UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE TECNOLOGIA E GEOCIÊNCIAS
DEPARTAMENTO DE ENGENHARIA BIOMÉDICA
GRADUAÇÃO EM ENGENHARIA BIOMÉDICA

SÉRGIO DE VASCONCELOS FILHO

Predição de remissão de pacientes oncológicos: uma abordagem baseada em ensembles

Recife

2022

SÉRGIO DE VASCONCELOS FILHO

Predição de remissão de pacientes oncológicos: uma abordagem baseada em ensembles

Trabalho apresentado ao curso de graduação em Engenharia Biomédica do departamento de Engenharia Biomédica da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de bacharel em Engenharia Biomédica.

Área de Concentração: Engenharia Biomédica

Orientador: Wellington Pinheiro dos Santos

Coorientador: Fernando Maciano de Paula Neto

Recife

2022

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

de Vasconcelos Filho, Sérgio.
Predição de remissão de pacientes oncológicos: uma abordagem baseada em ensembles / Sérgio de Vasconcelos Filho. - Recife, 2022.
56 : il., tab.

Orientador(a): Wellington Pinheiro dos Santos
Cooorientador(a): Fernando Maciano de Paula Neto
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Tecnologia e Geociências, Engenharia Biomédica - Bacharelado, 2022.

1. Inteligência Artificial. 2. Aprendizagem de Máquina. 3. Câncer. 4. XGBoost. 5. SEER. I. Pinheiro dos Santos, Wellington. (Orientação). II. Maciano de Paula Neto, Fernando. (Coorientação). III. Título.

620 CDD (22.ed.)

SÉRGIO DE VASCONCELOS FILHO

Predição de remissão de pacientes oncológicos: uma abordagem baseada em ensembles

Trabalho apresentado ao curso de graduação em Engenharia Biomédica do departamento de Engenharia Biomédica da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de bacharel em Engenharia Biomédica.

Aprovado em __/__/____.

BANCA EXAMINADORA

_____

Prof. Dr. Wellington Pinheiro dos Santos,

Departamento de Engenharia Biomédica da UFPE

_____

Profa. MSc. Juliana Carneiro Gomes,

Universidade de Pernambuco

_____

Profa. MSc. Maíra Araújo de Santana,

Universidade de Pernambuco

À minha família, Ana e Camilla, por me ensinar a viver. A mim mesmo, por não ter desistido nos piores momentos.

AGRADECIMENTOS

À minha mãe, Ana, que mesmo com todas as intempéries de ser uma mãe solo, conseguiu ensinar seus dois filhos a serem gente. Sem humanidade não somos nada.

À minha irmã, Camilla, por ser um exemplo para mim de resiliência, compaixão e altruísmo. Sou o que sou hoje também por causa de você, minha irmã.

A Alice, por ser minha pessoa.

A Giullia e Nathália, por fazermos um trio imbatível no tempo caótico da pandemia e de final de curso. Eu definitivamente descarrilharia se não fosse por vocês duas.

A Arlen, Victor, Ana e Felipe, companheiros do McFritos, o melhor grupo possível para atravessar as adversidades de um curso de engenharia e a primeira fase da vida adulta.

A Gabriel, por me fazer ver-me por inteiro quando ninguém mais pôde.

A Karla, por sempre me escutar com atenção e por falar o que é necessário, mesmo que seja doloroso.

Ao MegaZord, o grupo pelo qual conseguimos destruir os maiores vilões de um fim de graduação de engenharia.

A Cristine, pela paciência por me ensinar e pela sabedoria em me conduzir pelos caminhos da pesquisa acadêmica, pelo qual me abriu as portas das oportunidades que serei eternamente grato.

A Fernando, por acreditar no meu potencial de unir a área da computação e saúde. Graças à sua confiança que cheguei até aqui.

A Wellington, por sempre ter um olhar amplo e crítico sobre a universidade, até para além dos muros. Sua visão me inspira a utilizar a engenharia biomédica sempre para o interesse comum, jamais para qualquer outro.

Ao departamento de Engenharia Biomédica, em especial Rangel, a personificação da solicitude e boa vontade à frente da secretaria da graduação.

RESUMO

Câncer é uma das principais doenças não infecciosas que leva o ser humano à óbito, ceifando mais de 10 milhões de vidas todos os anos. Abordagens utilizando inteligência artificial são propostas, inclusive aquelas que inserem em suas metodologias algoritmos de aprendizagem de máquina. Não obstante, abordagens que usam combinação de algoritmos desempenham uma resposta mais eficiente comparada aos algoritmos em separado. Dessa forma, o trabalho se utiliza de uma base de dados do Estados Unidos, o Surveillance, Epidemiology, and End Results (SEER), que possui mais de 140 atributos sobre o paciente e 10 milhões de registros de tumores primários, com o intuito de predizer a remissão do dito cujo. Isso auxiliará tanto o diagnóstico como o prognóstico do paciente, sendo proposto então um sistema que auxilie a decisão médica. O eXtreme Gradient Boosting foi o algoritmo utilizado como modelo de aprendizado de máquina, tendo apenas sua melhor iteração extraída como hiperparâmetro otimizado. Após a divisão da base em treino, validação e teste para cada um dos 10 agrupamentos de tumores específicos, observou-se uma performance estatisticamente melhor combinando o modelo genérico, treinado na base inteira, e os modelos específicos, treinados em subconjuntos da base completa. Espera-se contribuir para a predição de remissão de câncer no contexto dessa base de dados, tendo em vista que nenhum outro trabalho propôs utilizar a base inteira e todos os anos disponíveis. Além disso, espera-se incentivar outros engenheiros biomédicos a trabalhar na área de ciência de dados, já que as contribuições nessa temática podem ser surpreendentes.

Palavras-chaves: Inteligência Artificial. Aprendizagem de Máquina. Câncer. XGBoost. SEER. Base de dados.

ABSTRACT

Cancer is one of the main non-infectious diseases that leads to human death, reaping more than 10 million lives every year. Approaches using artificial intelligence are proposed, including those that include machine learning algorithms in their methodologies. Nevertheless, approaches that use a combination of algorithms outperforms results compared to separate algorithms. In this way, this work uses a database from the United States, the Surveillance, Epidemiology, and End Results (SEER), that has more than 140 attributes about the patient and 10 million primary tumor records, in order to predict the remission of these patients. This will help both the diagnosis and the prognosis of them, being then proposed a system that helps the medical decision. The eXtreme Gradient Boosting algorithm was used as a machine learning model, with only its best iteration extracted as an optimized hyperparameter. After dividing the base into training, validation and testing for each of the 10 clusters of specific tumors, a statistically better performance was observed by combining the generic model, trained on the entire base, and the specific models, trained on subsets of the full dataset. It is expected to contribute to the prediction of cancer remission in the theme of this database, considering that no other work proposed to use the entire base and all the years available. In addition, it is hope to encourage other biomedical engineers to work in the field of data science, as the contributions in this area can be surprising.

Keywords: Artificial Intelligence. Machine Learning. Cancer. XGBoost. SEER. Databases.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de Máquina
AUC	*Area Under the Curve*
CID	Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde
CID-O	Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde para Oncologia
IA	Inteligência Artificial
IDH	Índice de Desenvolvimento Humano
NCI	*National Cancer Institute*
OMS	Organização Mundial da Saúde
ROC	*Receiver Operating Characteristics*
SEER	*Surveillance, Epidemiology, and End Results*
XGBoost	*eXtreme Gradient Boosting*

SUMÁRIO

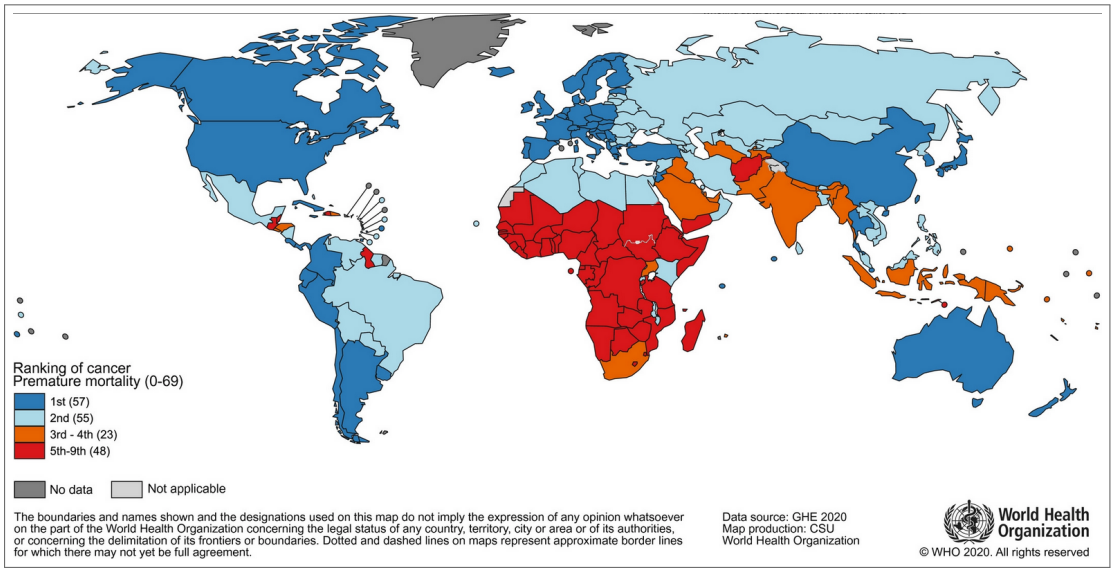
1 INTRODUÇÃO

1.1 MOTIVAÇÃO

Dados recentes continuam demonstrando que o câncer é uma das doenças que mais causam óbitos mundialmente, ficando atrás apenas de doenças cardiovasculares (BRAY et al., 2021). De fato, a enfermidade discutida é a primeira ou a segunda principal causa de morte em pessoas com menos 70 anos em 112 países, sendo o terceiro ou quarto em mais 23 (WORLD HEALTH ORGANIZATION, 2021).

Tendo essa informação distribuída geograficamente, é possível obter a Figura 1. Assim, destaca-se o ranking na África, continente onde se localiza muitos países não desenvolvidos: mesmo com uma clara correlação entre o Índice de Desenvolvimento Humano (IDH) e uma baixa expectativa de vida (UNITED NATIONS, 2020), percebe-se a severidade de casos da doença. No geral, essas informações sumarizam que mais de 10 milhões de vidas são perdidas todos os anos por causa dessa moléstia (SUNG et al., 2021).

Figura 1 – Ranking geográfico da mortalidade prematura por câncer



Fonte: WORLD HEALTH ORGANIZATION (2021)

Além do mais, prospecções indicam que, em países com baixo IDH, a prevalência dessa enfermidade irá praticamente dobrar até 2040 (SUNG et al., 2021). Não obstante, tais regiões são as mais atingidas pelas mudanças climáticas, poluição e corrupção (UNITED NATIONS, 2020), abalando ainda mais seus já frágeis sistemas de saúde.

Portanto, é primordial e urgente reforçar tais sistemas de saúde com tecnologias de apoio à decisão médica, de forma a acelerar e tornar mais assertivo o diagnóstico e prognóstico clínico. Para isso, abordagens utilizando Inteligência Artificial (IA), mais especificamente AM, são feitas na tentativa de predizer informações de saúde relacionadas ao câncer (BHINDER et al., 2021).

1.2 JUSTIFICATIVA

As metodologias empregadas em pesquisas que utilizam IA para essa problemática do câncer comumente se limitam a um tipo específico de tumor (CHENG; LANG, 2020; DOPPALAPUDI; QIU; BADR, 2021; LUO et al., 2022), carecendo de técnicas que generalizem para agrupamentos maiores da referida doença. Assim, torna-se evidente a necessidade do estudo de métodos que consigam abarcar variados tipos de tumores, com o intuito de gerar um sistema mais robusto e amplo para apoio à decisão clínica.

Somado a isso, pesquisas indicam que utilizar uma combinação de algoritmos de AM para um mesmo problema pode melhorar a eficiência da resposta, comparado com os mesmos algoritmos de modo isolado (NASEEM et al., 2022; GIANNUZZI et al., 2022). Assim, o presente trabalho propõe unificar a proposta generalista de câncer com a combinação de algoritmos de AM para inferir informações que apoiem o diagnóstico e o prognóstico em pacientes oncológicos.

1.3 OBJETIVO

Utilizar algoritmos de AM em uma base de dados contendo informações sobre pacientes com câncer para extrair características gerais e específicas de suas enfermidades para predizer a remissão do dito cujo.

1.4 OBJETIVOS ESPECÍFICOS

Para alcançar o objetivo geral do trabalho, é necessário:

1. Realizar o pré-processamento da base de dados de câncer, com o intuito de gerar os atributos e as variáveis-alvo para o treinamento dos algoritmos de AM;

2. Dividir a base de dados de maneira estratificada em relação a grupos específicos de tumor, de modo a obter amostras semelhantes entre si no que tange à aspectos morfológicos do câncer;

3. Separar os dados em conjuntos de treino, validação e teste e garantir a inexistência de registros duplicados entre eles;

4. Montar a *pipeline* de treinamento dos algoritmos de AM com o devido ajuste de hiperparâmetros e extração de métricas de desempenho;

5. Combinar os algoritmos treinados na base completa e nos subgrupos específicos;

6. Avaliar a qualidade dos resultados obtidos e verificar se há relevância estatística na abordagem utilizada.

1.5 ORGANIZAÇÃO DO TRABALHO

O presente trabalho está organizado como se segue: inicialmente, no capítulo 2, tratar-se-ão fundamentos sobre os conceitos utilizando ao decorrer do trabalho. Mais especificamente, a seção 2.1 discorrer-se-á brevemente sobre definições básicas de câncer, classificações internacionais e estadiamentos e a seção 2.2 será descrito sobre aprendizagem supervisionada, formas de divisão de bases de dados e algoritmos de AM que serão utilizados no presente trabalho.

