência Internacional de Pesquisa em Câncer (IARC), estimou-se, para o ano de 2020,

aproximadamente 370.000 novos casos de câncer de boca e por volta de 240.000 mortes por câncer de boca e orofaringe no mundo [IARC 2020]. No Brasil, existe uma alta incidência de câncer oral e, para este tipo, ocupa o segundo lugar na América Latina [Warnakulasuriya 2009] (com média de mais de 8200 casos no período de 2010 a 2019[1]), sendo estimados, para cada ano do triênio 2020-2022, mais de 15.000 novos casos [INCA 2020]. Adicionalmente, o conceito de Determinantes Sociais em Saúde (DSS) aponta que as condições de vida e trabalho dos indivíduos e de grupos populacionais estão relacionadas com a situação de saúde. Fatores, como: condições sociais, econômicas, culturais, étnico-raciais, psicológicas e comportamentais podem influenciar a ocorrência de problemas de saúde e seus fatores de risco na população [Fiocruz 2021]. Nesse sentido, a literatura aponta que grande parte dos indivíduos acometidos por câncer de boca possuem baixos níveis socioeconômicos [Dantas et al. 2016; Conway et al. 2008; Groome et al. 2011]. Além disso, condições socioeconômicas estão relacionadas à maior exposição aos fatores de risco, uma vez que grupos socialmente desprivilegiados tendem a apresentar maior contato com o álcool e o tabaco [Borges et al. 2009].

No Brasil o INCA é o responsável pelos Registros Hospitalares de Câncer (RHC) que são centros de coleta, armazenamento, processamento e análise de informações de pacientes atendidos em uma unidade hospitalar, com diagnóstico confirmado de câncer [INCA 2019b]. As informações obtidas pelos RHC's ajudam no monitoramento da assistência apresentada ao paciente. A principal função do RHC é clínica, sendo utilizado no acompanhamento e avaliação dos trabalhos realizados nos hospitais, incluindo os resultados no tratamento do câncer [INCA 2019b]. A utilização de bases de dados, como a do RHC, envolve múltiplos desafios, como a alta complexidade dos dados, o grande volume de informações, bem como problemas relacionados à qualidade dos dados armazenados. Assim, técnicas de mineração de dados podem ser exploradas para a descoberta de padrões e geração de conhecimento ao passo que buscam minimizar os problemas citados. Neste sentido, estudos que utilizam a mineração de dados podem auxiliar em tomadas de decisões referentes ao rastreamento de doenças, ao direcionamento de programas preventivos e à oferta de tratamentos especializados, podendo também contribuir para estimativas do risco de morte na população. Neste contexto, este trabalho propõe a exploração da base de dados do RHC, especificamente sobre câncer da cavidade oral e a construção e validação experimental de modelos preditivos de óbito para apoio no processo de tomada de decisão. Salvo melhor juízo, este é o primeiro trabalho a conduzir, em escala nacional brasileira, experimentos baseados em aprendizagem de máquina para predição de óbitos de pacientes diagnosticados com câncer de boca.

2. TRABALHOS RELACIONADOS

Considerando a gravidade dos problemas e desafios associados ao câncer, alguns trabalhos têm sido realizados com foco na descoberta de conhecimento em bases de dados, mas ainda de modo incipiente, a partir de conjuntos de dados limitados ou com análises pouco rigorosas. Desta forma, à medida que grandes bases de dados tornam-se disponíveis, o potencial dos métodos modernos de mineração de dados e aprendizado de máquina pode ser explorado em experimentos cuidadosos e em larga escala.

