UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Uso do Processamento de Linguagem Natural para extração de dados nos registros de câncer

André Luiz Pinto Santos

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

André Luiz Pinto Santos

Uso do Processamento de Linguagem Natural para extração de dados nos registros de câncer

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Roney Lira de Sales Santos

Versão original

São Carlos

2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

S237u

Santos, André Luiz Pinto
Uso do Processamento de Linguagem Natural para extração de dados nos registros de câncer / André Luiz Pinto Santos; orientador Roney Lira de Sales Santos. -- São Carlos, 2024.
61 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. Processamento de Linguagem Natural. 2. Registro de câncer. 3. Aprendizado de máquina. I. Santos, Roney Lira de Sales, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

André Luiz Pinto Santos

Use of Natural Language Processing to extract data in cancer registries

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Prof. Dr. Roney Lira de Sales Santos

Original version

São Carlos

2024

Dedico este trabalho à minha sobrinha, Liz, que acaba de chegar a este mundo.
Que você seja muito bem-vinda e cresça em uma sociedade livre de misoginia.
Que as tecnologias existentes e vindouras sejam suas aliadas na construção de um mundo mais justo e próspero para todos.

AGRADECIMENTOS

A Deus que me deu sabedoria e saúde para enfrentar os desafios acadêmicos e profissionais diariamente.

Aos meus pais, Elizabete e Emanuel, e aos meus irmãos, Vinícius e Emanuelle, que compreendem as minhas ausência nesses momentos de formação. Amo vocês!

Aos pacientes em tratamento oncológico do Hospital de Amor desafiados por esse diagnóstico, mas esperançosos pelo dia da cura.

Ao meu orientador, Prof. Dr. Roney Santos, que me fez evoluir na área de PLN e aos demais professores do MBA em Big Data e IA.

Aos profissionais de saúde do Hospital de Amor, meus colegas de trabalho, que doam suas vidas para cuidar de outras vidas.

"Não tenho medo da humanidade das máquinas, mas da desumanidade dos homens."
Milton Chamarelli Filho

RESUMO

PINTO, A. L. **Uso do Processamento de Linguagem Natural para extração de dados nos registros de câncer**. 2024. 61 p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

Na saúde, o Processamento de Linguagem Natural (PLN) tem se mostrado uma ferramenta poderosa na extração e estruturação de dados clínicos a partir de textos não estruturados, tornando esse processo mais rápido, preciso e menos sujeito a erros humanos. Nos registros de câncer são coletados, manualmente, dados essenciais para avaliar o perfil epidemiológico, a qualidade da assistência prestada e para o cumprimento de obrigações legais, a partir de laudos e notas clínicas. Com a implantação do Registro Eletrônico de Saúde (RES) no Hospital de Amor, tornou-se viável a utilização de textos para treinar modelos de PLN que possam auxiliar na extração de variáveis coletadas pelos registradores. O objetivo do presente trabalho foi treinar algoritmos de PLN que auxiliem na extração de variáveis coletadas manualmente nos registros de câncer a partir de textos de laudos de biópsias e narrativas clínicas. Os documentos foram pré-processados, vetorizados com TF-IDF(*Term Frequency-Inverse Document Frequency*) e BioBERTpt, e treinados usando algoritmos de aprendizado de máquina (regressão logística, *random forest*, *Support Vector Machine*, Naive Bayes e *gradient boosting*) para extrair informações sobre malignidade, topografia, morfologia e estadiamento de tumores. Os melhores modelos alcançaram acurácias superiores a 92%, 96% e 88% na extração de malignidade, topografia e morfologia, respectivamente. No entanto, os modelos enfrentaram dificuldades na extração de estadiamento devido à ausência dessa informação nas evoluções médicas, evidenciando uma limitação na qualidade das narrativas clínicas. A implementação desses modelos na rotina dos registros de câncer pode otimizar o trabalho dos registradores, melhorar a qualidade dos dados e reduzir o tempo de disponibilização das informações. Além disso, ao identificar lacunas na documentação clínica, os modelos também podem contribuir para a melhoria da qualidade das informações textuais no RES. A disponibilização desses modelos poderá beneficiar outros hospitais, otimizando a coleta de dados e possibilitando decisões clínicas e políticas públicas mais assertivas no controle do câncer.

Palavras-chave: Processamento de Linguagem Natural. Registro de câncer. Aprendizado de máquina. Qualidade de dados.

ABSTRACT

PINTO, A. L. **Use of Natural Language Processing to extract data in cancer registries**. 2024. 61 p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

In the field of health, Natural Language Processing (NLP) has proven to be a powerful tool for extracting and structuring clinical data from unstructured texts, making this process faster, more accurate and less susceptible to human error. In cancer registries, essential data is collected manually to assess the epidemiological profile, the quality of care provided and to comply with legal obligations, based on clinical reports and notes. With the implementation of the Electronic Health Record (EHR) at Hospital de Amor, it has become feasible to use texts to train NLP models that can help extract variables collected by registry professionals. The aim of this study was to train NLP algorithms to help extract variables collected manually in cancer registries from texts of biopsy reports and clinical narratives. The documents were pre-processed, vectorized with TF-IDF (Term Frequency-Inverse Document Frequency) and BioBERTpt, and trained using machine learning algorithms (logistic regression, random forest, Support Vector Machine, Naive Bayes and gradient boosting) to extract information on malignancy, topography, morphology and tumour staging. The best models obtained an accuracy above 92%, 96% and 88% in the extraction of malignancy, topography and morphology, respectively. However, the models had difficulties in extracting staging due to the absence of this information in medical records, showing a limitation in the quality of clinical narratives. Implementing these models in routine cancer registries could optimize the work of registry professionals, improve data quality and reduce the time it takes to make information available. In addition, by identifying lacunas in clinical documentation, the models can also contribute to improving the quality of textual information in the RES. The availability of these models could benefit other hospitals, optimizing data collection and enabling more assertive clinical decisions and public policies in cancer control.

Keywords: Natural Language Processing. Cancer registry. Machine learning. Data quality.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AP	Anatomopatológico
BERT	sigla do inglês, *Bidirectional Encoder Representations for Transformers*
CACON	Centros de Alta Complexidade em Oncologia
ETL	Extração, Transformação e Carregamento, sigla do inglês, *Extract, Transform and Load*)
IA	Inteligência Artificial
IHQ	Imunohistoquímico
INCA	Instituto Nacional de Câncer
HA	Hospital de Amor
LIS	Sistema de Informação Laboratorial, sigla do inglês, *Laboratory Information System*
MS	Ministério da Saúde
PACS	Sistema de Arquivamento e Compartilhamento de Imagens, sigla do inglês, *Picture Archive and Communication System*
PLN	Processamento de Linguagem Natural
PLNc	Processamento de Linguagem Natural clínico
RCBP	Registro de Câncer de Base Populacional
RES	Registros Eletrônicos de Saúde
RHC	Registro Hospitalar de Câncer
RIS	Sistema de Informação de Radiologia, sigla do inglês, *Radiology Information System*
SHAP	do inglês, *SHapley Additive exPlanations*
SVM	sigla do inglês, *Support Vector Machine*
TF-IDF	do inglês,*Term Frequency–Inverse Document Frequency*
UNACON	Unidades de Alta Complexidade em Oncologia

XAI Inteligência Artificial Explicável, sigla do inglês, *Explainable Artificial Intelligence*

SUMÁRIO

1 INTRODUÇÃO

A aplicação de algoritmos de Processamento de Linguagem Natural (PLN) tem permitido a automação de tarefas de extração de dados em diversas áreas. Na saúde o PLN pode ser usado para extrair e estruturar dados clínicos e laboratoriais a partir de textos não estruturados, presentes no prontuário do paciente. O uso de técnicas de PLN pode transformar o processo de coleta de dados nos serviços de saúde em uma tarefa mais rápida, precisa e menos suscetível a erros humanos.

Diariamente, nos Registros Hospitalares de Câncer (RHCs), pessoas revisam manualmente textos de laudos de exames e notas clínicas para coletar dados que permitam avaliar o perfil epidemiológico da instituição, a qualidade da assistência prestada e para o cumprimento dos deveres legais. Os RHCs são estruturas obrigatórias nos serviços de tratamento oncológicos habilitados pelo Ministério da Saúde. De forma semelhante, os Registros de Câncer de Base Populacional (RCBPs) realizam coleta de dados em diversos serviços de uma determinada região (cidade, região de saúde ou estado). No Brasil, existem mais de 360 registros de câncer, entre RHCs e RCBPs.

Com o aumento do uso de Registro Eletrônico de Saúde (RES) os hospitais passaram a gerar grandes volumes de textos em meio digital. Esses dados podem ser usados para treinar algoritmos de PLN que auxiliarão na extração de variáveis corriqueiramente coletadas. O Hospital de Amor (HA), maior hospital oncológico do país com atendimento 100% gratuito, anualmente produz milhões de textos de laudos de biópsias e narrativas clínicas. O HA também possui um dos maiores registros do país o que propicia um ambiente ideal para o treinamento e validação de modelos robustos de PLN para RHCs e RCBPs. Apesar de existirem algumas iniciativas pelo mundo, no Brasil não há publicações sobre uso de PLN no contexto dos registros de câncer.

O câncer é um grupo heterogêneo de doenças que têm em comum o crescimento descontrolado de células anormais, com potencial maligno. Topograficamente, o câncer pode se originar em várias partes do corpo, como mama, pele, pulmão, entre outros. O exame anatomopatológico é fundamental para o diagnóstico do câncer e é realizado por meio da análise de amostras de tecido tumoral retiradas durante a biópsia. Nesses fragmentos de tecido, são avaliadas as características morfológicas e moleculares das células para identificar o subtipo específico da doença. A combinação desses resultados com exames clínicos e de imagem possibilita o estadiamento do câncer, que indica a extensão da doença no corpo, ajudando na escolha do tratamento mais adequado e na previsão do comportamento clínico da doença.

Dessa maneira, o objetivo do presente trabalho foi treinar algoritmos de PLN

que auxiliem na extração de variáveis coletadas manualmente nos registros de câncer, a partir de textos disponíveis no RES. Para isso, foram treinados modelos capazes de extrair informações de malignidade, topografia, morfologia e estadiamento de tumores a partir de textos de laudos de exames anatomopatológicos e narrativas clínicas. Em seguida, foram apresentados e discutidos os resultados dos modelos treinados, bem como suas limitações e futuras aplicações.

Os modelos de PLN treinados nesse estudo apresentaram resultados satisfatórios. Se implementados na rotina dos registros de câncer, esse modelos podem não apenas otimizar o trabalho dos registradores, mas também melhorar a qualidade dos dados e reduzir o tempo de disponibilização das informações coletadas.

2 REFERENCIAL TEÓRICO

2.1 Câncer: conceito e classificações

O câncer é um conjunto de doenças caracterizadas pelo crescimento desordenado e disseminação de células anormais que podem invadir tecidos adjacentes e até se espalhar para órgãos distantes (NCI, 2007). O diagnóstico do câncer envolve realização de exames clínicos, laboratoriais, de imagem e anatomopatológico (AP). Os APs são realizados por meio de análises microscópicas de biópsias (pequenos fragmentos do tecido tumoral) que identifica a presença de células malignas. Esses mesmos fragmentos podem ser usados para realizar imunohistoquímica (IHQ) e teste moleculares, exames complementares ao AP, que auxiliam na identificação do subtipo do câncer.

