MIGUEL DÍAZ ITURRY

AVALIAÇÃO DA QUALIDADE DE DADOS DE LAUDOS DE PROCEDIMENTOS PARA A CODIFICAÇÃO AUTOMÁTICA DE DIAGNÓSTICOS SECUNDÁRIOS

São Paulo
2022

MIGUEL DÍAZ ITURRY

AVALIAÇÃO DA QUALIDADE DE DADOS DE LAUDOS DE PROCEDIMENTOS PARA A CODIFICAÇÃO AUTOMÁTICA DE DIAGNÓSTICOS SECUNDÁRIOS

Dissertação apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do título de Mestre em Ciências.

São Paulo
2022

MIGUEL DÍAZ ITURRY

AVALIAÇÃO DA QUALIDADE DE DADOS DE LAUDOS DE PROCEDIMENTOS PARA A CODIFICAÇÃO AUTOMÁTICA DE DIAGNÓSTICOS SECUNDÁRIOS

Versão Corrigida

Dissertação apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do título de Mestre em Ciências.

Área de Concentração:

Engenharia de Computação

Orientador:

Prfa. Dra. Solange Nice Alves de Souza

São Paulo
2022

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, __2___ de ____Agosto_____ de ___2022___

Assinatura do autor: _____

Assinatura do orientador: _____

Catalogação-na-publicação

Diaz, Miguel
Avaliação da qualidade de dados de laudos de procedimentos para a codificação automática de diagnósticos secundários / M. Diaz -- versão corr. -- São Paulo, 2022.
72 p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais.

1.Mineração de dados I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Computação e Sistemas Digitais II.t.

AGRADECIMENTOS

A minha esposa pela sua paciência, ajuda e companhia neste projeto e na vida. Aos meus pais, irmãos e irmãs pelo apoio moral. Aos amigos e amigas que me brindaram conselhos e bons momentos no transcurso do curso.

À Profa. Solange pela oportunidade profissional e a orientação no desenvolvimento da pesquisa. À Profa. Marcia e a Dra. Suzana pela colaboração e ajuda. Ao CNPq pela bolsa de estudos.

RESUMO

O pagamento referento ao custo de um paciente se baseia no consumo dos recursos utilizados nos procedimentos realizados e itens consumidos pelo paciente, bem como no código principal e secundário da doença diagnosticada para o paciente. Atualmente a classificação da doença é feita utilizando a Classificação Internacional de Doenças (CID). Nesse processo, a identificação dos códigos de diagnósticos secundários é realizada por especialistas que revisam os diagnósticos lançados pelo médico no resumo de alta, no prontuário do paciente e nos resultados presentes nos laudos de procedimentos. Tal processo, em geral, é manual e, devido à grande quantidade de laudos, é uma tarefa cansativa e suscetível a erros. Adicionalmente a essa dificuldade, pesquisas prévias identificaram problemas na Qualidade dos Dados (QD), o que prejudica na identificação dos códigos de doenças. O presente trabalho avalia e trata a qualidade dos dados para a aplicação de algoritmos de Aprendizagem de Máquina (AM) para a codificação automática de doenças relativas ao capítulo de neoplasias da CID, as quais são identificadas nos laudos de exames anatomopatológicos. As dimensões de QD identificadas pelos problemas encontrados nos laudos foram a *acessibilidade* e a *acurácia* dos textos. A outra contribuição do trabalho é a proposta de um modelo hierárquico para melhoria do desempenho do modelo de Aprendizagem de Máquina (AM) aplicado para a codificação automática dos diagnósticos secundários. Os resultados mostraram que o modelo hierárquico incrementa em 15% o desempenho do modelo clássico, comprovando que explorar a organização do padrão CID para a criação de modelos é uma vantagem na codificação automática. Adicionalmente, demostrou-se que o tratamento dos textos e o balanceamento das classes melhoram o desempenho dos modelos de codificação e tornam-os mais robustos à distribuição dos códigos.

Palavras-Chave – laudo médico, qualidade de dados, aprendizagem de máquinas, Classificação Internacional de Doenças (CID), codificação automática.

ABSTRACT

The reimbursement concerning a patient's cost is based in the medical resources employed, as well as in the principal and secondary disease code. Currently the classification of the diseases is based on the International Classification of Diseases (ICD). In this process, the identification of secondary disease codes is done by specialists that review the diagnoses written by the doctor in the patient's medical records and laboratory results. Such process is done manually and, because of the large quantity of medical records, it is a tiring task and susceptible to errors. Additionally to this difficulty, previous researches identified problems in the Data Quality (DQ), which jeopardize the identification of disease codes. In the present research, the DQ is assessed and treated to develop a Machine Learning (ML) model for automatic neoplasm disease coding found in the anatomopathological reports. The DQ dimenssions identified in the problems are text *accessibility* and *accuracy*. A contribution in the present project is a hierarchical model to improve the performance of the ML model for disease coding. The results showed that the hierarchical model outperforms by a 15% the performance of a classic model, proving that employing the organization of the ICD standard is a leverage in the automatic coding. Additionally, it was demostrated that text treatment and class balancing improve the performance of the models and make them more robust to the dataset class distribution.

Keywords – medical record, data quality, machine learning, International Classification of Diseases (ICD), automatic coding.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O All Patients Refined Diagnosis Related Groups (APR-DRG) é um sistema de agrupamento que utiliza dados demográficos e informações clínicas de alta de pacientes internados em hospitais, com o objetivo de mensurar o consumo de recursos hospitalares e a complexidade assistencial. Esse sistema é muito utilizado no exterior, porém é recente no Brasil (OSMO, 2017). Alguns hospitais privados, como o Hospital Sírio-Libanês, Hospital Israelita Albert Einstein e Hospital do Coração (HCor), estão testando esse modelo com o objetivo de ajustar as análises de suas populações por perfil de gravidade e risco de óbito e, dessa forma, melhorar as negociações com as fontes pagadoras.

Atualmente, todo o pagamento se baseia no consumo dos recursos utilizados nos procedimentos realizados e itens consumidos, bem como no código principal da doença, já que no país não existe um modelo padronizado para definição da complexidade dos tratamentos dos pacientes. Existem tabelas de referências que precificam esses procedimentos e são utilizadas pelos hospitais e operadoras para embasar modelos de pagamento, porém não há ajustes por complexidade. O APR-DRG busca mitigar tais problemas, uma vez que esse sistema utiliza o conjunto de dados do paciente (idade, peso ao nascimento se criança, tempo em ventilação mecânica, gênero, tempo de permanência no hospital, etc.) atrelado ao conjunto de códigos de diagnósticos (principal e secundários) e ao conjunto de procedimentos para definir a complexidade do atendimento. Os pacientes são classificados em um grupo clínico de maior relevância, subclassificados em um dos quatro níveis de severidade da doença e em uma das quatro classes de risco de mortalidade (OSMO, 2017). Adicionalmente, o APR-DRG permite a avaliação do desempenho e facilita a gestão das instituições, departamentos e corpo médico, pois ajusta indicadores de resultados de acordo a complexidade do tratamento dos pacientes (GARTNER et al., 2015; OSMO, 2017).

Os códigos dos diagnósticos (principal e secundários) empregados pelo APR-DRG são baseados na Classificação Internacional de Doenças (CID). A CID é o padrão de classificação de diagnósticos adotado e mantido pela Organização Mundial da Saúde (OMS), que define o universo de doenças, distúrbios, lesões e outras condições de saúde relacionadas, listadas de forma abrangente e hierárquica (GARTNER et al., 2015; OSMO, 2017).

Isso permite fácil armazenamento, recuperação e análise de informações para tomada de decisões, compartilhamento e comparação de informações de saúde entre hospitais, regiões e países[1].

A codificação das doenças deve ser feita com cuidado, uma vez que, quando mal documentados, causam erros na classificação de pacientes (GARTNER et al., 2015). A codificação de doenças secundárias, em geral, é manual e baseada nos registros clínicos e em informações dos laudos de procedimentos, o que leva a um grande esforço e requer tempo para sua execução, além de ser propensa a erros (AZAM et al., 2020; XIE; XING, 2018; XU et al., 2018). Adicionalmente, foram identificados problemas na qualidade dos dados nos registros clínicos usados para a codificação. Veras e Martins (1994) afirmam que existe baixa confiabilidade nos diagnósticos secundários devido a dificuldades relacionadas a ausência ou ambiguidade de algumas informações em notas médicas. Roos, Sharp e Wajda (1989) demostraram que existem problemas na concordância entre dados dos paciente devido às várias fontes de dados que geralmente são encontradas nos hospitais. De Coster et al. (2006) observou problemas nos dados administrativos por falta de detalhe clínico.

Abordou-se o problema da automação da codificação de doenças secundárias com o emprego de distintos algoritmos de Aprendizagem de Máquina (AM) e mineração de textos (AZAM et al., 2020; LAURíA; MARCH, 2011; XIE; XING, 2018; XU et al., 2018; ZHONG; GAO; YI, 2018). No entanto, são poucos os que estudam o impacto da má qualidade dos dados nas predições dos algoritmos. Lauría e March (2011) consideram o problema, mas avaliam somente a dimensão acurácia dos dados, não havendo relatos na literatura que abordem o tratamento de outras dimensões da qualidade de dados.

O tratamento da qualidade dos dados clínicos e das informações dos laudos de procedimentos são tarefas importantes para a codificação das doenças secundárias, tendo grande impacto na classificação realizada pelo APR-DRG, refletindo no gerenciamento financeiro e negociação do pagamento junto às operadoras.

Dado o contexto, este trabalho busca melhorar a qualidade de dados clínicos utilizados para a codificação de doenças, bem como automatizá-la, auxiliando na implantação do APR-DRG. Como base de dados serão utilizados dados do departamento oncológico do HCor, porém os resultados podem servir como base para realizar a codificação automática de diferentes capítulos da CID e em diferentes hospitais.

[1] https://www.who.int/classifications/icd

1.1 Objetivos

O objetivo principal da pesquisa é realizar a codificação automática de doenças, utilizando como base de dados os laudos de procedimentos anatomopatologicos.

Como objetivos específicos se tem:

- Avaliar a qualidade dos textos dos laudos para melhorar o desempenho do modelos de classificação.
- Realizar o balanceamento das classes para o treinamento dos modelos.

Como resultado, espera-se obter um modelo que codifique automaticamente as doenças, evitando erros humanos como interpretações errôneas de laudos, erros de digitação e identificação equivocada do código, muitas vezes ocasionados pelo grande fluxo de pacientes. Outras vantagens seriam a redução do tempo de obtenção dos códigos secundários de doenças e economia de recurso humano das instituições.

1.2 Metodologia

O desenvolvimento do projeto foi dividido nas seguintes etapas:

1. **Escrita de projeto para o Comitê de Ética em Pesquisa (CEP).** Para o desenvolvimento do modelo de codificação automática e a sua validação, foram empregados dados de pacientes internados no HCor. Por serem dados confidenciais, o projeto de pesquisa foi submetido à Plataforma Brasil para a sua avaliação e autorização do CEP. A pesquisa com o uso dos dados foram iniciados somente após o recebimento da autorização pelo CEP com CAAE 28400820.3.0000.0060.
2. **Aquisição de dados.** Os dados fornecidos pelo hospital HCor correspondem a pacientes oncológicos que receberam alta hospitalar a partir de janeiro de 2019. Esses dados foram extraídos do Sistema de Informação do Hospital -Hospital Information System (HIS)- pelo departamento de epidemiologia e são constituídos pela tabela de internações de pacientes, os laudos de anatomopatologia e a tabela do APR-DRG, a qual contem os códigos das doenças.
3. **Levantamento bibliográfico.** Desenvolveu-se duas revisões da literatura seguindo as diretrizes sugeridas em Kitchenham (2004). A primeira delas teve como objetivo

identificar as dimensões de QD empregadas para a avaliação de registros médicos e as atividades elaboradas para mitigar os problemas encontrados. O foco da segunda revisão foi levantar os modelos de AM mais comuns na codificação automática de diagnósticos e as métricas para avaliar seu desempenho.

4. **Preparação e ajustes dos conjuntos de dados.** As atividades realizadas foram: (i) anonimização de dados, retirando informações pessoais dos pacientes, (ii) união dos laudos de anatomopatologia com seus respectivos códigos da doença, (iii) exclusão de registros duplicados e com códigos inexistentes na CID, (iv) balanceamento de classes, gerando laudos simulados em classes menos frequentes e removendo laudos das classes mais frequentes.

5. **Melhora da qualidade dos textos de laudos médicos.** Aplicação de uma estratégia *data-driven* para correção de erros ortográficos.

6. **Tratamento dos textos de laudos médicos.** Adequação do texto para o modelo de classificação com atividades comumente empregados na literatura de AM.

7. **Seleção e desenho da arquitetura do modelo.** Com base nos resultados da revisão da literatura sobre modelos de AM para codificação automática e análises dos dados, definiu-se o algoritmo e a arquitetura do modelo a ser treinado.

8. **Ajuste dos hiper-parâmetros do modelo.** Seleção dos valores ótimos para hiper-parâmetros do modelo, validando-os com múltiplas amostras dos dados e cálculo de métricas de desempenho.

9. **Validação dos resultados.** O desempenho do modelo foi avaliado utilizando as métricas levantadas na revisão da literatura, aliado ao emprego de experimentos que permitiram contrastar o efeito do tratamento de texto no treinamento.

10. **Escrita e submissão de artigos.** Com os resultados da revisão da literatura sobre a avaliação e tratamento da QD em registros médicos, publicou-se o artigo Iturry et al. (2021), apresentado na 16ª Conferência Ibérica de Sistemas e Tecnologias de Informação (CISTI). Adicionalmente, com os resultados finais da codificação automática e tratamento da QD, foi aceito o artigo a ser apresentado no Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS).

11. **Escrita da qualificação e do texto final da dissertação.** Como parte do Programa de Pós-Graduação em Engenharia Elétrica (PPGEE), apresentaram-se

os resultados parciais da pesquisa em um exame de qualificação em novembro de 2020 e elaborou-se o texto com os resultados finais para a dissertação final do curso.

2 CONTEXTUALIZAÇÃO TEÓRICA

Neste capítulo são apresentados os principais conceitos utilizados para o desenvolvimento deste trabalho. Assim, discorre-se sobre AM, algoritmos de AM mais usados para a classificação de textos, medidas de desempenho desses algoritmos, conceitos de mineração de textos, algoritmos para o pré-processamento de textos, conceitos de QD e dimensões de QD.

2.1 Aprendizagem de Máquina (AM)

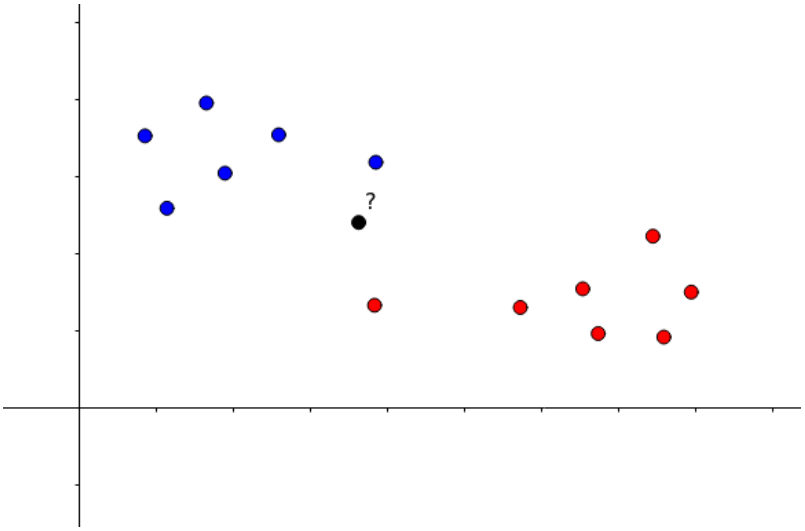
AM é o campo da Inteligencia Artificial (IA) que se destina em construir programas computacionais que *aprendam* pela experiência (MITCHELL, 1997). Os algoritmos usados em AM são agrupados em aprendizagem supervisionada e aprendizagem não supervisionada (EISENSTEIN, 2018; GOODFELLOW; BENGIO; COURVILLE, 2016; HASTIE; TIBSHIRANI; FRIEDMAN, 2009; JAMES et al., 2013).

A aprendizagem supervisionada consiste em, dado um conjunto de dados de entrada (X) com saída conhecida (Y), encontrar uma função f de tal modo que $Y \simeq f(X)$. Os dados de entrada podem ser chamados também de preditores e são representados por uma matriz n x p (GOODFELLOW; BENGIO; COURVILLE, 2016; HASTIE; TIBSHIRANI; FRIEDMAN, 2009; JAMES et al., 2013).

O processo de ajuste da função f é chamado de *treinamento*, no qual dividem-se os dados em dois conjuntos: treinamento e teste. Ajustam-se os parâmetros da função empregando o primeiro conjunto (treinamento) e aplicando algoritmos que procuram minimizar o erro da predição ou optimizar alguma função objetivo. O conjunto de teste é empregado no modelo resultante para avaliar seu desempenho, aplicando métricas que serão apresentadas na Seção 2.1.4.

Um exemplo de aprendizagem supervisionada pode ser observado na Figura 1. Nessa, o modelo é treinado usando as informações das coordenadas dos pontos azuis e vermelhos para predizer a cor de qualquer outro ponto no mesmo plano.