Especificamente para câncer oral, [Tseng et al. 2015] buscaram identificar a diferença entre os sintomas de pacientes que morreram ou sobreviveram. Foi utilizada uma base de dados de pesquisa oncológica de um centro médico no sul de Taiwan, incluindo 673 casos diagnosticados entre 1996 e 2008, sendo que 426 sobreviveram e 247 não. Os autores utilizaram os algoritmos de árvores de decisão e redes neurais artificiais e a técnica de validação cruzada com 10 *folds*. Os autores aplicaram os algoritmos de classificação para predizer sobrevida específica da doença em 5 anos e sobrevida livre de doença em 5 anos. Além dos modelos preditivos, o uso da árvore de decisão permitiu a obtenção de regras considerando os sintomas dos pacientes e sua associação com as taxas de sobrevida específica e de sobrevida livre da doença. Os autores também aplicaram descoberta de grupos com o algoritmo *K-means* para determinar as diferenças entre os sintomas de quem sobreviveu e quem não. A análise dos

[1]Segundo dados obtidos do Registro Hospitalar de Câncer - `https://irhc.inca.gov.br/RHCNet/`

resultados indicou uma baixa taxa de sobrevivência ao câncer oral em pacientes que o *pN*, (condição do linfonodos), foi *N2b*, (múltiplos locais de metástase de linfonodo ipsilateral com tumores <= 6 cm), o nível de metástase no linfonodo entre I e III, ocorrência de metástase no linfonodo igual a T4, estágio do câncer igual a IV e diferenciação de células como moderada. Apesar de apresentar resultados promissores, o trabalho descrito analisou a eficácia a partir de medidas de avaliação pouco específicas não permitindo uma análise detalhada dos resultados.

Em [Sharma and Om 2013] foram explorados dados de câncer de boca de pacientes atendidos em um Centro de Cuidados Terciários entre 2004 e 2009. Com o propósito de predizer a sobrevivência dos pacientes, foram utilizados os algoritmos *Single Tree*, *Decision Tree Forest* e *TreeBoost*, com a utilização de 20% dos registros para teste e 80% para treino. Os três modelos treinados obtiveram 100% de acurácia, contudo além do conjunto limitado de dados (N=1024), não é possível estimar a robustez em relação à utilização de diferentes dados de treinamento e teste dos modelos.

Considerando outro tipo de câncer, o trabalho de [Delen 2009] usou dados de pacientes com câncer de próstata do *SEER Program* do Instituto Nacional de Câncer dos Estados Unidos, com o propósito de predizer a capacidade de sobrevivência por 60 dias ou mais após o diagnóstico. Foram utilizados 120.000 registros e 77 variáveis, onde foram realizadas limpeza dos dados e preparação para deixar os dados mais homogênios e também a remoção de registros que o autor considerou ser insubstituível. Utilizando validação cruzada (10 *folds*), os melhores resultados foram alcançados com o *SVM*, com acurácia de 92,85%, sensibilidade de 94,23% e especificidade de 75,72%. Em outro contexto, [Salmi and Rustam 2019] utilizaram dados do *I-Islam Hospital Bandung Idonesia* de pacientes diagnosticados com câncer de cólon. A base incluía apenas 209 registros com 7 atributos, tendo sido avaliado apenas o algoritmo *Naive Bayes* para determinar a presença ou não do câncer. Utilizando 80% dos dados para treinamento, obteve-se acurácia de 94,25%, precisão de 100% e recall de 94%. Apesar das altas taxas de eficácia, o processo experimental pode ser considerado pouco rigoroso, por basear-se em um conjunto bastante limitado de dados (N=209).

Considerando os trabalhos anteriores no contexto de pacientes com câncer, observa-se que de modo geral são reportados resultados promissores. Contudo, apresentam limitações importantes, *e.g.*, baseando-se em um número bastante limitado de pacientes ou atributos, metodologia de validação experimental pouco rigorosa ou análises com medidas de avaliação que possuem limitações e podem sofrer enviesamento pelo desbalanceamento das amostras para cada classe. Além disso, não foram realizados ou apresentados processos de otimização dos modelos. Assim, esse trabalho propõe o uso de métodos de aprendizado de máquina para a descoberta de conhecimento em uma base de dados de câncer, especificamente relacionada a pacientes com câncer de boca. Contudo, para maior confiabilidade nos resultados, é proposto um processo experimental robusto, incluindo tratamento criterioso dos dados, além da utilização de protocolos de validação rigorosos e análise detalhada da eficácia dos modelos preditivos e suas limitações.