A Classificação Internacional de Doenças para Oncologia (CID-O) é usada para codificar a topografia e a morfologia das neoplasias (tumores malignos e benignos) (OMS, 2005). O código topográfico indica a localização de origem de um tumor. Os tumores (malignos, benignos, *in situ* ou incertos) são registrados com códigos topográficos da CID-O como: C50 (tumor na mama), C34 (tumor no pulmão), C44 (tumor na pele), C61 (tumor na próstata), entre outros (Anexo A).

Já o código morfológico registra o tipo de célula que se tornou neoplásica e o seu comportamento biológico (Anexo A). Quanto ao comportamento, um tumor pode crescer num local e não ter potencial de disseminação (benigno); pode ser maligno e estar limitado ao local de origem (não invasivo ou *in situ*); pode invadir os tecidos adjacentes (maligno); ou pode estar disseminado e começando a crescer em outros locais (metastático) (OMS, 2005).

Os tumores malignos ainda são classificados quanto a sua extensão anatômica utilizando o sistema TNM. Esse sistema categoriza o carcinoma *in situ* como estadio 0, tumores localizados no órgão de origem como estadio I e II, tumores com disseminação local extensa como estadio III e tumores com metástase a distância como estadio IV (UICC, 2023).

Atualmente, o câncer já é segunda doença que mais causa de mortes no mundo. No Brasil, o Instituto Nacional de Câncer (INCA) estima que haverá cerca de 704 mil novos casos de câncer em 2024 (INCA, 2022). Globalmente, as estimativas apontaram que em 2022 tivemos aproximadamente 20 milhões de novos casos de câncer e quase 10 milhões de mortes atribuíveis à doença (GLOBOCAN, 2022). Essas estimativas são baseadas em dados enviados pelos de registros de câncer no Brasil e no mundo.

2.2 Registros de câncer

No Brasil, os serviços de tratamento oncológico devem seguir algumas regras estabelecidas em portaria para serem habilitados pelo Ministério da Saúde (MS) (Brasil, 2023). Uma das obrigações legais desses serviços é "a coleta, armazenamento, análise e divulgação de forma sistemática e contínua das informações dos pacientes atendidos e acompanhados no hospital, repassando os dados para o Instituto Nacional de Câncer"(Brasil, 2023). O INCA recebe dados como: perfil sociodemográfico de cada paciente, topografia, morfologia e estadiamento do câncer, datas do diagnóstico e do início do tratamento, estado atual da doença, entre outras informações (INCA, 2012).

Para que isso seja possível, em de cada serviço de oncologia deve ser implementado um Registro Hospitalar de Câncer (RHC), onde esses dados são coletados partir das informações registradas no prontuário do paciente (Brasil, 2023). Portanto, os RHCs são estruturas que têm como principal função fornecer informações sobre a atenção ao paciente com câncer nos hospitais habilitados, tendo finalidade administrativa e de avaliação da assistência oncológica (Bray; Parkin, 2009).

O trabalho realizado nos RHCs também facilita a coleta de dados feita pelos Registros de Câncer de Base Populacional (RCBPs). Diferente dos RHCs que registram casos de um único serviço de oncologia, os RCBPs registram todos os casos de cânceres de uma população residente numa determinada área geográfica (ex.: uma cidade, uma região de saúde ou um estado). Com essa informação é possível calcular estimativas de incidência de câncer por ano numa região. São as estimativas de casos novos que permitem planejamento dos programas de prevenção e controle do câncer e o estabelecimento de Unidades ou Centros de Alta Complexidade em Oncologia (UNACONs ou CACONs) em uma determinada região.

Os RHCs e RCBPs são formados por profissionais chamados de registradores que, geralmente, coletam dados de forma manual tornando moroso o processo de consolidação e disponibilização das informações. Além disso, a alta rotatividade desses profissionais atrapalham a continuidade dos serviços, pois a formação de um registrador é demorada. O baixo financiamento é outro fato que atrapalha a sustentabilidade dos registros. O uso de tecnologias pode ajudar na redução de custos dos registros de câncer e aumentar da produtividade, mantendo a qualidade dos serviços.

2.3 Registro eletrônico de saúde (RES)

Desde os primórdios da medicina existe a preocupação dos profissionais de saúde em registrar dados relacionados as características dos seus pacientes (Dalianis, 2018). Essas informações podem ser registradas em documentos físicos ou eletrónicos conhecidos como prontuário médico. A resolução nº 1.638/2002 do Conselho Federal de Medicina define

como o prontuário médico "o documento único constituído de um conjunto de informações, sinais e imagens registradas, geradas a partir de fatos, acontecimentos e situações sobre a saúde do paciente e a assistência a ele prestada, de caráter legal, sigiloso e científico, que possibilita a comunicação entre membros da equipe multiprofissional e a continuidade da assistência prestada ao indivíduo"(CFM, 2002).

Com a digitalização, o prontuário também passou a ser chamado de Registro Eletrônico de Saúde (RES). O RES é uma coleção de informações de usuários de um serviço de saúde armazenadas em meio digital. Assim como o prontuário de papel, no RES são coletadas informações sobre o histórico de saúde do paciente, como diagnósticos, medicamentos em uso, exames, alergias, imunizações e planos terapêuticos (NCI, 2011). Popularmente, o RES também é chamado Prontuário Eletrônico do Paciente (PEP). Pesquisadores da área acreditam que o RES é uma das inovações mais significativas introduzidas na área da saúde nas últimas décadas (Nelson; Staggers, 2017).

O RES é formado por vários módulos como: sistema de prescrição eletrônica de medicamentos, sistema de apoio à decisão clínica, sistemas departamentais (pronto-socorro, centro cirúrgico, ambulatório e financeiro), sistema de arquivamento e compartilhamento de imagens (PACS, do inglês, *Picture Archive and Communication System*), sistema de informação de radiologia (RIS, do inglês, *Radiology Information System*), sistema de informação laboratorial (LIS, do inglês, *Laboratory Information System*) e sistema de documentação clínica (Colicchio, 2020).

No sistema de documentação clínica são registrados as narrativas ou notas clínicas que são textos descritivos que documentam e comunicam apresentações clínicas, impressões do paciente, detalhes de procedimentos e tomada de decisões (Rosenbloom *et al.*, 2011). Os laudos de exames, presentes no LIS, também possuem textos livres relativos interpretação de resultados dos parâmetros analisados. Embora o texto livre seja eficaz na comunicação entre profissionais de saúde, se torna um desafio pesquisar, resumir ou analisar esses dados não estruturados para geração de indicadores de qualidade da assistência ou para fins secundários, como pesquisa (Yim *et al.*, 2016).

2.4 Processamento de Linguagem Natural (PLN) e suas aplicações nas áreas biomédicas

O PLN é um campo de pesquisa que investiga e propõe métodos e sistemas de processamento computacional das línguas faladas pelos humanos, também conhecidas como linguas naturais (Caseli; Nunes, 2024). O PLN também é considerado uma subárea de Inteligência Artificial (IA). Com o avanço dos algoritmos de IA para PLN, a área da saúde vem sendo beneficiada com o surgimento de ferramentas que coletam dados de narrativas clínicas ou textos de laudos de exames. Há pelo menos 30 anos, pesquisadores utilizam abordagens de PLN para automatizar a extração de informação em larga escala,

a partir de texto clínicos, estruturando os dados fenotípicos dos pacientes para realização de pesquisa ou para tomadas de decisões clínicas e gerenciais (Savova *et al.*, 2019).

No domínio biomédico, o PLN é também conhecido com BioPLN ou PLN clínico (PLNc). Com o aumento do poder computacional e o surgimento de novos algoritmos nos últimos anos, o PLNc vem demostrando suas potencialidades em diversas especialidades médicas. No entanto, existem muitos desafios relacionados a variabilidade linguística, a abundância de terminologia médica, abreviações, sinônimos, jargões e inconsistências ortográficas predominantes em textos clínicos (Savova *et al.*, 2019).

Kreimeyer e colegas (2017) revisam os sistemas de PLNc mais populares como: MetaMap (mapeamento de conceito) (Aronson; Lang, 2010); Apache cTAKES (componentes clássicos de PNL, mapeamento de conceito, entidades e atributos, relações, temporalidade) (Savova *et al.*, 2010); YTex (entidade e atributos) (Garla *et al.*, 2011); anotador OBO (mapeamento de conceito) (Jackson *et al.*, 2021); TIES (ligação de relatórios de patologia a dados de banco de tecidos); MedLEE (entidades e atributos, relações) (Friedman, 2000); CLAMP (entidades e atributos) (Soysal *et al.*, 2018); e NOBRE (entidades e atributos) (Tseytlin *et al.*, 2016).

Na oncologia, as abordagens de PLNc vêm sendo aplicada para extrair temporalidade (Strötgen; Gertz, 2013; Lin *et al.*, 2016; Lin *et al.*, 2019), características de tumores (Savova *et al.*, 2017; Qiu *et al.*, 2018; Gao *et al.*, 2018), resposta a tratamentos (Bergquist *et al.*, 2017), toxicidade a drogas (Hong *et al.*, 2020), progressão da doença, metástases, além de auxiliar no recrutamento de pacientes para pesquisa clínica (Shivade *et al.*, 2016; Zhang; Demner-Fushman, 2017; Bustos; Pertusa, 2018). Essas são as mesmas variáveis corriqueiramente coletadas em registros de câncer pelo mundo.

Em alguns países já existem diversas iniciativas usando PLN no contexto dos registros hospitalares de câncer (Merriman *et al.*, 2021; Hochheiser *et al.*, 2023). Na era dos dados, cresce as demandas impostas aos registradores por de informações adicionais e envio de relatórios no menor tempo possível para agências de governo (Merriman *et al.*, 2021). Para atender tal demanda, cada vez mais os registradores vão precisar aumentar produtividade com auxílio de tecnologias com PLN. Apesar de já ser realidade em alguns países, no Brasil não há publicações do uso dessa tecnologia na rotina dos registradores.

2.5 *Pipeline* de desenvolvimento de modelos de PLN

2.5.1 Pré-processamento

A primeira etapa do *pipeline* de construção de modelos de IA em PLN é pré-processamento dos *corpora* (conjunto de textos) (Patel, 2020). Nessa etapa pode haver a normalização dos textos (ex.: converter o texto em minúsculo); a remoção de marcadores (ex.: tags HTML), caracteres especiais e espaços em branco; e a substituição de abreviaturas

e siglas.

Em seguida, pode ser necessário os processos de lematização e "stemização". Na lematização as palavras são substituídas pela sua forma básica ou normal (ex.: “cuidamos” para “cuidar”) e na "stemização"são extraídos apenas os radicais das palavras, removendo os afixos (ex.: “cuidar” para “cuid”). Na sequência, remove-se as *stopword* que são palavras muito frequentes e com pouco ou nenhum significado (ex.: "a", "de", "um", "para", entre outros).

2.5.2 Representação textual

A representação textual é a conversão dos textos em vetores numéricos para serem utilizados pelos algoritmos de Aprendizado de Máquina (AM). Na vetorização do texto podem ser usadas técnicas como *Bag-of-Words*, TF-IDF (*Term Frequency-Inverse Document Frequency*), ou *embeddings* (*Word2Vec* e *BERT*).