Figura 1: Exemplo aprendizagem supervisionada.

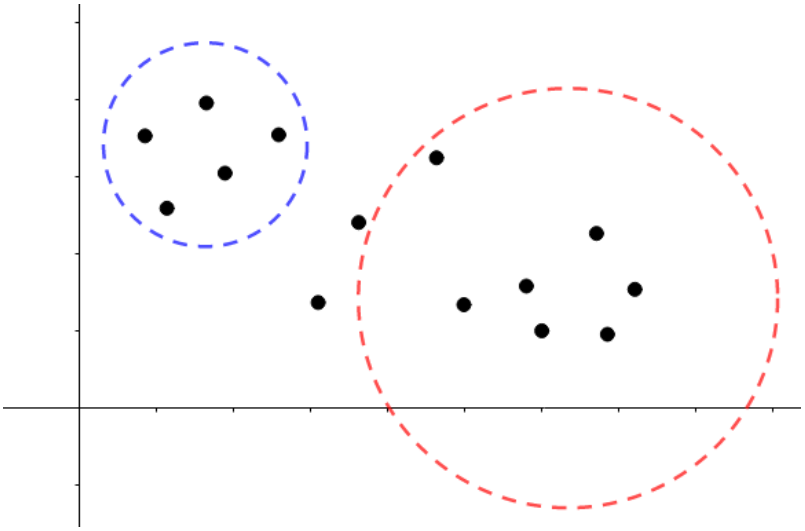


A tarefa de aprendizagem supervisionada é denominada *classificação* para saídas Y categóricas, conhecidas como rótulos. Já para as saídas numéricas, denominadas valores, tem-se a tarefa chamada *regressão*. O exemplo dos pontos azuis e vermelhos corresponde a uma tarefa de classificação.

Dentre os algoritmos que são usados nas tarefas de classificação, podem-se citar: K Vizinhos Mais Próximos (*K-Nearest Neighbours*), *Naive Bayes*, Árvores de Decisão (*Decision Trees*), Support Vector Machine (SVM)) e Redes Neurais (AGGARWAL, 2018b; JAMES et al., 2013). Para tarefas de regressão são usados os modelos de Regressão Linear, Lasso Regression, Ridge Regression, Regressão Polinomial e Redes Neurais (AGGARWAL, 2018b; JAMES et al., 2013). Dependendo dos dados de entrada e dos recursos computacionais disponíveis, cada algoritmo pode ser mais ou menos adequado.

Aprendizagem não supervisionada tem um conjunto de dados de entrada (X), sem saída conhecida, e tem como objetivo encontrar relações entre os dados. Como exemplo, pode-se encontrar grupos por meio da técnica de *clustering*. Na Figura 2, ilustra-se um exemplo desse tipo de aprendizagem, no qual o modelo gera grupos de pontos considerando a posição e distancias entre eles. Esse tipo de aprendizagem não faz do escopo deste trabalho e portanto, não é detalhado aqui.

Figura 2: Exemplo aprendizagem não supervisionada.



A tarefa de codificação de doenças pode ser formulada como uma tarefa de classificação multi-rotulada, a qual consiste em um problema com três ou mais classes de saída (rótulos) (BOUTELL et al., 2004; READ et al., 2009; SCHAPIRE; SINGER, 2000; TSOUMAKAS; KATAKIS, 2007; TSOUMAKAS; KATAKIS; VLAHAVAS, 2008; ZHANG; ZHOU, 2014).

Nas seções que seguem, apresentam-se os algoritmos mais utilizados na literatura para a classificação de textos.

2.1.1 Regressão Logística

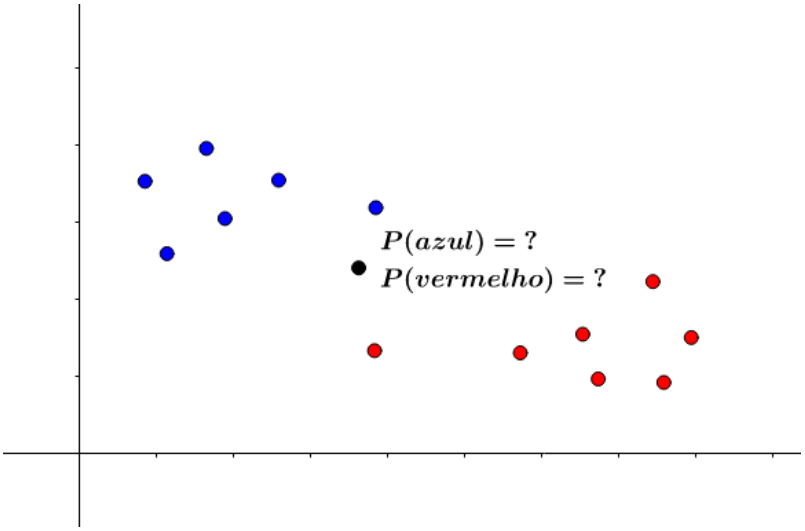
A Regressão Logística realiza a classificação dos dados, calculando a probabilidade de acontecer cada rótulo. Para o treinamento, aplica-se o método da máxima verossimilhança (JAMES et al., 2013; BISHOP, 2006).

A função de probabilidade de cada rótulo é dado pela Equação 2.1.

$$\hat{y} = \frac{e^{\beta_0 + \sum_{i=1}^{p} \beta_i x_i}}{1 - e^{\beta_0 + \sum_{i=1}^{p} \beta_i x_i}} \tag{2.1}$$

No exemplo de classificação de cores, ilustrado na Figura 3, o modelo calcula a probabilidade do ponto preto pertencer ao grupo de pontos azuis e de pertencer ao grupo de vermelhos, escolhendo a cor com maior probabilidade.

Figura 3: Exemplo da regressão logística.



Para a classificação dos laudos médicos, cada ponto representa um laudo, cada eixo representa uma palavra no vocabulário e cada cor representa um código da doença. Como resultado, tem-se um espaço com centenas (ou milhares) de eixos (dimensões) e centenas de "cores"possíveis (rótulos).

A Regressão Logística apresenta dois problemas: (i) são sensíveis aos preditores usados no treinamento, o que significa que a inclusão de um preditor novo no conjunto de dados de entrada pode levar a um resultado muito diferente, que pode ser o caso de inclusão de uma palavra no vocabulário dos laudos (JAMES et al., 2013). (ii) o método de máxima verossimilhança em problemas linearmente separáveis pode conduzir ao sobreajuste (*overfitting*) do modelo, que significa que o modelo tem um bom desempenho com os dados usados para treinamento, mas não com os dados de teste (BISHOP, 2006). Os textos também são suscetíveis ao segundo problema porque devido a sua grande quantidade de preditores faz com que sejam linearmente separáveis no espaço multi-dimensional.

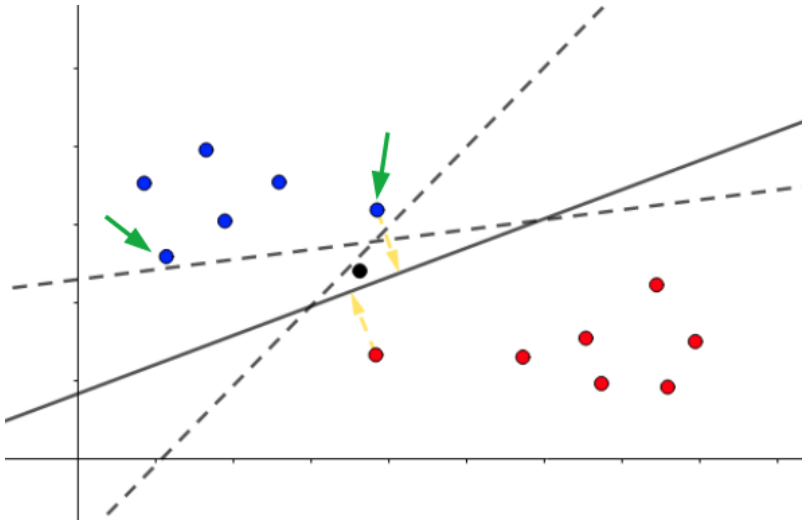
A vantagem da Regressão Logística é que é um modelo fácil de treinar e de interpretar, o que no contexto da codificação de diagnósticos, permitiria saber quais palavras influem mais na predição.

2.1.2 Support Vector Machine (SVM)

SVM é um classificador que busca separar o espaço p-dimensional por hiperplanos, de tal forma que pontos de classes diferentes não ocupem o mesmo subespaço. No treinamento, o modelo tenta maximizar a distância do ponto mais próximo a qualquer hi-

perplano (margem), objetivo que se conhece como hiperplano de máxima margem (AGGARWAL, 2018b; JAMES et al., 2013). Na Figura 4, ilustra-se um exemplo da construção dos hiperplanos. As três linhas pretas representam possíveis margens que separam as classes: as duas linhas pontilhadas são descartadas por estarem próximas aos pontos azuis sinalizados com setas verdes; a linha contínua é selecionada pelo modelo por maximizar as distâncias representadas pelas setas amarelas.

Figura 4: Exemplo da SVM.



Existem tarefas de classificação que não podem ser separadas por hiperplanos. Nesses casos, o classificador tenta suavizar a função objetivo, permitindo que alguns pontos fiquem do lado incorreto das margens. O problema de otimização do SVM pode ser formulado como apresenta-se nas Equações 2.2 - 2.5:

$$\text{maximizar}_{\beta_1,\beta_2,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n} M \tag{2.2}$$

Sujeito a

$$\sum_{j=1}^{p} \beta_j^2 = 1 \tag{2.3}$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i) \tag{2.4}$$

$$\epsilon_i \geq 0, \quad \sum_{i=1}^{n} \epsilon_i \leq C \tag{2.5}$$

Onde

- $\epsilon_1, \epsilon_2, ..., \epsilon_n$ são variáveis de folga, as quais permitem que algumas observações estejam no lado incorreto da margem,
- M é a largura da margem, e
- C é o parâmetro de ajuste.

A partir da resolução do problema, demonstra-se que o classificador linear pode ser representado pela Equação 2.6 (JAMES et al., 2013), o qual tem um bom desempenho para situações nas quais a fronteira entre as duas classes é linear.

$$\hat{y} = \beta_0 + \sum_{i \in S} \alpha_i \langle x, x_i \rangle \tag{2.6}$$

Para abordar casos com fronteiras não lineares, a melhor opção é ampliar o espaço de preditores com *kernels* (JAMES et al., 2013; GOODFELLOW; BENGIO; COURVILLE, 2016), os quais permitem treinar modelos não lineares, modificando a equação como mostra a Equação 2.7,

$$\hat{y} = \beta_0 + \sum_{i \in S} \alpha_i K(x, x_i) \tag{2.7}$$

onde a função K pode tomar diferentes formas dependendo do *kernel*.

Para tarefas de classificação de textos, o modelo de AM precisa ser capaz de lidar com dados de entrada esparsos e com grande dimensionalidade. SVM tem mostrado ser uma boa alternativa nessas situações; especificamente o SVM com *kernel* linear, o qual reduz o viés como a variança do modelo (AGGARWAL, 2018a).

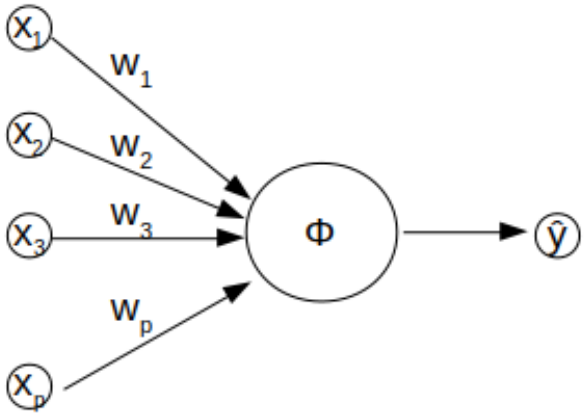
2.1.3 Redes Neurais

Uma Rede Neural é um modelo computacional de AM composto por um ou vários nós conectados entre si. Cada nó, chamado neurônio, aplica uma função nas entradas que recebe e passa os resultados aos neurônios posteriores; em cada neurônio, as entradas são afetadas pelos pesos das conexões. No processo de aprendizagem, vão se modificando os pesos dentro da rede, até produzir as respostas esperadas.

O poder de aprendizagem de uma rede neural depende da quantidade de neurônios que a compõem e a forma como são combinados. A quantidade de dados de entrada deve ser grande o suficiente para se conseguir ajustar a grande quantidade de pesos das conexões (AGGARWAL, 2018b).

Perceptron é a rede neural mais simples, a qual é composta por um neurônio, como apresentado na Figura 5.

Figura 5: Desenho esquemático de um Perceptron.



onde $x = [x_1, x_2, x_3, ..., x_p]$ é a entrada, $W = [w_1, w_2, w_3, ..., w_p]$ são os pesos das conexões e Φ é a função de ativação da rede.

A saída da rede é dada pela Equação 2.8:

$$\hat{y} = \Phi(W \cdot x) \tag{2.8}$$

As funções Φ mais frequentes para saídas reais são as funções linear e ReLU (*Rectified Linear Unit*). Já para as saídas binárias, são as funções sigmóide, signo e tangente hiperbólica (AGGARWAL, 2018b).

Para o ajuste dos pesos das conexões, a rede neural busca minimizar o erro da predição, dado pela função de perda. Geralmente, essa função de perda para saídas reais é uma perda ao quadrado simples (*simple squared loss*) e uma perda de entropia cruzada (*cross-entropy loss*), para saídas de múltiplas classes (AGGARWAL, 2018b).

O Perceptron tem bom desempenho para problemas linearmente separáveis, mas para problemas mais complexos, precisa-se de redes neurais com maior número de unidades e camadas. Assim, as redes mais usadas para tarefas de classificação de imagens e textos são: *Multilayer Network*, Convolutional Neural Network (CNN) e Long Short Term Memory (LSTM).

Uma das desvantagens dos modelos de Redes Neurais é que precisam de uma grande quantidade de dados para o seu ajuste.

2.1.4 Medidas de avaliação

Os modelos de AM podem ser avaliados por diferentes métricas e sua adequação dependerá do tipo do dado da variável alvo (JAMES et al., 2013; GOODFELLOW; BENGIO; COURVILLE, 2016). Para tarefas de classificação com dois rótulos -*Positivo* e *Negativo*-, pode-se elaborar uma tabela com os possíveis resultados do modelo, como apresentado na Tabela 1, também conhecida como *Matriz de Confusão*.

Tabela 1: Resultados esperados da classificação - Matriz de Confusão.

		Y	
		Positivo	Negativo
$\hat{\mathbf{y}}$	Positivo	Verdadeiro Positivo (VP)	Falso Positivo (FP)
	Negativo	Falso Negativo (FN)	Verdadeiro Negativo (VN)

A partir da Matriz de Confusão, podem-se calcular as métricas denominadas *acurácia*, *precisão*, *recall* e *F1-score*.

A *acurácia* é a fração de rótulos que foram corretamente previstos, sendo calculada pela Equação 2.9.

$$Ac = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.9}$$

A *precisão* do modelo, calculada pela Equação 2.10, representa a fração de rótulos positivos previstos corretamente.

$$p = \frac{VP}{VP + FP} \tag{2.10}$$

O *recall* é a proporção de rótulos positivos que foram previstos pelo modelo, sendo calculado pela Equação 2.11.

$$r = \frac{VP}{VP + FN} \tag{2.11}$$

O *F1-score* é a média harmônica da *precisão* e do *recall*, o qual permite avaliar o modelo equilibrando a *precisão* e o *recall*.

$$F1_{score} = \frac{2(p * r)}{p + r} \tag{2.12}$$

Em tarefas de classificação multi-rotuladas, as métricas são calculadas para cada classe, considerando as restantes como rótulo negativo. O valor final de cada métrica é calculado pela média dos valores obtidos e são denominados *macro precisão*, *macro recall* e *macro F1-score*. Por exemplo, assumindo os resultados apresentados na Tabela 2 para uma classificação de cores,

Tabela 2: Matriz de Confusão para múltiplas classes.

		Y			Total
		Azul	Vermelho	Amarelo	
	Azul	16	3	0	19
$\hat{\mathbf{y}}$	Vermelho	1	4	1	6
	Amarelo	4	3	19	26
	Total	21	10	20	51

as métricas são calculadas da seguinte forma:

$$p_{azul} = \frac{16}{16+3+0} = 0,8421$$

$$p_{vermelho} = \frac{4}{1+4+1} = 0,6667$$

$$p_{amarelo} = \frac{19}{4+3+19} = 0,7308$$

$$\text{macro precisão} = p = \text{média}(p_{azul}, p_{vermelho}, p_{amarelo}) = 0,7465$$

$$r_{azul} = \frac{16}{16+1+4} = 0,7619$$

$$r_{vermelho} = \frac{4}{3+4+3} = 0,40$$

$$r_{amarelo} = \frac{19}{0+1+19} = 0,95$$

$$\text{macro recall} = r = \text{media}(r_{azul}, r_{vermelho}, r_{amarelo}) = 0,7040$$

$$F1_{azul} = \frac{2*p_{azul}*r_{azul}}{p_{azul}+r_{azul}} = 0,80$$

$$F1_{vermelho} = \frac{2*p_{vermelho}*r_{vermelho}}{p_{vermelho}+r_{vermelho}} = 0,50$$

$$F1_{amarelo} = \frac{2*p_{amarelo}*r_{amarelo}}{p_{amarelo}+r_{amarelo}} = 0,8261$$

$$\text{macro F1} = F1 = \text{média}(F1_{azul}, F1_{vermelho}, F1_{amarelo}) = 0,7087$$

A vantagem desse cálculo é que as métricas dão o mesmo peso tanto para as classes

menos frequentes quanto para as mais frequentes (EISENSTEIN, 2018). Assim, no exemplo apresentado, todas as cores são consideradas no cálculo da média, mesmo tendo uma menor quantidade de observações vermelhas.