3. METODOLOGIA

A Figura 1 apresenta as etapas seguidas no desenvolvimento deste trabalho incluindo a obtenção e preparação dos dados, bem como o processo de validação experimental nos modelos preditivos desenvolvidos. O *RHC* é uma base de dados pública que reúne registros de todos os pacientes com diagnóstico de câncer no Brasil, sendo os registros obtidos a partir dos prontuários [Ministerio da Saude do Brasil 2019]. Essa base pode ser acessada por meio do Integrador *RHC*[2], sendo possível fazer consultas analíticas simples e o *download* de todos os registros da base. Com dados referentes desde o ano de 1985, os registros possuem 45 atributos, incluindo dados clínicos, socioeconômicos, escolaridade, data de primeira consulta, de diagnóstico, de início de tratamento, além de informações relacionadas ao tipo do tumor e estadiamento. Junto com essas informações, a base de dados também

[2]`https://irhc.inca.gov.br/RHCNet/`

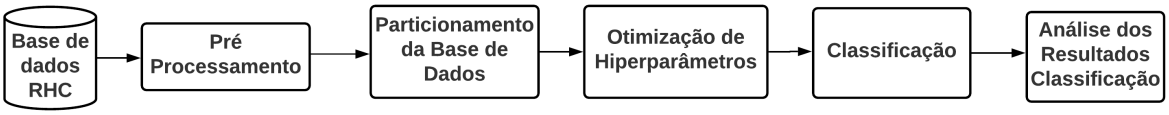

Fig. 1. Workflow geral seguido neste trabalho para a predição de óbitos de pacientes com câncer de boca.

oferece dados do estado brasileiro de nascimento, faixa etária, estado da doença ao final do primeiro tratamento no hospital, histórico de consumo de bebidas alcoólicas, data de diagnóstico e da consulta, escolaridade, informações do tratamento de um paciente, entre outros [Ministerio da Saude do Brasil 2019]. Para nossos experimentos, os dados foram coletados do RHC no dia 7 de novembro de 2019 e foram considerados os registros disponíveis desde o ano de 1985. A Tabela I apresenta os atributos da base e uma breve descrição.

Um atributo importante é a localização primária do câncer (LOCTUDET), composto de três dígitos de C00 à C80. Neste trabalho selecionamos apenas os registros com localização primária na região da boca (C00 a C09). A base de dados inicialmente coletada, contava com 3.446.425 registros, dentre esses 104.635 com referentes a C00 a C09, representando cerca de 3% do total. Outro critério de seleção baseou-se no atributo *Tipo Histológico* (TIPOHIST) para incluir apenas tumores do tipo carcinoma de células escamosas, visto que estes representam cerca de 90% de todos os casos de câncer de boca [Neville et al. 2020]. Com os dados selecionados, realizou-se a análise descritiva de modo a gerar uma visão geral dos dados. A Tabela II mostra as características para os principais atributos relacionados a determinantes sociais. Essa análise permitiu também identificar inconsistências em determinados atributos, como a ocorrência de valores espúrios, e.g., "3" para atributo SEXO e "Z" para atributo ESTADIAG, que são valores incompatíveis com o atributo.

Table I. Atributos disponíveis para os registros na base de dados do RHC.