O TF-IDF é uma técnica que se baseia na atribuição de peso a cada palavra com base em sua frequência num documento e sua raridade em outros (Salton; Buckley, 1988). Ou seja, termos frequentes em documentos específicos e raros nos demais recebem pesos mais altos. Dessa maneira, os vetores são construídos usando a equação 2.1.

$$TFIDF = \frac{\text{nº do termo x no documento}}{\text{total de termos do documento}} \times \log\left(\frac{\text{nº de documentos}}{\text{nº de documentos contendo x}}\right) \quad (2.1)$$

Os *word embeddings* são representações densas e de valores contínuo de palavras que capturam relações semânticas entre elas. Em 2018, o Google introduziu o modelo BERT (*Bidirectional Encoder Representations from Transformers*), que utiliza o componente *encoder* para criar *embeddings* contextuais na representação de textos (Devlin *et al.*, 2019). Em 2020, foi publicado o BioBERTpt que enriqueceu o BERT para a língua portuguesa com aprendizado com notas clínicas e literatura biomédica (Schneider *et al.*, 2020).

2.5.3 Treinamento dos modelos

Após a vetorização, os dados estão pronto para serem usados para treinar de modelos de AM. Para resolver os problemas de PNL são usados tanto modelos mais clássicos como regressão logística, *random forest*, SVM (do inglês, *Support Vector Machine*), Naive Bayes *gradient boosting*, quanto algoritmos mais recentes de aprendizado profundo, como rede neural recorrente e rede neural convolucional (Zhou *et al.*, 2022).

O método mais tradicional para resolver problemas de classificação é regressão logística (Berkson, 1944). A regressão logística é um método estatístico utilizado para modelar a probabilidade de um evento binário. Embora o nome sugira uma relação com a regressão linear, a principal diferença é que a regressão logística mapeia a saída de uma

combinação linear das características através de uma função logística (ou sigmoide) para produzir um valor entre 0 e 1, que pode ser interpretado como uma probabilidade.

O SVM também um modelo muito usado para classificação e busca encontrar o hiperplano ideal que separa as classes de dados no espaço de características, maximizando a margem entre as classes mais próximas (Cortes; Vapnik, 1995). O SVM é eficaz em espaços de alta dimensionalidade e pode ser ajustado para problemas lineares e não lineares.

O algoritmo Naive Bayes é um classificador probabilístico baseado no teorema de Bayes com a suposição de independência condicional entre as características (Thompson; Duda; Hart, 1974). Esse pressuposto de independência entre os atributos simplifica o cálculo das probabilidades, tornando o algoritmo eficiente e fácil de implementar.

O *random forest* é uma técnica de *ensemble* que combina múltiplas árvores de decisão para melhorar a precisão da predição e evitar o sobreajuste (*overfitting*), um problema comum em árvores de decisão individuais. Cada árvore no *ensemble* é construída usando uma amostra aleatória dos dados e uma subamostra das características (Breiman, 2001).

Outra técnica de *ensemble* bastante utilizada é o *gradient boosting*. Esse método cria modelos sequencialmente, onde cada novo modelo é treinado para corrigir os erros dos modelos anteriores. Especificamente, ele ajusta modelos simples (geralmente árvores de decisão) a um resíduo dos erros do modelo anterior, combinando-os para formar um modelo mais forte (Friedman, 2001).

2.5.4 Métricas de avaliação de modelos de classificação

Métricas como acurácia, precisão, *recall* e *f1-score* são fundamentais para comparar o desempenho de modelos de classificação durante o treinamento e ajudam a entender melhor como o modelo está lidando com os diferentes tipos de erros. Uma revisão sistemática, aponta que a maiorias dos trabalhos de PLN no contexto da oncologia usaram acurácia e *recall* para avaliar os algoritmos (Gholipour *et al.*, 2023).

A acurácia mede a proporção de previsões corretas em relação ao total de previsões feitas. É calculada como o número de previsões corretas dividido pelo número total de amostras (Equação 2.2). Em conjuntos de dados desbalanceados, a acurácia pode ser enganosa, pois um modelo que simplesmente prevê a classe majoritária terá alta acurácia, mas não necessariamente um bom desempenho.

$$\text{Acurácia} = \frac{\text{Verdadeiros Positivos} + \text{Verdadeiros Negativos}}{\text{Total de amostras}} \tag{2.2}$$

A precisão mede a proporção de verdadeiros positivos entre todas as previsões positivas feitas pelo modelo, ou seja, indica a exatidão das previsões positivas (Equação

2.3). Em situações onde o impacto do falso positivo é alto, a precisão se torna a métrica mais relevante.

$$\text{Precisão} = \frac{\text{Verdadeiros Positivos}}{\text{Verdadeiros Positivos} + \text{Falsos Positivos}} \tag{2.3}$$

O *recall* ou revocação mede a proporção de verdadeiros positivos entre todos os casos que são realmente positivos (Equação 2.4). Essa métrica indica a capacidade do modelo de identificar corretamente todos os casos positivos. O *recall* é crucial em situações onde perder um positivo (falso negativo) tem consequências graves, como nos testes de rastreio de doenças.

$$\textit{Recall} = \frac{\text{Verdadeiros Positivos}}{\text{Verdadeiros Positivos} + \text{Falsos Negativos}} \tag{2.4}$$

O *f1-score* é a média harmônica entre a precisão e o *recall* (Equação 2.5). Ele fornece um único valor que representa um balanço entre os dois, sendo útil quando é necessário um equilíbrio entre precisão e *recall*. O *f1-score* é útil em problemas onde há um desbalanceamento entre classes, e onde tanto a precisão quanto o *recall* são importantes.

$$\textit{F1-score} = \frac{2 \times \text{Precisão} \times \textit{Recall}}{\text{Precisão} + \textit{Recall}} \tag{2.5}$$

2.5.5 Explicabilidade e interpretabilidade

A inteligência artificial explicável (XAI, do inglês, *Explainable Artificial Intelligence*) é definida como estratégias que permitem aos usuários compreenderem e interpretarem previsões feitas pelos modelos de IA. Na área da saúde a XAI pode acelerar a adoção de sistemas de IA por reduzir desconfiança dos profissionais, trazendo mais transparência nas tomadas de decisão realizadas pelos modelos (Band *et al.*, 2023).

Uma das técnicas para tornar os modelos de IA mais interpretáveis é o SHAP (*SHapley Additive exPlanations*) (Lundberg; Lee, 2017). O SHAP foi proposto para explicar de forma gráfica a importância das variáveis para previsão feita pelo modelo. No gráfico, quanto mais positivo for o valor de SHAP (ou, *SHAP value*) de uma variável, mais ela impactou positivamente para aquela predição e quanto mais negativo, maior foi o impacto negativo na predição.

3 PROPOSTA E METODOLOGIA

O Hospital de Amor (HA), além de ser o maior hospital oncológico do país, possui RHC e RCBP com qualidade reconhecida. Por isso, o HA pode ser um laboratório ideal para treinamento e validação de modelos de PLN que auxiliem a rotina dos registros de câncer de todo o país.

Nesse contexto, propõem-se treinar modelos de PLN para extrair 4 importantes varáveis categóricas, coletadas manualmente pelos registradores a partir de textos de laudos de exames e notas clínicas. Essas variáveis foram: (1) malignidade da amostra (ex.: neoplasia maligna ou doença benigna), (2) topografia do câncer (câncer de mama, de pele, de pulmão, entre outros), (3) morfologia do câncer (carcinoma basocelular, escamocelular, escamocelular *in situ*, ductal infiltrante, adenocarcinoma, entre outros) e (4) estadiamento da doença (ex.: estadio 0, I, II, III, IV, X ou Y).

Para treinar os modelos de classificação foram utilizados textos de laudos de exames anatomopatológico (AP) e imunohistoquímico (IHQ) realizados a partir de biopsia de tumores analisados no HA de Barretos. Os textos dos laudos foram extraídos do banco *Oracle* do LIS. Após a extração, os dados foram tratados para eliminar inconsistências e pseudoanonimizados com a criação de identificadores fictícios. Inicialmente, foram realizadas analises exploratórias para avaliar a frequência relativa simples e acumulada das categorias (Figura 1).

Em seguida, foram amostrados 5.000 laudos de neoplasias malignas e 5.000 laudos de doenças benigna para treinar os modelos de magnilidade, categorizando os laudos em duas classes. Para treinar os modelos de predição de topografia, amostramos 15.000 laudos, sendo 3.000 de cada um dos três cânceres mais prevalentes (mama, pele e pulmão) e 3.000 de outros cânceres, estabelecendo quatro categorias. Para os modelos de classificação de morfologia, foram amostrados 1.000 laudos das patologias mais comuns (carcinoma basocelular, carcinoma ductal infiltrante, carcinoma escamocelular, adenocarcinoma, carcinoma escamocelular *in situ* e outras morfologias), formando 6 categorias (Figura 1).

A etapa de pré-processamento dos textos foi realizada com o auxílio de funções do pacote *sklearn* (Pedregosa *et al.*, 2011). Dividiu-se as amostras em treino (80%) e teste (20%) usando a função *train test split*, foram removidas as *stop-words* usando o *get stop words* e os textos foram vetorizados com a função *TF-IDF Vectorizer*. Com o objetivo de utilizar ferramentas mais modernas de transformação, foram gerados *embedings* com modelos pré-treinados baseados em BERT. Para realizar essas transformações, reduziu-se o número de amostras pelo alto custo computacional dessas aplicações. Foi utilizado o BioBERTpt (Schneider *et al.*, 2020), um modelo treinado com narrativas clínicas em