2.2 Aprendizagem de Máquina em textos

Os modelos abordados nas seções anteriores são formulados para entradas de dados numéricos. Para introduzir textos nos modelos, é necessário representar as palavras de uma maneira interpretável pelos programas computacionais. Dentre as formas de representar um texto, as mais populares na literatura de classificação de texto são Bag of Words (BOW), Term Frequency - Inverse Document Frequency (TF-IDF) e *Word Embeddings*.

2.2.1 Bag of Words (BOW)

O texto é representado por um vetor, sendo que cada palavra é uma dimensão do vetor. Os valores do vetor correspondem ao número de vezes que cada palavra se repete no texto (AGGARWAL, 2018a; EISENSTEIN, 2018). Na Figura 6 ilustra-se um exemplo de um laudo representado como um BOW.

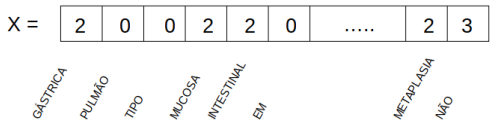
Figura 6: Exemplo BOW.

(a) Texto do laudo

Mucosa gástrica de tipos fúndico e pilórico com revestimento foveolar discretamente tortuoso. A lâmina própria é expandida por leve infiltrado linfoplasmocitário, com ocasionais eosinófilos, sem participação de neutrófilos. Não há atrofia ou metaplasia intestinal. Não foram identificados bacilos com morfologia de H. pylori.

Mucosa gástrica apresentando alongamento hiperplásico irregular de fovéolas, sem displasia. As glândulas exibem discreta ectasia. Observa-se discreto infiltrado inflamatório linfoplasmocitário. Não há atrofia ou metaplasia intestinal. Não foram identificados bacilos com morfologia própria de Helicobacter pylori

(b) Representação

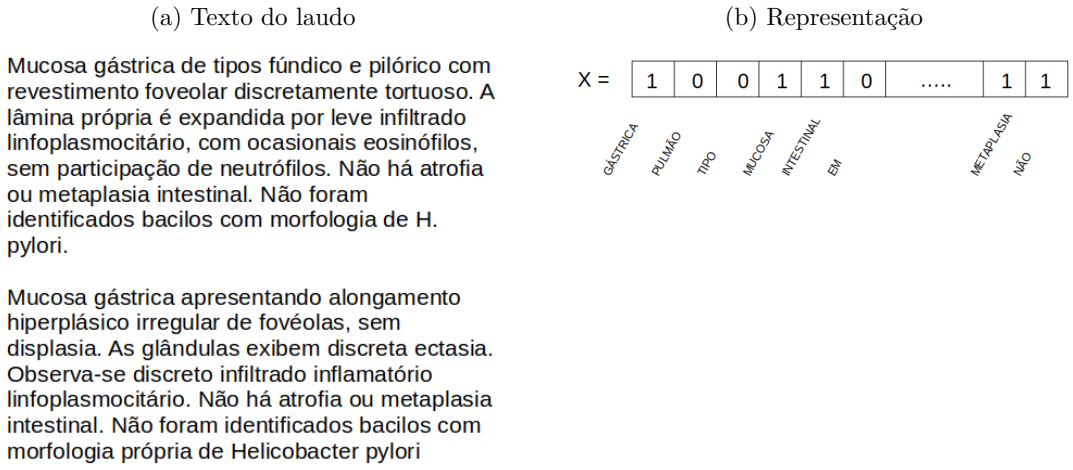


As vantagens do BOW são a simplicidade da representação e a eficiência computacional. As desvantagens são: (i) perda da ordem das palavras no texto, resultando na perda do significado semântico; (ii) grande dimensionalidade do vetor devido ao número elevado de palavras no vocabulário; (iii) contagem elevada de palavras muito frequentes,

como artigos e conectores, que tendem a não ser representativas no texto (AGGARWAL, 2018a; EISENSTEIN, 2018).

Uma modificação do BOW é substituir a contagem de cada palavra por um valor binário, o qual indicará se a palavra está presente ou não no texto, como pode-se observar na Figura 7.

Figura 7: Exemplo BOW binário.



2.2.2 Term Frequency - Inverse Document Frequency (TF-IDF)

A representação TF-IDF é similar ao BOW, no entanto busca solucionar o problema ocasionado pela alta frequência de palavras não representativas.

A frequência de cada palavra no vetor é normalizada pela proporção de textos em que a palavra está presente (AGGARWAL, 2018a; EISENSTEIN, 2018). Assim, o valor da palavra i no vetor é dado pela Equação 2.13,

$$\text{TF-IDF}_i = \frac{f_i}{log(1 + \frac{n_i}{n})} \tag{2.13}$$

onde:

- f_i é o número de vezes que a palavra aparece no texto;
- n_i é o número de textos em que a palavra aparece;
- n é o número total de textos.

Na Figura 8 ilustra-se a representação de um texto usando TF-IDF.

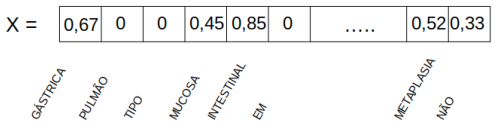
Figura 8: Exemplo TF-IDF.

(a) Texto do laudo

Mucosa gástrica de tipos fúndico e pilórico com revestimento foveolar discretamente tortuoso. A lâmina própria é expandida por leve infiltrado linfoplasmocitário, com ocasionais eosinófilos, sem participação de neutrófilos. Não há atrofia ou metaplasia intestinal. Não foram identificados bacilos com morfologia de H. pylori.

Mucosa gástrica apresentando alongamento hiperplásico irregular de fovéolas, sem displasia. As glândulas exibem discreta ectasia. Observa-se discreto infiltrado inflamatório linfoplasmocitário. Não há atrofia ou metaplasia intestinal. Não foram identificados bacilos com morfologia própria de Helicobacter pylori

(b) Representação

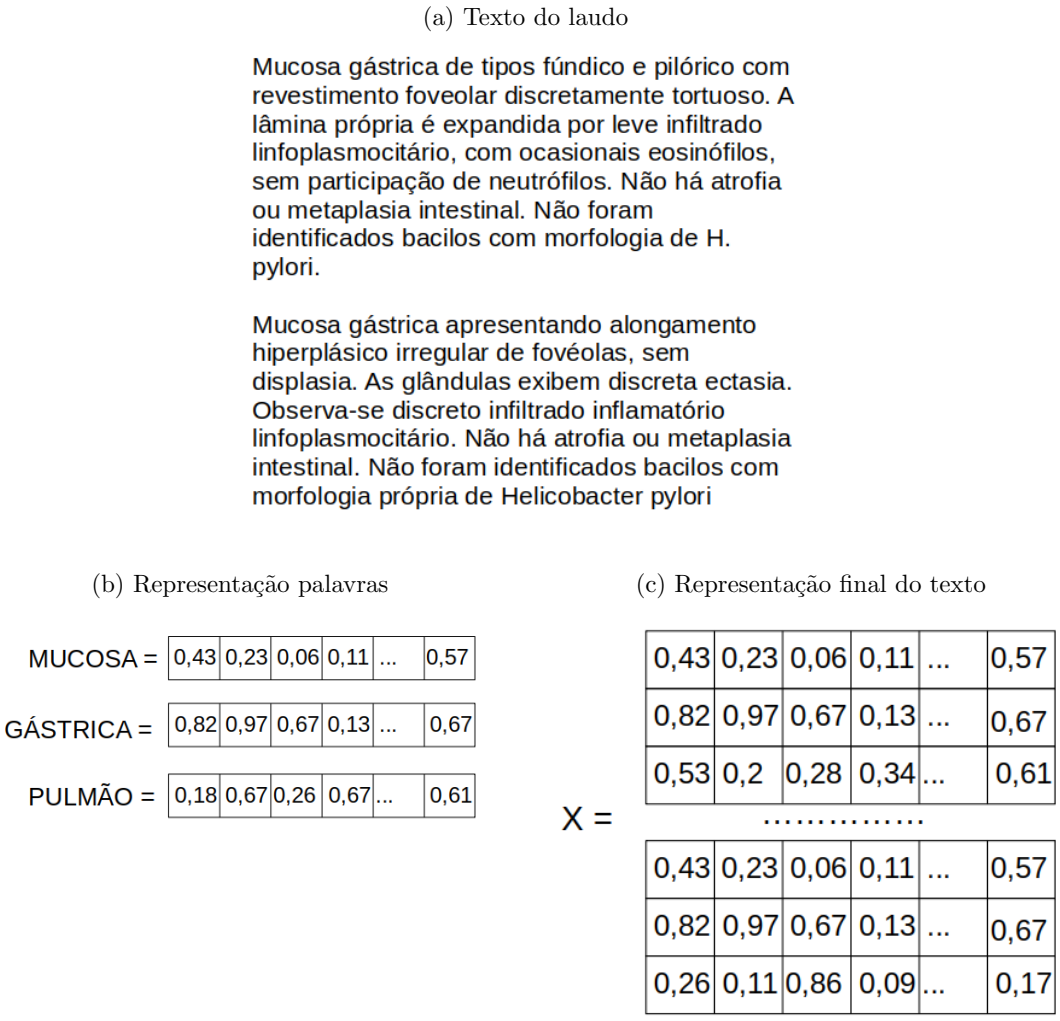


2.2.3 *Word Embeddings*

Word Embeddings consiste de uma solução para o problema da perda do significado semântico (AGGARWAL, 2018a), utilizando redes neurais de uma camada oculta para encontrar a relação entre as palavras de um contexto. Representa-se cada palavra como um vetor, tentando capturar as características do contexto da palavra. Assim, palavras relacionadas são representadas por vetores vizinhos em um hiperespaço (AGGARWAL, 2018a; EISENSTEIN, 2018; YOUNG et al., 2018).

Na Figura 9 (a) ilustra-se a representação de um texto empregando *Word Embeddings*. Na Figura 9 (b) observa-se a representação de cada palavra como um vetor e, na Figura 9 (c), observa-se a representação final do texto, mantendo a ordem das palavras.

Figura 9: Exemplo de Word Embeddings.



Duas das implementações mais populares são o modelo CBOW e o modelo Skip-Gram, ambos implementados usando redes neurais com uma camada oculta (AGGARWAL, 2018a; EISENSTEIN, 2018; YOUNG et al., 2018). No modelo CBOW, treina-se a rede neural a fim de predizer uma palavra a partir das palavras ao seu redor. Por outro lado, no modelo Skip-Gram, fixa-se uma palavra e tenta-se predizer as palavras ao seu redor (MIKOLOV et al., 2013).

2.3 Qualidade dos Dados (QD)

Qualidade é definida como a totalidade de características de um produto para satisfazer as necessidades estabelecidas. Também é chamada *fitness for use* ou "adequada para

o uso pretendido". Assim, a QD está ligada ao uso dos dados e a sua avaliação deve basear-se nos requisitos do usuário e no contexto da aplicação (BARBOSA et al., 2019).

A qualidade do dado impacta na informação gerada a partir dos dados e, por conseguinte, na eficiência e efetividade das organizações e negócios que utilizam tal informação. Considerando a relevância, a natureza e a variedade de tipos de dados e sistemas de informação, além da variedade de tipos de problemas associado aos dados, o tratamento da QD é complexo e multifacetado (BATINI et al., 2009).

2.3.1 Dimensões de QD

A QD é definida em termos de dimensões de QD, as quais são utilizadas para identificar os problemas nos dados. Não existe consenso na literatura quanto a quantidade de dimensões, nem quanto ao significado das dimensões de QD, pois elas dependem do domínio dos dados (FRANCISCO et al., 2017; BATINI et al., 2009; BATINI; SCANNAPIECO, 2016). Por exemplo, para dados estruturados em tabelas, define-se a dimensão *completude* como a quantidade de valores ausentes para colunas específicas. Por outro lado, em dados textuais, a *completude* não pode ser medida, pois não possuem estrutura alguma que indique quais valores (palavras) deveriam estar presentes.

Dentre as dimensões de qualidade para textos, as mais comuns são: *acurácia, legibilidade* e *acessibilidade* (BATINI; SCANNAPIECO, 2016), detalhadas a seguir.

A ***acurácia*** é definida como a distância entre palavras do texto e palavras de um vocabulário de referência (dicionário). Uma das métricas empregadas calcula a distância como a quantidade de caracteres que devem ser removidos, aumentados ou substituídos em uma palavra para ela pertencer ao dicionário (BATINI; SCANNAPIECO, 2016; BATINI et al., 2009). Por exemplo, na Figura 10, a distância resultante é a quantidade de operações realizadas para que a primeira palavra seja igual à segunda: remover hífen em vermelho, substituir ***o*** (em azul) por ***ó*** (com acento) e acrescentar na última posição a letra ***o***.

Figura 10: Distância de palavras.

*distância(*anatomo-patologic_, anatomopatológico*) = 3*

Em casos com muitas palavras nos textos e no vocabulário, o cálculo dessa métrica pode ser computacionalmente custoso. Consequentemente, a alternativa para medir a acurácia é calcular a proporção de palavras corretas no texto (BATINI; SCANNAPIECO, 2016).

A ***legibilidade*** em textos refere-se a facilidade do texto para "ser lido". Algumas das métricas empregadas para esta dimensão é o índice Gunning Fog (GF), o índice de legibilidade automatizado ARI (*automated readibility index*) e *Facilidade de Leitura Flesh*. As fórmulas são dadas pelas Equações 2.14, 2.15 e 2.16 (BATINI; SCANNAPIECO, 2016; GRAESSER et al., 2004; ALUISIO et al., 2014).

$$GF = 0,4 * \left(\frac{n\_palavras}{n\_oracoes} + 100 * \frac{n\_palavras\_complexas}{n\_palavras} \right) \tag{2.14}$$

$$ARI = 4,71 * \frac{n\_caracteres}{n\_palavras} + 0,5 * \frac{n\_palavras\_complexas}{n\_oracoes} \tag{2.15}$$

$$Flesh = 206,835 - 1,015 * \frac{n\_palavras}{n\_oracoes} - 84,6 * \frac{n\_silabas}{n\_palavras} \tag{2.16}$$

Tanto o índice GF e o ARI consideram *palavras complexas* aquelas com mais de duas sílabas. Valores altos do GF e ARI indicam textos com menor legibilidade. Por outro lado, um valor baixo da métrica Flesh indica uma menor legibilidade.

De acordo com as três métricas, textos com palavras ou orações muito compridas são menos legíveis.

A ***acessibilidade*** está relacionada com a capacidade do leitor em entender o texto, sendo então, dependente do usuário para ao qual o texto é dirigido. Assim, os textos médicos que contém palavras específicas da área da saúde são entendidos por pessoas com os conhecimentos da área (BATINI; SCANNAPIECO, 2016). Para essa dimensão, não foram encontradas métricas na bibliografia.

2.3.2 Estratégias para melhoria da QD

Data-driven e *process-driven* são estratégias adotadas para a melhoria da QD. A estratégia *data-driven* está associada ao tratamento direto do dado a partir de atividades corretivas, de acordo com os problemas identificados. *Process-driven* busca identificar a causa raíz dos problemas de QD identificados. Isso, em geral, leva a identificação dos processos de criação/captação e consumo dos dados, culminando em um possível redesenho desses processos para mitigar a causa dos problemas (BATINI et al., 2009; BATINI; SCANNAPIECO, 2016; GLOWALLA; SUNYAEV, 2014).

A melhora da QD pode ser a curto ou longo prazo. A estratégia de *process-driven*

age na fonte de dados gerando resultados a longo prazo, mas sua aplicação comumente representa um custo alto. Os resultados de *data-driven* são a curto prazo e, geralmente, mais simples de implantar e com menor custo. Porém, quando as atividades de melhoria são repetidas frequentemente, o custo pode ser superior ao do *process-driven* (BATINI et al., 2009; BATINI; SCANNAPIECO, 2016; GLOWALLA; SUNYAEV, 2014).

Nesta pesquisa, a estratégia seguida foi a *data-driven*, pois o objetivo foi tratar o dado para viabilizar a codificação automática das doenças. A estratégia *process-driven* deve ser perseguida para mitigar a raíz dos problemas de QD identificados nos laudos. No entanto, a estratégia *process driven* está fora do escopo deste trabalho por envolver mudanças nos processos da instituição, as quais necessitam de decisões mais estruturais.

As atividades comuns de *data-driven* são descritas a seguir:

- Aquisição de novos dados com maior qualidade a fim de substituir os valores existentes com problemas.
- Substituição de valores por um padrão definido (padronização). Um exemplo é converter as abreviações em suas palavras completas (seu significado).
- Localização e correção de erros, que em textos podem ser erros de escrita.
- Aplicação de restrições de integridade (consistência) para detecção de problemas e correção automática. Por exemplo, pode-se validar que no endereço registrado de um paciente, o estado seja consistente com a cidade e, em caso não ser, corrigi-lo.