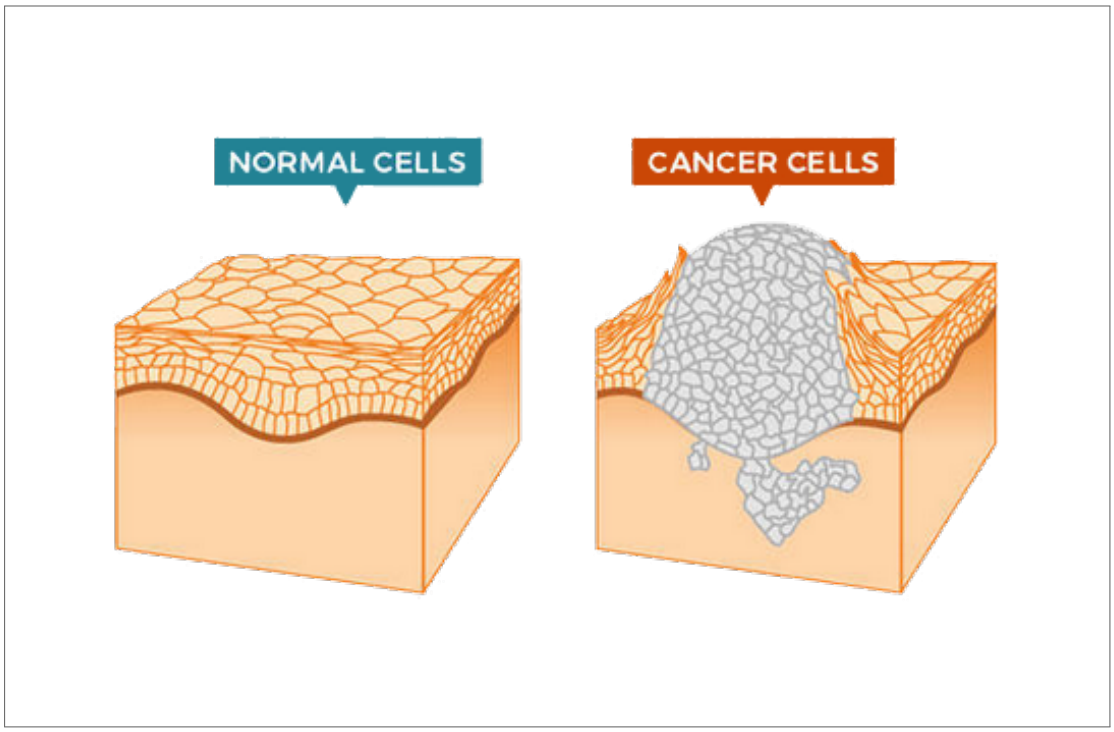
A seguir, o capítulo 3 trata sobre o conjunto de dados utilizado pelo trabalho concomitantemente com os algoritmos citados no capítulo anterior para alcançar os objetivos do dito cujo. Assim, também é abordado nesse capítulo algumas técnicas de pré-processamento de dados e trabalhos anteriores utilizando essa base com algoritmos de AM. Além disso, discute sobre a metodologia utilizada no trabalho, de que forma a base e os algoritmos citados foram manipulados de forma a se obter os resultados, sendo esses apresentados no capítulo 4 e discutidos no capítulo 5. Por fim, o capítulo 6 evidencia as conclusões decorridas do presente trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 NEOPLASIA

Neoplasia é o termo utilizado para indicar um crescimento anormal de células. Caso esse crescimento seja localizado e essas células não sejam muito diferenciadas do tecido original, é designado o nome de neoplasia benigna. Caso haja invasão a outros tecidos e grande diferenciação, chama-se neoplasia maligna ou, popularmente, câncer (NACIONAL CANCER INSTITUTE, 2021). Uma representação gráfica disso pode ser vista na Figura 2.

Figura 2 – Representação de células normais e cancerígenas



Fonte: NACIONAL CANCER INSTITUTE (2021)

As causas para essa doença ainda não são completamente entendidas, mas já se sabe que ela é resultado da genética do indivíduo com: carcinogênicos físicos, como radiações ionizantes, químicos, como álcool e componentes do cigarro, e biológicos, como vírus e bactérias (WORLD HEALTH ORGANIZATION, 2022a).

Além disso, já se sabe que o risco de desenvolver câncer aumenta proporcionalmente com o envelhecer. Assim, ao se somar outros fatores de risco, como uso de tabaco, dietas desbalanceadas, sedentarismo e poluição, chances maiores da aparição de um tumor são reforçadas (WORLD HEALTH ORGANIZATION, 2022a).

Não obstante, países de baixa renda são mais afetados por infecções crônicas, pelas quais algumas são carcinogênicas. Dessa maneira, aproximadamente 13 % dos diagnósticos globais de câncer em 2018 foram atribuídos a infecções cancerígenas, incluindo papilomavírus humano e vírus das hepatites B e C (MARTEL et al., 2020).

A seguir, é descrito na seção 2.1.1 a classificação utilizada mundialmente para agrupar e estudar as neoplasias. Na mesma linha de raciocínio, na seção 2.1.2, descreve-se brevemente sobre outros tipos de agrupamento, levando em conta aspectos morfológicos e gravidade do câncer.

2.1.1 Classificação Internacional de Doenças

A Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde (CID) determina a classificação e codificação das doenças e uma ampla variedade de sinais, sintomas, achados anormais, denúncias, circunstâncias sociais e causas externas de danos e/ou doença. Ela é publicada pela Organização Mundial da Saúde (OMS) e é usada globalmente para estatísticas de morbilidade e de mortalidade, sistemas de reembolso e de decisões automáticas de suporte em medicina.

O sistema foi desenhado para permitir e promover a comparação internacional da coleção, processamento, classificação e apresentação do tipo de estatísticas das enfermidades (WORLD HEALTH ORGANIZATION, 2022b). Na sua décima edição, a CID possui um capítulo específico para neoplasias, dividindo-as em mais de 140 grupos (WORLD HEALTH ORGANIZATION, 2019).

De modo a estender o sistema já existente, a OMS criou a Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde para Oncologia (CID-O). Em sua terceira edição, essa classificação mais específica inclui categorias morfológicas do tumor, aumentando o nível de detalhe da diferenciação (WORLD HEALTH ORGANIZATION, 2013).

2.1.2 Estadiamento

O estadiamento do câncer advém da constatação de que as taxas de sobrevida são diferentes quando a doença está restrita ao órgão de origem ou quando ela se estende a outros órgãos (MINISTÉRIO DA SAÚDE, 2021).

O estadiamento pode ser resumido em três classificações: o tamanho do tumor, indicado pela letra T, se há envolvimento de linfonodos regionais, sumarizado pela letra N, e se há

presença de metástases distantes, indicado pela letra M, sendo assim conhecido como Estadiamento TNM (GREENE et al., 2006).

O Quadro 1 detalha os níveis da classificação T. Dessa forma, percebe-se uma correlação entre o aumento da numeração e da gravidade do tumor. De modo semelhante, os Quadros 2 e 3 detalham os níveis das classificações N e M, respectivamente. Nota-se uma menor cardinalidade na classe M, indicando a severidade do câncer quando há a presença de metástases distantes.

Quadro 1 – Classificação T do Estadiamento TNM: tamanho do tumor

Categoria	Descrição
TX	Tumor primário não pode ser avaliado.
T0	Sem evidência de tumor primário.
Tis	Carcinoma *in situ*.
T1, T2, T3, T4	Avanço crescente do tamanho e/ou extensão do local do tumor primário.

Fonte: GREENE et al. (2006)

Quadro 2 – Classificação N do Estadiamento TNM: linfonodos regionais

Categoria	Descrição
NX	Linfonodos regionais impossibilitados de serem avaliados.
N0	Linfonodos regionais sem metástase.
N1, N2, N3	Envolvimento crescente dos linfonodos regionais.

Fonte: GREENE et al. (2006)

Quadro 3 – Classificação M do Estadiamento TNM: metástase distante

Categoria	Descrição
MX	Metástase distante não pôde ser avaliada.
M0	Sem metástase distante.
M1	Presença de metástase distante.

Fonte: GREENE et al. (2006)

2.2 APRENDIZAGEM DE MÁQUINA

A AM é um ramo da Ciência da Computação, mais especificamente dentro da área de IA, voltada para o entendimento de métodos que se apoiam em informações para aumentar a performance em determinadas tarefas (MITCHELL, 1997). Dessa maneira, esses algoritmos constroem um modelo baseado em dados amostrais para realizar decisões sem serem explicitamente programados para isso (KOZA et al., 1996).

Esses algoritmos são utilizados para uma grande variedade de aplicações, como em medicina, filtragem de e-mail, reconhecimento de fala e visão computacional, onde é difícil ou inviável desenvolver algoritmos convencionais para realizar as tarefas necessárias (HU et al., 2020). Além disso, também é utilizado para previsão do clima (BOCHENEK; USTRNUL, 2022), controle de aviões (MUR et al., 2022) e até mesmo estabilização de fusões nucleares (DEGRAVE et al., 2022).

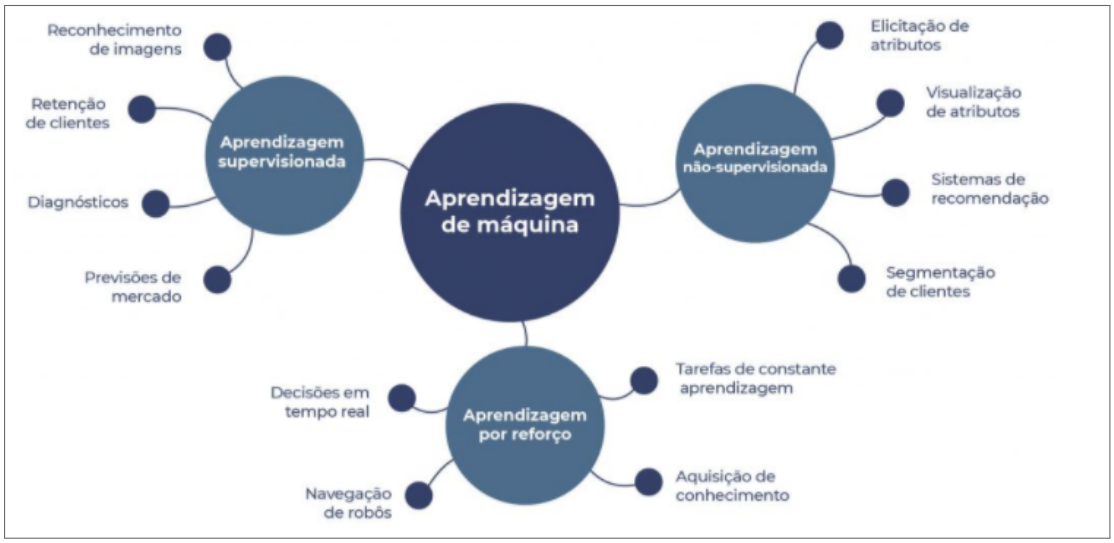
Dentro da área de AM existem três grandes subgrupos, a aprendizagem supervisionada, não supervisionada e a aprendizagem por reforço. A primeira, como será mais detalhada na seção 2.2.1, lida com os problemas que possuem registros com variáveis de entrada e uma resposta desejada, a variável de saída. O intuito é conseguir com que o algoritmo extraia as características relevantes do problema para aplicá-las em registros futuros (RUSSELL; NORVIG, 2009).

Em contraste com o aprendizado supervisionado, o não supervisionado não requere a informação da variável de saída, sendo aplicado em problemas de representação de moléculas (JAEGER; FULLE; TURK, 2018) e física de partículas (ANDREASSEN et al., 2019), por exemplo. Já a aprendizagem por reforço é o intermédio das duas anteriormente citadas, utilizando apenas um sinal positivo ou negativo (o reforço) para indicar se a ação geradora do estímulo foi benéfica ou não de acordo com o problema proposto. Esse tipo de aprendizado é visto em agentes autômotos (MNIH et al., 2013) e na robótica (KOBER; BAGNELL; PETERS, 2013). Uma sumarização desses tipos de aprendizado de máquina pode ser visto na Figura 3.

Assim sendo, a seção 2.2.1 refere-se em específico ao aprendizado supervisionado. É através desse tipo de aprendizado em que as predições propostas na seção 1.4 serão criadas. A seguir, detalha-se as bases de dados utilizados necessárias para tal aprendizagem ocorrer, na seção 2.2.2. Prosseguindo, a seção 2.2.3 e 2.2.4 tratam de fundamentos teóricos que embasam o algoritmo abordado em 2.2.5, pelo qual esse presente trabalho utilizará. Por fim, na seção 2.2.6, discorre-se sobre indicadores de desempenho dos modelos gerados pelos algoritmos de

AM.

Figura 3 – Sumarização das áreas de AM



Fonte: Elaborada pelo autor (2022)

2.2.1 Aprendizagem Supervisionada

Algoritmos de aprendizagem supervisionada constroem um modelo matemático através de um conjunto de dados que contém tanto as informações de entrada como as informações de saída (RUSSELL; NORVIG, 2009). Dessa forma, esse conjunto é chamado de base de treinamento e é constituído por uma coleção de amostras de treino. Cada amostra possui uma ou mais entradas e sua respectiva saída, essa última conhecida também como variável alvo ou dependente. Caso esse atributo dependente represente categorias, é definido que a tarefa é do grupo de problemas de classificação. Caso a variável alvo represente um conjunto contínuo, o problema é de regressão (FREEDMAN, 2009).

Na visão matemática, cada amostra do conjunto é representanda por um vetor, e o conjunto por completo, por uma matriz. Por meio da otimização iterativa de uma função objetivo, o algoritmo de aprendizado supervisionado extrai uma função que pode ser usada para prever a saída associada a novas entradas (MOHRI; ROSTAMIZADEH; TALWALKAR, 2012).

Uma função otimizada permitirá que o algoritmo determine a saída para entradas que não faziam parte dos dados de treinamento. Diz-se que um algoritmo que melhora a precisão de suas saídas ou previsões ao longo do tempo aprendeu a realizar essa determinada tarefa (MITCHELL, 1997).

2.2.2 Conjuntos de Treino, Validação e Teste

Para o correto uso de algoritmos de AM, é necessário conjuntos distintos de uma base de dados. Em particular, três coleções são utilizadas: base de treino, validação e teste. O primeiro é a base pelo qual o algoritmo irá ser treinado, ou seja, exemplos pelos quais o modelo gerado extrairá as características relevantes do problema tratado. De acordo com a resposta atual do algoritmo aos exemplos de entrada e a resposta verdadeira, os parâmetros do modelo são ajustados (RIPLEY, 2008).

Seguidamente, o modelo ajustado é usado para prever as respostas para as observações em um segundo conjunto de dados, esse chamado de base de validação (JAMES et al., 2013). Ele fornece uma avaliação imparcial de um modelo treinado no conjunto de dados de treinamento enquanto ajusta os seus hiperparâmetros[1]. A referida base pode ser usada para regularização do modelo por interrupção antecipada, interrompendo o treinamento quando o erro no conjunto de dados de validação aumenta, pois isso é um sinal de ajuste excessivo ao conjunto de dados de treinamento, comumente chamado de *overfitting* (PRECHELT, 2012).