Atributo	**Descrição**	**Atributo**	**Descrição**
ALCOOLIS	Histórico de consumo de bebida alcoólica	IDADE	Idade na primeira consulta
ANOPRIDI	Ano do diagnóstico	INSTRUC	Escolaridade
ANTRI	Ano da triagem	LATERALI	Lateralidade do tumor
BADIAGSP	Base mais importante para o diagnóstico do tumor para SP	LOCALNAS	Local de nascimento
BASMAIMP	Base mais importante para o diagnóstico do tumor	LOCTUDET	Localização primária (Categoria 3d)
CLIATEN	Clínicas do primeiro atendimento - entrada do paciente	LOCTUPRI	Localização primária detalhada (Subcategoria 4d)
CLITRAT	Clínica de início do tratamento	LOCTUPRO	Localização provável do tumor primário
CNES	Número do CNES do Hospital	MAISUMTU	Ocorrência de mais um tumor primário
DATAINITRT	Data do início do primeiro tratamento específico para o tumor, no hospital	MUUH	Município da unidade hospitalar
DATAOBITO	Data do óbito	OCUPACAO	Ocupação principal
DATAPRICON	Data da 1^a consulta	ORIENC	Origem do encaminhamento
DIAGANT	Diagnóstico e tratamento anteriores	OUTROESTA	Outros estadiamentos clínicos do tumor
DTDIAGNO	Data do primeiro diagnóstico	PRITRATH	Primeiro tratamento recebido no hospital
DTINITRT	Ano do início do primeiro tratamento específico para o tumor, no hospital	PROCEDEN	Código do Município de procedência (residência)
DTPRICON	Ano da 1^a consulta	RACACOR	Raça/cor
DTTRIAGE	Data da triagem	RZNTR	Principal razão para a não realização do tratamento antineoplásico no hospital
ESTADIAG	Estadiamento clínico do tumor (TNM) - Grupo	SEXO	Sexo
ESTADIAM	Estadiamento clínico do tumor (TNM)	TABAGISM	Histórico de consumo de tabaco
ESTADRES	UF de procedência (residência)	TIPOHIST	Tipo histológico do tumor primário
ESTCONJ	Estado conjugal atual	TNM	Codificação do estádio clínico segundo classificação TNM
ESTDFIMT	Estado da doença ao final do primeiro tratamento no hospital	TPCASO	Tipo de caso
EXDIAG	Exames relevantes para o diagnóstico e planejamento da terapêutica do tumor	UFUH	UF da unidade hospitalar
HISTFAMC	Histórico familiar de câncer		

Table II. Análise Descritiva dos atributos da base de dados do RHC. A soma das frequências não equivalem ao tamanho total da base (72.039) devido à ausência de valores para alguns os registros. A análise completa está disponível em: `https://doi.org/10.6084/m9.figshare.13072640.v2`

Variáveis	N	(%)	Variáveis	N	(%)
Sexo			Estado Conjugal		
Masculino	55.127	76,5	Casado	26.869	51,0
Feminino	16.909	23,5	Solteiro	14.062	26,7
Faixa etária			Viúvo	6.731	12,8
Idoso (> 45 anos)	64.174	89,1	Separado	4.203	8,0
Jovem (<= 45 anos)	7.855	10,9	União Consensual	816	1,50
Raça cor			Histórico Familiar de Câncer		
Branca	26.080	48,4	Não	16.664	58,9
Parda	23.636	43,8	Sim	11.610	41,1
Preta	3.766	7,0	Alcoolismo		
Amarela	367	0,7	Sim	20.967	52,0
Indígena	64	0,1	Ex-Consumidor	6.106	15,2
Nível de instrução			Não avaliado	808	2,0
Fundamental	36.786	51,1	Tabagismo		
Nenhuma	9.907	18,2	Sim	28.570	66.6
Médio	5.929	8,20	Nunca	7.702	17,9
Superior	1.797	2,50	Ex-Consumidor	6.121	14,3
			Não avaliado	535	1,2

3.1 Pré-processamento dos dados

Para tornar os dados consistentes e adequados para os algoritmos de aprendizagem de máquina, algumas etapas de pré-processamento foram necessárias. Em cada etapa, foram consideradas diferentes técnicas e seus impactos analisados para a escolha das mais eficazes.

3.1.1 *Seleção e engenharia de atributos.* Nessa etapa foi realizada a seleção de atributos, buscando-se descartar aqueles que não incorporavam informações úteis ao processo preditivo do óbito, são redundantes com outros ou trazem apenas informações administrativas, como datas. Essa etapa foi realizada junto a profissionais da área de saúde que auxiliaram na escolha dos atributos a serem utilizados. Assim, um conjunto de atributos foi descartado, sendo eles: PROCEDEN, ANOPRIDI, ANTRI, CLIATEN, CLITRAT, CNES, DATAPRICON, DTINITRT, DTPRICON, DTTRIAGE, EXDIAG, LATERALI, LOCTUPRO, MUUH, OUTROESTA, TNM, VALOR_TOT, TIPOHIST.