2.4 Tratamento de dados para AM

Os modelos de aprendizagem supervisionada dependem de dados de exemplo para ajustar seus parâmetros e assim realizar predições. Portanto, se o treinamento dos modelos é realizado com dados com baixa qualidade ou não adequados, o desempenho diminui. Os desafios identificados com maior frequência na literatura de classificação de textos são apresentados a seguir:

- ***Label noise***: ocorre quando existem rótulos errados nos dados empregados para o treinamento (GUPTA et al., 2021; WHANG; LEE, 2020; JAIN et al., 2020). Essa condição faz com que o modelo cometa erros na predição, atribuindo o rótulo errado em entradas similares. O *label noise* pode ser detectado implementando regras de validação para os dados ou comparando-os com uma fonte de dados confiáveis. Uma

vez detectados os rótulos errados, esses devem ser corrigidos ou descartados para o treinamento do modelo.

- **Desbalanceamento de classes**: refere-se a diferença da frequência dos rótulos presentes nos dados de treinamento. Na aprendizagem, muitos modelos procuram minimizar o erro nas predições. Como consequência, na saída do modelo, são gerados os rótulos mais frequentes e ignorados os menos frequentes. Para melhorar o desempenho dos modelos, são aplicadas técnicas de re-amostragem nos dados de treinamento, removendo os rótulos com maior número de ocorrências e aumentando a contagem dos rótulos menos frequentes (GUPTA et al., 2021; JAIN et al., 2020).

- ***Noise data***: são tókens nos dados de entrada que não contribuem com informação relevante para realizar a previsão dos rótulos (ALNAJRAN et al., 2018). Tóken refere-se à unidades do texto que incluem palavras, números, datas e símbolos. A presença de *noise data* ocasiona que o modelo não atribua o peso adequado a preditores chaves no aprendizagem. Em classificação de textos, os números, datas, acrônimos, símbolos e palavras conhecidas como *stop-words* (artigos, preposições e outros), são, no geral, causadores desse problema. O tratamento realizado nos textos para redução do *noise data* consiste em identificar e excluir os tókens não relevantes.

- **Complexidade do texto**: está relacionada com a dificuldade de entender o texto, tanto para humanos como para modelos de AM (COLLINS; ROZANOV; ZHANG, 2018). O grau de complexidade do texto em modelos é influenciada pelo comprimento do texto, tamanho do vocabulário e ambiguidades nas palavras (COLLINS; ROZANOV; ZHANG, 2018; SUN et al., 2018). Esse problema afeta as dimensões de *legibilidade* e de *acessibilidade*, ou seja, um texto de alta complexidade refletirá em baixa *legibilidade* e baixa *acessibilidade*. Portanto, a alta complexidade do texto pode ser considerado como um problema de qualidade. Uma das atividades para mitigar esse problema é a aplicação da técnica de *stemming*, a qual reduz as palavras a seus radicais, diminuindo, assim, o tamanho do vocabulário e problemas de ambiguidades ocasionados pelas inflexões.

Vale ressaltar que, excluir tókens, como abreviações, pode levar a exclusão de informações importantes (ex. Hist. Pilórico) e representaria um problema de QD. Nesse caso, seria mais adequado substituir as abreviações pelas palavras completas. Outro ponto importante a destacar é que, a técnica de *stemming*, do ponto de vista da QD, pode representar um problema ao reduzir as palavras ao seu radical, já que, em alguns casos,

reduz ao mesmo radical palavras com significados diferentes. Uma das soluções para esse problema é a aplicação da técnica de processamento de linguagem natural, chamada *POS-tagging* (AGGARWAL, 2018a), a qual adiciona a informação da categoria da palavra ao radical. Assim, é possível diferenciar se a palavra original era um nome, verbo ou artigo.

A aplicação de técnicas de processamento de linguagem natural saem do escopo deste projeto, pois precisam de modelos complexos pré-treinados com uma grande quantidade de textos similares aos empregados na pesquisa.

3 REVISÃO DA LITERATURA SOBRE QUALIDADE DE DADOS EM REGISTROS CLÍNICOS

Como parte da metodologia adotada para o desenvolvimento deste trabalho, foi realizada uma revisão da literatura, apresentada em Iturry et al. (2021), cujo objetivo foi identificar os problemas de QD em registros clínicos e as soluções propostas.

Neste capítulo, apresenta-se uma síntese dessa revisão, destacando-se o método empregado e os principais resultados. Maiores detalhes, como formas de avaliação, dentre outros, podem ser encontrados no artigo citado.

Para a execução da revisão, seguiu-se o método desenvolvido por Kitchenham (2004). Nele, o autor propõe diretrizes para realizar uma Revisão Sistemática da Literatura (RSL) em Engenharia de Software. Diferentemente da área da saúde, na computação, não se considera dados empíricos, no sentido adquirir dados de uma população alvo. No caso da computação, dados empíricos se referem ao uso de *software* ou *hardware*. Frequentemente, a RSL baseia-se apenas em dados teóricos.

O método de Kitchenham (2004) é divido em três fases: planejamento, execução e apresentação dos resultados. Na fase de planejamento, definiram-se os objetivos, as questões da pesquisa, as bases de dados para a pesquisa de artigos e os critérios de inclusão e exclusão dos artigos. Já na fase de execução, selecionaram-se os artigos, de acordo com os critérios de inclusão e exclusão. Finalmente, as questões da pesquisa foram respondidas e reportaram-se as conclusões da revisão na fase de apresentação dos resultados.

3.1 Principais aspectos da revisão da literatura

Para identificar os métodos de avaliação e melhoria da QD em registros clínicos e laudos de procedimentos, foram elaboradas duas questões de pesquisa (QP1 e QP2), descritas a seguir:

- *QP1*: Quais dimensões são usadas para qualificar os problemas de QD identificados

nos registros médicos?

- *QP2*: Quais técnicas são empregadas para mitigar os problemas de QD identificados nos registros médicos?

Em Electronic Health Record (EHR) e Electronic Medical Record (EMR) é possível obter informações do histórico médico e de tratamento de pacientes. Assim, para encontrar artigos para responder a QP1 e QP2, formulou-se as strings de busca:

- *"data quality"AND "ehr"*;
- *"data quality"AND "emr"*

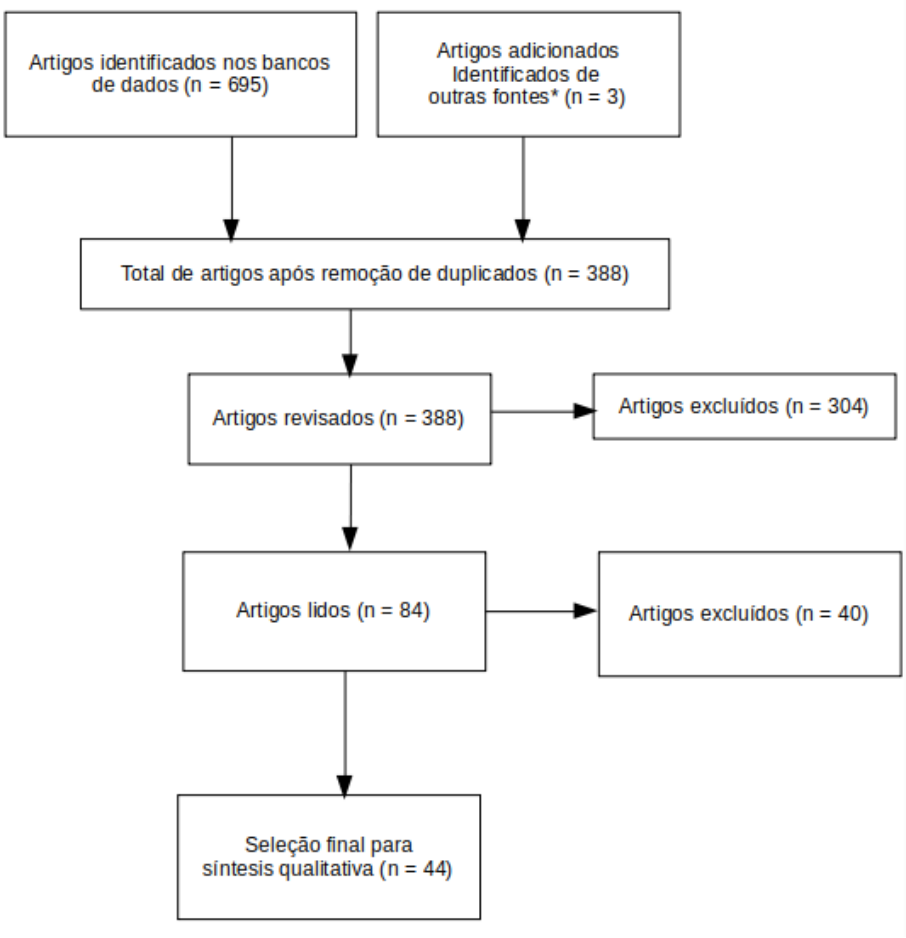
Para a seleção de artigos, foram utilizadas as seguintes bases de dados de pesquisa: ACM, IEEE, Scopus, Web of Science e PubMed. Essas bases foram escolhidas por sua creditabilidade e por reunirem artigos das áreas de conhecimento em computação e da área médica. Os critérios de inclusão e exclusão dos artigos são apresentados na Tabela 3.

Tabela 3: Critérios de inclusão e exclusão.

Critério de inclusão	Critério de exclusão
I1 O artigo emprega alguma dimensão para a avaliação da QD em registros clínicos	**E1** O artigo não trabalha com registros clínicos
I2 O artigo elabora alguma atividade de melhoria QD em registros clínicos	**E2** O artigo foi publicado antes de 2015

Inicialmente, foram encontrados 695 artigos e 3 artigos adicionais de revisões de literatura similares a essa. Na Figura 11, ilustra-se o processo completo da seleção dos artigos.

Figura 11: Fluxo da seleção de artigos.



Na Tabela 4, observa-se que uma grande porcentagem (95%) dos artigos utilizam dimensões de QD para avaliar os dados. No entanto, a partir da Tabela 5, verifica-se que apenas 20,45% empregam alguma técnica para a melhoria da qualidade.

Tabela 4: Artigos que empregam dimensões de QD.

Emprega dimensões	Artigos	%	Referência
sim	42	95,45	Abiy et al. (2018), Ali et al. (2018), Badr (2019), Bae et al. (2015), Carsley et al. (2018), Daniel et al. (2019), Deng et al. (2016), Dziadkowiec et al. (2016), Ehsani-Moghaddam, Martin e Queenan (2019), Estiri et al. (2018, 2019), Feder (2018), Ford et al. (2016), Fox et al. (2018), León-Chocano et al. (2015, 2016), Hartzband e Jacobs (2016), Johnson et al. (2015), Kahn et al. (2016), Lee et al. (2015), Lee, Grobe e Tiro (2016), Lee, Weiskopf e Pathak (2017), Lingren et al. (2018), Liu, Zowghi e Talaei-Khoei (2020), Medhanyie et al. (2017), Muthee et al. (2018), Nobles et al. (2015), Noselli et al. (2017), Puttkammer et al. (2016, 2017), Reimer, Milinovich e Madigan (2016), Scholte et al. (2016), Singer et al. (2016), Skyttberg et al. (2017), Sollie et al. (2017), Staff, Roberts e March (2016), Taggart, Liaw e Yu (2015), Terry et al. (2019), Toftdahl et al. (2018), Tu et al. (2015), Weiskopf et al. (2017), Wennberg et al. (2019)
não	2	4,55	Hart e Kuo (2017), Jones et al. (2018)

Tabela 5: Artigos que aplicam atividades de melhoria da QD.

Aplica melhoria	Artigos	%	Referência
sim	9	20,45	Badr (2019), Daniel et al. (2019), Dziadkowiec et al. (2016), Ehsani-Moghaddam, Martin e Queenan (2019), León-Chocano et al. (2015, 2016), Hart e Kuo (2017), Skyttberg et al. (2017), Taggart, Liaw e Yu (2015)
não	35	79,55	Abiy et al. (2018), Ali et al. (2018), Bae et al. (2015), Carsley et al. (2018), Deng et al. (2016), Estiri et al. (2018, 2019), Feder (2018), Ford et al. (2016), Fox et al. (2018), Hartzband e Jacobs (2016), Johnson et al. (2015), Jones et al. (2018), Kahn et al. (2016), Lee et al. (2015), Lee, Grobe e Tiro (2016), Lee, Weiskopf e Pathak (2017), Lingren et al. (2018), Liu, Zowghi e Talaei-Khoei (2020), Medhanyie et al. (2017), Muthee et al. (2018), Nobles et al. (2015), Noselli et al. (2017), Puttkammer et al. (2016, 2017), Reimer, Milinovich e Madigan (2016), Scholte et al. (2016), Singer et al. (2016), Sollie et al. (2017), Staff, Roberts e March (2016), Terry et al. (2019), Toftdahl et al. (2018), Tu et al. (2015), Weiskopf et al. (2017), Wennberg et al. (2019)

3.2 Resultados da revisão da literatura

Na Figura 12, apresentam-se as diferentes dimensões de QD, com o respectivo percentual de artigos que as abordam. Pode-se observar que, 11 das 15 dimensões foram empregadas em menos de 20% dos artigos. As 4 dimensões, abordadas em mais de 20% dos artigos avaliados, são detalhadas na Tabela 6. Nessa, mostra-se as diferentes definições encontradas na literatura, com os respectivos artigos associados à cada definição.

Figura 12: Porcentagem de artigos que aplicam cada dimensão de QD.

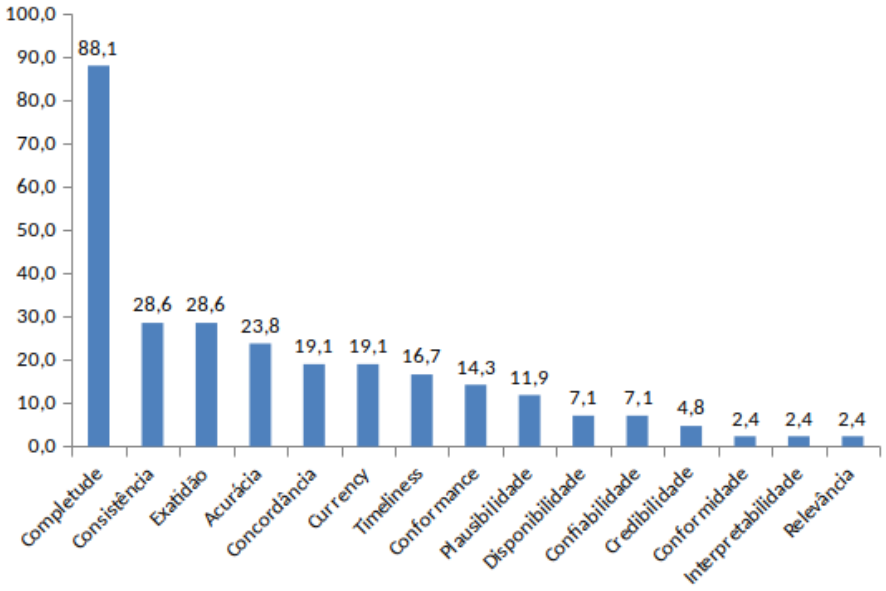


Tabela 6: Definições das dimensões de QD.

Dimensão	Definição	Referência
Completude	Presença (ou ausência) de um valor	Bae et al. (2015); Daniel et al. (2019); Fox et al. (2018); Johnson et al. (2015); Muthee et al. (2018); Nobles et al. (2015); Noselli et al. (2017); Scholte et al. (2016); Taggart, Liaw e Yu (2015); Terry et al. (2019); Sollie et al. (2017)
	Presença de um valor que é obrigatório	Ali et al. (2018); Badr (2019); Carsley et al. (2018); Ehsani-Moghaddam, Martin e Queenan (2019); Feder (2018); León-Chocano et al. (2015); Kahn et al. (2016); Lee, Weiskopf e Pathak (2017); Liu, Zowghi e Talaei-Khoei (2020); Puttkammer et al. (2016); Skyttberg et al. (2017); Weiskopf et al. (2017)
Exatidão	O valor é válido, razoável e não anormal	Noselli et al. (2017); Skyttberg et al. (2017); Taggart, Liaw e Yu (2015)

Tabela 6 Continuação

	O valor não tem erros e é igual a um valor padrão ou o valor real	Badr (2019); León-Chocano et al. (2016); Johnson et al. (2015); Scholte et al. (2016); Sollie et al. (2017); Terry et al. (2019); Weiskopf et al. (2017)
Consistência	Os dados satisfazem regras e restrições	León-Chocano et al. (2015); Johnson et al. (2015)
	O valor é correto	Fox et al. (2018); Nobles et al. (2015)
	Os valores dos dados são iguais em diferentes bancos de dados	Ali et al. (2018)
	Não existem diferenças entre itens que representam um mesmo objeto	Ehsani-Moghaddam, Martin e Queenan (2019)
Acurácia	Os dados são precisos, livres de erros e iguais a valores reais	Ali et al. (2018); Badr (2019); Carsley et al. (2018); Ehsani-Moghaddam, Martin e Queenan (2019); Feder (2018); Fox et al. (2018)
	Os valores são consistentes nas relações entre sim	Puttkammer et al. (2016)

Na Tabela 7, exibe-se as atividades de melhoria da QD realizadas por cada artigo.