Por fim, o conjunto de dados de teste é um conjunto de dados usado para fornecer uma avaliação imparcial de um modelo ajustado utilizando as bases de treinamento e validação. Somado a isso, decidir os tamanhos e estratégias para a divisão das coleções de dados em conjuntos de treinamento, teste e validação depende muito do problema e das informações disponíveis (JAMES et al., 2013).

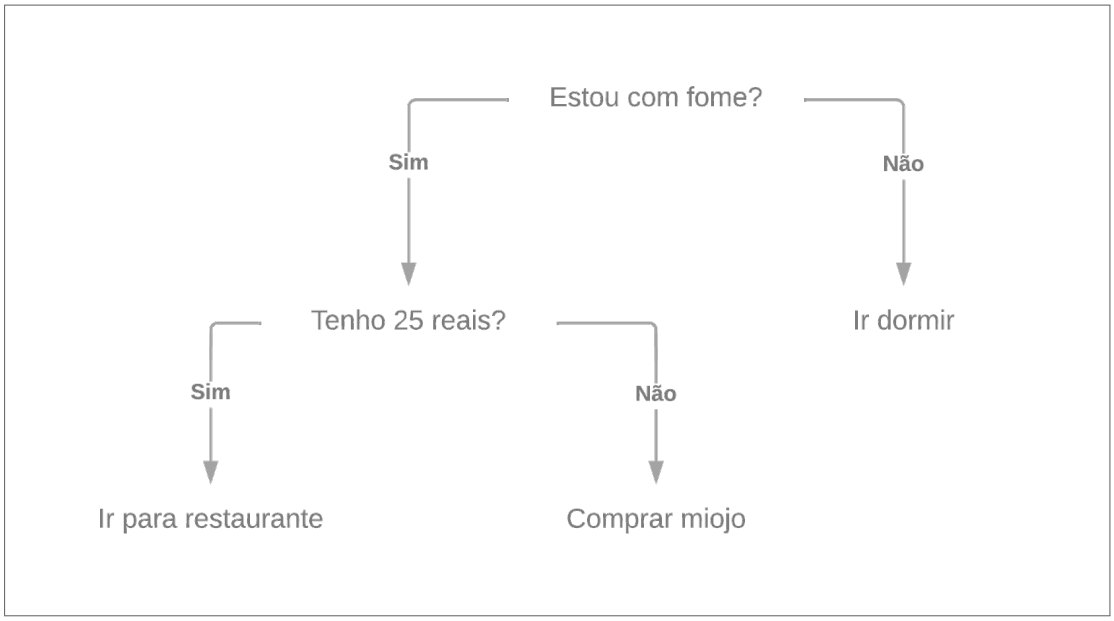
2.2.3 Árvore de Decisão

Essa seção possui forte embasamento no trabalho de Breiman, criador do algoritmo da árvore de decisão (BREIMAN, 2017). O referido algoritmo é uma das abordagens de modelagem preditiva usadas em estatística, mineração de dados e AM. Ele é utilizado para ir das observações sobre um item (representado nos ramos) até as conclusões sobre o valor alvo do item (representado nas folhas). Um exemplo de árvore pode ser visto na Figura 4. Esse exemplo deixa claro a utilização de decisões binárias, como "estou com fome?" e "tenho 25 reais?", e os resultados tomadas a depender da situação atual. A mesma lógica é aplicada em dados tabelares: atributos são selecionados do conjunto de dados e decisões binárias são realizadas,

[1] Hiperparâmetro é um parâmetro cujo valor é usado para controlar o processo de aprendizagem. Diferentemente de um parâmetro comum, o hiperparâmetro não pode ser inferidos durante o processo de aprendizado tendo em vista que ele se refere à seleção do modelo (YANG; SHAMI, 2020).

sendo o atributo selecionado para o próximo ramo aquele que minimize a impureza de Gini, detalhada na seção 2.2.3.1. A depender da amostra atual, diferentes caminhos são trilhados, atribuindo robustez ao modelo de AM.

Figura 4 – Exemplo de árvore de decisão



Fonte: Elaborada pelo autor (2022)

Os modelos em que a variável alvo pode assumir um conjunto discreto de valores são chamados de árvores de classificação; nessas estruturas de árvore, as folhas representam rótulos de classe e ramos representam conjunções de recursos que levam a esses rótulos de classe. As árvores de decisão em que a variável dependente pode assumir valores contínuos são chamadas de árvores de regressão. Esse tipo de modelo estão entre os algoritmos de AM mais populares devido à sua inteligibilidade e simplicidade (BREIMAN, 2017).

Uma árvore é construída dividindo o conjunto fonte, constituindo o nó raiz da árvore, em subconjuntos – que constituem os filhos sucessores. A divisão é baseada em um conjunto de regras de divisão baseadas em características de classificação (SHALEV-SHWARTZ; BEN-DAVID, 2014). Um exemplo dessas regras se baseia na Impureza de Gini, detalhado na seção a seguir. Esse processo é repetido em cada subconjunto derivado recursivamente, e ela é concluída quando o subconjunto em um nó tem todos os mesmos valores da variável alvo ou quando a divisão não adiciona mais valor às previsões.

Esse processo de indução de árvores de decisão de cima para baixo é um exemplo de algoritmo guloso[2] (QUINLAN, 1986) e é de longe a estratégia mais comum para aprender

[2] Qualquer algoritmo que segue a heurística de solução de problemas de fazer a escolha localmente ótima

árvores de decisão a partir de dados (ROKACH; MAIMON, 2005).

2.2.3.1 Impureza de Gini

A impureza de Gini é uma medida de frequência de quanto um elemento escolhido aleatoriamente do conjunto seria rotulado incorretamente se fosse rotulado aleatoriamente de acordo com a distribuição de rótulos no subconjunto (BREIMAN, 2017). A impureza Gini pode ser calculada utilizando (2.1) abaixo:

$$I_G(p) = \sum_{i=1}^{J} \left(p_i \sum_{k \neq i} p_k \right) \tag{2.1}$$

O termo p_i refere-se à fração dos itens no referido conjunto de dados rotulados com a categoria i e o somatório de com o termo $p_k = 1 - p_i$ sendo a probabilidade do erro em categorizar corretamente o item. J se refere ao conjunto de classes possíveis na base, ou seja, $i \in \{1, 2, \ldots, J\}$. Com base nessas informações, é possível reescrever (2.1) da maneira exibida em (2.2):

$$\begin{aligned} I_G(p) &= \sum_{i=1}^{J} \left(p_i \sum_{k \neq i} p_k \right) \\ &= \sum_{i=1}^{J} \left[p_i \left(1 - p_i \right) \right] \\ &= \sum_{i=1}^{J} \left(p_i - p_i^2 \right) \\ &= \sum_{i=1}^{J} p_i - \sum_{i=1}^{J} p_i^2 \\ &= 1 - \sum_{i=1}^{J} p_i^2 \end{aligned} \tag{2.2}$$

Dessa maneira, a impureza de Gini pode ser utilizada para avaliar o ganho de informação da ávore. Ela atinge seu mínimo, o valor zero, quando todos os casos do nó se enquadram em uma única categoria.

em cada estágio (GUTIN; YEO; ZVEROVICH, 2002).

2.2.4 *Gradient Boosting*

Gradient Boosting é uma técnica de AM que fornece um modelo de previsão na forma de um conjunto de modelos fracos, que normalmente são árvores de decisão (PIRYONESI; EL-DIRABY, 2020; HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Tratando-se agora sobre o algoritmo em si, discorre-se que em muitos problemas de aprendizado supervisionado existe uma variável de saída y e um vetor de variáveis de entrada x, relacionadas entre si com alguma distribuição probabilística. O objetivo é encontrar alguma função $\hat{F}(x)$ que melhor aproxime a variável de saída dos valores das variáveis de entrada. Isso é formalizado introduzindo alguma função de perda $L(y, F(x))$ e minimizando-a como mostra (2.3):

$$\hat{F} = \arg\min_{F} \mathbb{E}_{x,y}[L(y, F(x))] \tag{2.3}$$

O método de *Gradient Boosting* assume um valor real de y e busca uma aproximação $\hat{F}(x))$ na forma de uma soma ponderada de funções $h_i(x)$ de uma classe $\mathcal{H}$, chamados aprendizes de base, como é mostrado em (2.4):

$$\hat{F}(x) = \sum_{i=1}^{M} \gamma_i h_i(x) + \text{const.} \tag{2.4}$$

Como citado na seção 2.2.2 é utilizado a base de dados de treinamento para que o método encontre uma aproximação de $\hat{F}(x)$ que minimize o valor médio da função de perda nesse conjunto de dados. Isso é feito iniciando com um modelo $F_0(x)$ e incrementalmente expandido através de um algoritmo guloso mostrado em (2.5) e (2.6):

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma) \tag{2.5}$$

$$F_m(x) = F_{m-1}(x) + \arg\min_{h_m \in \mathcal{H}} \left[\sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + h_m(x_i)) \right] \tag{2.6}$$

onde $h_m \in \mathcal{H}$ é a função do aprendiz de base.

Infelizmente, escolher a melhor função h em cada passo para uma função de perda arbitrária L é um problema de otimização computacionalmente inviável em geral. Portanto, abordagem

é restringida a uma versão simplificada do problema: A ideia é aplicar o conceito de gradiente descendente. Dessa maneira, a tarefa é encontrar um mínimo local da função de perda iterando em $F_m(x)$ (LAMBERS, 2011). Portanto, movendo uma pequena quantidade γ de modo que a aproximação linear permaneça válida, produz (2.7):

$$F_m(x) = F_{m-1}(x) - \gamma \sum_{i=1}^{n} \nabla F_{m-1} L(y_i, F_{m-1}(x_i)) \tag{2.7}$$

Nesse contexto, o referido algoritmo tipicamente é utilizando em árvores de decisão com uma quantidade fixada de aprendizes. Para este caso especial, é proposto uma modificação no método de aumento de gradiente que melhora a qualidade de ajuste de cada aluno de base (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

O aumento de gradiente genérico na m-ésima etapa ajustaria uma árvore de decisão $h_m(x)$ para pseudo-resíduos. Seja J_m o número de suas folhas. Desse modo, a árvore particiona o espaço de entrada em J_m regiões disjuntas $R_{1m}, \ldots, R_{J_m m}$ e prevê um valor constante em cada região. Usando a notação de indicador[3], a saída de $h_m(x)$ para a entrada x pode ser escrita como em (2.8):

$$h_m(x) = \sum_{j=1}^{J_m} b_{jm} \mathbf{1}_{R_{jm}}(x) \tag{2.8}$$

onde b_{jm} é o valor previsto na região R_{jm}.

Então os coeficientes b_{jm} são multiplicados por algum valor γ_m, escolhido usando busca de linha para minimizar a função de perda, e o modelo é atualizado como é exibido em (2.9) e (2.10):

$$F_m(x) = F_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} \mathbf{1}_{R_{jm}}(x) \tag{2.9}$$

$$\gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i + \gamma) \tag{2.10}$$

[3] Função que mapeia elementos do subconjunto para o valor um e todos os outros elementos para o valor zero (FOLLAND, 1999).

2.2.5 *eXtreme Gradient Boosting*

eXtreme Gradient Boosting (XGBoost) é um sistema de AM escalável para árvore de decisão enriquecida com *Gradient Boosting*, comumente chamada em inglês por *tree boosting*. Seu diferencial se destaca na possibilidade de ser mais econômico no que tange ao processamento de bilhões de amostras, comparados a sistemas existentes (CHEN; GUESTRIN, 2016). Dessa maneira, essa seção explicita resumidamente os principais destaques do XGBoost em relação a outros algoritmos de AM.

Tendo em vista que em alguns cenários os dados não podem ser carregados totalmente na memória, ou até mesmo em sistemas distribuídos, o sistema disponibiliza uma aproximação para o algoritmo guloso do *Gradient Boosting*. Dessa maneira, o algoritmo do XGBoost calcula os valores de divisão dos nós baseado nos quartis da distribuição da referida coluna. As variáveis contínuas então são mapeadas nesses agrupamentos e procura-se a melhor solução entre os agrupamentos propostos baseado nas estatísticas agregadas (CHEN; GUESTRIN, 2016).

Outro diferencial desse sistema é a sua maneira própria em lidar com os dados de entrada. Com o objetivo de diminuir o custo da ordenação de dados, essas informações são armazenadas em blocos na memória do sistema. Os dados armazenados neles são comprimidos por coluna e sorteados. Não obstante, o XGBoost também inovam no acesso ao cache, persistência dos dados em disco e a construção de uma arquitetura ponta-a-ponta de AM (CHEN; GUESTRIN, 2016).

2.2.6 Métricas de Desempenho

Para avaliar se o algoritmo de AM está, de fato, aprendendo a tarefa, conforme definido na seção 2.2.1, é necessário medir a performance de desempenho da resposta do modelo gerado pelo referido algoritmo. Assim, essa seção destina-se a detalhar as principais métricas de desempenho para avaliação de modelos de classificação binária (POWERS, 2008). Portanto, descreve-se primeiramente os termos em comum que aparecerão no decorrer da seção:

1. Exemplos positivos (ou apenas positivos) - P: exemplos que pertencem à classe positiva;
2. Exemplos negativos (ou apenas negativos) - N: exemplos que pertencem à classe negativa;

3. Verdadeiros positivos - VP: exemplos positivos que o modelo de decisão classificou-os como exemplos da classe positiva;
4. Verdadeiros negativos - VN: exemplos negativos que o modelo de decisão classificou-os como exemplos da classe negativa;
5. Falsos positivos - FP: exemplos negativos que o modelo de decisão classificou-os como exemplos da classe positiva;
6. Falsos negativos - FN: exemplos positivos que o modelo de decisão classificou-os como exemplos da classe negativa.

2.2.6.1 Acurácia

A métrica da acurácia dita sobre o quão próximo ou distante um determinado conjunto de medições está do seu valor real. Em linguagem matemática é mostrado em (2.11):

$$\text{Acurácia} = \frac{\text{VP} + \text{VN}}{\text{P} + \text{N}} \tag{2.11}$$

2.2.6.2 Precisão

A precisão mede o quão próximo ou distante as medições estão entre si. Matematicamente, em (2.12):

$$\text{Precisão} = \frac{\text{VP}}{\text{VP} + \text{FP}} \tag{2.12}$$

2.2.6.3 Sensibilidade

Sensibilidade pode ser definida como a porcentagem dos registros da classe positiva corretamente atribuídos. De outra forma, em (2.13):

$$\text{Sensibilidade} = \frac{\text{VP}}{\text{VP} + \text{FN}} \tag{2.13}$$

2.2.6.4 F1

A métrica F1 nada mais do que uma média geométrica entre a sensibilidade e a precisão. Ela é comumente utilizada tendo em vista a sua correlação com as duas métricas supracitadas, visto matematicamente em (2.14):

$$\text{F1} = \frac{\text{Sensibilidade} \cdot \text{Precisão}}{\text{Sensibilidade} + \text{Precisão}} \tag{2.14}$$

2.2.6.5 AUC-ROC

A curva característica do receptor, do inglês ROC, é uma curva de probabilidade. Dessa maneira, ele mede o quanto o classificador acertou versus a quantidade de vezes que o mesmo classificador errou.