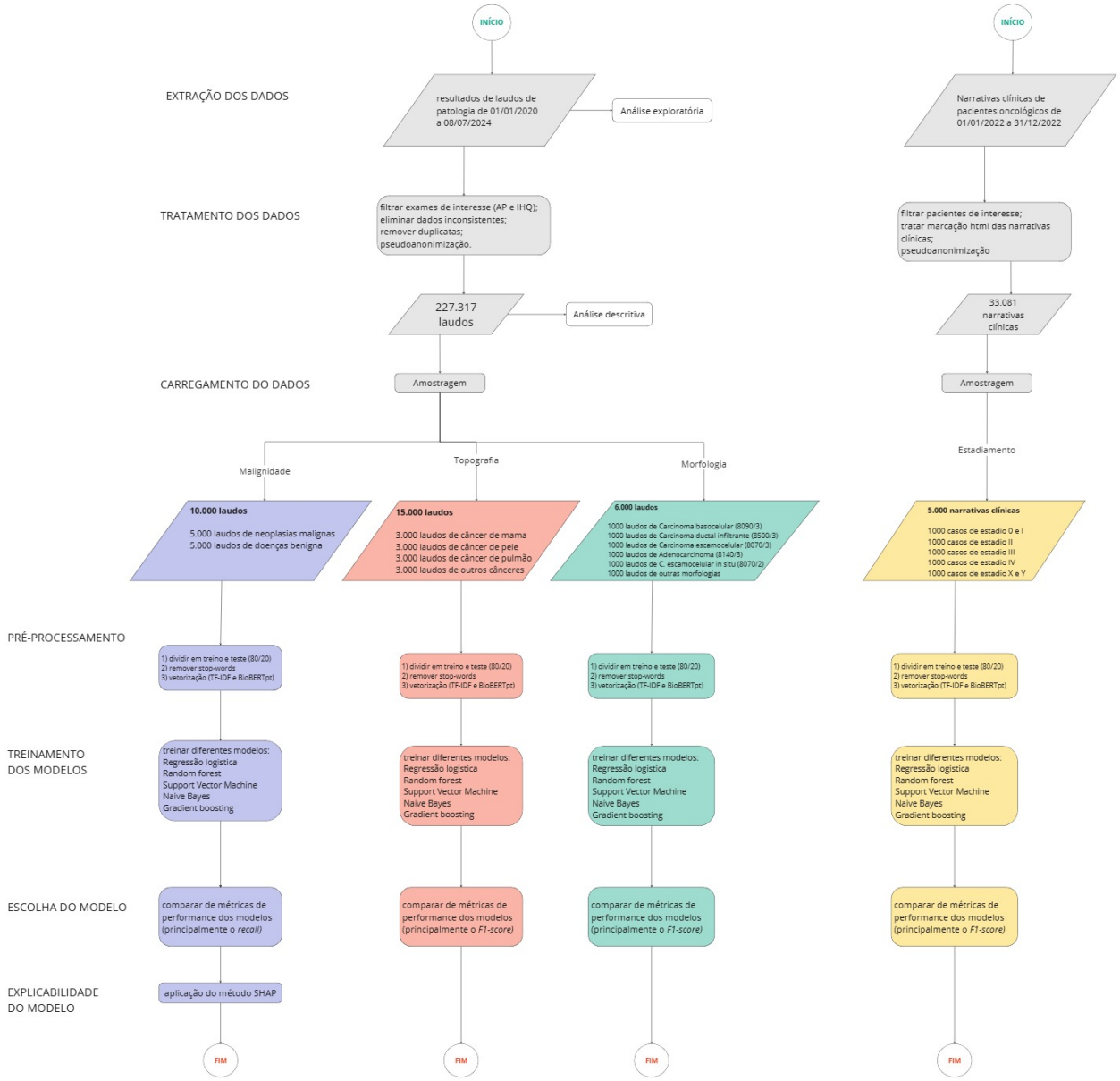

Figura 1 – Fluxograma da metodologia

português disponível no *Hugging Face*.

Na etapa de treinamento, optou-se por utilizar 5 modelos de classificação com abordagens distintas: regressão logística, *random forest*, *Support Vector Machine* (SVM), *Naive Bayes* e *gradient boosting*. No treinamento, foram aplicados os hipérmetros *default* das funções do *sklearn*. Para concluir, utilizou-se funções as *classification report* e matrizes de confusão para comparar de métricas de performance entre os modelos e eleger a melhor estratégia. Na seção seguinte serão apresentados os resultados dos experimentos realizados.

4 AVALIAÇÃO EXPERIMENTAL

4.1 Análises exploratórias

O objetivo das análises exploratórias foi definir a categorização da variáveis escolhidas, avaliando a frequência relativa simples e acumulada dos dados extraídos dos nossos bancos internos do LIS e do RHC. A frequência acumulada nos permitiu selecionar as categorias responsável por 70% ou 80% do trabalho realizado pelos registradores do RHC (Figura 1).

4.1.1 Análise exploratória das variáveis malignidade e topografias

Dos laudos extraídos do banco do LIS, 61,7% têm como diagnóstico doenças benignas e 20,2% são cânceres de mama (C50), pele (C44) e pulmão (C34) (Figura 2). Isso quer dizer que treinando um modelo para malignidade com 2 classes (benigno ou maligno) e outro modelo para topografia com 4 classes (mama, pele, pulmão e outros) resolvemos 81,9% do volume de laudos analisados pelos registradores.

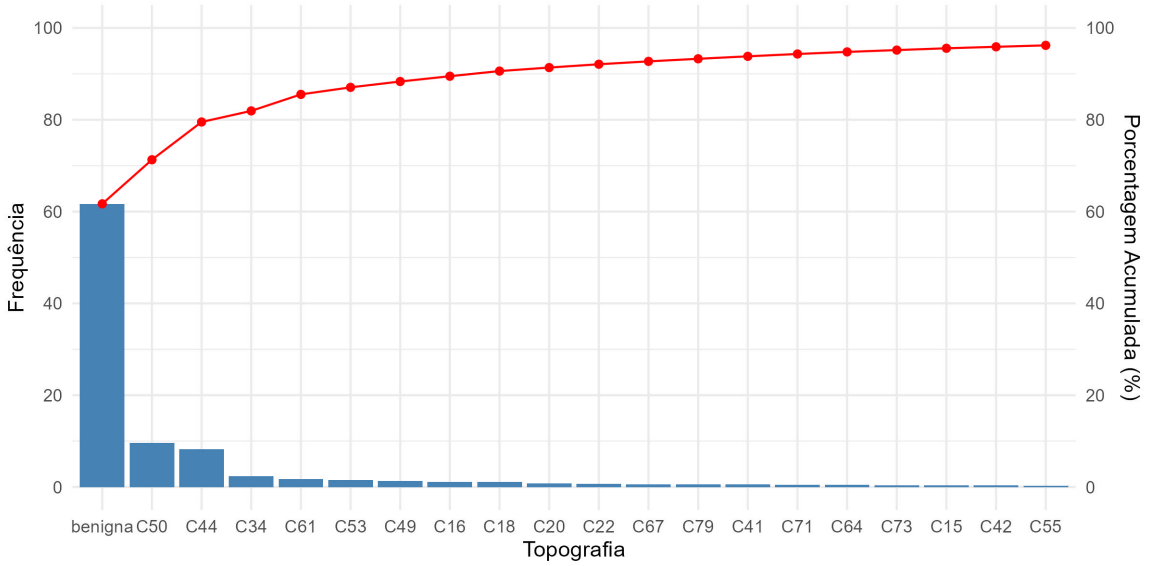

Figura 2 – Frequência relativa simples e acumulada das topografias

4.1.2 Análise exploratória da variável morfologias

Usando a mesma lógica para as morfologias, observamos que 71,1% do volume de laudos concentram apenas 5 tipos de morfologias (Figura 3). Dessa maneira, criando 6 classes (80903 - carcinoma basocelular, 80703 - escamocelular, 80702 - escamocelular *in situ*, 85003 - ductal infiltrante, 81403 - adenocarcinoma e outro), podemos reduzir para 28,9% o volume de laudos que serão analisados manualmente pelos registradores.

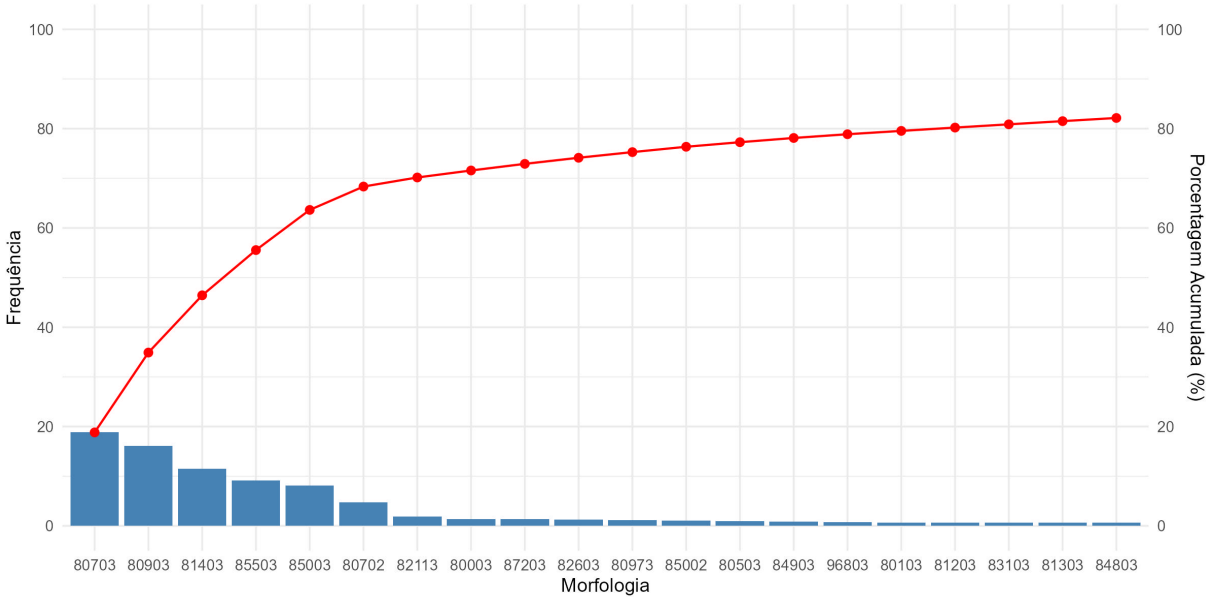

Figura 3 – Frequência relativa simples e acumulada das morfologias

4.1.3 Análise exploratória da variável estadiamento

Enquanto mais de 30% das narrativas clínicas extraídas eram de casos de cânceres estadio I, os estadio 0, X e Y apresentavam menos de 10% (Figura 4). Os estadio 0 e I são clincicamente semelhantes por isso decidimos formar uma única classe denominada "0 e I"(UICC, 2023). O mesmo aconteceu com o X e Y. Desse forma, temos 5 classes para treinar os modelos de classificação.

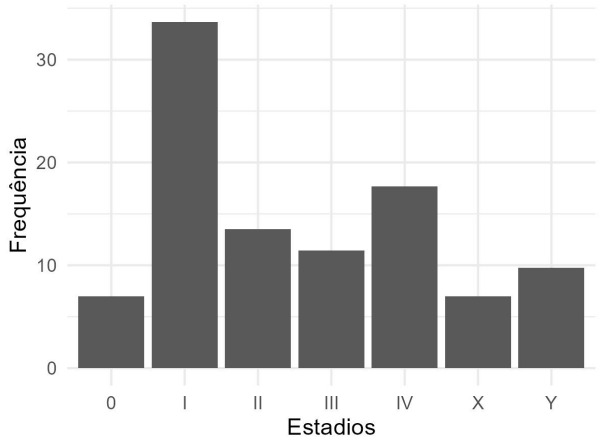

Figura 4 – Frequência relativa dos estadios

4.2 Resultados e Discussão

4.2.1 Modelo de classificação de malignidade

A avaliação de malignidade nos registros de câncer é a primeira etapa do processo de coleta de dados. Se o registrador não identificar todos os casos malignos, o número de casos novos de câncer registrados pelo hospital vai estar subestimado. Por isso, para ter utilidade para os registros hospitalares, o modelo de classificação de malignidade deve ter o menor número de falso negativos, ou seja, o maior *recall* para classe "maligno". Dos modelos treinados, o *random forest* foi o único que apresentou um *recall* de 95% para classe "maligno"(Tabela 1).