3.3 Conclusões da revisão da literatura

A revisão da literature permitiu concluir que, não há consenso quanto as definições das dimensão de QD. No entanto, verificou-se que, quatro dimensões são frequentemente encontradas na literatura. São elas: completude, exatidão (*correctness*), consistência e acurácia.

Apesar de identificar problemas de QD, apenas um quinto dos artigos relatam sobre

Tabela 7: Tratamento da QD.

Artigo	Governança de Dados			Limpeza de Dados	
	Padronização	Treinamento de *staff*	Melhora de *software*	Localização de erro	Correção de erro
León-Chocano et al. (2015)	x				x
León-Chocano et al. (2016)	x				x
Dziadkowiec et al. (2016)				x	x
Skyttberg et al. (2017)	x	x	x		
Hart e Kuo (2017)		x		x	
Ehsani-Moghaddam, Martin e Queenan (2019)				x	x
Taggart, Liaw e Yu (2015)		x			
Daniel et al. (2019)		x	x		

o tratamento dos dados. Desses, nove artigos detalharam o tratamento de dados a fim de mitigar os problemas de QD. Conforme apresentado na Tabela 7, as classificações foram agrupadas em atividades de *Governança de dados* e de *Limpeza de dados* realizadas por cada artigo. Alguns dos artigos mencionados relatam sobre o treinamento de equipes de saúde, para o uso dos dados (TAGGART; LIAW; YU, 2015; SKYTTBERG et al., 2017; HART; KUO, 2017; DANIEL et al., 2019). Os demais artigos concentram-se na correção da QD em registros clínicos.

Todos os artigos selecionados na revisão da literatura baseiam-se em dados estruturados. É importante ressaltar que não foram encontrados artigos que avaliem a QD em textos médicos ou em laudos de procedimentos, os quais são considerados como dados não estruturados, revelando a importância deste trabalho.

4 REVISÃO DA LITERATURA SOBRE CODIFICAÇÃO AUTOMÁTICA DE DOENÇAS

A codificação automática de doenças é realizada a partir de modelos de classificação de textos médicos de diagnósticos que atribuam o correspondente código da doença. Os benefícios da aplicação de modelos são: (i) rapidez na codificação de grande quantidade de textos médicos, e (ii) diminuição de erros na identificação do código da doença.

Para o desenvolvimento do modelo de codificação, sendo essa uma das principais tarefas deste trabalho, realizou-se uma revisão da literatura, conforme método proposto em Kitchenham (2004), o qual foi apresentado brevemente no capítulo anterior.

Neste capítulo são apresentados o processo da revisão bibliográfica e os resultados obtidos, nos quais embasou-se o desenvolvimento do modelo de codificação.

4.1 Planejamento e da revisão bibliográfica

O objetivo da revisão bibliográfica foi identificar trabalhos prévios que realizaram a codificação automática de doenças, com o intuito de guiar a escolha do algoritmo de AM a ser empregado e as métricas para sua avaliação. Assim, as seguintes questões de pesquisa (QP) foram elaboradas:

- *QP1*: Quais são os algoritmos de AM empregados para a codificação automática de doenças?
- *QP2*: Quais são as métricas empregadas para avaliar o desempenho dos modelos?

Os termos chaves utilizados para a busca dos artigos foram: "*machine learning*", "*text classification*", "*diagnosis coding*"e "*icd coding*"(ICD é a abreviação em inglês de CID). Assim, formularam-se as seguintes *strings* de busca:

- "*machine learning*"AND "*diagnosis coding*"

- "*text classification*" AND "*diagnosis coding*"
- "*machine learning*" AND "*icd coding*"
- "*text classification*" AND "*icd coding*"

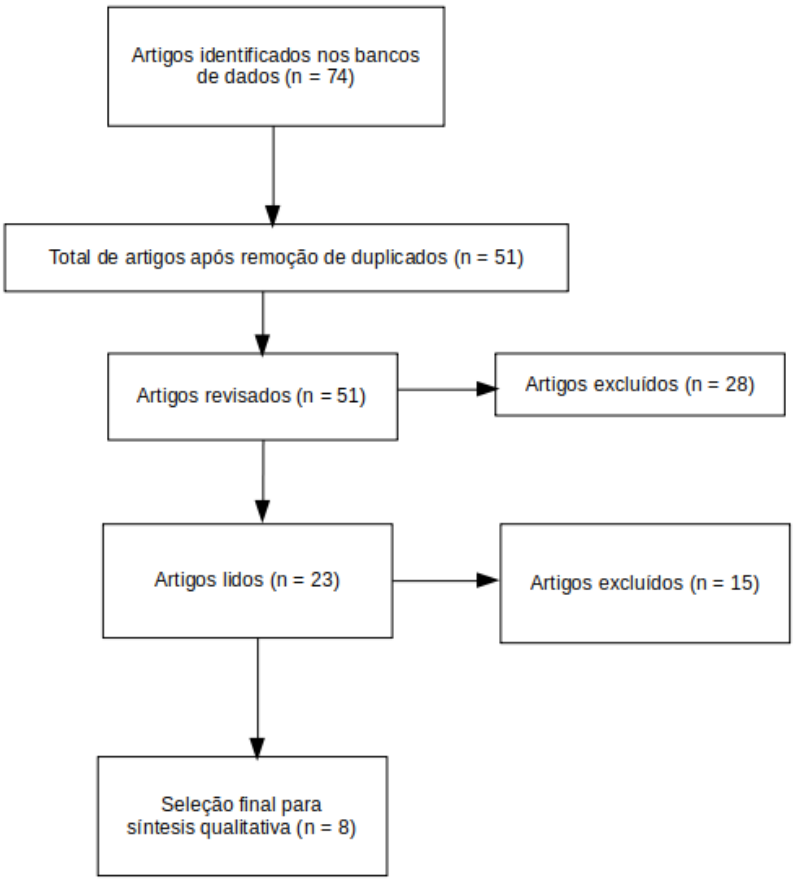
Os critérios de inclusão e exclusão dos artigos são apresentados na Tabela 8.

Tabela 8: Critérios de inclusão e exclusão.

Critério de inclusão	Critério de exclusão
I1 O artigo apresenta modelos de AM para a codificação automática de diagnósticos **I2** As doenças são codificadas com base no CID	**E1** O artigo foi publicado antes de 2010

Escolheram-se as seguintes bases de dados para a busca dos artigos: ACM, IEEE, Scopus, Web of Science e PubMed. Na Figura 13 ilustra-se o processo de seleção, inclusão e exclusão dos artigos.

Figura 13: Fluxo da seleção de artigos.



4.2 Resultados e conclusões da revisão bibliográfica

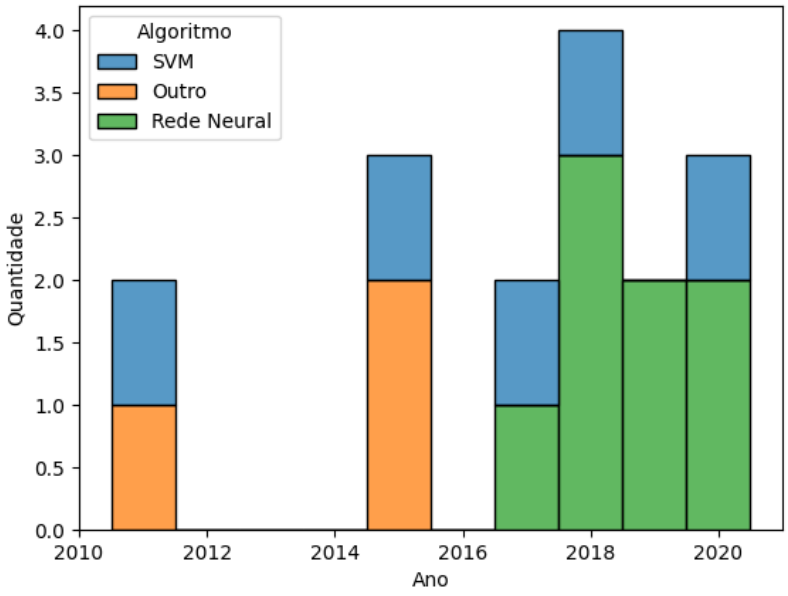
Os modelos de AM empregados na amostra de artigos selecionada foram: Regressão Logística, Naive Bayes, SVM, Multilayer Neural Network, CNN e LSTM. A Tabela 9 apresenta o(s) modelo(s) empregado(s) em cada artigo. É possível observar que o SVM é mais frequentemente utilizado.

Tabela 9: Algoritmos empregados.

Paper	Regressão Logística	Naive Bayes	SVM	Multilayer Network	CNN	LSTM
Lauría e March (2011)		x	x			
Kavuluru, Rios e Lu (2015)	x	x	x			
Lee e Muis (2017)			x	x		
Xu et al. (2018)					x	x
Zhong, Gao e Yi (2018)			x			
Xie e Xing (2018)						x
Li et al. (2019)				x	x	
Azam et al. (2020)			x		x	x

Na Figura 14 ilustra-se a distribuição da aplicação dos modelos por ano. Nota-se que o SVM tem sido bastante empregado nos últimos 10 anos. Também observa-se que as Redes Neurais têm sido muito utilizadas nos últimos anos. Tal fato pode ser explicado pela melhoria da eficiência computacional, permitindo treinar redes complexas, à exemplo das convolucionais e as recorrentes, aliado à manipulação de grandes quantidades de dados.

Figura 14: Modelos por ano.



Na Tabela 10 apresenta-se as métricas empregadas para a avaliação dos modelos, sendo as mais comuns: precisão, *recall* e F1-*Score*.

Tabela 10: Métricas empregadas.

Paper	Acurácia	Precisão	*Recall*	F-*score*
Lauría e March (2011)	x			
Kavuluru, Rios e Lu (2015)		x	x	x
Lee e Muis (2017)		x	x	x
Xu et al. (2018)				x
Zhong, Gao e Yi (2018)		x	x	x
Xie e Xing (2018)		x	x	
Li et al. (2019)		x	x	x
Azam et al. (2020)	x	x	x	x

5 DESCRIÇÃO DO EXPERIMENTO E RESULTADOS

Para codificação automática das doenças, utilizaram-se dados hospitalares referentes a registros clínicos de pacientes oncológicos. Classifica-se a pesquisa como observacional retrospectiva. Os dados utilizados nos experimentos são referentes a pacientes maiores de 18 anos, que foram admitidos no hospital por pelo menos 24 horas e que receberam um código de doença principal ou secundário, relacionado ao capítulo de neoplasias do CID-10. Os dados utilizados foram extraídos de laudos de procedimentos realizados para esse perfil de paciente. Teve-se acesso apenas a dados anonimizados, após autorização de comitê de ética (autorização do CEP com CAAE 28400820.3.0000.0060).

Neste capítulo apresentam-se inicialmente conceitos relativos à neoplasia, laudos médicos, codificação de doenças e estrutura dos códigos da CID, dando um contexto dos dados empregados. São descritos também os desafios encontrados para codificação automática e as soluções implementadas no tratamento dos textos de laudos e do conjunto de dados para o treinamento do modelo. Finalmente, expõe-se a arquitetura do modelo treinado e os resultados dos testes desenvolvidos para avaliação de seu desempenho.

5.1 Neoplasia e Laudos médicos

Neoplasia é o crescimento celular não controlado, também conhecido como tumor. É denominado como *tumor benigno* quando o crescimento é organizado e lento e, como *tumor maligno (câncer)*, quando o crescimento é rápido e pode invadir tecidos vizinhos. A ciência que estuda as neoplasias é chamada de oncologia (KUMAR et al., 2005; INCA, 2020; FUNDAP, 2012).

Para classificar os tumores é necessário conhecer sua localização, denominada *peça anatômica*, e informações macro e microscópicas, como o tamanho do tumor, a aparência física, a produção de substâncias e as manifestações clínicas (KUMAR et al., 2005; INCA, 2020; FUNDAP, 2012). Após a análise do tumor, todos os resultados são registrados em um documento conhecido como *laudo de procedimentos*.

Os laudos de oncologia empregados neste trabalho estão organizados nas seções: dados clínicos, peças, exame macroscópico, exame microscópico e conclusão do diagnóstico. Um exemplo resumido do conteúdo de um laudo é apresentado na Figura 15.

Figura 15: Exemplo resumido de um laudo médico.

LAUDO ANATOMOPATOLÓGICO

DADOS CLÍNICOS

Produto de biópsia de tumor em hipofaringe.
Tem exame intraoperatório: 1019 - 000706.

PEÇAS

1 - FARINGE - BIÓPSIA DE LESÃO (AMOSTRA I)
2 - FARINGE - BIÓPSIA DE LESÃO (AMOSTRA II)

EX. MICROSCÓPICO

O exame histológico revela:
1- Cortes de mucosa de faringe, revestida por epitélio escamoso estratificado com arquitetura em camadas preservada, alterações regenerativas do epitélio e erosões com deposição de malha fibrinoleucocitária. O córion subepitelial apresenta infiltrado inflamatório predominantemente linfomononuclear e vasos capilares ectásicos.
2- Cortes evidenciando neoplasia epitelial formada por células com núcleos volumosos e irregulares, com citoplasma amplo, por vezes mostrando intensa queratinizaçâo anormal. As células neoplásicas formam pequenos blocos neoplásicos que infiltram o estroma, suscitando reação desmoplásica.
s.r.l.

EX. MACROSCÓPICO

O espécime, previamente fixado em formaldeído 4% tamponado com fosfato básico de sódio ph 7.2-7.4, consta de:
1- Quatro fragmentos irregulares de tecido, elásticos e acastanhados, medindo de 0,2 a 0,4cm.
2- Quatro fragmentos irregulares de tecido, elásticos e acastanhados, medindo de 0,1 a 0,2cm.
* Todo material é submetido a exame histológico assim designado:
1 FARINGE - BIÓPSIA DE LESÃO (AMOSTRA I)
1-A - (4fti) 1-A-1 - H-E
2 FARINGE - BIÓPSIA DE LESÃO (AMOSTRA II)
2-A - (4fti) 2-A-1 - H-E

CONCLUSÃO

1- PROCESSO INFLAMATORIO CRONICO COM COMPONENTE DE EROSÃO EM MUCOSA DE FARINGE (AMOSTRA I DE LESÃO DE HIPOFARINGE).
2- CARCINOMA DE CÉLULAS ESCAMOSAS QUERATINIZANTE INVASIVO EM AMOSTRA II DE LESÃO DE HIPOFARINGE.
s.r.l.

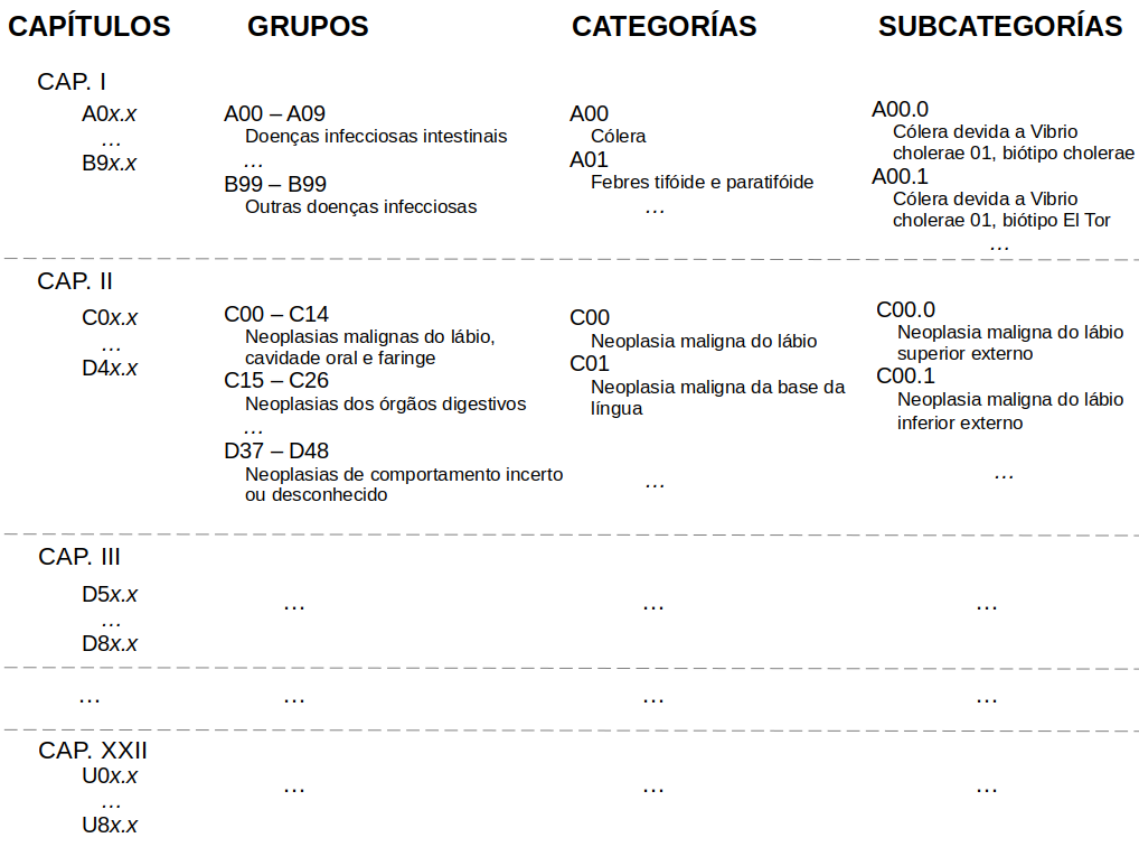
5.2 Codificação dos diagnósticos secundários

Para a classificação de doenças, o hospital faz uso da décima versão da CID (CID-10). O código CID-10 apresenta quatro caracteres, sendo o primeiro uma letra e os restantes, números. Os códigos são ordenados compreendendo a faixa *A000* a *Z999*.