O atributo referente ao primeiro tratamento recebido no hospital (PRITRATH), por ser multivalorado, foi convertido para múltiplos atributos binários. Os múltiplos valores do atributo são representados originalmente por: 1) Nenhum; 2) Cirurgia; 3) Radioterapia; 4) Quimioterapia; 5) Hormonoterapia; 6) Transplante de medula óssea; 7) Imunoterapia; 8) Outras; 9) Sem informação. Então, se um determinado paciente foi submetido a Cirurgia, Radioterapia e Quimioterapia, o atributo é representado com "234", sendo que cada número indica um procedimento realizado. Então, foram criados os atributos binários CIRURGIA, RADIOTERAPIA, QUIMIOTERAPIA, HORMONOTERAPIA, TRANSMEDUOSSEA (transplante de medula óssea), IMUNOTERAPIA, OUTROS e SEMINFORMACAO. Cada atributo criado recebe o valor 1 (um), para informar que o procedimento foi realizado, e 0 (zero) caso o procedimento não tenha sido realizado.

O atributo REGIAO foi criado a partir do atributo UF da unidade hospitalar (UFUH) para indicar a região do tratamento: NO (Norte), NE (Nordeste), CO (Centro Oeste), SE (Sudeste) e SUL. Adicionalmente, sendo a IDADE o único atributo numérico, aplicou-se a discretização para a adequação dos dados para a predição, utilizando-se 10 intervalos de igual tamanho. Como variável alvo do processo preditivo foi introduzido o atributo OBITO (1 - sim; 0 - não), definindo-se com "sim" aqueles registros para os quais havia data de óbito informada e "não" para os demais. Adicionalmente, a variável "Estado da doença ao final do tratamento" (ESTDFIMT) foi descartada no processo de aprendizado devido à alta correspondência com a variável de interesse (OBITO), o que levaria à uma associação preditiva indesejada.

3.1.2 *Limpeza da base de dados.* Conforme verificado em etapas anteriores, alguns registros possuíam valores ausentes para muitos atributos, tornando-os inadequados para utilização no processo de treinamento dos preditores. Assim, buscando-se descartar registros com muitos atributos vazios, mas mantendo o maior número possível para os experimentos, decidiu-se por descartar registros com mais do que 8 atributos vazios. Isso resultou na remoção de 11.271 registros. Foram descartados também registros que possuíam data de óbito como 88/88/8888 ou 99/99/9999 (quantidade de registros removidos 1.272). Outra inconsistência que levou à remoção de 116 registros foi entre a data do óbito e o atributo estado da doença ao final do primeiro tratamento (ESTDFIMT), ocorrendo quando o atributo ESTDFIMT indicava que o paciente foi a óbito, mas o atributo DATAOBITO não apresentava à data do óbito. O cruzamento entre essas duas variáveis auxiliou na verificação da validade da variável DATAOBITO. Ao final dessa etapa, foram mantidos 72.039 registros para os experimentos. Por fim, considerando que o processo de limpeza descartou apenas registros com grandes quantidades de atributos vazios, diferentes técnicas de tratamento foram avaliadas, sendo elas: a) imputação com o valor mais frequente; b) imputação por predição baseada em *k-Nearest Neighbors* (k-NN) com k=1,3,5; e c) *one-hot-encoding*. Com o *one-hot-encoding*, cada possível valor para o atributo torna-se um novo atributo binário, sendo valores ausentes representados como "*false*" em todos estes novos atributos.