Tabela 1 – Métricas de performances dos modelos treinados para malignidade

Modelo	**Tumor**	**Precisão**	**Recall**	**F1-score**	**Acurácia**
Regressão logística	benigno	0,93	0,91	0,92	0,92
	maligno	0,91	0,93	0,92	
Random forest	benigno	**0,95**	**0,89**	**0,92**	**0,92**
	maligno	**0,90**	**0,95**	**0,92**	
SVM	benigno	0,94	0,91	0,92	0,92
	maligno	0,91	0,94	0,92	
Naive Bayes	benigno	0,92	0,81	0,86	0,87
	maligno	0,83	0,93	0,87	
Gradient boosting	benigno	0,93	0,91	0,92	0,92
	maligno	0,91	0,93	0,92	

As matrizes de confusão mostram que *random forest* foi o modelo que mais classificou corretamente laudos "malignos"(938), errando em apenas 5% (50 laudos) (Figura 5). Na análise desses 50 laudos, observou-se que 7 laudos haviam sido rotulados erroneamente por humano e 4 laudos eram inconclusivos para malignidade. Treze laudos eram de cânceres mais raros como as neoplasias hematológicas (8), melanoma (4) e primário oculto (1). Provavelmente, esses erros aconteceram pelo baixo número de exemplos dessas doenças no treinamento do modelo.

Na análise de SHAP, observa-se que os termos “carcinoma”, "adenocarcinoma"e "invasivo"são os que mais influenciam positivamente para que o modelo classifique o laudo como neoplasia maligna (Figura 6). Esses termos, de fato dizem respeito a morfologia e ao comportamento de tumores malignos, mostrando que o modelo tomou decisões baseadas numa lógica do "mundo real".

Por outro lado, as palavras "pylori", "intraepitelial"e "pólipo"foram as que mais contribuíram para o modelo classificar o laudo como doença benigna (Figura 6). O termo "pylori"vem de *Helicobacter pylori* uma bactéria que coloniza a mucosa do estômago podendo causar gastrite, uma doença benigna. As palavras "intraepitelial"e "pólipo"estão presentes em laudos de lesões precursora de colo de útero e colorretal, respectivamente, mas que ainda

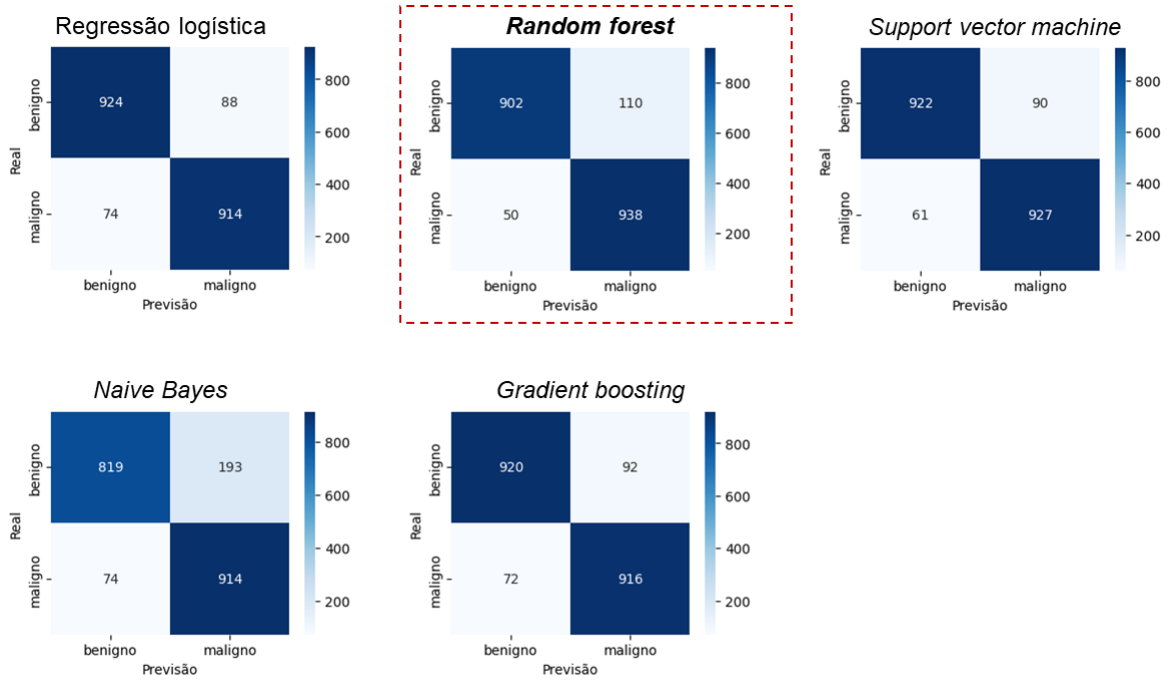

Figura 5 – Matrizes de confusão da predição dos modelos treinados para malignidade

não são considerados cânceres. Da mesma maneira, a explicabilidade/interpretabilidade trazida pelo SHAP evidencia que o modelo estar tomando decisões que vão ao encontro das premissas biomédicas.

4.2.2 Modelo de classificação de topografias

Para que um modelo de classificação de topografias possa auxiliar na rotina dos registros de câncer, é necessário que ele acerte o maior número de classificações possível. Os poucos erros cometidos poderão ser corrigidos por humanos em etapas posteriores, sem tanto prejuízo para o processo. Por isso, nesse caso, como as classes estão balanceadas, usou-se acurácia para escolher o melhor modelo, e como critério de desempate foram usados o *f1-score* e o *recall*.

Todos os modelos treinados para extrair as topografias apresentaram acurácia acima de 94% e *recall* e *f1-score* superiores a 90% em todas as classes (Tabela 2). Os modelos de SVM e regressão logística performaram de maneira semelhante com acurácia de 96% e *f1-scores* também idênticos (Tabela 2). O modelo SVM foi escolhido por apresentar um *recall* superior na classe "outros".

Nas matrizes de confusão também é possível observar que a grande maioria dos erros de classificação aconteceram envolvendo a classe "outros"(Figura 7). Esse resultado era esperado por se tratar de uma classe muito heterogênea composta de dezenas de topografias distintas. Trinta e quarto laudos da classe "pulmão"foram classificados como

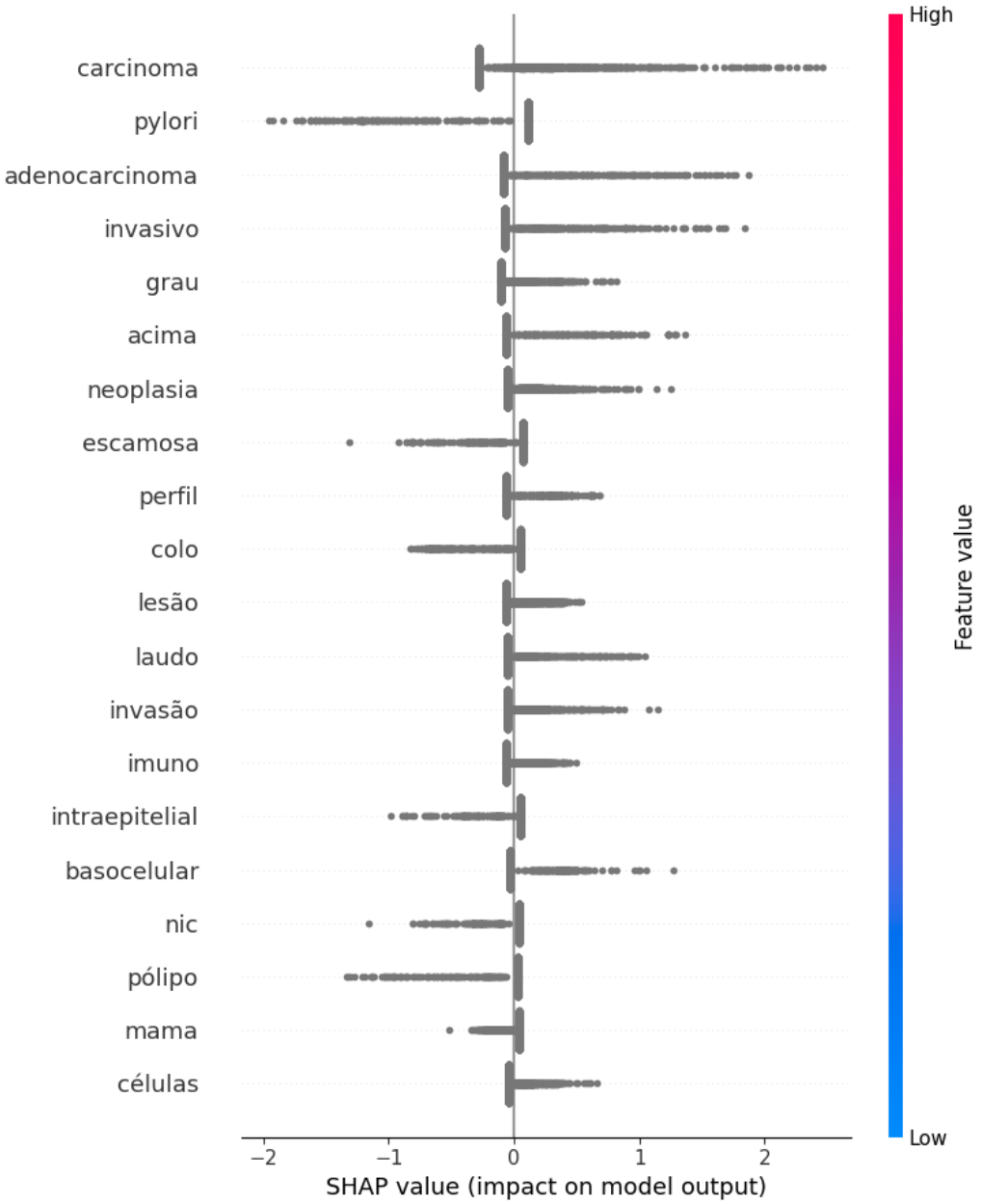

Figura 6 – Análise de explicabilidade para o modelo de malignidade

"outros". Ao analisar esses laudos, observa-se que 18 foram erros humanos de rotulagem. Oito laudos não apresentavam informações suficiente e somente em 10 casos o modelo, de fato, errou.

No caso dos 33 erros classificando a topografia "pele"como "outros", observou-se 13 erros humanos de rotulagem ou eram laudos com informação insuficiente. Os 20 laudos classificados equivocadamente pelo modelo são de câncer de pele do tipo melanoma, que por ser mais raro, foram apresentadas poucas amostras durante o treinamento. Não foi possível realizar análise de explicabilidade do modelo por limitações de poder computacional.