Para plotar o gráfico ROC utiliza-se a Taxa dos Verdadeiros Positivos no eixo vertical, definida pela divisão entre a quantidade de Verdadeiros Positivos e o total de exemplos Positivos. De maneira semelhante, o eixo horizontal do gráfico utiliza a Taxa dos Falsos Positivos, definida pela divisão entre a quantidade de Falsos Positivos e o total de exemplos Negativos.

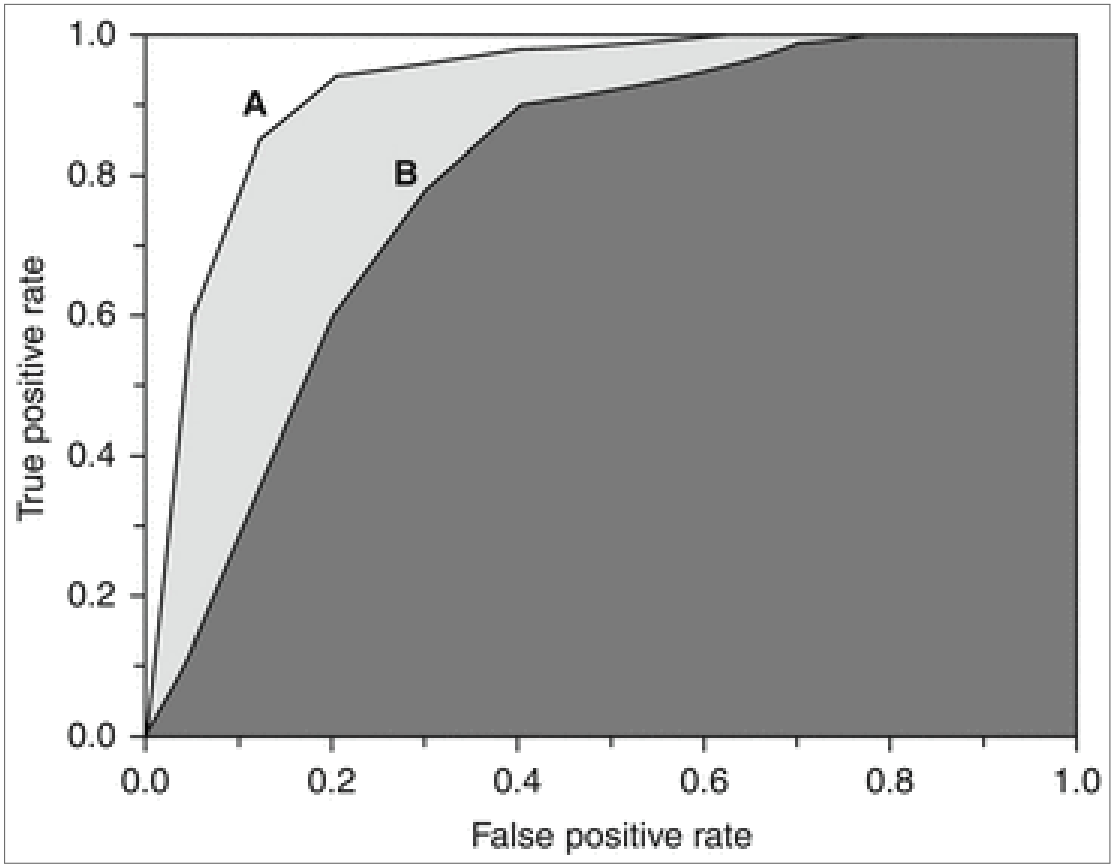
Um exemplo dessa curva pode ser vista na Figura 5. Nele percebe-se claramente que o modelo A é mais performático do que o modelo B, já que a área que a curva do modelo A ocupa é maior do que a área da curva do modelo B.

Para obter valores numéricos, então, é utilizado a área sob a curva, do inglês *Area Under the Curve* (AUC). Para tanto, é utilizado conceitos de cálculo numérico (MOORTHY; SANKAR, 2019). Assim, a métrica é conhecida pela junção das duas siglas, AUC-ROC.

2.2.6.6 Índice Kappa

O índice Kappa de Cohen (κ) é uma estatística que é usada para medir a confiabilidade interexaminador (e também a confiabilidade intraexaminador) para itens categóricos. Geralmente é considerado uma medida mais robusta do que o simples cálculo percentual de concordância, pois (κ) leva em consideração a possibilidade de a concordância ocorrer por acaso (MCHUGH, 2012). Em classificação binária, possui a formulação de (2.15):

Figura 5 – Exemplo de um gráfico ROC



Fonte: MELO (2013)

$$\kappa = \frac{2 \cdot (\mathrm{VP} \cdot \mathrm{VN} - \mathrm{FP} \cdot \mathrm{FN})}{(\mathrm{VP}+\mathrm{FP}) \cdot (\mathrm{FP}+\mathrm{VN}) + (\mathrm{VP}+\mathrm{FN}) \cdot (\mathrm{FN}+\mathrm{VN})} \tag{2.15}$$

dessa forma, seu valor pode ir desde -1 até 1, sendo -1 a predição perfeitamente errada, 0 equivalente à um modelo randômico e 1 a predição perfeitamente certa.

3 MATERIAIS E MÉTODOS

3.1 BASE DE DADOS *SURVEILLANCE, EPIDEMIOLOGY, AND END RESULTS*

O programa de Vigilância, Epidemiologia e Resultados Finais, do inglês SEER, do *National Cancer Institute* (NCI) é uma fonte de informações epidemiológicas sobre as taxas de incidência e sobrevivência de câncer nos Estados Unidos. O programa coleta e publica a incidência de câncer e dados de sobrevivência da população, cobrindo aproximadamente 34,6 % da população do país (NACIONAL CANCER INSTITUTE, 2022).

Os registros do programa SEER rotineiramente coletam dados sobre dados demográficos do paciente, local e morfologia do tumor primário, estágio no diagnóstico, primeiro curso de tratamento e acompanhamento para status de vida. O programa é a única fonte abrangente de informações de base populacional nos Estados Unidos que inclui o estágio do câncer no momento do diagnóstico e dados de sobrevida do paciente (NACIONAL CANCER INSTITUTE, 2022).

A seguir, na seção 3.1.1, é detalhada a organização dos dados e os atributos presentes na base disponível do SEER. Após isso, na seção 3.1.2, explicita-se técnicas de pré-processamento de dados tabelares e suas aplicações no contexto descrito. Por fim, em 3.1.3, há uma sumarização de trabalhos correlacionados utilizando a referida base de dados com técnicas de IA e AM para suporte à decisão clínica nessa área de neoplasias.

3.1.1 Organização e Atributos

Todos os dados da base são disponibilizados publicamente através de um termo de responsabilidade. Por questões éticas as informações sensíveis do paciente são criptografadas e representadas apenas como uma chave de valor único, sendo possível apenas reconhecer o registro de diferentes tumores de um mesmo paciente. Esse banco é concedido através de arquivos de texto, agrupados por regiões do corpo humano. Assim, os tumores primários são divididos em relação à região das mamas, sistema respiratório, digestório, linfático, reprodutor masculino, reprodutor feminino, urinário, colorretal e outros tumores que não se enquadraram nas categorias anteriores (NACIONAL CANCER INSTITUTE, 2022). O Quadro 4 mapeia as siglas que serão utilizadas no decorrer do trabalho e os grupos as quais elas equivalem. Procurou-se manter as siglas originais na base de dados para facilitar o entendimento.

Quadro 4 – Siglas das bases de dados do SEER

Sigla	Descrição
ALL	Base completa.
BREAST	Seios.
COLRECT	Cólon e reto.
DIGOTHR	Outros do aparelho digestivo.
FEMGEN	Genitália feminina.
LYMYLEUK	Linfoma de todos os locais e leucemia.
MALEGEN	Genitália masculina.
OTHER	Todos os outros sites.
RESPIR	Aparelho respiratório.
URINARY	Aparelho urinário.

Fonte: NACIONAL CANCER INSTITUTE (2022)

No quesito das colunas, há um total de 147 divididas entre informações de CID, sistema TNM, órgãos afetados por metástase, além de informações do paciente, como estado civil, idade no momento do diagnóstico, estado e cidade de residência, usufruto de plano de saúde, entre outros. Ao somar todos os anos disponíveis, de 1975 até 2016, o banco possui um total de mais de 10 milhões de registros, ou seja, informações de tumores, benignos ou malignos (NACIONAL CANCER INSTITUTE, 2020).

3.1.2 Técnicas de Pré-processamento de dados

3.1.2.1 Tratamento de Valores Faltantes

É muito comum em bases de dados relacionais e não-relacionais a existências de campos com valores ausentes (STEKHOVEN; BUHLMANN, 2011). Entretanto, para a correta utilização de um algoritmo de AM, é necessário a presença de todas as informações dos atributos fornecidos ao algoritmo.

Dessa forma, uma das maneiras mais simples de resolver isso é através de imputação de dados utilizando atributos estatísticos da referida variável. Para atributos de natureza contínua, utiliza-se a média. De forma semelhante, é utilizado a moda nos casos em que as variáveis são de origem categórica ou binária. Conforme explicitado na seção 2.2.2, para evitar a adição de viés, as referidas estatísticas são calculadas utilizando apenas as informações da base de treino.

3.1.2.2 Tratamento de Atributos Categóricos

Ainda discorrendo sobre o tratamento de dados necessários para o correto funcionamento de algoritmos de AM, evidencia-se agora sobre a questão dos atributos categóricos. É pela natureza matemática e estatística dos algoritmos de AM que eles aceitam apenas entradas de atributos no espaço do contínuo (PEDREGOSA et al., 2011), dessa maneira é preciso variáveis que representam categorias para o supracitado espaço.

A maneira mais simples de convertê-las é através de sua representação em variáveis *dummies* (GUJARATI, 2003). Isso é feito criando N novos atributos, sendo N a quantidade de categorias distintas. Assim, cada novo atributo é mapeado para a sua categoria correspondente, e, assim, recebe o valor 1 se o referido exemplo for da referida categoria, ou 0, caso contrário.

3.1.2.3 Normalização de Variáveis Contínuas

Continuando com o argumento da normalização das variáveis, também é necessário transformar variáveis contínuas para um intervalo unitário. Para isso, comumente é utilizado o escore z. Tais transformações tem como embasamento a distribuição normal (SPIEGEL; STEPHENS, 2007). O cálculo pode ser visto em (3.1):

$$z = \frac{x - \bar{x}}{S} \tag{3.1}$$

onde z é a variável contínua normalizada, x é a variável contínua original, $\bar{x}$ é a média da população amostral da variável e S é o desvio padrão dessa mesma população.

3.1.3 Trabalhos Correlacionados

Os estudos analisados focaram em apenas um tipo de câncer específico ou um grupo muito restrito de tumores (KARHADE et al., 2018; DOPPALAPUDI; QIU; BADR, 2021; CHENG; LANG, 2020; LUO et al., 2022; THIO et al., 2018; BERGQUIST et al., 2017), já que o intuito foi avaliar o poder de discernimento de algoritmos de IA específicos em cenários restritos no que tange à grupos de tumores variados.

Para o tratamento dos valores faltantes, alguns estudos (KARHADE et al., 2018; DOPPALAPUDI; QIU; BADR, 2021; CHENG; LANG, 2020; LUO et al., 2022; THIO et al., 2018) sugerem a utilização do algoritmo MissForest (STEKHOVEN; BUHLMANN, 2011), enquanto os outros utilizam a abordagem mais simples da utilização do valor médio para variáveis contínuas e a moda para as variáveis categóricas.

No que tange ao tratamento das variáveis para o treinamento dos algoritmos, a abordagem do One Hot Encoder para as categóricas foi consenso entre as pesquisas analisadas, enquanto para as variáveis contínuas, algumas utilizando a transformada pelo escore Z (CHENG; LANG, 2020), enquanto as outras não realizavam tratamento algum.

Já sobre os algoritmos de aprendizado de máquina em si, os estudos utilizavam XGBoost (CHENG; LANG, 2020; LUO et al., 2022), árvore randômica (DOPPALAPUDI; QIU; BADR, 2021; CHENG; LANG, 2020; LUO et al., 2022; THIO et al., 2018; BERGQUIST et al., 2017), regressão logística (DOPPALAPUDI; QIU; BADR, 2021; CHENG; LANG, 2020; BERGQUIST et al., 2017), máquina de vetores de suporte (KARHADE et al., 2018; BERGQUIST et al., 2017) e redes neurais (KARHADE et al., 2018; DOPPALAPUDI; QIU; BADR, 2021; CHENG; LANG, 2020; THIO et al., 2018). A maioria dos estudos analisados também utilizaram a abordagem da validação cruzada com 10 divisões com treino, validação e teste, com o objetivo de escolher os melhores hiperparâmetros (KARHADE et al., 2018; CHENG; LANG, 2020; BERGQUIST et al., 2017). Entretanto, o método da otimização dos hiperparâmetros escolhido não fora mencionado em nenhuma das pesquisas.

Por fim, as métricas de desempenho escolhidas para avaliar problemas de classificação foram: AUCROC, C-index e Brier score, além das usuais, como acurácia, precisão, sensibilidade, especificidade e F1-score (KARHADE et al., 2018; DOPPALAPUDI; QIU; BADR, 2021; CHENG; LANG, 2020; LUO et al., 2022; THIO et al., 2018; BERGQUIST et al., 2017).

3.2 METODOLOGIA

Inicialmente foi realizado um estudo sobre o dicionário de dados disponibilizado ao público sobre a base SEER (NACIONAL CANCER INSTITUTE, 2020). Notoriamente, todo o ambiente de experimentação foi realizado através de uma máquina com sistema operacional Ubuntu 22.04, CPU Intel Core i9-10900F, 24GB de memória RAM e GPU NVIDIA RTX 3060. Além do mais, em todo o trabalho se utilizou a linguagem de programação Python (ROSSUM; JR, 1995).