3.2 Modelos Preditivos e Configurações Experimentais

Os algoritmos utilizados para construção dos modelos preditivos foram: *Naive Bayes*, Árvores de decisão, *Random Forest*, utilizando a biblioteca *Scikit Learn*[3], e *XGBoost*[4]. A base de dados foi particionada em treino/validação e teste. Considerando-se o desbalanceamento do conjunto em relação à variável alvo, inicialmente foram selecionados aleatoriamente 3.000 registros de óbitos (aprox. 10% da base) e 3.000 de não-óbito, sendo estes reservados para o teste final dos modelos. Os demais 26.501 registros de óbitos foram utilizados no treinamento, juntamente com outros 26.501 registros de não-óbito selecionados aleatoriamente. Para otimização de hiperparâmetros foi utilizada a técnica de *gridsearch* sobre os dados de treinamento. Os valores foram escolhidos a partir de experimentos exploratórios preliminares e avaliados com o uso do *cross-validation* usando 5 *folds*. As Tabelas III, IV e V apresentam, respectivamente, os valores avaliados para cada um dos hiperparâmetros.

Table III. Conjunto de parâmetros de entrada para para a árvore de decisão no *gridsearch*.

Criterion	Max depth	Min samples split
gini	2; 10-100, passo 10	100-1000, passo 100; 700-940, passo 20

Table IV. Conjunto de parâmetros de entrada para a *Random Forest* no *gridsearch*.

Criterion	Max depth	Min Samples Split	Number of Estimators
gini	20; 40; 80; 120	2-8, passo 2; 10-90, passo 10; 100-200, passo 20; 100-1000, passo 100	2; 10-90, passo 10; 100-900, passo 100

Table V. Conjunto de parâmetros de entrada para o *XGBoost* no *gridsearch*.

GAMMA	Learning Rate	Max depth	Min Child Weight	Number of Estimators
0,2; 0,4; 0,6	0,1; 0,01; 0,001	10; 20; 30; 40; 80; 120	1; 2; 3	2; 4; 8; 20-80, passo 20

4. RESULTADOS

No processo de otimização dos hiperparâmetros a partir da técnica de *grid search* considerou-se a área sobre a curva ROC como medida alvo para seleção das melhores configurações. Este processo foi realizado considerando cada um dos métodos de tratamento de dados ausentes descritos na Seção 3 e os resultados são apresentados na Tabela VI. De modo geral, todos os métodos permitiram valores de eficácia próximos para os algoritmos de *Random Forest* e *XGBoost*. Contudo, o tratamento baseado em *one-hot-encoding* foi consistentemente superior para todos os algoritmos. Assim, o *one-hot-encoding*

[3]https://scikit-learn.org/stable/
[4]https://xgboost.readthedocs.io/

foi utilizado na etapa final de avaliação, e as melhores configurações de cada algoritmo foram: Árvore de Decisão (criterion=gini, *max depth*=10 e *min samples split*=840), *Random Forest* (*criterion*=gini, *max depth*=20, *min samples split*=2 e *n-estimators*=70), *XGBoost* (*gamma*=0,4, *learning rate*=0,01, *max depth*=10, *min child weight*=1 e *n-estimators*=600).

A partir das melhores configurações descobertas para cada algoritmo, os modelos finais foram treinados a partir de todo do conjunto de treinamento e avaliados com o conjunto de teste inicialmente separado. A Tabela VII apresenta os resultados para várias medidas de avaliação com o conjunto final de testes. Os valores em negrito indicam os melhores resultados em cada medida. As respectivas matrizes de confusão são apresentadas na Figura 2. Além disso, comparando-se os resultados do treinamento com validação cruzada (Tabela VI) e da avaliação final com dados previamente separados (Tabela VII), verifica-se que, em termos de *AUC-ROC*, na etapa de testes foram alcançados resultados superiores àqueles da fase de treinamento. Isso sugere que os modelos construídos alcançaram um estado de generalização satisfatório para predição a partir de novos dados.