Tabela 2 – Métricas de performances dos modelos treinados para topografias

Modelo	Topografias	Precisão	Recall	F1-score	Acurácia
Regressão logística	outros	0,92	0,94	0,93	0,96
	mama	0,98	0,98	0,98	
	pele	0,98	0,96	0,97	
	pulmão	0,97	0,96	0,97	
Random forest	outros	0,89	0,91	0,90	0,94
	mama	0,98	0,98	0,98	
	pele	0,98	0,94	0,96	
	pulmão	0,93	0,95	0,94	
SVM	outros	**0,92**	**0,95**	**0,93**	**0,96**
	mama	**0,98**	**0,98**	**0,98**	
	pele	**0,98**	**0,96**	**0,97**	
	pulmão	**0,96**	**0,96**	**0,97**	
Naive Bayes	outros	0,89	0,90	0,90	0,94
	mama	0,96	0,98	0,97	
	pele	0,97	0,95	0,96	
	pulmão	0,94	0,92	0,93	
Gradient boosting	outros	0,89	0,93	0,91	0,95
	mama	0,97	0,98	0,97	
	pele	0,98	0,95	0,96	
	pulmão	0,96	0,94	0,95	

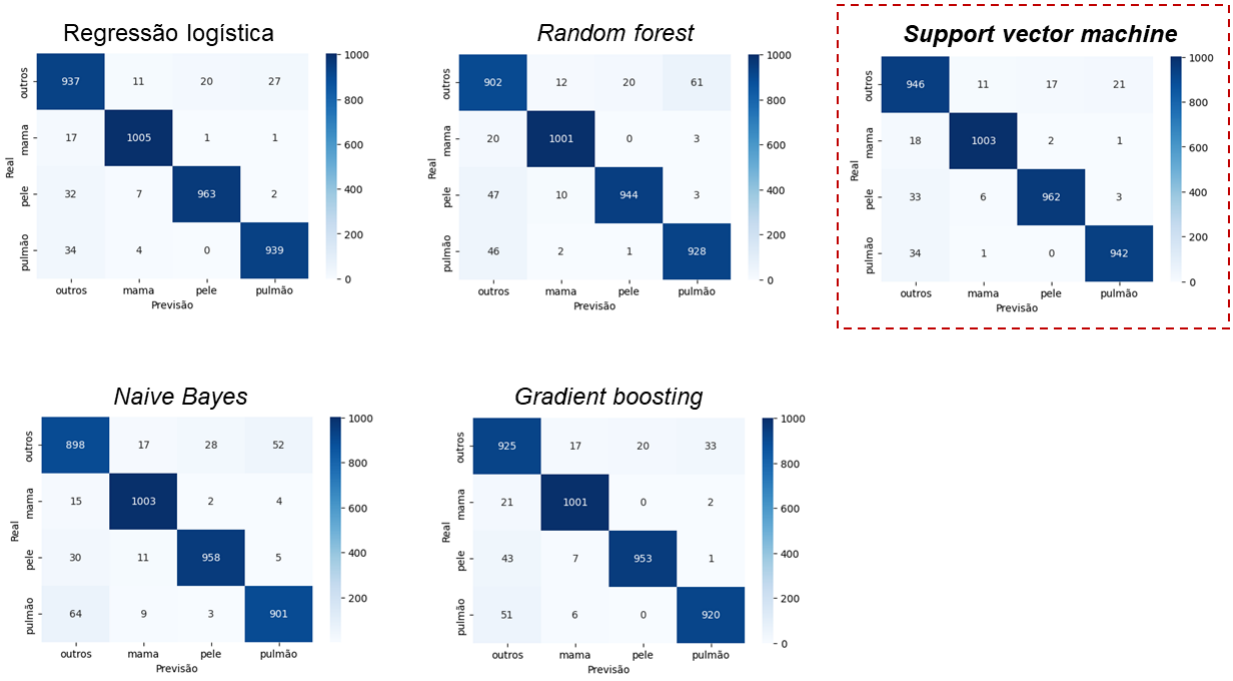

Figura 7 – Matrizes de confusão da predição dos modelos treinados para topografia

4.2.3 Modelo de classificação de morfologias

Assim como no caso anterior, para auxiliar os registradores, esse modelo precisa acertar o máximo possível, mas o erro também pode ser corrigido em etapas futuras do processo. Por isso, elegemos o modelo com maior acurácia e os melhores *f1-score*.

O modelo *gradient boosting* apresentou a melhor acurácia (88%) e os melhores *f1-scores* em todas as classes (Tabela 3). Por termos 6 categorias diferentes, era esperando que a acurácia não fosse tão alta como nos modelos de malignidade e topografia. O *f1-score* foi o menor na classe "outras"(76%), acredita-se que heterogeneidade do grupo (Tabela 3).

Tabela 3 – Métricas de performances dos modelos treinados para morfologias

Modelo	**Morfologias**	**Precisão**	**Recall**	**F1-score**	**Acurácia**
Regressão logística	outras	0,78	0,70	0,74	0,85
	80903	0,87	0,95	0,91	
	85003	0,91	0,95	0,93	
	80703	0,88	0,78	0,83	
	81403	0,78	0,86	0,82	
	80702	0,90	0,88	0,89	
Random forest	outras	0,78	0,70	0,74	0,85
	80903	0,87	0,94	0,91	
	85003	0,89	0,97	0,93	
	80703	0,88	0,76	0,81	
	81403	0,80	0,83	0,83	
	80702	0,91	0,94	0,93	
SVM	outras	0,79	0,75	0,77	0,86
	80903	0,86	0,96	0,91	
	85003	0,93	0,95	0,94	
	80703	0,87	0,76	0,81	
	81403	0,80	0,87	0,84	
	80702	0,90	0,87	0,88	
Naive Bayes	outras	0,80	0,55	0,66	0,73
	80903	0,66	0,90	0,76	
	85003	0,84	0,99	0,91	
	80703	0,65	0,48	0,55	
	81403	0,73	0,87	0,79	
	80702	0,73	0,63	0,68	
Gradient boosting	outras	**0,79**	**0,74**	**0,76**	**0,88**
	80903	**0,94**	**0,93**	**0,93**	
	85003	**0,94**	**0,95**	**0,95**	
	80703	**0,89**	**0,84**	**0,87**	
	81403	**0,80**	**0,86**	**0,83**	
	80702	**0,89**	**0,94**	**0,91**	

Na matriz de confusão, observa-se que 20 laudos rotulados com a classe "81403"(adenocarcinoma) foram classificados como "outras"morfologias (Figura 8). Ao analisar cada um desses laudos, observa-se que houve erro humano de rotulagem em 11 dos 20 documentos.

Nos demais, o modelo classificou erroneamente, provavelmente, por se tratar de outros tipos adenocarcinomas. Não foi possível realizar análise de explicabilidade do modelo por limitações de poder computacional.

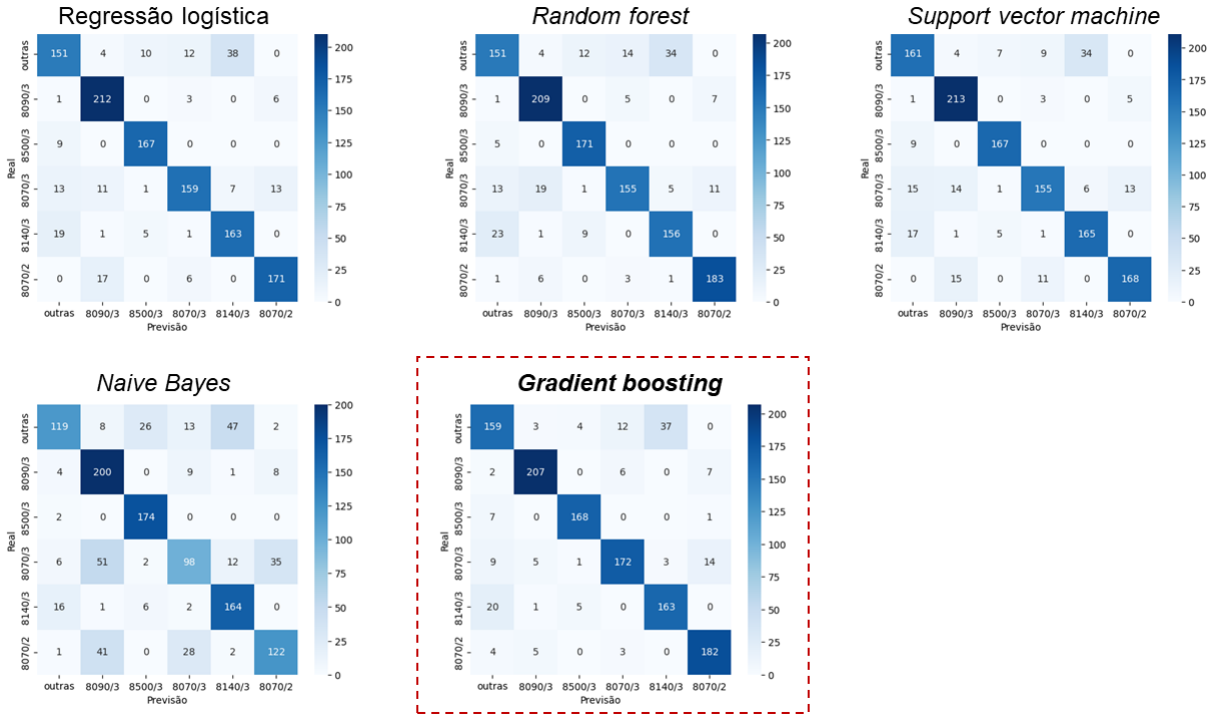

Figura 8 – Matrizes de confusão da predição dos modelos treinados para morfologia

4.2.4 Modelo de classificação de estadiamento

As performances dos modelos de classificação de estadiamento ficaram bem abaixo dos demais descritos neste trabalho (Tabela 4 e Figura 9). O modelo com melhor desempenho atingiu apenas 68% de acurácia. Apesar da baixa performance, esse modelo pode ter uma utilidade no contexto dos registros de câncer. Um dos objetivos dos RHCs é avaliar a qualidade das informações contidas nas narrativas clínicas. Ao analisar esses documentos, observa-se que poucas evoluções médicas apresentam explicitamente o estadio da doença, o que pode explicar a baixa acurácia desses modelos.

4.2.5 Modelos de classificação usando o BioBERTpt

Devido o alto custo computacional da transformação dos documentos em *word embedding* usando BioBERTpt, o tamanho das amostras foi reduzido de cinco a dez vezes, o que inviabiliza a comparações com os modelos acima. De qualquer forma, os resultados estão disponíveis no apêndice A (Tabelas 5, 6, 7 e 8).

Tabela 4 – Métricas de performances dos modelos treinados para estadios

Modelo	Estadios	Precisão	Recall	F1-score	Acurácia
Regressão logística	0 e I	0,81	0,78	0,79	0,65
	II	0,73	0,69	0,71	
	III	0,54	0,51	0,53	
	IV	0,50	0,50	0,50	
	X e Y	0,64	0,75	0,69	
Random forest	0 e I	0,72	0,74	0,73	0,61
	II	0,67	0,67	0,67	
	III	0,47	0,46	0,46	
	IV	0,51	0,46	0,48	
	X e Y	0,66	0,70	0,68	
Support Vector Machine	0 e I	0,80	0,78	0,79	0,63
	II	0,70	0,66	0,68	
	III	0,52	0,47	0,49	
	IV	0,48	0,50	0,49	
	X e Y	0,64	0,73	0,68	
Naive Bayes	0 e I	1,00	0,43	0,60	0,55
	II	0,85	0,56	0,67	
	III	0,45	0,32	0,37	
	IV	0,39	0,58	0,47	
	X e Y	0,47	0,85	0,60	
Gradient boosting	0 e I	**0,76**	**0,78**	**0,77**	**0,68**
	II	**0,74**	**0,68**	**0,71**	
	III	**0,56**	**0,57**	**0,56**	
	IV	**0,58**	**0,59**	**0,59**	
	X e Y	**0,75**	**0,77**	**0,76**	

4.2.6 Considerações finais

De modo geral, os *pipelines* de análise desenvolvidos neste trabalho se mostraram promissores quanto a capacidade de auxiliar na extração de informações, principalmente, dos textos de laudos de AP.

Os modelos treinados para as variáveis malignidade, topografia e morfologia apresentaram resultados satisfatórios. As acurácias foram de 92%, 96% e 88%, respectivamente, com *f1-score* e *recall* também altos. A performance seria ainda melhor se corrigidos os erros de rotulagem.