Após isso, a etapa de engenharia de atributos foi realizada. Assim sendo, a primeira etapa

foi a modelagem dos atributos. Eles foram construídos conforme o dicionário supracitado e características relevantes ao problema proposto, utilizando a biblioteca Spark 3.2.1 (ZAHARIA et al., 2010).

Seguindo a *pipeline* do treinamento dos algoritmos de AM, houve a divisão da base nos subconjuntos citados na seção 3.1.1 da seguinte maneira: inicialmente a base completa foi dividida em três subconjuntos, as bases de treinamento, validação e teste conforme a seção 2.2.2 detalha. Houve a divisão percentual em 50 % para a base de treinamento e 25 % para os conjuntos de validação e teste. Após isso, os exemplos de tumores específicos para cada divisão de tumor são filtrados em cada uma das três bases, totalizando assim um conjunto de treino, validação e teste para os nove subgrupos de tumores, além dos três originais da base completa. Uma sumarização dessa etapa pode ser vista na Figura 6 Também foi realizado a exclusão de registros duplicados das bases e entre elas. Toda essa etapa foi feita através das bibliotecas Pandas 1.4.2 (MCKINNEY, 2010), Numpy 1.22.3 (HARRIS et al., 2020), e Scikit-Learn 1.0.2 (PEDREGOSA et al., 2011).

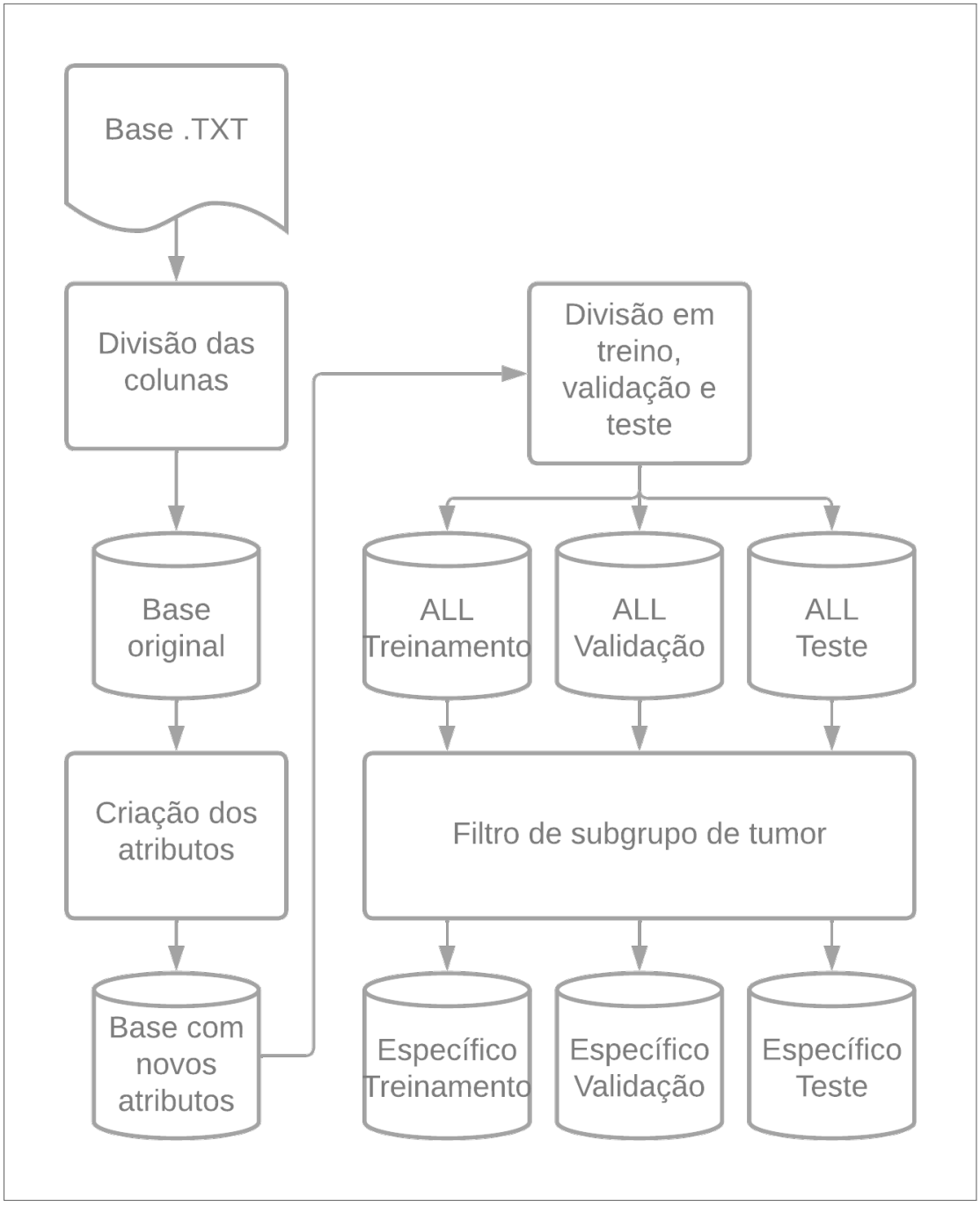
A seguir, houve o tratamento dos valores nulos, a conversão dos atributos categóricos em variáveis *dummies* e a normalização das variáveis contínuas, conforme discorrido na seção 3.1.2. Conforme a seção 2.2.2 descreve, utilizou-se apenas as informações presentes na base de treinamento para realizar as devidas transformações, com o intuito de diminuir o viés.

Assim, todo o ambiente foi montado para o devido treinamento dos algoritmos de AM. Essa fase é descrita como se segue: foram utilizados os hiperparâmetros padrão da biblioteca que implementa o XGBoost, na versão 1.5.1, conforme descrito na seção 2.2.5, para o treinamento utilizando as bases de treino de cada um dos dez conjuntos obtidos. O *early stopping* foi configurando utilizando a respectiva base de validação de cada conjunto, com uma espera de 50. Para cada modelo foi executado um total de trinta vezes modificando apenas a semente de cada iteração, com o intuito de verificar a robustez estatística de seus resultados. Avaliou-se, então, os modelos utilizando as métricas de classificação citadas na seção 2.2.6.

Por fim, uniu-se os conjuntos de treino e validação e utilizou-se a média da melhor iteração através do *early stop* de cada modelo da etapa anterior para um novo treinamento. Essa etapa é necessária para a avaliação final dos modelos, em um cenário de uso real. Nessa etapa também foram avaliadas as métricas citadas citadas na seção 2.2.6.

A última etapa foi a combinação do modelo geral com os modelos específicos. Como todos os modelos foram treinados com os mesmos atributos, a união deles foi possível apenas combinando suas respostas em probabilidade. Essa combinação levou em conta a média das

Figura 6 – Diagrama de divisão das bases



Fonte: Elaborada pelo autor (2022)

probabilidades para a classe positiva. Assim, com os novos resultados, avaliou-se a performance da mesma maneira como foi feita anteriormente.

4 RESULTADOS

Esse capítulo destina-se aos resultados obtidos aplicando a metodologia supracitada. Os 147 atributos foram gerados a partir dos arquivos disponibilizados após o preenchimento do termo de compromisso com o NCI sobre o correto uso dos dados. Logo após isso houve a manipulação desses atributos originais para a formação dos novos, todos baseados no dicionário de dados da base (NACIONAL CANCER INSTITUTE, 2020). Nesse processo houve a remoção de valores para simbolizar a ausência de informação, do inglês *Not a Number*, além da concatenação de colunas que representavam uma mesma informação. No final desse processo, 28 variáveis foram criadas, sendo elas sumarizadas no Quadro 5.

Após a etapa de engenharia de atributos houve a montagem da *pipeline* da divisão das bases para o treinamento, validação e teste. A base original possui um total de 10 455 432 exemplos, pelos quais 883 740 não possuíam a informação da variável alvo disponível. Sendo assim, eles foram excluídos.

Utilizando a estratégia de 50 % da base para treinamento e 25 % para validação e teste, foram obtidas um total de 4 783 574, 2 392 576 e 2 395 542 de registros para as respectivas bases. A diferença de amostras entre os conjuntos de validação e de teste se justificam pelo erro de divisão de bases inerente à arquitetura distribuída em que o Spark é construído (ZAHARIA et al., 2010).

Após isso, houve a remoção dos exemplos duplicados entre as bases. Assim, foram excluídas 29 064, 8 295 e 8 145 das bases de treino, validação e teste, respectivamente. De modo semelhante, foram removidas os exemplos comuns entre as bases, tornando-as conjuntos disjuntos. Dessa maneira, foram removidas 40 362 e 12 583 das bases de treino e validação, respectivamente. Nenhuma amostra foi removida da base de teste com o intuito de preservá-la o máximo possível.

Com esse tratamento inicial, foram obtidas as três bases que serão rotuladas com a sigla ALL, já que eles não possuem nenhum filtro de tumor. Seguindo com essa linha de raciocínio de filtro, eles foram aplicados nas três bases supracitadas para obter os grupos definidos no Quadro 4. Assim, gerou-se as outras nove bases faltantes. Uma sumarização da quantidade de registro em cada agrupamento tumoral e tipo de base, além da porcentagem da classe positiva, pode ser vista na Tabela 1.

A seguir, sumarizam-se as informações de imputação de valores faltantes dos atributos

Quadro 5 – Atributos gerados para o treinamento dos modelos

Atributo	Tipo	Descrição
STATE	Categórico	Estado de residência do paciente.
MARITAL_STATUS	Categórico	Estado civil do paciente.
RACE	Categórico	Raça do paciente.
IS_FEMALE	Binário	Se o paciente é do sexo feminino.
AGE_AT_DX	Contínuo	Idade do paciente no momento do diagnóstico.
HISTORIC_TUMOR_MALIGN	Contínuo	Quantidade de tumores malignos anteriores àquele diagnóstico do paciente.
HISTORIC_TUMOR_BENIGN	Contínuo	Quantidade de tumores benignos anteriores àquele diagnóstico.
NODES_POSITIVE	Contínuo	Quantidade de linfonodos examinados com presença de células tumorais.
NODES_EXAMINED	Contínuo	Quantidade de linfonodos examinados.
TUMOR_SIZE	Contínuo	Tamanho do tumor, em milímetros.
ICD	Categórico	CID do tumor.
LATERAL	Categórico	Lateralidade do tumor.
BEHAVIOR_TYPE	Categórico	Comportamento maligno ou benigno do tumor.
DIFFERENTIATION_TYPE	Categórico	Tipo de diferenciação de célula tumoral de sua origem.
CLASS_T	Categórico	Estadiamento T do tumor.
CLASS_N	Categórico	Estadiamento N do tumor.
CLASS_M	Categórico	Estadiamento M do tumor.
SUMM_STAGE	Categórico	Estágio do tumor.
FIRST_TUMOR	Binário	Indica se é o primeiro tumor do paciente.
CS_SCHEMA	Categórico	Localização específica do tumor.
SURGERY_PRIMARY_SITE	Categórico	Se houve e que tipo de cirurgia foi realizada no tumor primário.
SURGERY_OTHER_SITE	Categórico	Se houve e que tipo de cirurgia foi realizada em outros locais do corpo.
REASON_NO_SURGERY	Categórico	O motivo da não realização de procedimento cirúrgico.
INSURANCE	Categórico	Tipo de plano de saúde do paciente.
RADIATION_THERAPY	Categórico	Se houve e que tipo de radioterapia foi realizada no paciente.
RADIATION_SEQUENCE	Categórico	Ordem cronológica da cirurgia e/ou radioterapia.
DEATH	Alvo	Indica se o paciente sobreviveu menos do que 60 meses da data do diagnóstico.

Fonte: Elaborada pelo autor (2022)

Tabela 1 – Distribuição dos registros de acordo com os agrupamentos tumorais e tipo de base

Colunas	Treino		Validação		Teste	
	N°	% Pos.	N°	% Pos.	N°	% Pos.
ALL	4 714 148	58,02	2 371 698	58,13	2 387 397	58,24
BREAST	790 849	40,14	396 000	40,09	396 032	40,09
COLRECT	483 551	58,82	242 721	58,82	242 533	58,94
DIGOTHR	374 950	87,69	189 046	87,76	189 007	87,72
FEMGEN	298 048	50,92	148 909	51,23	149 507	51,11
LYMYLEUK	382 584	62,54	193 258	62,72	195 513	63,01
MALEGEN	644 687	38,20	326 166	38,27	331 007	38,27
OTHER	804 992	56,85	406 272	57,17	411 560	57,66
RESPIR	603 629	86,69	303 778	86,74	305 823	86,86
URINARY	330 858	56,27	165 548	56,18	166 415	56,29

Fonte: Elaborada pelo autor (2022)

através das ferramentas descritas na seção 3.1.2. Tais dados podem ser vistos na Tabela 2.

De modo semelhante, a Tabela 3 condiz sobre a quantidade de categorias de cada atributo categórico para sua conversão utilizando o algoritmo *one hot encoder*.

Assim, as Tabelas 4 e 5 discorrem sobre a média e o desvio padrão, respectivamente, dos atributos contínuos para suas posteriores normalizações.

A seguir, houve o momento do treinamento dos modelos utilizando as bases de treino e validação para a extração da melhor iteração dos algoritmos de XGBoost. Isso resultou na Tabela 6. Mesmo sendo repetido trinta vezes conforme o capítulo 3 cita, todas as métricas do trabalho possuíram desvio padrão igual a zero. Assim, para reduzir redundâncias, omitiu-se tais informações do restante do texto.

Por fim, as métricas de desempenho são explicitadas. As Figuras 7, 8, 9, 10, 11, 12 representam, respectivamente, as métricas de acurácia, precisão, sensibilidade, F1, AUCROC e Índice Kappa.