CID-10 apresenta as doenças divididas em 22 capítulos, que por sua vez subdividem-se em grupos, categorias e subcategorias. Para organizar os códigos de doenças, emprega-se

uma estrutura hierárquica nos seus caracteres. A partir da letra e do primeiro dígito, sinaliza-se o capítulo; do segundo dígito, obtém-se o grupo e a categoria; do último dígito, obtém-se a subcategoria. Na Figura 16 ilustra-se a organização do CID-10, apresentando alguns códigos como exemplos.

Figura 16: Estrutura hierárquica dos códigos da CID.



Os códigos correspondentes à área de Neoplasia pertencem ao Capítulo II, que compreende a faixa de *C000* até *D489*. Esse capítulo organiza-se nos seguintes grupos:

- **(C00 - C14)** - Neoplasias malignas do lábio, cavidade oral e faringe.
- **(C15 - C26)** - Neoplasias malignas dos órgãos digestivos.
- **(C30 - C39)** - Neoplasias malignas do aparelho respiratório e dos órgãos intratorácicos.
- **(C40 - C41)** - Neoplasias malignas dos ossos e das cartilagens articulares.
- **(C43 - C44)** - Melanoma e outras(os) neoplasias malignas da pele.
- **(C45 - C49)** - Neoplasias malignas do tecido mesotelial e tecidos moles.

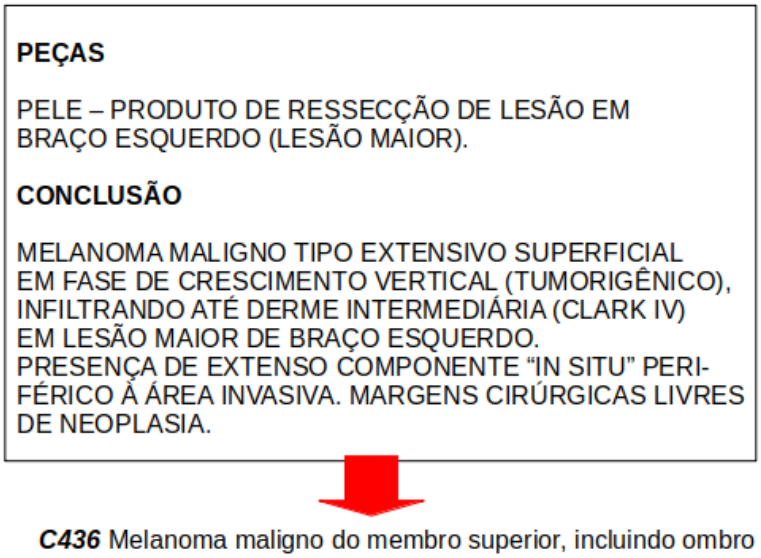
- **(C50 - C50)** - Neoplasias malignas da mama.
- **(C51 - C58)** - Neoplasias malignas dos órgãos genitais femininos.
- **(C60 - C63)** - Neoplasias malignas dos órgãos genitais masculinos.
- **(C64 - C68)** - Neoplasias malignas do trato urinário.
- **(C69 - C72)** - Neoplasias malignas dos olhos, do encéfalo e de outras partes do sistema nervoso central.
- **(C73 - C75)** - Neoplasias malignas da tireoide e de outras glândulas endócrinas.
- **(C76 - C80)** - Neoplasias malignas de localizações mal definidas, secundárias e de localizações não especificadas.
- **(C81 - C96)** - Neoplasias [tumores] malignas(os), declaradas ou presumidas como primárias, dos tecidos linfático, hematopoético e tecidos correlatos.
- **(C97 - C97)** - Neoplasias malignas de localizações múltiplas independentes (primárias).
- **(D00 - D09)** - Neoplasias [tumores] *in situ*.
- **(D10 - D36)** - Neoplasias [tumores] benignas(os).
- **(D37 - D48)** - Neoplasias [tumores] de comportamento incerto ou desconhecido.

Pode-se observar que a maioria dos grupos estão relacionados com a localização anatômica do tumor. Portanto, a informação da peça constante no laudo anatomopatológico é fundamental para a detecção da doença.

Atualmente, a codificação de doenças é manual e feita por especialistas do hospital. Os códigos principais são dados pelo médico no momento da consulta do paciente. Os secundários são dados após a alta dos pacientes, por especialistas que revisam os resultados dos laudos anatomopatológicos, especificamente as seções de peças e conclusões. Essa é uma tarefa repetitiva e, dependendo do volume de laudos a serem avaliados, muito cansativa e propensa a enganos. Assim, aplicando-se o método proposto neste trabalho para codificação automática de doenças, contribuiria-se para maior rapidez na avaliação dos laudos e, consequentemente, na melhora da produtividade dos profissionais envolvidos com essa tarefa.

Para melhor entendimento da codificação, a Figura 17 exibe um exemplo de um laudo anatomopatológico e o código identificado.

Figura 17: Exemplo codificação CID.



5.3 Preparação dos dados e tratamento dos textos

Os laudos oncológicos são gerados e armazenados em formato PDF no banco de dados do HCor. Para seu emprego por diferentes sistemas computacionais, o hospital desenvolveu um *software* que cria uma tabela com o conteúdo dos arquivos, separando as seções do laudo em colunas e inserindo o código hospital do paciente e a data do procedimento. A partir de preenchimento manual, é criada outra tabela que possui os códigos de doenças e data de atendimento, atrelado ao código do paciente. Assim, a união de ambas tabelas permite relacionar os laudos com os códigos de doenças.

Uma vez que, apenas em um atendimento podem ser realizados vários procedimentos e identificados múltiplos códigos de doenças, inclusive com CID de outros capítulos, foi necessário efetuar uma validação dos registros, excluindo os registros não pertencentes ao capítulo de neoplasia. Também foram corrigidos manualmente os códigos descontinuados na CID. Ambos tratamentos de dados é conhecido como redução de *label noise*. Como resultado, foram obtidos 115 laudos, com seus respectivos códigos de doenças.

Tratamento do texto

Para tratar os textos, empregaram-se técnicas comumente utilizadas para AM, tendo em vista os problemas de QD citados no capítulo 2. Assim, para o tratamento dos textos de laudos de procedimentos foram empregadas as técnicas a seguir:

1. Tokenização;
2. Redução de *noise data*;
3. Avaliação da acurácia e melhoria da acessibilidade do texto;
4. Diminuição da complexidade do texto.

Tokenização: O primeiro passo no processo de tratamento foi a tokenização, o qual consiste em separar um texto em unidades chamadas tókens. Os tókens podem ser números, palavras ou datas, sendo separados pelo delimitador "|". Na Figura 18(a) exibe-se o texto do laudo e na Figura 18(b) mostra-se o resultado da tokenização.

Figura 18: Exemplo Tokenização.

(a) Texto do laudo

Mucosa gástrica apresentando alongamento hiperplásico irregular de fovéolas, sem displasia.
Observa-se discreto infiltrado inflamatório linfo-plasmocitário (Ver nota).

(b) Tókens

Mucosa | gástrica | apresentando | alongamento | hiperplásico | irregular | de | fovéolas | sem | displasia | Observa-se | discreto | infiltrado | inflamatório | linfo-plasmocitário | ver | nota

Redução de *Noise Data*: Após a tokenização, removeram-se todos os tókens que não são consideradas como palavras, com o intuito de reduzir o *noise data*. Consideram-se como palavras aqueles tókens que contém apenas letras do alfabeto português, além daquelas que contém um dos seguintes caracteres: **-** (hífen), ***0*** (zero), ***1*** (número um) ou ***6*** (número seis), devido à presença de palavras compostas que contém hífen e palavras que na conversão do PDF para tabela, foram alteradas. Essas alterações consistiram de substituições de algumas letras pelos números 0, 1 ou 6. Exemplos dessas palavras alteradas podem ser observadas na Tabela 11.

Tabela 11: Palavras alteradas no processo de transformação do texto PDF para tabelas.

Em laudo	Correto
a0rtic	aórtica
adenov1rus	adenovírus
cutane0	cutâneo
distr0fica	distrofica
fundo-pil6rica	fundo-pilórica

Avaliação e correção da Acurácia: A *acurácia* dos textos foi medida calculando-se a média da porcentagem de palavras corretas por cada laudo. Para esse propósito, empregou-se a ferramenta *hunspell*[1], a qual detecta palavras mal escritas e, para algumas, sugere palavras similares para a correção.

Para melhorar a *acurácia* e a acessibilidade do texto para aplicar o modelo de AM, desenvolveu-se um programa que seleciona uma das sugestões de *hunspell* para substituir a palavra errada. Para tal, o programa implementa uma heurística que calcula a distância das palavras, ponderando de diferentes maneiras os erros de acentos, hifens e os números identificados na conversão do PDF (ver Tabela 11). Por exemplo, para a correção da palavra ***a0rtic***, *hunspell* sugere as seguintes palavras: 'artice', 'articida', 'articito', 'articite', 'articula', 'aórtico', 'urtica', 'aritica', 'anórtico', 'agírtico', 'pértica', 'artético'. A heurística seleciona como correta a opção ***aórtico*** por considerar o ***0*** similar à ***ó*** e calcular a distância igual a 1. Após a correção do texto aplicando o programa desenvolvido, melhorou-se a acurácia em 5%. Na Tabela 12 apresenta-se a *acurácia* do texto antes e após a correção.

Tabela 12: Acurácia dos laudos.

	Acurácia
Antes da correção	92,19 %
Após correção	97,06 %

Através da redução de *noise data* e a melhora na acurácia, tornou-se o texto mais *acessível*, já que obteve-se um texto com vocabulário bem definido e sem tókens que contribuíssem com pouca informação para o modelo.

Diminuição da complexidade do texto: Os fatores que aumentam a **complexidade do texto** dos laudos, dificultando a codificação automática, são:

- o **vocabulário extenso**, que reflete-se em uma grande quantidade de preditores nos dados empregados para treinar o modelo, complicando a detecção das palavras chaves para a codificação;
- as **diferentes inflexões de palavras com o mesmo significado**, como por exemplo, fúndic**o** e fúndic**a**. Os modelos não são capazes de discernir que ambas palavras

[1]http://hunspell.github.io/ é um verificador ortográfico de vários *software* reconhecidos como LibreOffice, OpenOffice, Mozilla Firefox, Google Chrome e Thunderbird

representam o mesmo conceito, ao menos que seja realizado um pré-processamento. Assim, palavras escritas com diferentes inflexões aumentam o tamanho do vocabulário e perdem peso ao serem ponderadas no treinamento.

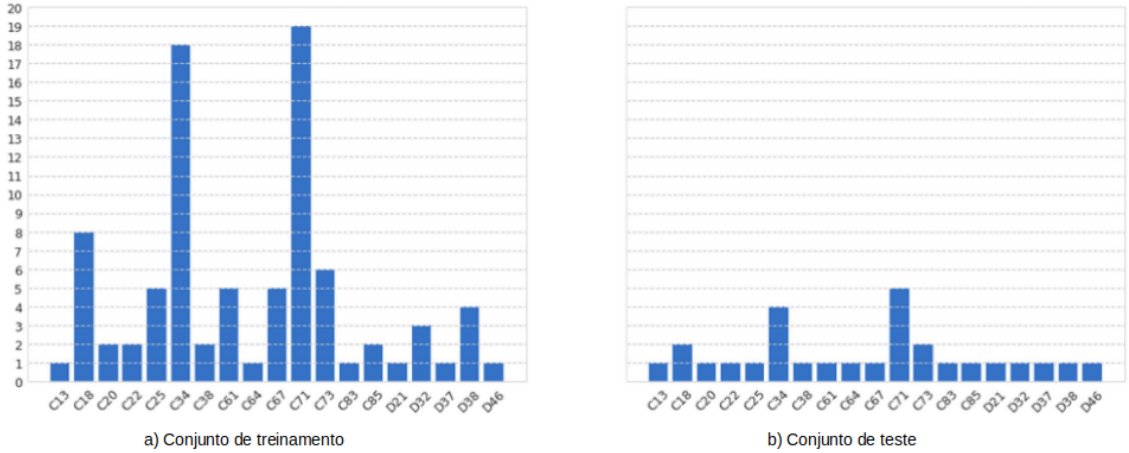
- as **orações longas**, nas quais as relações entre as palavras ganham importância e diminuem a **legibilidade** do texto. Existem algoritmos de AM que capturam essas relações nos textos e, com uma grande quantidade de dados, conseguem melhores resultados que os algoritmos que não consideram tais relações.

Os problemas do vocabulário extenso e diferenças de inflexão foram tratados realizando o *stemming* das palavras, empregando também a ferramenta *hunspell*. O *stemming* é um tratamento comumente usado nas tarefas de classificação de texto, o qual reduz as palavras aos seus radicais. Dessa maneira, consegue-se diminuir o vocabulário e eliminar as diferenças nas inflexões. O tratamento de orações longas e melhora da *legibilidade* necessita de uma grande quantidade de dados, o que não foi possível obter neste trabalho. Assim, não realizaram-se esses tratamentos e empregou-se um modelo de classificação que não considera as relações entre palavras de uma mesma frase.

5.4 Elaboração dos experimentos

Os 115 laudos codificados foram divididos em dois conjuntos: (i) conjunto de treinamento (80% dos laudos); (ii) conjunto de teste (20% dos laudos). Na separação dos conjuntos, procurou-se manter a mesma distribuição dos rótulos em cada amostra, obrigando a ter pelo menos um rótulo de cada código em ambos conjuntos. Como resultado, obtiveram-se os dois conjuntos como apresentados na Figura 19.

Figura 19: Conjuntos de treinamento e teste.



Desbalanceamento de Classes

Um dos problemas encontrados nos dados de treinamento foi a distribuição não uniforme dos rótulos do Capítulo II do CID-10, o que afetaria a fase de treinamento do modelo. Assim, o problema do *desbalanceamento de classes* foi mitigado aplicando-se técnicas de sub-amostragem e sobre-amostragem.

A sobre-amostragem consiste no incremento de rótulos menos frequentes, usando dados de outras fontes ou gerando dados artificiais. Por outro lado, a sub-amostragem consiste na exclusão aleatória de registros das classes que são muito frequentes na amostra. Ambas técnicas são realizadas de forma a obter uma distribuição balanceada (ou uniforme). A sub-amostragem é um processo menos complicado, uma vez que não requer dados externos que apresentem a mesma estrutura dos dados disponíveis. Para a avaliação das distribuições obtidas, normalmente emprega-se algum indicador estatístico, como o teste de uniformidade Chi-quadrado. Devido à baixa quantidade de laudos disponíveis, neste trabalho aplicaram-se ambas as técnicas.

Para gerar dados artificiais, simularam-se laudos a partir da descrição dos códigos de doenças do CID-10, através de um programa desenvolvido em colaboração com especialistas. Para isso, extraiu-se o nome das peças constante no CID-10 e o conteúdo da conclusão do laudo foi escrito usando o texto da descrição do código da doença. Por exemplo, com o código *C17 - Neoplasia maligna do intestino delgado*, foi simulado um laudo com

- **Peça:** Intestino delgado e
- **Conclusão:** Neoplasia maligna do intestino delgado.

Os passos do processo de re-amostragem executado é descrito a seguir:

1. Criar o conjunto de dados S, unindo os laudos de treinamento reais com os laudos simulados;
2. Calcular a frequência de cada código no conjunto S;
3. Calcular o Chi-quadrado, comparando a distribuição de S com uma distribuição uniforme,

   3.1. se o p-*value* $\leq \alpha = 0,1$, os dados estão balanceados e o processo é finalizado[2].

[2]Onde α é o nível de significância.