O SVM e os dois métodos de *ensembles* (*random forest* e *gradient boosting*) foram os algoritmos de AM que mais se destacaram. Por outro lado, o Naive Bayes apresentou as piores performances em todos os cenários.

O modelo de estadiamento, mesmo apresentando métricas de performance mais baixas, também se mostrou útil por evidenciar a ausência de uma informação muito importante nas narrativas clínica: o estadio da doença.

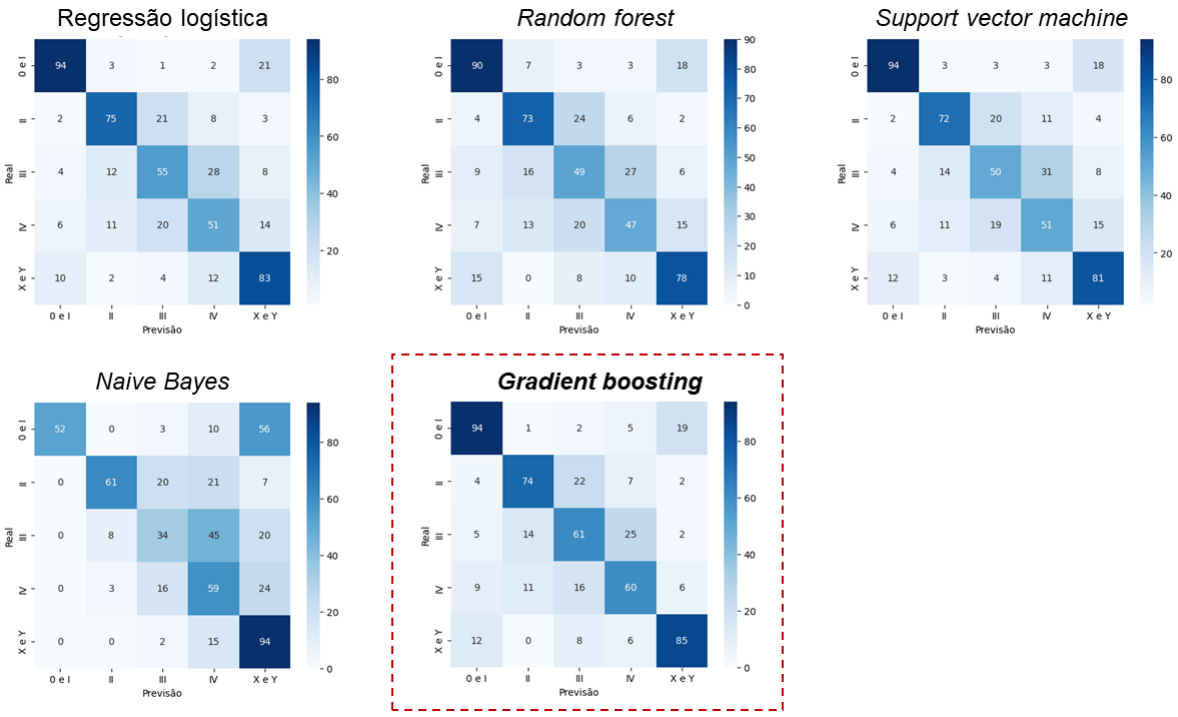

Figura 9 – Matrizes de confusão da predição dos modelos treinados para estadio

5 CONCLUSÃO

O presente trabalho desenvolveu um *pipeline* (pré-processamento, representação textual e aplicação de modelos de aprendizado de máquina) com potencial de executar as tarefas de extração de variáveis, hoje, coletadas manualmente. Os melhores modelos extraíram informações de malignidade, topografia e morfologia a partir de textos de laudos com acurácias superiores a 92%, 96% e 88%, respetivamente.

Os modelos testados não foram capazes de extrair o estadio da doença das narrativas clínicas. Apesar da relevância do estadiamento para a tomada de decisão terapêutica, a informação não está presente em grande parte das notas clínicas. Isso quer dizer que esses modelos podem ser usados para apontar melhorias na qualidade das informações textuais contidas nos registros eletrônicos de saúde.

Como trabalhos futuros, pretende-se desenvolver uma interface gráfica e disponibilizar esses modelos para hospitais que, assim como o Hospital de Amor, oferecem tratamentos oncológicos 100% gratuito, otimizando o trabalho de centenas de registradores de hospitais públicos e filantrópicos desse país. Com informação de qualidade, os serviços de oncologia podem tomar decisões mais assertivas, melhorando a assistência, e os gestores públicos podem planejar melhores políticas para controle do câncer a nível populacional.

REFERÊNCIAS

ARONSON, A. R.; LANG, F.-M. An overview of MetaMap: historical perspective and recent advances. **Journal of the American Medical Informatics Association: JAMIA**, v. 17, n. 3, p. 229–236, 2010. ISSN 1527-974X.

BAND, S. S. *et al.* Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. **Informatics in Medicine Unlocked**, v. 40, p. 101286, jan. 2023. ISSN 2352-9148. Disponível em: https://www.sciencedirect.com/science/article/pii/S2352914823001302.

BERGQUIST, S. L. *et al.* Classifying Lung Cancer Severity with Ensemble Machine Learning in Health Care Claims Data. **Proceedings of Machine Learning Research**, v. 68, p. 25–38, ago. 2017. ISSN 2640-3498.

BERKSON, J. Application of the Logistic Function to Bio-Assay. **Journal of the American Statistical Association**, [American Statistical Association, Taylor & Francis, Ltd.], v. 39, n. 227, p. 357–365, 1944. ISSN 0162-1459. Disponível em: https://www.jstor.org/stable/2280041.

BRASIL. **PORTARIA SAES/MS Nº 688**. 2023. Disponível em: https://bvsms.saude.gov.br/bvs/saudelegis/saes/2023/prt0688_30_08_2023.html.

BRAY, F.; PARKIN, D. M. Evaluation of data quality in the cancer registry: principles and methods. Part I: comparability, validity and timeliness. **European Journal of Cancer (Oxford, England: 1990)**, v. 45, n. 5, p. 747–755, mar. 2009. ISSN 1879-0852.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32, out. 2001. ISSN 1573-0565. Disponível em: https://doi.org/10.1023/A:1010933404324.

BUSTOS, A.; PERTUSA, A. Learning Eligibility in Cancer Clinical Trials Using Deep Neural Networks. **Applied Sciences**, Multidisciplinary Digital Publishing Institute, v. 8, n. 7, p. 1206, jul. 2018. ISSN 2076-3417. Disponível em: https://www.mdpi.com/2076-3417/8/7/1206.

CASELI, H. M.; NUNES, M. G. V. (ed.). **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. 2. ed. BPLN, 2024. ISBN 978-65-00-95750-1. Disponível em: https://brasileiraspln.com/livro-pln/2a-edicao/.

CFM. **RESOLUÇÃO Nº 1.638**. 2002. Disponível em: https://pesquisa.in.gov.br/imprensa/jsp/visualiza/index.jsp?jornal=1&pagina=184&data=09/08/2002.

COLICCHIO, T. K. **Introdução à informática em saúde: Fundamentos, aplicações e lições aprendidas com a informatização do sistema de saúde americano**. 1ª edição. ed. [*S.l.: s.n.*]: Artmed, 2020.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, n. 3, p. 273–297, set. 1995. ISSN 1573-0565. Disponível em: https://doi.org/10.1007/BF00994018.

DALIANIS, H. **Clinical Text Mining: Secondary Use of Electronic Patient Records**. 1st ed. 2018 edição. ed. [*S.l.: s.n.*]: Springer, 2018.

DEVLIN, J. *et al.* **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. [*S.l.*], 2019. ArXiv:1810.04805 [cs] type: article. Disponível em: http://arxiv.org/abs/1810.04805.

FRIEDMAN, C. A broad-coverage natural language processing system. **Proceedings of the AMIA Symposium**, p. 270–274, 2000. ISSN 1531-605X. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243979/.

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. **The Annals of Statistics**, v. 29, n. 5, out. 2001. ISSN 0090-5364. Disponível em: https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-5/Greedy-function-approximation-A-gradient-boosting-machine/10.1214/aos/1013203451.full.

GAO, S. *et al.* Hierarchical attention networks for information extraction from cancer pathology reports. **Journal of the American Medical Informatics Association**, v. 25, n. 3, p. 321–330, mar. 2018. ISSN 1527-974X. Disponível em: https://doi.org/10.1093/jamia/ocx131.

GARLA, V. *et al.* The Yale cTAKES extensions for document classification: architecture and application. **Journal of the American Medical Informatics Association: JAMIA**, v. 18, n. 5, p. 614–620, 2011. ISSN 1527-974X.

GHOLIPOUR, M. *et al.* Extracting cancer concepts from clinical notes using natural language processing: a systematic review. **BMC Bioinformatics**, v. 24, n. 1, p. 405, 2023. ISSN 1471-2105. Disponível em: https://doi.org/10.1186/s12859-023-05480-0.

GLOBOCAN. **Cancer Today**. 2022. Disponível em: https://gco.iarc.who.int/today/.

HOCHHEISER, H. *et al.* DeepPhe-CR: Natural Language Processing Software Services for Cancer Registrar Case Abstraction. **medRxiv: The Preprint Server for Health Sciences**, p. 2023.05.05.23289524, out. 2023.

HONG, J. C. *et al.* Natural language processing for abstraction of cancer treatment toxicities: accuracy versus human experts. **JAMIA Open**, v. 3, n. 4, p. 513–517, dez. 2020. ISSN 2574-2531. Disponível em: https://doi.org/10.1093/jamiaopen/ooaa064.

INCA. **Notas técnicas - Integrador RHC**. 2012. Disponível em: https://irhc.inca.gov.br/files/Notas_tecnicas_final.pdf.

INCA. **Estimativa 2023: Incidência de Câncer no Brasil**. [*S.l.: s.n.*], 2022.

JACKSON, R. *et al.* OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. **Database**, v. 2021, p. baab069, set. 2021. ISSN 1758-0463. Disponível em: https://doi.org/10.1093/database/baab069.

KREIMEYER, K. *et al.* Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. **Journal of Biomedical Informatics**, v. 73, p. 14–29, set. 2017. ISSN 1532-0464. Disponível em: https://www.sciencedirect.com/science/article/pii/S1532046417301685.

LIN, C. *et al.* Multilayered temporal modeling for the clinical domain. **Journal of the American Medical Informatics Association**, v. 23, n. 2, p. 387–395, mar. 2016. ISSN 1067-5027. Disponível em: https://doi.org/10.1093/jamia/ocv113.

LIN, C. *et al.* A BERT-based Universal Model for Both Within- and Cross-sentence Clinical Temporal Relation Extraction. *In*: RUMSHISKY, A. *et al.* (ed.). **Proceedings of the 2nd Clinical Natural Language Processing Workshop**. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 2019. p. 65–71. Disponível em: https://aclanthology.org/W19-1908.

LUNDBERG, S.; LEE, S.-I. **A Unified Approach to Interpreting Model Predictions**. [*S.l.*], 2017. ArXiv:1705.07874 [cs, stat] type: article. Disponível em: http://arxiv.org/abs/1705.07874.

MERRIMAN, K. W. *et al.* Evolution of the Cancer Registrar in the Era of Informatics. **JCO Clinical Cancer Informatics**, Wolters Kluwer, n. 5, p. 272–278, dez. 2021. Disponível em: https://ascopubs.org/doi/10.1200/CCI.20.00123.