Considerando medidas mais gerais, como *F1*, Acurácia e *AUC-ROC*, apesar de próximos, o método baseado em *Random Forest* alcançou resultados ligeiramente superiores, indicando melhor eficácia geral no processo preditivo para óbitos e não óbitos. Apesar disso, dada a criticidade do contexto da aplicação e os impactos dos diferentes tipos de erro (falsos positivos e falsos negativos), é necessário fazer análises mais detalhadas da eficácia de cada algoritmo. Por exemplo, apesar de não alcançar os melhores resultados gerais, o método de árvore decisão foi aquele com o menor número de falsos positivos (indicação incorreta de óbito), apenas 232 (7,7%) de 3000. Neste sentido, é importante destacar que além de terem alcançado resultados superiores para as medidas mais gerais, o métodos baseados em *Random Forest* e *XGBoost* também resultaram em baixas quantidades de falsos positivos (maior precisão), sugerindo sua melhor adequação geral nestes aspectos. Por outro lado, considerando os sérios impactos da ocorrência de falsos positivos (indicação incorreta de não óbito), o método baseado em *Naive Bayes* foi aquele com a menor quantidade de falsos negativos (maior recall), apenas 640 (21,3%) de 3000. Assim, apesar dos baixos valores de precisão, este foi o algoritmo que mais acertadamente identificou o risco de óbito para os pacientes, que poderiam então ter tratamento diferenciado, mesmo que isso demandasse um maior número de pacientes sendo desnecessariamente definidos com de maior risco e gerando maior demanda para o serviço de saúde.

Table VI. Melhores AUC-ROC alcançados pelos algoritmos de classificação via *GridSearch* com validação cruzada.

Algoritmos	One-Hot Encoding	Mais Frequente	knn-1	knn-3	knn-5
Naive Bayes	**0,7100**	0,6400	0,500	0,500	0,500
Árvores de Decisão	**0,7859**	0,7471	0,7297	0,7284	0,7331
Random Forest	**0,7900**	0,7808	0,7700	0,769	0,7721
XGBoost	**0,7891**	0,7848	0,7774	0,7754	0,7772

Table VII. Resultados da predição para base de testes utilizando as melhores configurações dos algoritmos.

Algoritmo	Precisão	Recall	F1	Acurácia	AUC ROC
Naive Bayes	0,6933	**0,7863**	0,7370	0,7193	0,7100
Árvores de decisão	**0,8961**	0,6667	0,7645	0,7947	0,7947
Random Forest	0,8831	0,7000	**0,7810**	**0,8037**	**0,8037**
XGBoost	0,8773	0,6937	0,7748	0,7983	0,7983

A) Árvore de Decisão

Classe real	**Classe predita** Não óbito	Óbito
Não óbito	2768	232
Obito	1000	2000

B) *Random Forest*

Classe real	**Classe predita** Não óbito	Óbito
Não óbito	2722	278
Obito	900	2100

C) *XGBoost*

Classe real	**Classe predita** Não óbito	Óbito
Não óbito	2709	291
Óbito	919	2081

D) *Naive Bayes*

Classe real	**Classe predita** Não óbito	Óbito
Não óbito	1956	1044
Óbito	640	2360

Fig. 2. Matrizes de confusão para cada algoritmo utilizado na classificação.

5. CONCLUSÃO

Este trabalho explorou técnicas de engenharia de dados e aprendizagem de máquina para predição de óbitos em pacientes com câncer de boca. Foram utilizados dados reais em escala nacional brasileira e conduzido um processo de validação experimental amplo e rigoroso. Os resultados indicaram a eficácia da técnica de *one-hot-encoding* para o tratamento do problema de valores ausentes, um dos maiores desafios da aplicação aqui abordada. Em termos do poder de predição de óbitos, os resultados indicaram eficácia bastante promissora, especialmente considerando dados originalmente com problemas de qualidade e consistência. Adicionalmente, verificou-se que apesar da melhor eficácia geral ter sido alcançada com o algoritmo de *Random Forest*, a minimização de falsos positivos e falsos negativos foi alcançada, respectivamente, por árvores de decisão e *Naive Bayes*, que são considerados mais simples e menos custosos computacionalmente. Estes resultados demonstram o potencial da engenharia de dados e modelos preditivos baseados em aprendizagem de máquina no auxílio à tomada de decisão a partir de grandes bases de dados. Destaca-se ainda que o processo proposto para predição de óbitos por câncer de boca pode ser adaptado para outros tipos de câncer com dados disponíveis no RHC e submetidos ao processo de validação. Por fim, novos experimentos podem ser conduzidos, por exemplo, com maiores volumes de dados, outros algoritmos de aprendizado ou métodos de fusão, especialmente considerando o desafio de minimização simultânea de erros dos tipos falsos positivos e falsos negativos.