A categoria que representa "Treino - Específico" condiz sobre o primeiro treinamento, o necessário para obter a melhor iteração de cada modelo para cada agrupamento. Já o "Teste - Específico" simboliza o treinamento final, ou seja, aquele em que se utiliza a base de treino e validação como treinamento para o modelo e o uso da base de teste como avaliação final. A questão é que esse rótulo significa a performance do modelo treinado em sua base específica. No caso específico da base "ALL", esse rótulo representa o concatenamento dos resultados dos algoritmos específicos. O rótulo "Teste - Genérico" simboliza a performance do modelo

Tabela 2 – Valor imputado para cada atributo

Atributo	ALL	BREAST	COLRECT	DIGOTHR	FEMGEN	LYMYLEUK	MALEGEN	OTHER	RESPIR	URINARY
STATE	6	6	6	6	6	6	6	6	6	6
MARITAL_STATUS	2	2	2	2	2	2	2	2	2	2
RACE	1	1	1	1	1	1	1	1	1	1
ICD	C50	C50	C18	C25	C54	C42	C61	C44	C34	C67
LATERAL	0	2	0	0	0	0	0	0	1	0
BEHAVIOR_TYPE	3	3	3	3	3	3	3	3	3	3
DIFFERENTIATION_TYPE	2	2	2	3	2	6	2	2	3	2
CLASS_T	88	18	30	88	12	88	18	88	40	10
CLASS_N	88	10	10	88	10	88	88	88	20	88
CLASS_M	10	10	10	10	10	10	10	10	10	10
SURGERY_PRIMARY_SITE	0	22	30	0	50	98	0	0	0	27
SURGERY_OTHER_SITE	0	0	0	0	0	0	0	0	0	0
REASON_NO_SURGERY	0	0	0	1	0	1	0	0	1	0
INSURANCE	3	3	3	3	3	3	3	3	3	3
RADIATION_THERAPY	0	0	0	0	0	0	0	0	0	0
RADIATION_SEQUENCE	0	0	0	0	0	0	0	0	0	0
SUMM_STAGE	1	1	1	7	1	7	1	1	7	1
AGE_AT_DX	64	61	68	67	60	61	66	59	68	67
HISTORIC_TUMOR_MALIGN	0	0	0	0	0	0	0	0	0	0
HISTORIC_TUMOR_BENIGN	0	0	0	0	0	0	0	0	0	0
TUMOR_SIZE	129	21	45	49	56	983	22	145	42	47
NODES_POSITIVE	0	1	1	1	0	0	0	0	0	0
NODES_EXAMINED	4	6	11	3	6	0	2	3	2	1

Fonte: Elaborada pelo autor (2022)

treinado com a base "ALL". Não obstante, o rótulo "Teste - Ensemble" consiste no uso do *soft voting* do modelo específico e do modelo geral.

Tabela 3 – Quantidade de categorias dos atributos categóricos

Atributo	ALL	BREAST	COLRECT	DIGOTHR	FEMGEN	LYMYLEUK	MALEGEN	OTHER	RESPIR	URINARY
STATE	13	13	13	13	13	13	13	13	13	13
MARITAL_STATUS	6	6	6	6	6	6	6	6	6	6
RACE	4	4	4	4	4	4	4	4	4	4
ICD	70	1	4	10	8	69	4	62	7	5
LATERAL	6	4	5	5	5	6	5	6	5	5
BEHAVIOR_TYPE	4	2	3	3	3	1	3	4	2	2
DIFFERENTIATION_TYPE	8	4	4	4	4	8	4	8	4	4
CLASS_T	23	13	9	14	21	15	16	18	11	14
CLASS_N	14	13	9	6	9	8	8	11	7	4
CLASS_M	4	1	3	3	4	3	4	4	3	1
SURGERY_PRIMARY_SITE	73	47	33	45	55	56	25	47	37	31
SURGERY_OTHER_SITE	6	6	6	6	6	6	6	6	6	6
REASON_NO_SURGERY	6	6	6	6	6	6	6	6	6	6
INSURANCE	4	4	4	4	4	4	4	4	4	4
RADIATION_THERAPY	9	9	9	9	9	9	9	9	9	9
RADIATION_SEQUENCE	8	8	8	8	8	7	8	8	8	8
SUMM_STAGE	8	7	7	7	7	6	8	8	7	7

Fonte: Elaborada pelo autor (2022)

Tabela 4 – Média dos atributos contínuos para a normalização

Atributo	ALL	BREAST	COLRECT	DIGOTHR	FEMGEN	LYMYLEUK	MALEGEN	OTHER	RESPIR	URINARY
AGE_AT_DX	64	61	68	67	60	61	66	59	68	67
HISTORIC_TUMOR_MALIGN	0	0	0	0	0	0	0	0	0	0
HISTORIC_TUMOR_BENIGN	0	0	0	0	0	0	0	0	0	0
TUMOR_SIZE	129	21	45	49	56	983	22	145	42	47
NODES_POSITIVE	0	1	1	1	0	0	0	0	0	0
NODES_EXAMINED	4	6	11	3	6	0	2	3	2	1

Fonte: Elaborada pelo autor (2022)

Tabela 5 – Desvio padrão dos atributos contínuos para a normalização

Atributo	ALL	BREAST	COLRECT	DIGOTHR	FEMGEN	LYMYLEUK	MALEGEN	OTHER	RESPIR	URINARY
AGE_AT_DX	15	14	14	14	15	20	12	19	12	14
HISTORIC_TUMOR_MALIGN	0	0	0	0	0	0	0	1	0	0
HISTORIC_TUMOR_BENIGN	0	0	0	0	0	0	0	0	0	0
TUMOR_SIZE	208	22	27	35	41	52	15	216	26	32
NODES_POSITIVE	2	3	3	2	2	0	0	2	1	1
NODES_EXAMINED	7	7	10	7	9	0	5	7	5	4

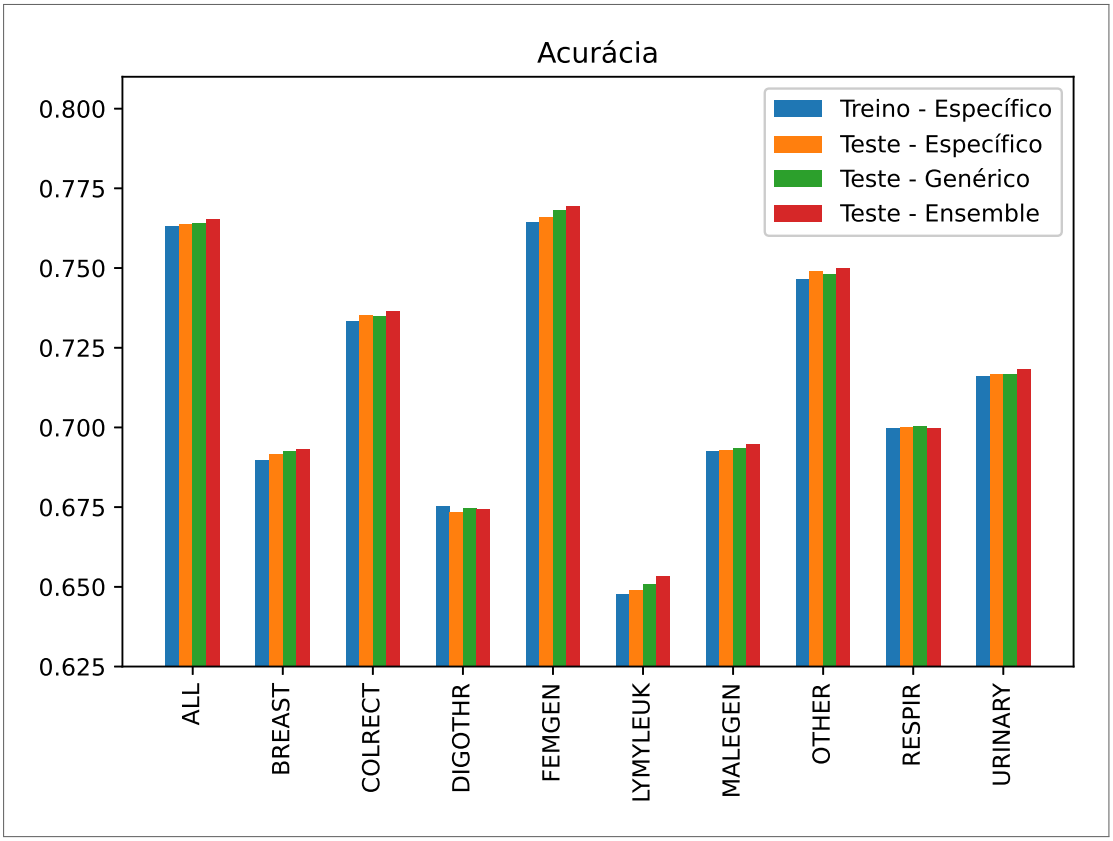
Fonte: Elaborada pelo autor (2022)

Tabela 6 – Melhor iteração de acordo com cada grupamento tumoral

Agrupamento	Melhor Iteração
ALL	1227
BREAST	290
COLRECT	160
DIGOTHR	155
FEMGEN	220
LYMYLEUK	173
MALEGEN	287
OTHER	452
RESPIR	123
URINARY	145

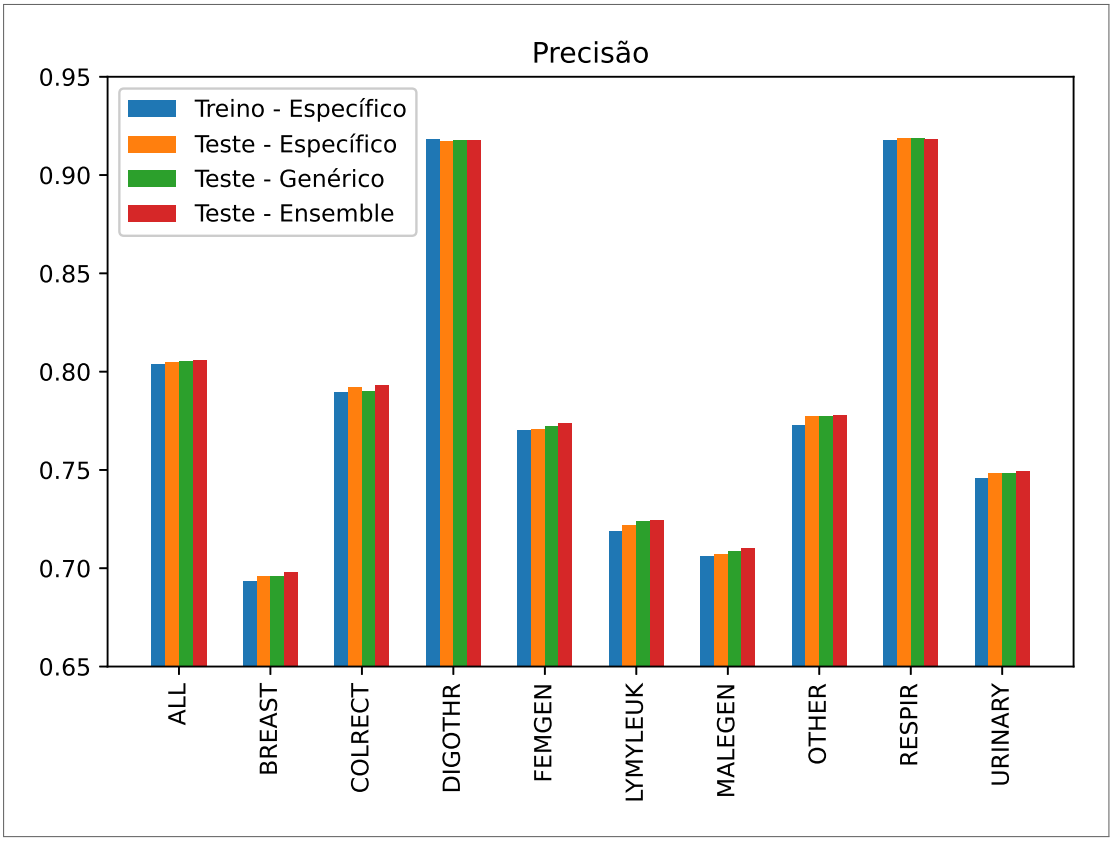
Fonte: Elaborada pelo autor (2022)

Figura 7 – Acurácia dos modelos



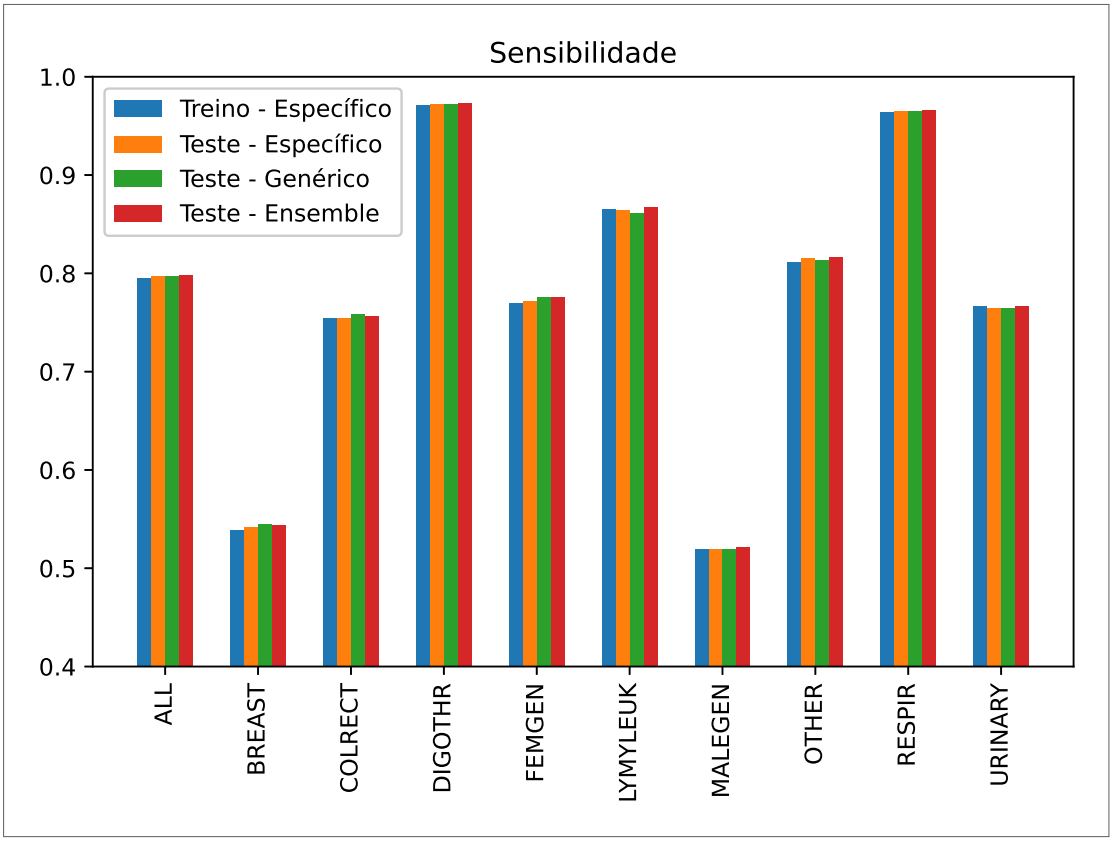
Fonte: Elaborada pelo autor (2022)