NCI. **What Is Cancer? - NCI**. 2007. Disponível em: https://www.cancer.gov/about-cancer/understanding/what-is-cancer.

NCI. **Definition of electronic health record - NCI Dictionary of Cancer Terms - NCI**. 2011. Disponível em: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/electronic-health-record.

NELSON, R.; STAGGERS, N. **Health Informatics: An Interprofessional Approach**. 2º edição. ed. St. Louis, Missouri: Mosby, 2017. ISBN 9780323402316.

OMS. **Classificação Internacional de Doenças para Oncologia**. 3. ed. [*S.l.: s.n.*], 2005.

PATEL, S. **NLP Pipeline: Building an NLP Pipeline, Step-by-Step**. 2020. Disponível em: https://suneelpatel18.medium.com/nlp-pipeline-building-an-nlp-pipeline-step-by-step-7f0576e11d08.

PEDREGOSA, F. *et al.* Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

QIU, J. X. *et al.* Deep Learning for Automated Extraction of Primary Sites From Cancer Pathology Reports. **IEEE Journal of Biomedical and Health Informatics**, v. 22, n. 1, p. 244–251, 2018. ISSN 2168-2208. Disponível em: https://ieeexplore.ieee.org/document/7918552/authors#authors.

ROSENBLOOM, S. T. *et al.* Data from clinical notes: a perspective on the tension between structure and flexible documentation. **Journal of the American Medical Informatics Association: JAMIA**, v. 18, n. 2, p. 181–186, 2011. ISSN 1527-974X.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. **Information Processing & Management**, v. 24, n. 5, p. 513–523, jan. 1988. ISSN 0306-4573. Disponível em: https://www.sciencedirect.com/science/article/pii/0306457388900210.

SAVOVA, G. K. *et al.* Use of Natural Language Processing to Extract Clinical Cancer Phenotypes from Electronic Medical Records. **Cancer Research**, v. 79, n. 21, p. 5463–5470, nov. 2019. ISSN 0008-5472. Disponível em: https://doi.org/10.1158/0008-5472.CAN-19-0579.

SAVOVA, G. K. *et al.* Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. **Journal of the American Medical Informatics Association**, v. 17, n. 5, p. 507–513, set. 2010. ISSN 1067-5027. Disponível em: https://doi.org/10.1136/jamia.2009.001560.

SAVOVA, G. K. *et al.* DeepPhe: A Natural Language Processing System for Extracting Cancer Phenotypes from Clinical Records. **Cancer Research**, v. 77, n. 21, p. e115–e118, out. 2017. ISSN 0008-5472. Disponível em: https://doi.org/10.1158/0008-5472.CAN-17-0615.

SCHNEIDER, E. T. R. *et al.* BioBERTpt - A Portuguese Neural Language Model for Clinical Named Entity Recognition. *In*: RUMSHISKY, A. *et al.* (ed.). **Proceedings of the 3rd Clinical Natural Language Processing Workshop**. Online: Association for Computational Linguistics, 2020. p. 65–72. Disponível em: https://aclanthology.org/2020.clinicalnlp-1.7.

SHIVADE, C. *et al.* Automatic data source identification for clinical trial eligibility criteria resolution. **AMIA ... Annual Symposium proceedings. AMIA Symposium**, v. 2016, p. 1149–1158, 2016. ISSN 1942-597X.

SOYSAL, E. *et al.* CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines. **Journal of the American Medical Informatics Association**, v. 25, n. 3, p. 331–336, mar. 2018. ISSN 1527-974X. Disponível em: https://doi.org/10.1093/jamia/ocx132.

STRöTGEN, J.; GERTZ, M. Multilingual and cross-domain temporal tagging. **Language Resources and Evaluation**, v. 47, n. 2, p. 269–298, jun. 2013. ISSN 1574-0218. Disponível em: https://doi.org/10.1007/s10579-012-9179-y.

THOMPSON, M.; DUDA, R. O.; HART, P. E. Pattern Classification and Scene Analysis. *In*: **Leonardo**. [*S.l.: s.n.*], 1974. v. 7, n. 4, p. 370. ISSN 0024094X. Disponível em: https://www.jstor.org/stable/1573081?origin=crossref.

TSEYTLIN, E. *et al.* NOBLE – Flexible concept recognition for large-scale biomedical natural language processing. **BMC Bioinformatics**, v. 17, n. 1, p. 32, jan. 2016. ISSN 1471-2105. Disponível em: https://doi.org/10.1186/s12859-015-0871-y.

UICC. **TNM Classificação de tumores malignos**. 8. ed. [*S.l.: s.n.*], 2023.

YIM, W.-w. *et al.* Natural Language Processing in Oncology: A Review. **JAMA Oncology**, v. 2, n. 6, p. 797–804, 2016. ISSN 2374-2437. Disponível em: https://doi.org/10.1001/jamaoncol.2016.0213.

ZHANG, K.; DEMNER-FUSHMAN, D. Automated classification of eligibility criteria in clinical trials to facilitate patient-trial matching for specific patient populations. **Journal of the American Medical Informatics Association: JAMIA**, v. 24, n. 4, p. 781–787, jul. 2017. ISSN 1527-974X.

ZHOU, S. *et al.* CancerBERT: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. **Journal of the American Medical Informatics Association : JAMIA**, v. 29, n. 7, p. 1208–1216, mar. 2022. ISSN 1067-5027. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9196678/.

APÊNDICES

APÊNDICE A – MODELOS USANDO O BIOBERTPT

Tabela 5 – Métricas de performances dos modelos treinados para malignidade usando o BioBERTpt

Modelo	Tumor	Precisão	Recall	F1-score	Acurácia
Regressão logística	benigno	0,85	0,85	0,85	0,84
	maligno	0,83	0,83	0,83	
Random forest	benigno	0,80	0,81	0,80	0,79
	maligno	0,78	0,77	0,77	
SVM	benigno	0,82	0,79	0,81	0,80
	maligno	0,78	0,81	0,79	
Gradient boosting	benigno	**0,88**	**0,83**	**0,85**	**0,85**
	maligno	**0,82**	**0,87**	**0,85**	

Tabela 6 – Métricas de performances dos modelos treinados para topografias usando o BioBERTpt

Modelo	Topografias	Precisão	Recall	F1-score	Acurácia
Regressão logística	outros	**0,83**	**0,89**	**0,86**	**0,92**
	mama	**0,98**	**0,98**	**0,98**	
	pele	**0,93**	**0,94**	**0,94**	
	pulmão	**0,93**	**0,86**	**0,89**	
Random forest	outros	0,71	0,83	0,76	0,85
	mama	0,93	0,94	0,93	
	pele	0,95	0,90	0,92	
	pulmão	0,87	0,77	0,82	
SVM	outros	0,74	0,91	0,82	0,88
	mama	0,99	0,96	0,97	
	pele	0,96	0,91	0,94	
	pulmão	0,91	0,77	0,83	
Gradient boosting	outros	0,75	0,85	0,79	0,87
	mama	0,95	0,91	0,92	
	pele	0,93	0,94	0,94	
	pulmão	0,88	0,80	0,84	

Tabela 7 – Métricas de performances dos modelos treinados para morfologia usando o BioBERTpt

Modelo	**Morfologias**	**Precisão**	**Recall**	**F1-score**	**Acurácia**
Regressão logística	outras	**0,58**	**0,61**	**0,60**	**0,59**
	80903	**0,8**	**0,60**	**0,69**	
	85003	**0,83**	**0,79**	**0,81**	
	80703	**0,24**	**0,29**	**0,26**	
	81403	**0,57**	**0,71**	**0,63**	
	80702	**0,62**	**0,54**	**0,58**	
Random forest	outras	0,52	0,48	0,5	0,54
	80903	0,69	0,45	0,55	
	85003	0,73	0,84	0,78	
	80703	0,25	0,35	0,29	
	81403	0,39	0,53	0,45	
	80702	0,71	0,50	0,59	
SVM	outras	0,62	0,57	0,59	0,42
	80903	0,33	0,75	0,46	
	85003	0,79	0,58	0,67	
	80703	0,15	0,18	0,16	
	81403	0,38	0,35	0,36	
	80702	0,75	0,12	0,21	
Gradient boosting	outras	0,50	0,52	0,51	0,48
	80903	0,65	0,65	0,65	
	85003	0,90	0,47	0,62	
	80703	0,23	0,29	0,26	
	81403	0,40	0,59	0,48	
	80702	0,53	0,42	0,47	

Tabela 8 – Métricas de performances dos modelos treinados para estadiamento usando o BioBERTpt

Modelo	**Estadios**	**Precisão**	**Recall**	**F1-score**	**Acurácia**
Regressão logística	0 e I	**0,63**	**0,43**	**0,51**	0,43
	II	**0,46**	**0,79**	**0,58**	
	III	**0,19**	**0,40**	**0,26**	
	IV	**0,53**	**0,33**	**0,41**	
	X e Y	**0,38**	**0,33**	**0,36**	
Random forest	0 e I	0,50	0,29	0,36	0,39
	II	0,45	0,71	0,56	
	III	0,20	0,50	0,29	
	IV	0,54	0,29	0,38	
	X e Y	0,38	0,38	0,38	
SVM	0 e I	0,50	0,07	0,12	0,18
	II	0,44	0,29	0,35	
	III	0,13	1,00	0,22	
	IV	0,00	0,00	0,00	
	X e Y	0,25	0,08	0,12	
Gradient boosting	0 e I	0,50	0,29	0,36	0,35
	II	0,46	0,79	0,58	
	III	0,10	0,30	0,15	
	IV	0,50	0,29	0,37	
	X e Y	0,35	0,25	0,29	

ANEXOS

ANEXO A – DESCRIÇÃO DE CÓDIGOS DO CID-O

Tabela 9 – Códigos e descrições de topografias

Código	Descrição da topografia
C15	Esôfago
C16	Estomago
C18	Cólon
C20	Reto
C22	Fígado e vias biliares intrahepáticas
C34	Brônquios e pulmões
C41	Ossos e cartil. Articulares de outras localizações
C42	Sistema hematopoiético e reticuloendotelial
C44	Pele
C49	Tecido conjuntivo, subcutâneo e outros tecidos moles
C50	Mama
C53	Colo do útero
C55	Útero, SOE
C61	Próstata
C64	Rim
C67	Bexiga
C71	Encéfalo
C73	Glândula tiróide

Tabela 10 – Códigos e descrições de morfologias

Código	Descrição de morfologia
80703	Carcinoma escamocelular, SOE
80903	Carcinoma basocelular, SOE
81403	Adenocarcinoma, SOE
85503	Carcinoma de células acinosas
80702	Carcinoma escamocelular in situ, SOE
82113	Adenocarcinoma tubula
80003	Neoplasia maligna
87203	Melanoma maligno, SOE
82603	Adenocarcinoma papilar, SOE
80973	Carcinoma basocelular nodular
85002	Carcinoma intraductal não infiltrante, SOE
80503	Carcinoma papilar, SOE
84903	Carcinoma de células em anel de sinete
96803	Linfoma maligno de células grandes b, difuso, SOE
80103	Carcinoma, SOE
81203	Carcinoma de células transicionais, SOE
83103	Adenocarcinoma de células claras, SOE
81303	Carcinoma papilar de células transicionais
84803	Adenocarcinoma mucinoso