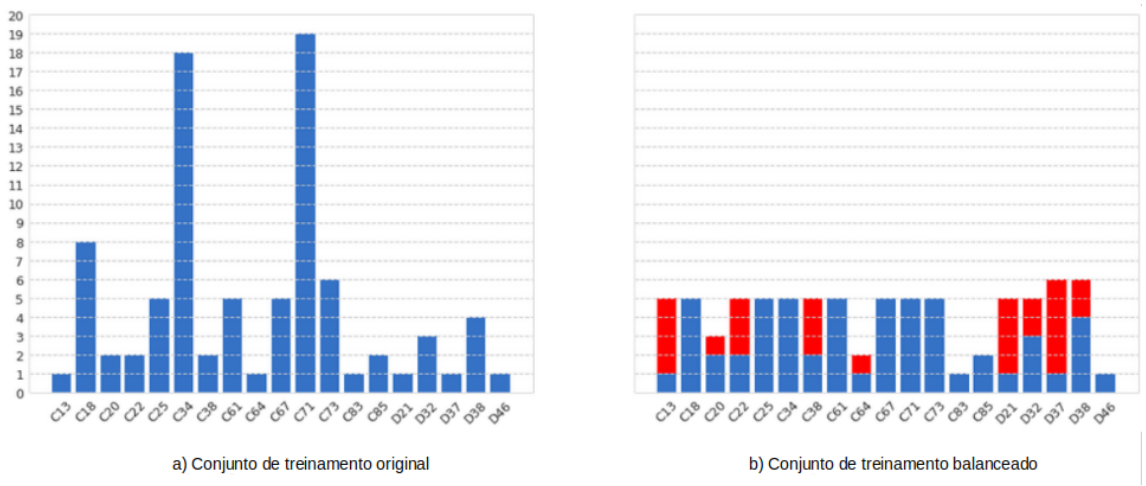
3.2. se p-*value* $> \alpha = 0,1$, escolher a classe mais frequente e remover laudos pseudo-aleatoriamente, até que se tenha a frequência esperada, dada por $1/n_{codigos}$,

4. Repetição do processo a partir item 2.

No passo 3.2, inicialmente é realizada a a remoção dos laudos simulados e posteriormente, se necessário, remove-se laudos reais. Essa ordem é aplicada porque os laudos reais são mais representativos para o treinamento.

Na Figura 20 ilustra-se o resultado do balanceamento. No lado esquerdo, tem-se a distribuição real do conjunto de treinamento, na qual observam-se classes como a C34 e C71 que aparecem na amostra com maior frequência. No lado direito, tem-se a distribuição após o balanceamento, na qual os laudos simulados são representados pelos quadrados vermelhos.

Figura 20: Balanceamento das classes.



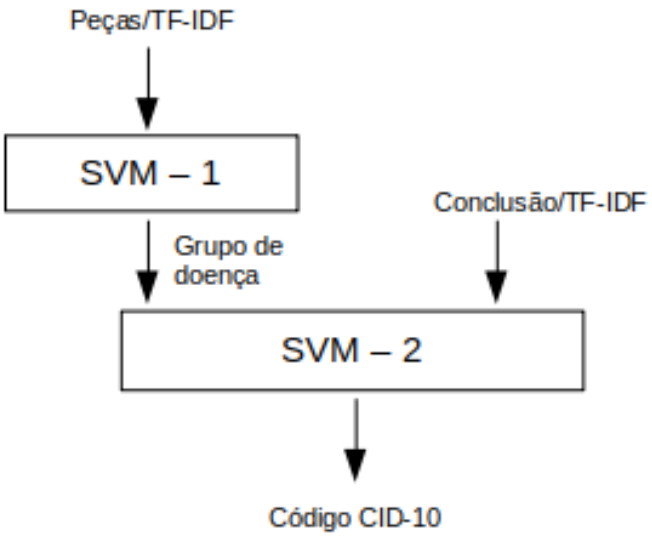
Modelos de AM

O algoritmo empregado para a codificação automática foi a SVM, devido à sua robustez para problemas de classificação de texto. Essa escolha foi pautada nos resultados da revisão da literatura.

Assim, a partir da aplicação de SVM, desenvolveram-se dois modelos. O primeiro, denominado **SVM-clássica**, foi treinado para predizer os primeiros três caracteres do código da doença, fazendo uso dos textos das conclusões representados como vetores de TF-IDF. O segundo modelo foi formado por duas SVM, organizadas sequencialmente, permitindo explorar a estrutura hierárquica dos códigos do CID. A Figura 21 ilustra a arquitetura desse modelo, denominado **SVM-hierárquica**. Como pode-se observar na

Figura 21, a primeira SVM (**SVM-1**) recebe como entrada os vetores TF-IDF dos textos das peças e produz como saída o grupo da doença. Já a segunda SVM (**SVM-2**) recebe o grupo da saída da SVM-1, representado como um vetor de valores binários, atrelado ao vetor TF-IDF das conclusões, gerando como saída os três primeiros caracteres código final da doença.

Figura 21: SVM hierárquica.

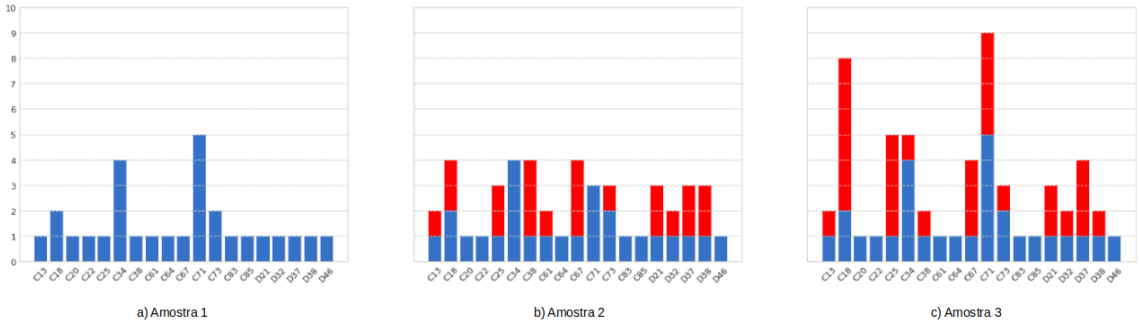


Com o objetivo de validar a robustez do modelo quanto à distribuição do conjunto de teste, produziram-se três tipos de amostras:

- **Amostra 1:** manteve-se a distribuição original dos laudos. Neste cenário, valida-se a capacidade do modelo responder aos casos típicos do hospital.
- **Amostra 2:** utilizou-se a distribuição balanceada. É um cenário utópico, o qual permite medir o desempenho dos modelos sem viés nos códigos.
- **Amostra 3:** gerou-se uma distribuição aleatória, excluindo alguns registros do conjunto teste e incluindo outros dos laudos simulados que não foram empregados no treinamento. O propósito foi validar a capacidade do modelo responder com dados desbalanceados que são diferentes dos dados originais.

Na Figura 22 ilustra-se as distribuições das três amostras. Tanto na amostra 2 quanto na amostra 3, incluíram-se laudos simulados, os quais são representados pelos quadrados vermelhos.

Figura 22: Amostras para teste.



As métricas empregadas para a avaliação dos modelos foram a **macro precisão, macro *recall*** e **macro F1-*score***, por serem capazes de avaliar o comportamento dos modelos para a predição de múltiplas classes, atribuindo o mesmo peso a cada uma delas.

5.5 Avaliação dos resultados

Com o objetivo de analisar o efeito do tratamento do texto nos modelos, compararam-se o desempenho dos modelos treinados com três versões dos textos, explicadas a seguir:

- ***Txt-v1***: textos reduzidos a tókens e sem *noise data*;
- ***Txt-v2***: textos reduzidos a tókens, sem *noise data* e com correção de acurácia;
- ***Txt-v3***: textos reduzidos a tókens, sem *noise data*, com correção de acurácia e *stemming*.

O ajuste do hiper-parâmetro C e o *kernel* dos modelos foi realizado aplicando-se *grid search* com validação cruzada. Treinaram-se múltiplos modelos variando o valores de C e do *kernel*, dividindo o conjunto de treinamento em 5 partes (*folds*). Para cada combinação de hiper-parâmetros, calcularam-se as métricas de desempenho em cada *fold* e selecionou-se a combinação que alcançava um valor médio de **macro F1-score** mais alto. Os valores testados para os hiper-parâmetros são apresentados a seguir:

- *Kernel*: Linear e *Radial Basis Function* (RBF).
- C: [0,1; 0,2; 0,3; ... ; 0,9]∪[1; 2; 3; ... ; 9]∪[10; 20; 30; ... ; 90]∪[100; 200; 300; ... ; 900].

Na Tabela 13 apresenta-se as combinações de hiper-parâmetros selecionadas para cada versão do texto.

Tabela 13: Hiper-parâmetros selecionados por modelo.

Modelo	Hiper-parâmetro	Txt-v1	Txt-v2	Txt-v3
SVM-clássica	C	3	2	5
	Kernel	Linear	Linear	Linear
SVM-hierárquica Level 1	C	3	2	2
	Kernel	Linear	Linear	Linear
SVM-hierárquica Level 2	C	2	2	2
	Kernel	Linear	Linear	Linear

Na Tabela 14 apresentam-se as métricas dos modelos para cada amostra e para cada etapa do tratamento do texto. Evidencia-se que, em todos os casos, os modelos têm melhor desempenho quando treinados com textos nos quais foram aplicados maior quantidade de tratamentos (Txt-v3).

Tabela 14: Desempenho dos modelos.

Modelo	Amostra	Métrica	Tratamento do texto		
			Txt-v1	Txt-v2	Txt-v3
SVM-clássica	Amostra 1	Precisão	0,4474	0,4684	**0,7412**
		Recall	0,4947	0,5474	**0,7895**
		F1	0,4472	0,4754	**0,7404**
	Amostra 2	Precisão	0,4302	0,4551	**0,6883**
		Recall	0,4254	0,4781	**0,6096**
		F1	0,3840	0,4169	**0,6174**
	Amostra 3	Precisão	0,4478	0,4611	**0,6715**
		Recall	0,4056	0,4845	**0,6235**
		F1	0,3676	0,3967	**0,6125**
SVM-hierárquica	Amostra 1	Precisão	0,5921	0,5921	**0,8947**
		Recall	0,6579	0,6579	**0,9211**
		F1	0,6000	0,6000	**0,8930**
	Amostra 2	Precisão	0,5623	0,5623	**0,7751**
		Recall	0,6009	0,6009	**0,7500**
		F1	0,5654	0,5654	**0,7214**
	Amostra 3	Precisão	0,5372	0,5372	**0,7570**
		Recall	0,5700	0,5700	**0,7727**
		F1	0,5186	0,5186	**0,7207**

Nas Figuras 23, 24 e 25 ilustram-se a comparação dos modelos SVM-clássico e SVM-hierárquico. Observa-se que o modelo hierárquico obtém melhores resultados, incrementando em até 0,15 as métricas.

Figura 23: Desempenho dos modelos com a Amostra 1.

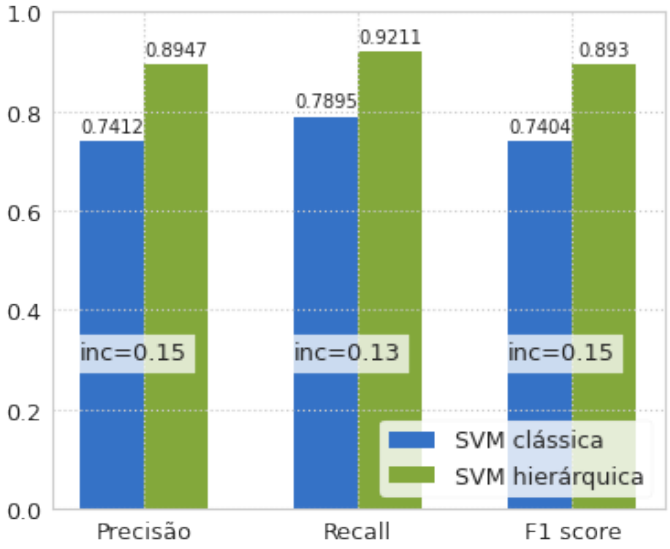


Figura 24: Desempenho dos modelos com a Amostra 2.

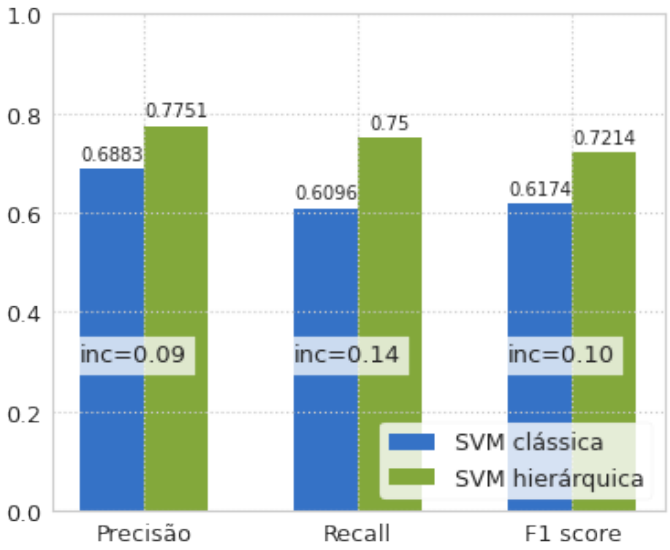
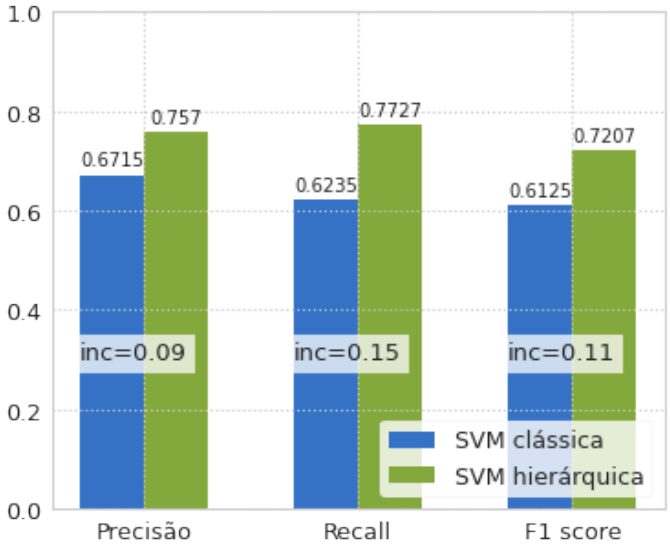


Figura 25: Desempenho dos modelos com a Amostra 3.



Como esperado, o tratamento que resulta em Txt-v3, aplicado aos textos com ***noise data***, baixa **acurácia** e alta **complexidade**, impacta positivamente no desempenho dos modelos. O comportamento constante nas métricas dos modelos nas diferentes amostras demostram que, o balanceamento efetuado nos dados são eficazes para evitar o viés dos modelos para rótulos muito frequentes. Finalmente, comprova-se que, explorar a estrutura dos códigos para o desenvolvimento de um modelo hierárquico melhora significativamente o desempenho da classificação.

6 CONSIDERAÇÕES FINAIS

6.1 Conclusões

CID é um padrão que organiza as doenças em grupos, categorias e subcategorias, empregando para essa finalidade um sistema de códigos constituídos por quatro caracteres (uma letra e três números). Esse código de doença, atrelados a dados clínicos e demográficos do paciente, são utilizados para mensurar a complexidade dos procedimentos.

Atualmente, o processo da codificação de doenças de diagnósticos secundários é realizado por especialistas, que revisam os laudos de procedimentos médicos, lista de medicamentos e outras informações com este objetivo. Devido à grande quantidade de pacientes e dados médicos, essa codificação secundária é uma tarefa que demanda muito tempo dos profissionais especializados de saúde, repetitiva e propensa a erros. A fim de diminuir a carga desse tipo de trabalho sobre os profissionais responsáveis e, assim, melhorar a produtividade dos mesmos, este trabalho propôs a codificação automática para identificar os CID secundários. Para tal, pesquisou-se as técnicas de AM que pudessem ser aplicadas ao problema. Optou-se, então, pelo SVM, devido ao seu bom desempenho quando aplicado em baixa quantidade de dados, conforme evidenciado na literatura. Assim, desenvolveu-se um modelo de AM que identifica os códigos de doenças correspondentes ao capítulo de neoplasia da CID-10, tendo como entrada informações dos laudos anatomopatológicos. Esse modelo realiza a codificação de doença de forma automática, gerando resultados com uma precisão e *recall* acima de 70%.

Pesquisas prévias sobre a codificação automática afirmaram que, um dos principais desafios no desenvolvimento do modelo foi causado pela baixa QD. Portanto, inicialmente, foi elaborada uma revisão da literatura para identificar os problemas de QD de registros médicos, na qual concluiu-se que não existe consenso nas atividades de avaliação e melhora da QD. A melhoria da QD pode ser de curto ou longo prazo, aplicando estratégias de *data-driven* e/ou *process-driven*. Neste trabalho, optou-se pela estratégia de *data-driven*, que modifica diretamente o valor dos dados com atividades corretivas. Vale mencionar que, para o tratamento da causa-raíz dos problemas de QD (*process-driven*), seria necessário

intervir nos processos do hospital, o que está fora do escopo deste trabalho. Para complementar, pesquisou-se bibliografia relacionada com problemas de qualidade de textos e atividades de tratamento de dados para desenvolvimento de modelos de classificação. Como resultado, verificou-se as dimensões para a qualidade de textos mais estudadas na literatura são: *acurácia*, *legibilidade* e *acessibilidade*. Para a melhoria da *acurácia* e *acessibilidade* foram aplicadas as atividades de correção de erros nos textos dos laudos e exclusão de *noise data*, respectivamente. Já a *legibilidade*, que está relacionada com o tamanho das frases do texto (orações longas), não foi tratada, dado que o modelo desenvolvido recebe como entrada palavras e não frases. Assim, essa dimensão não impacta no desempenho do modelo utilizado.