Figura 8 – Precisão dos modelos



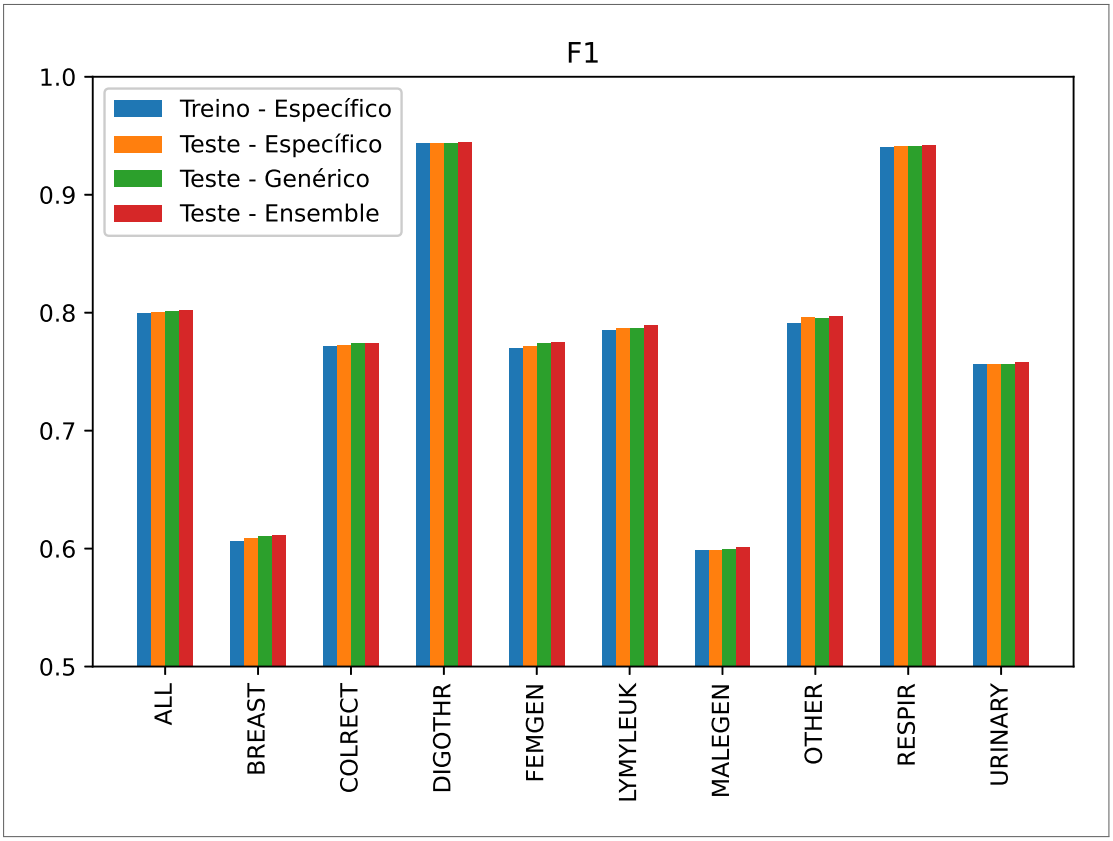
Fonte: Elaborada pelo autor (2022)

Figura 9 – Sensibilidade dos modelos



Fonte: Elaborada pelo autor (2022)

Figura 10 – Métrica F1 dos modelos

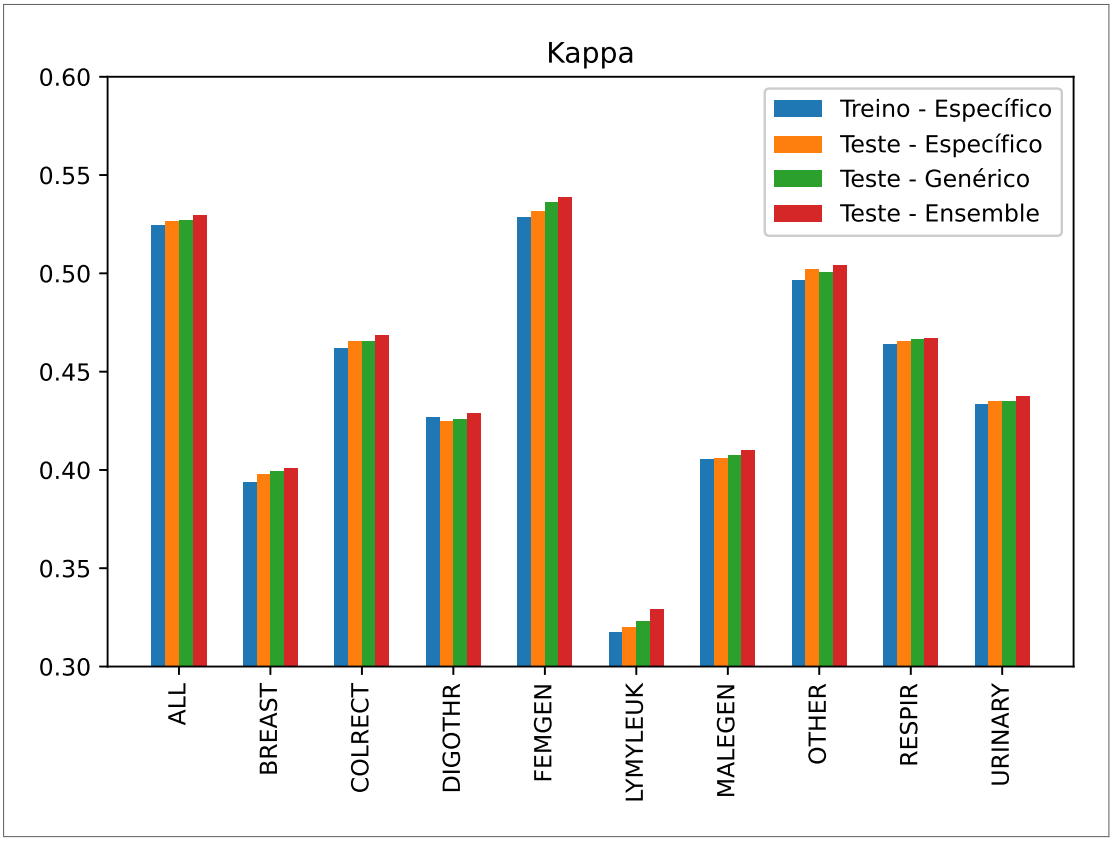


Fonte: Elaborada pelo autor (2022)

Figura 11 – AUCROC dos modelos

Fonte: Elaborada pelo autor (2022)

Figura 12 – Índice Kappa dos modelos



Fonte: Elaborada pelo autor (2022)

5 DISCUSSÃO

Inicia-se a discussão tratando sobre a questão dos atributos originais da base de dados. Entende-se que o NCI possui um trabalho árduo desde 1975 coletando esses dados de pacientes de todos os Estados Unidos, entretanto constrói-se uma crítica sobre o particionamento de algumas variáveis em relação ao tempo. É possível diminuir o espalhamento de informação da base de dados com um trabalho lento, porém recompensatório de limpeza e reestruturação da base de dados. Outra crítica em cima do banco de dados original é sobre a forma de disponibilização ao usuário final. Recentemente não é mais possível obter a base com o formato de texto, tendo agora que recorrer a *software* próprio do NCI para extrair as mesmas informações. Além de tudo, alguns atributos utilizados nesse trabalho não são mais disponíveis, como informações de radioterapia e quimioterapia, dificultando o trabalho de estatísticos e cientistas de dados no geral.

A próxima discussão está em torno da tolerância de erro do Spark no momento de divisão de bases de dados. Apesar de ser uma plataforma madura, esse controle ainda não está disponível e seria proveitoso obter tal manuseio, mesmo que a custo de poder computacional.

Sobre os valores imputados, a Tabela 2, percebe-se que a maioria dos atributos possuiu um valor único imputado. Isso dá indícios de que há valores muito comuns, que perpassam todos os agrupamentos. Nota-se, entretanto, que alguns agrupamentos obtiveram o valor "não se aplica" nos estadiamentos TNM. Estudos detalhados são exigidos para extrair alguma relação de causa e efeito. Ademais, percebe-se o valor da idade no momento do diagnóstico imputado. Como é uma variável contínua, a imputação levou em conta a média do atributo. Assim, infere-se que os mais acometidos de câncer é a população idosa, acima dos 60 anos. Um ponto de destaque final sobre a imputação dos atributos é sobre o tamanho do tumor para o grupo "LYMYLEUK". Por se tratar de tumores primários em linfonodos ou sobre leucemia, é razoável interligar a relação causa e consequência, tendo em vista que o sistema linfático recobre todo o corpo humano.

Já sobre a Tabela 3, percebe-se que há atributos com uma quantidade fixa de categorias, enquanto outros variam com a mudança de agrupamento. Notoriamente e obviamente o atributo que trata do CID e de cirurgia no tumor primário variam de acordo com o agrupamento. Esse último é variável tendo em vista que cada agrupamento possui seus conjuntos de cirurgias próprios.

A Tabela 4 corrobora com a argumentação supracitada sobre o maior acometimento do câncer na 3° idade. Percebe-se também que a média dos históricos de tumor maligno e benigno são zero, significando que a média da população da base de treino teve o tumor pela primeira vez. Somado com a informação da Tabela 5, reafirma-se os argumentos supracitados. Era de se esperar da Tabela 6 de que o modelo do grupo "ALL" obteria uma iteração limite maior do que os demais, tendo em vista a maior quantidade de exemplos e, consequentemente, possibilidades de extração de características mais generalistas.

Alcança-se, então, os gráficos das métricas dos modelos. Da métrica da acurácia, apenas o agrupamento "RESPIR" não obteve uma melhora na resposta com a combinação dos modelos. Sobre a precisão, apenas os agrupamentos "DIGOTHR" e "RESPIR" não obteve uma melhora da combinação dos modelos comparado com os modelos em separado. Entretanto, discorre-se sobre a alta performance desses modelos na métrica da precisão, tendo em vista que esses dois subgrupos foram os únicos a obterem um valor acima de 0.9.

A métrica de sensibilidade demonstrou respostas distintas no que tange à combinação dos modelos. Entretanto, aponta-se que o *ensemble* de algoritmos ou performou tão melhor quanto os algoritmos em isolado ou melhor, nunca inferior. Somado a isso, infere-se que a métrica F1 tenha melhorado no que tange à combinação dos modelos por causa da métrica de precisão, tendo em vista a melhora da métrica de precisão nesse mesmo cenário. A métrica da AUCROC demonstrou uma melhora em todos os agrupamentos, reforçando o argumento de que a combinação melhora o resultado final dos modelos. Por fim, o índice Kappa também melhorou em todos os agrupamentos no que tange ao *ensemble* de algoritmos, enfatizando o poder da combinação desses modelos generalistas e específicos.

Levanta-se a hipótese de que o modelo específico aprimorou o modelo generalista tendo em vista que o modelo específico extraiu características próprias do subgrupo o qual foi treinando. Assim, o modelo generalista obtém uma pontuação geral, enquanto o modelo específico refina essa resposta por ter sido treinado apenas em amostras específicas daquele tipo de tumor.

É importante citar também as dificuldades enfrentadas. A principal e maior delas foi no que tange ao tratamento dos atributos categóricos. Como foi possível ver no Quadro 5, a maioria das variáveis são categóricas, e a abordagem utilizando o *one hot encoder* aumenta expressivamente o uso de memória para cada adição de atributo categórico. Dessa maneira, a solução encontrada foi trabalhar com um subconjunto dos atributos disponíveis, mesmo que isso impacte na performance dos modelos.

Uma outra dificuldade apresentada foi a fundamentação teórica dos algoritmos de AM.

Argumenta-se de que o curso de Engenharia Biomédica na Universidade Federal de Pernambuco pode aproveitar mais da interdisciplinaridade que a própria Engenharia Biomédica tem a oferecer e possuir trilhas de conhecimento para além da Engenharia Clínica, atual foco de carreira do curso na Universidade citada. Assim, abre-se um leque maior de possibilidade pelo qual o estudante pode cursar, não necessitando buscar conhecimento, inclusive de AM, em disciplinas de outros cursos.

6 CONCLUSÃO

Algoritmos de AM são utilizados como sistemas de apoio a decisão médica, auxiliando profissionais de saúde no diagnóstico e prognóstico de seus pacientes. Foi proposto a utilização desses algoritmos na área oncológica, com o intuito de predizer a remissão do paciente. Os modelos de AM foram treinados em subconjuntos da base de dados original, divididos por tipos de tumores, além de um modelo treinado com o conjunto completo. Após a combinação do algoritmo treinado no conjunto completo e os modelos dos subconjuntos percebeu-se o aumento, mesmo que discreto, da performance de todas as métricas de desempenho medidas. Assim, compreende-se como alcançado o objetivo deste trabalho, inclusive todos os objetivos específicos. Mesmo sendo difícil o diagnóstico e prognóstico por câncer atualmente, mais ainda em países de baixa renda, esforços são feitos para o suporte à decisão clínica. Dessa forma, é possível utilizar os algoritmos produzidos por esse trabalho para inferir novas informações da doença do paciente oncológico, de maneira tal que auxilie-o até sua remissão.

Por fim, destaca-se as possibilidades de aprimoramento desse trabalho. Como já supracitado, novos estudos são necessários para abordar os atributos que não foram incluídos nessa pesquisa. Como a base original possui mais de 140 atributos, é de se esperar que muitas outras variáveis tinham sido ignoradas no processo de criação desse trabalho. Soma-se a isso a contribuição de novidade desse trabalho. Como já citado nos trabalhos anteriores, nenhuma pesquisa utilizando essa base de dados utilizou-a de maneira completa, com todos os anos e todos os tumores disponíveis.

Espera-se não só contribuir para a área da saúde com algoritmos de IA mas também inspirar futuras gerações de engenheiros biomédicos a adentrar na área de ciência de dados, tendo em vista que é uma temática que terá grande impacto na área da saúde, de maneira a ser ator relevante para um mundo melhor.

REFERÊNCIAS

ANDREASSEN, A.; FEIGE, I.; FRYE, C.; SCHWARTZ, M. D. JUNIPR: a framework for unsupervised machine learning in particle physics. *The European Physical Journal C*, Springer Science and Business Media LLC, v. 79, n. 2, fev. 2019. Disponível em: <https://doi.org/10.1140/epjc/s10052-019-6607-9>.

BERGQUIST, S. L.; BROOKS, G. A.; KEATING, N. L.; LANDRUM, M. B.; ROSE, S. Classifying lung cancer severity with ensemble machine learning in health care claims data. *Proc Mach Learn Res*, v. 68, p. 25–38, ago. 2017.