REFERENCES

Borges, D., Sena, M., Ferreira, M., and Roncalli, Mortalidade por câncer de boca e condição sócio-econômica no brasil. *Cadernos de Saúde Pública* vol. 25(2), pp. 321–327, 2009.

Conway, D., Petticrew, M., Marlborough, H., Berthiller, J., Hasbibe, M., and Macpherson, L. Socioeconomic inequalities and oral cancer risk: a systematic review and meta-analysis of case-control studies. *International journal of câncer.* vol. 122:2811–2819, 2008.

Dantas, T., Silva, P., Sousa, E., Cunha, M., Aguiar, A., Costa, F., Mota, M., Alves, A., and Sousa, F. Influence of educational level, stage, and histological type on survival of oral cancer in a brazilian population: A retrospective study of 10 years observation. *Medicine (Baltimore)*, 2016.

Delen, D. Analysis of cancer data: a data mining approach. *The Journal of Knowledge Engineering, Expert Systems* 26 (1): 100–112, 2009.

Fiocruz. Determinantes sociais. https://pensesus.fiocruz.br/determinantes-sociais, 2021. acessado: 07-08-2021.

Groome, P., Rohland, S., Hall, S., Irish, J., Mackillop, M., and O'Sullivan, B. A population-based study of factors associated with early versus late stage oral cavity cancer diagnoses. *Oral oncology, 47(7):642-647.*, 2011.

IARC. Cancer tomorrow [internet]. http://gco.iarc.fr/tomorrow/graphicbar?type=1&population=900&mode=population&sex=0&cancer=39&age_group=value&apc_male= 0&apc_female=0, 2020. acessado: 07-08-2021.

INCA. O que é câncer. https://www.inca.gov.br/o-que-e-cancer, 2019a. acessado: 07-08-2021.

INCA. Registro hospitalar de câncer. https://www.inca.gov.br/numeros-de-cancer/registros-hospitalares-de-cancer-rhc, 2019b. acessado: 07-08-2021.

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, 2020.

INCA. Câncer de boca [internet]. https://www.inca.gov.br/tipos-de-cancer/cancer-de-boca, 2021. Acessado: 07-08-21.

Ministerio da Saude do Brasil. Manual de Bases Técnicas da Oncologia – SIA/SUS - Sistema de Informações Ambulatoriais. https://www.inca.gov.br/sites/ufu.sti.inca.local/files//media/document//manual-oncologia-26a-edicao.pdf, 2019. acessado: 09-08-2021.

Neville, B., Damm, D., Allen, C., and JE, J. B. *Patologia oral e Maxilofacial.* GEN Guanabara Koogan, Rio de Janeiro, 2020.

Salmi, N. and Rustam, Z. Naïve bayes classifier models for predicting the colon cancer. *IOP Conference Series: Materials Science and Engineering* vol. 546, pp. 052068, jun, 2019.

Sharma, N. and Om, H. Data mining models for predicting oral cancer survivability. *Network Modeling Analysis in Health Informatics and Bioinformatics* vol. 2, pp. 285–295, 2013.

Tseng, W., Chiang, W., Liu, S., Roan, J., and Lin, C. The application of data mining techniques to oral cancer prognosis. *Journal of Medical Systems* 39 (59), 2015.

Warnakulasuriya, S. Global epidemiology of oral and oropharyngeal cancer. *Oral Oncol* vol. 45(4-5), pp. 309–316, 2009.