BHINDER, B.; GILVARY, C.; MADHUKAR, N. S.; ELEMENTO, O. Artificial intelligence in cancer research and precision medicine. *Cancer discovery*, v. 11, n. 4, p. 900—915, April 2021. ISSN 2159-8274. Disponível em: <https://europepmc.org/articles/PMC8034385>.

BOCHENEK, B.; USTRNUL, Z. Machine learning in weather prediction and climate analyses—applications and perspectives. *Atmosphere*, Multidisciplinary Digital Publishing Institute, v. 13, n. 2, p. 180, 2022.

BRAY, F.; LAVERSANNE, M.; WEIDERPASS, E.; SOERJOMATARAM, I. The ever-increasing importance of cancer as a leading cause of premature death worldwide. *Cancer*, v. 127, n. 16, p. 3029–3030, 2021. Disponível em: <https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.1002/cncr.33587>.

BREIMAN, L. *Classification and Regression Trees*. London, England: CRC Press, 2017.

CHEN, T.; GUESTRIN, C. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2016. (KDD '16), p. 785–794. ISBN 978-1-4503-4232-2. Disponível em: <http://doi.acm.org/10.1145/2939672.2939785>.

CHENG, A.; LANG, J. Survival analysis of lymph node resection in ovarian cancer: A population-based study. *Frontiers in Oncology*, v. 10, 2020. ISSN 2234-943X. Disponível em: <https://www.frontiersin.org/article/10.3389/fonc.2020.00355>.

DEGRAVE, J.; FELICI, F.; BUCHLI, J.; NEUNERT, M.; TRACEY, B.; CARPANESE, F.; EWALDS, T.; HAFNER, R.; ABDOLMALEKI, A.; CASAS, D. de L. et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, Nature Publishing Group, v. 602, n. 7897, p. 414–419, 2022.

DOPPALAPUDI, S.; QIU, R. G.; BADR, Y. Lung cancer survival period prediction and understanding: Deep learning approaches. *International Journal of Medical Informatics*, v. 148, p. 104371, 2021. ISSN 1386-5056. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1386505620319079>.

MINISTÉRIO DA SAÚDE. *Estadiamento | INCA - Instituto Nacional de Câncer*. 2021. Disponível em: <https://www.inca.gov.br/estadiamento>.

NACIONAL CANCER INSTITUTE. *SEER RESEARCH PLUS DATA DESCRIPTION*. 2020. Disponível em: <https://seer.cancer.gov/data-software/documentation/seerstat/nov2020/TextData.FileDescription.pdf>.

NACIONAL CANCER INSTITUTE. *What Is Cancer? - NCII*. 2021. Disponível em: <https://www.cancer.gov/about-cancer/understanding/what-is-cancer>.

NACIONAL CANCER INSTITUTE. *Overview of the SEER Program*. 2022. Disponível em: <https://seer.cancer.gov/about/overview.html>.

UNITED NATIONS. *Human Development Report 2020*. 2020. ed. United Nations, 2020. Disponível em: <https://www.un-ilibrary.org/content/books/9789210055161>.

WORLD HEALTH ORGANIZATION. *International Classification of Diseases for Oncology*. World Health Organization, 2013. (Nonserial Publications). ISBN 9789240692121. Disponível em: <https://books.google.com.br/books?id=g83ljgEACAAJ>.

WORLD HEALTH ORGANIZATION. *ICD-10 Version:2019*. 2019. Disponível em: <https://icd.who.int/browse10/2019/en#/II>.

WORLD HEALTH ORGANIZATION. *Global health estimates: Leading causes of death*. 2021. Disponível em: <https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death>.

WORLD HEALTH ORGANIZATION. *Cancer*. 2022. Disponível em: <https://www.who.int/en/news-room/fact-sheets/detail/cancer>.

WORLD HEALTH ORGANIZATION. *Internacional Classification of Diseases (ICD)*. 2022. Disponível em: <https://www.who.int/classifications/classification-of-diseases>.

FOLLAND, G. B. *Real analysis*. 2. ed. Nashville, TN: John Wiley & Sons, 1999. (Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts).

FREEDMAN, D. A. *Statistical Models*. [S.l.: s.n.], 2009. ISBN 9781139477314.

GIANNUZZI, D.; MOTA, L. F. M.; PEGOLO, S.; GALLO, L.; SCHIAVON, S.; TAGLIAPIETRA, F.; KATZ, G.; FAINBOYM, D.; MINUTI, A.; TREVISI, E.; CECCHINATO, A. In-line near-infrared analysis of milk coupled with machine learning methods for the daily prediction of blood metabolic profile in dairy cattle. *Scientific Reports*, v. 12, n. 1, p. 8058, May 2022. ISSN 2045-2322. Disponível em: <https://doi.org/10.1038/s41598-022-11799-0>.

Purposes and principles of staging. In: GREENE, F. L.; COMPTON, C. C.; FRITZ, A. G.; SHAH, J. P.; WINCHESTER, D. P. (Ed.). *AJCC Cancer Staging Atlas*. New York, NY: Springer New York, 2006. p. 1–9. ISBN 978-0-387-33126-3. Disponível em: <https://doi.org/10.1007/0-387-33126-3_1>.

GUJARATI, D. N. *Basic Econometrics*. 4. ed. [S.l.]: McGraw-Hill Companies, 2003.

GUTIN, G.; YEO, A.; ZVEROVICH, A. Traveling salesman should not be greedy: domination analysis of greedy-type heuristics for the TSP. *Discrete Appl. Math.*, Elsevier BV, v. 117, n. 1-3, p. 81–86, mar. 2002.

HARRIS, C. R.; MILLMAN, K. J.; WALT, S. J. van der; GOMMERS, R.; VIRTANEN, P.; COURNAPEAU, D.; WIESER, E.; TAYLOR, J.; BERG, S.; SMITH, N. J.; KERN, R.; PICUS, M.; HOYER, S.; KERKWIJK, M. H. van; BRETT, M.; HALDANE, A.; RÍO, J. F. del; WIEBE, M.; PETERSON, P.; GÉRARD-MARCHANT, P.; SHEPPARD, K.; REDDY, T.; WECKESSER, W.; ABBASI, H.; GOHLKE, C.; OLIPHANT, T. E. Array programming with NumPy. *Nature*, Springer Science and Business Media LLC, v. 585, n. 7825, p. 357–362, set. 2020. Disponível em: <https://doi.org/10.1038/s41586-020-2649-2>.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. Boosting and additive trees. In: *The Elements of Statistical Learning*. New York, NY: Springer New York, 2009, (Springer series in statistics). p. 337–387.

HU, J.; NIU, H.; CARRASCO, J.; LENNOX, B.; ARVIN, F. Voronoi-based multi-robot autonomous exploration in unknown environments via deep reinforcement learning. *IEEE Transactions on Vehicular Technology*, v. 69, n. 12, p. 14413–14423, 2020.

JAEGER, S.; FULLE, S.; TURK, S. Mol2vec: Unsupervised machine learning approach with chemical intuition. *Journal of Chemical Information and Modeling*, v. 58, n. 1, p. 27–35, 2018. PMID: 29268609. Disponível em: <https://doi.org/10.1021/acs.jcim.7b00616>.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. *An introduction to statistical learning*. 1. ed. New York, NY: Springer, 2013. (Springer texts in statistics).

KARHADE, A. V.; THIO, Q.; OGINK, P.; KIM, J.; LOZANO-CALDERON, S.; RASKIN, K.; SCHWAB, J. H. Development of machine learning algorithms for prediction of 5-year spinal chordoma survival. *World Neurosurgery*, Elsevier BV, v. 119, p. e842–e847, nov. 2018. Disponível em: <https://doi.org/10.1016/j.wneu.2018.07.276>.

KOBER, J.; BAGNELL, J. A.; PETERS, J. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, v. 32, n. 11, p. 1238–1274, 2013. Disponível em: <https://doi.org/10.1177/0278364913495721>.

KOZA, J. R.; BENNETT, F. H.; ANDRE, D.; KEANE, M. A. Automated design of both the topology and sizing of analog electrical circuits using genetic programming. In: *Artificial Intelligence in Design '96*. Springer Netherlands, 1996. p. 151–170. Disponível em: <https://doi.org/10.1007/978-94-009-0279-4_9>.

LAMBERS, J. *Lecture 10 Notes*. 2011. Disponível em: <https://www.math.usm.edu/lambers/mat419/lecture10.pdf>.

LUO, L.; LIN, H.; LIN, B.; HUANG, F.; LUO, H. Risk factors and prognostic nomogram for patients with second primary cancers after lung cancer using classical statistics and machine learning. Research Square Platform LLC, jan. 2022. Disponível em: <https://doi.org/10.21203/rs.3.rs-1256731/v1>.

MARTEL, C. de; GEORGES, D.; BRAY, F.; FERLAY, J.; CLIFFORD, G. M. Global burden of cancer attributable to infections in 2018: a worldwide incidence analysis. *The Lancet Global Health*, Elsevier BV, v. 8, n. 2, p. e180–e190, fev. 2020. Disponível em: <https://doi.org/10.1016/s2214-109x(19)30488-7>.

MCHUGH, M. L. Interrater reliability: the kappa statistic. *Biochemia Medica*, Croatian Society for Medical Biochemistry and Laboratory Medicine, p. 276–282, 2012. Disponível em: <https://doi.org/10.11613/bm.2012.031>.

MCKINNEY Wes. Data Structures for Statistical Computing in Python. In: WALT Stéfan van der; MILLMAN Jarrod (Ed.). *Proceedings of the 9th Python in Science Conference*. [S.l.: s.n.], 2010. p. 56 – 61.

MELO, F. Area under the ROC curve. In: *Encyclopedia of Systems Biology*. Springer New York, 2013. p. 38–39. Disponível em: <https://doi.org/10.1007/978-1-4419-9863-7_209>.

MITCHELL, T. *Machine Learning*. New York, NY: McGraw-Hill Professional, 1997. (McGraw-Hill series in computer science).

MNIH, V.; KAVUKCUOGLU, K.; SILVER, D.; GRAVES, A.; ANTONOGLOU, I.; WIERSTRA, D.; RIEDMILLER, M. *Playing Atari with Deep Reinforcement Learning*. arXiv, 2013. Disponível em: <https://arxiv.org/abs/1312.5602>.

MOHRI, M.; ROSTAMIZADEH, A.; TALWALKAR, A. *Foundations of machine learning*. London, England: MIT Press, 2012. (Adaptive Computation and Machine Learning series).

MOORTHY, C. G.; SANKAR, G. U. *Numerical methods for calculus students*. [S.l.]: LAP Lambert Academic Publishing, 2019.

MUR, A.; TRAVÉ-MASSUYÈS, L.; CHANTHERY, E.; PONS, R.; RIBOT, P. A neural algorithm for the detection and correction of anomalies: Application to the landing of an airplane. *Sensors*, MDPI, v. 22, n. 6, p. 2334, 2022.

NASEEM, U.; RASHID, J.; ALI, L.; KIM, J.; HAQ, Q. E. U.; AWAN, M. J.; IMRAN, M. An automatic detection of breast cancer diagnosis and prognosis based on machine learning using ensemble of classifiers. *IEEE Access*, p. 1–1, 2022.

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.; DUCHESNAY Édouard. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, v. 12, n. 85, p. 2825–2830, 2011. Disponível em: <http://jmlr.org/papers/v12/pedregosa11a.html>.

PIRYONESI, S. M.; EL-DIRABY, T. E. Data analytics in asset management: Cost-effective prediction of the pavement condition index. *J. Infrastruct. Syst.*, American Society of Civil Engineers (ASCE), v. 26, n. 1, p. 04019036, mar. 2020.

POWERS, D. Evaluation: From precision, recall and f-factor to roc, informedness, markedness correlation. *Mach. Learn. Technol.*, v. 2, 01 2008.

PRECHELT, L. Early stopping — but when? In: *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, (Lecture notes in computer science). p. 53–67.

QUINLAN, J. R. Induction of decision trees. *Mach. Learn.*, Springer Science and Business Media LLC, v. 1, n. 1, p. 81–106, mar. 1986.

RIPLEY, B. D. *Pattern Recognition and Neural Networks*. Cambridge, England: Cambridge University Press, 2008.

ROKACH, L.; MAIMON, O. Top-down induction of decision trees classifiers—a survey. *IEEE Trans. Syst. Man Cybern. C Appl. Rev.*, Institute of Electrical and Electronics Engineers (IEEE), v. 35, n. 4, p. 476–487, nov. 2005.

ROSSUM, G. V.; JR, F. L. D. *Python reference manual*. [S.l.]: Centrum voor Wiskunde en Informatica Amsterdam, 1995.

RUSSELL, S.; NORVIG, P. *Artificial intelligence*. 3. ed. Upper Saddle River, NJ: Pearson, 2009.

SHALEV-SHWARTZ, S.; BEN-DAVID, S. *Understanding machine learning*. Cambridge, England: Cambridge University Press, 2014.

SPIEGEL, M. R.; STEPHENS, L. J. *Schaum's Outline of Statistics*. 4. ed. New York, NY: Schaum Outline Series, 2007. (Schaum's Outline Series).

STEKHOVEN, D. J.; BUHLMANN, P. MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics*, Oxford University Press (OUP), v. 28, n. 1, p. 112–118, out. 2011. Disponível em: <https://doi.org/10.1093/bioinformatics/btr597>.

SUNG, H.; FERLAY, J.; SIEGEL, R. L.; LAVERSANNE, M.; SOERJOMATARAM, I.; JEMAL, A.; BRAY, F. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, v. 71, n. 3, p. 209–249, 2021. Disponível em: <https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21660>.

THIO, Q. C. B. S.; KARHADE, A. V.; OGINK, P. T.; RASKIN, K. A.; BERNSTEIN, K. D. A.; CALDERON, S. A. L.; SCHWAB, J. H. Can machine-learning techniques be used for 5-year survival prediction of patients with chondrosarcoma? *Clinical Orthopaedics &amp Related Research*, Ovid Technologies (Wolters Kluwer Health), v. 476, n. 10, p. 2040–2048, set. 2018. Disponível em: <https://doi.org/10.1097/corr.0000000000000433>.

YANG, L.; SHAMI, A. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, Elsevier BV, v. 415, p. 295–316, nov. 2020.

ZAHARIA, M.; CHOWDHURY, M.; FRANKLIN, M. J.; SHENKER, S.; STOICA, I. Spark: Cluster computing with working sets. In: *Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing*. USA: USENIX Association, 2010. (HotCloud'10), p. 10.