Dos problemas encontrados nos dados, o *label noise* foi tratado através da exclusão e substituição de registros não válidos, com ajuda de profissionais da área. Para corrigir o *desbalanceamento de classes*, efetuaram-se atividades de re-amostragem (sub e sobre-amostragem), empregando laudos simulados que foram gerados a partir das descrições dos códigos CID-10. Desse modo, obteve-se um modelo robusto que tem o mesmo desempenho para diferentes tipos de distribuição dos dados. Adicionalmente, tratamentos de textos comuns na literatura de AM foram aplicados. Para a redução de *noise data*, efetuou-se a tokenização e exclusão de tokéns que não foram considerados como palavras corretas. Para diminuição da *complexidade do texto*, realizou-se o *stemming*. Esse tratamento reduz as palavras aos seus radicais e por conseguinte, diminui o tamanho do vocabulário, eliminando as inflexões.

Além do modelo SVM de classificação com arquitetura clássica, foi proposta uma arquitetura hierárquica, a qual explora o agrupamento lógico de códigos de neoplasias por localização do corpo humano e a estrutura particular dos laudos médicos. Esse modelo hierárquico foi composto por duas SVM organizadas sequencialmente. A primeira recebia como entrada a *peça* analisada no exame médico e produzia o grupo da doença. A segunda SVM recebia o grupo da doença, junto à conclusão constante no laudo e gerava o código final da doença. Ambos modelos, o clássico e o hierárquico, foram avaliados com as métricas de *macro precisão*, *macro recall* e *macro F1-score*. Essas métricas foram calculadas sobre três amostras de teste com diferentes distribuições, de forma a validar a robustez dos modelos. Os resultados mostraram um acréscimo de, em média, 12% nas métricas do modelo hierárquico em relação ao modelo clássico.

O desempenho dos modelos também foi avaliado sobre distintos níveis de tratamento dos textos. Observou-se que, como esperado, os modelos têm melhor desempenho quando treinados com textos nos quais foram aplicados maior quantidade de tratamentos (Txt-

v3).

Finalmente, comprovou-se que, a qualidade dos textos tem uma grande influência no desempenho do modelo e que, explorar estratégias relacionadas a particularidades dos dados podem ser uma alavanca na codificação automática.

6.2 Limitações e trabalhos futuros

A principal limitação da pesquisa foi a pouca quantidade de dados. Apesar da disponibilidade de 7000 laudos anatomopatológicos e 35000 de atendimentos, para o treinamento dos modelos foram utilizados apenas 115 laudos codificados. Consequentemente, a escolha do algoritmo de AM limitou-se à SVM, não sendo possível explorar as vantagens de modelos mais complexos, como as redes neurais.

Adicionalmente, o uso de documentos com formato PDF para o armazenamento dos laudos dificulta a extração dos textos e obriga ao uso de ferramentas de conversão de arquivos que introduzem problemas de qualidade. Assim, sugere-se capturar os resultados dos procedimentos em tabelas estruturadas ou textos semiestruturados antes do armazenamento em PDF.

Os resultados obtidos desta pesquisa, podem ser empregados no hospital para o tratamento automático do texto no momento da geração de laudos e como um sistema de apoio na codificação de doenças.

Como trabalho futuro, pode-se analisar a estrutura e agrupamento dos códigos de outros capítulos da CID e adequar o modelo hierárquico para a codificação automática. Além disso, tendo uma maior quantidade de dados, é possível desenvolver modelos mais complexos, como os modelos baseados em redes neurais, e incluir a dimensão de *legibilidade* na análise do impacto dos problemas de QD.

REFERÊNCIAS

ABIY, R. et al. A comparison of electronic records to paper records in Antiretroviral Therapy Clinic in Ethiopia: What is affecting the Quality of the Data? *Online Journal of Public Health Informatics*, v. 10, n. 2, set. 2018. ISSN 1947-2579. Disponível em: ⟨https://journals.uic.edu/ojs/index.php/ojphi/article/view/8309⟩.

AGGARWAL, C. C. *Machine Learning for Text*. Cham: Springer International Publishing, 2018. 1–493 p. ISBN 978-3-319-73530-6.

AGGARWAL, C. C. *Neural Networks and Deep Learning*. Cham: Springer International Publishing, 2018. 389–411 p. ISBN 978-3-319-94462-3.

ALI, S. et al. Data Quality: A Negotiator between Paper-Based and Digital Records in Pakistan's TB Control Program. *Data*, v. 3, n. 3, p. 27, jul. 2018. ISSN 2306-5729. Disponível em: ⟨http://www.mdpi.com/2306-5729/3/3/27⟩.

ALNAJRAN, N. et al. A heuristic based pre-processing methodology for short text similarity measures in microblogs. In: . IEEE, 2018. p. 1627–1633. ISBN 978-1-5386-6614-2. Proposes a methodology to preprocess Twitter text for similarity measurements. Disponível em: ⟨https://ieeexplore.ieee.org/document/8623003/⟩.

ALUISIO, S. et al. Readability assessment for text simplification. *ITL - International Journal of Applied Linguistics*, v. 165, p. 194–222, 12 2014. ISSN 0019-0829. Disponível em: ⟨http://www.jbe-platform.com/content/journals/10.1075/itl.165.2.04vaj⟩.

AZAM, S. S. et al. Cascadenet: An lstm based deep learning model for automated icd-10 coding. In: ARAI, K.; BHATIA, R. (Ed.). *Advances in Information and Communication*. Cham: Springer International Publishing, 2020. p. 55–74. ISBN 978-3-030-12385-7.

BADR, N. Guidelines for Health IT Addressing the Quality of Data in EHR Information Systems. In: *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies*. SCITEPRESS - Science and Technology Publications, 2019. p. 169–181. ISBN 978-989-758-353-7. Disponível em: ⟨http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0006941001690181⟩.

BAE, C. J. et al. The Challenges of Data Quality Evaluation in a Joint Data Warehouse. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, v. 3, n. 1, p. 12, maio 2015. ISSN 2327-9214.

BARBOSA, W. L. et al. Data quality problems identified in the bioclimatic data collection process - a survey. In: *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*. [S.l.: s.n.], 2019. p. 1–7.

BATINI, C. et al. Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, v. 41, n. 3, p. 1–52, jul. 2009. ISSN 0360-0300.

BATINI, C.; SCANNAPIECO, M. *Data and Information Quality*. Cham: Springer International Publishing, 2016. 16–1–16–20 p. (Data-Centric Systems and Applications). ISBN 978-3-319-24104-3.

BISHOP, C. M. *Pattern recognition and machine learning*. [S.l.]: Springer, 2006. ISSN 1613-9011. ISBN 978-0387-31073-2.

BOUTELL, M. R. et al. Learning multi-label scene classification. *Pattern Recognition*, v. 37, n. 9, p. 1757–1771, set. 2004. ISSN 0031-3203. Disponível em: ⟨https://linkinghub.elsevier.com/retrieve/pii/S0031320304001074⟩.

CARSLEY, S. et al. Completeness and accuracy of anthropometric measurements in electronic medical records for children attending primary care. *BMJ Health & Care Informatics*, v. 25, n. 1, p. 19–26, jan. 2018. ISSN 2632-1009.

COLLINS, E.; ROZANOV, N.; ZHANG, B. Evolutionary data measures: Understanding the difficulty of text classification tasks. In: . Association for Computational Linguistics, 2018. p. 380–391. ISBN 9781948087728. Disponível em: ⟨http://arxiv.org/abs/1811.01910http://aclweb.org/anthology/K18-1037⟩.

DANIEL, C. et al. Initializing a hospital-wide data quality program. The AP-HP experience. *Computer Methods and Programs in Biomedicine*, Elsevier B.V., v. 181, n. xxxx, p. 104804, nov. 2019. ISSN 0169-2607. Disponível em: ⟨https://linkinghub.elsevier.com/retrieve/pii/S0169260718306242⟩.

De Coster, C. et al. Identifying priorities in methodological research using ICD-9-CM and ICD-10 administrative data: report from an international consortium. *BMC Health Services Research*, v. 6, n. 1, p. 77, dez. 2006. ISSN 1472-6963.

DENG, X. et al. From descriptive to diagnostic analytics for assessing data quality: An application to temporal data elements in electronic health records. In: *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2016. p. 236–239. ISBN 978-1-5090-2455-1. Disponível em: ⟨http://ieeexplore.ieee.org/document/7455878/⟩.

DZIADKOWIEC, O. et al. Using a Data Quality Framework to Clean Data Extracted from the Electronic Health Record: A Case Study. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, v. 4, n. 1, p. 11, jun. 2016. ISSN 2327-9214.

EHSANI-MOGHADDAM, B.; MARTIN, K.; QUEENAN, J. A. Data quality in healthcare: A report of practical experience with the Canadian Primary Care Sentinel Surveillance Network data. *Health Information Management Journal*, p. 183335831988774, dez. 2019. ISSN 1833-3583.

EISENSTEIN, J. *Introduction to Natural Language Processing*. [S.l.]: The MIT Press, 2018.

ESTIRI, H. et al. A federated EHR network data completeness tracking system. *Journal of the American Medical Informatics Association*, v. 26, n. 7, p. 637–645, jul. 2019. ISSN 1527-974X. Disponível em: ⟨https://academic.oup.com/jamia/article/26/7/637/5423491⟩.

ESTIRI, H. et al. Exploring completeness in clinical data research networks with DQe-c. *Journal of the American Medical Informatics Association*, v. 25, n. 1, p. 17–24, jan. 2018. ISSN 1527-974X. Disponível em: ⟨https://academic.oup.com/jamia/article/25/1/17/4562678⟩.

FEDER, S. L. Data Quality in Electronic Health Records Research: Quality Domains and Assessment Methods. *Western Journal of Nursing Research*, v. 40, n. 5, p. 753–766, maio 2018. ISSN 0193-9459.

FORD, E. et al. What evidence is there for a delay in diagnostic coding of RA in UK general practice records? An observational study of free text. *BMJ Open*, v. 6, n. 6, p. e010393, jun. 2016. ISSN 2044-6055.

FOX, F. et al. A Data Quality Framework for Process Mining of Electronic Health Record Data. In: *2018 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 2018. p. 12–21. ISBN 978-1-5386-5377-7. Disponível em: ⟨https://ieeexplore.ieee.org/document/8419262/⟩.

FRANCISCO, M. M. C. et al. Total Data Quality Management and Total Information Quality Management Applied to Costumer Relationship Management. In: *Proceedings of the 9th International Conference on Information Management and Engineering - ICIME 2017*. New York, New York, USA: ACM Press, 2017. p. 40–45. ISBN 9781450353373. Disponível em: ⟨http://dl.acm.org/citation.cfm?doid=3149572.3149575⟩.

FUNDAP. *Livro do aluno oncologia*. 1ra edição. ed. Sao Paulo: TecSaúde, 2012. 272 p. ISBN 978-85-7285-137-4.

GARTNER, D. et al. Machine Learning Approaches for Early DRG Classification and Resource Allocation. *INFORMS Journal on Computing*, v. 27, n. 4, p. 718–734, nov. 2015. ISSN 1091-9856.

GLOWALLA, P.; SUNYAEV, A. Process-driven data quality management: A critical review on the application of process modeling languages. *J. Data and Information Quality*, Association for Computing Machinery, New York, NY, USA, v. 5, n. 1-2, sep 2014. ISSN 1936-1955. Disponível em: ⟨https://doi.org/10.1145/2629568⟩.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. ISBN 9780262035613.

GRAESSER, A. C. et al. Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, v. 36, p. 193–202, 5 2004. ISSN 0743-3808. Disponível em: ⟨http://link.springer.com/10.3758/BF03195564⟩.

GUPTA, N. et al. Data quality for machine learning tasks. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. New York, NY, USA: Association for Computing Machinery, 2021. (KDD '21), p. 4040–4041. ISBN 9781450383325. Disponível em: ⟨https://doi.org/10.1145/3447548.3470817⟩.

HART, R.; KUO, M. H. Better data quality for better healthcare research results-a case study. *Studies in Health Technology and Informatics*, v. 234, p. 161–166, 2017. ISSN 1879-8365.

HARTZBAND, D.; JACOBS, F. Deployment of Analytics into the Healthcare Safety Net: Lessons Learned. *Online Journal of Public Health Informatics*, v. 8, n. 3, p. 1–13, dez. 2016. ISSN 1947-2579. Disponível em: ⟨http://journals.uic.edu/ojs/index.php/ojphi/article/view/7000https://journals.uic.edu/ojs/index.php/ojphi/article/view/7000⟩.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning*. New York, NY: Springer New York, 2009. (Springer Series in Statistics). ISBN 978-0-387-84857-0.

INCA. *ABC do câncer : abordagens básicas para o controle do câncer*. 6ta edição. ed. Rio de Janeiro: Instituto Nacional de Câncer José Alencar Gomes da Silva, 2020. 112 p. ISBN 9788573183948.

ITURRY, M. D. et al. Data quality in health records: A literature review. In: *2021 16th Iberian Conference on Information Systems and Technologies (CISTI)*. IEEE, 2021. p. 1–6. ISBN 978-989-54659-1-0. ISSN 21660735. Disponível em: ⟨https://ieeexplore.ieee.org/document/9476536/⟩.

JAIN, A. et al. Overview and importance of data quality for machine learning tasks. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. New York, NY, USA: Association for Computing Machinery, 2020. (KDD '20), p. 3561–3562. ISBN 9781450379984. Disponível em: ⟨https://doi.org/10.1145/3394486.3406477⟩.

JAMES, G. et al. *An Introduction to Statistical Learning*. New York, NY: Springer New York, 2013. v. 103. (Springer Texts in Statistics, v. 103). ISBN 978-1-4614-7137-0.

JOHNSON, S. G. et al. A Data Quality Ontology for the Secondary Use of EHR Data. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, v. 2015, p. 1937–46, 2015. ISSN 1942-597X. Disponível em: ⟨http://www.ncbi.nlm.nih.gov/pubmed/26958293http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4765682⟩.

JONES, M. et al. An Optimization Program to Help Practices Assess Data Quality and Workflow With Their Electronic Medical Records: Observational Study. *JMIR Human Factors*, v. 5, n. 4, p. e30, dez. 2018. ISSN 2292-9495. Disponível em: ⟨http://humanfactors.jmir.org/2018/4/e30/⟩.

KAHN, M. G. et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, v. 4, n. 1, p. 18, set. 2016. ISSN 2327-9214.

KAVULURU, R.; RIOS, A.; LU, Y. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artificial Intelligence in Medicine*, Elsevier B.V., v. 65, p. 155–166, 10 2015. ISSN 09333657. Disponível em: ⟨http://dx.doi.org/10.1016/j.artmed.2015.04.007https://linkinghub.elsevier.com/retrieve/pii/S0933365715000482⟩.

KITCHENHAM, B. Procedures for Performing Systematic Literature Reviews. *Keele, UK, Keele Univ.*, v. 33, 2004.

KUMAR, V. et al. *Robbins and Cotran pathologic basis of disease*. 7th editio. ed. [S.l.]: Philadelphia, Pa.: Elsevier Saunders, 2005. 1525 p. ISBN 0721601871.

LAURíA, E. J. M.; MARCH, A. D. Combining bayesian text classification and shrinkage to automate healthcare coding. *Journal of Data and Information Quality*, v. 2, p. 1–22, 12 2011. ISSN 1936-1955. Disponível em: ⟨https://dl.acm.org/doi/10.1145/2063504.2063506⟩.

LEE, J. M.; MUIS, A. O. *Diagnosis Code Prediction from Electronic Health Records as Multilabel Text Classification: A Survey*. 2017. Disponível em: ⟨http://people.cs.pitt.edu/~jlee/papers/cp1_survey_jlee_amuis.pdf⟩.

LEE, K.; WEISKOPF, N.; PATHAK, J. A Framework for Data Quality Assessment in Clinical Research Datasets. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, v. 2017, p. 1080–1089, 2017. ISSN 1942-597X. Disponível em: ⟨http://www.ncbi.nlm.nih.gov/pubmed/29854176http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5977591⟩.

LEE, S. J. C.; GROBE, J. E.; TIRO, J. A. Assessing race and ethnicity data quality across cancer registries and EMRs in two hospitals. *Journal of the American Medical Informatics Association*, v. 23, n. 3, p. 627–634, maio 2016. ISSN 1067-5027.

LEE, Y. et al. Validation for Accuracy of Cancer Diagnosis in Electronic Medical Records Using a Text Mining Method. *Studies in Health Technology and Informatics*, v. 216, p. 882, 2015.

LEÓN-CHOCANO, R. García-de et al. Construction of quality-assured infant feeding process of care data repositories: Construction of the perinatal repository (Part 2). *Computers in Biology and Medicine*, Elsevier, v. 71, p. 214–222, abr. 2016. ISSN 0010-4825.

LEÓN-CHOCANO, R. García-de et al. Construction of quality-assured infant feeding process of care data repositories: definition and design (Part 1). *Computers in Biology and Medicine*, v. 67, p. 95–103, dez. 2015. ISSN 0010-4825. Disponível em: ⟨https://linkinghub.elsevier.com/retrieve/pii/S0010482515003340⟩.

LI, M. et al. Automated icd-9 coding via a deep learning approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, v. 16, p. 1193–1202, 7 2019. ISSN 1545-5963. Disponível em: ⟨https://ieeexplore.ieee.org/document/8320340/⟩.

LINGREN, T. et al. Electronic medical records as a replacement for prospective research data collection in postoperative pain and opioid response studies. *International Journal of Medical Informatics*, Elsevier, v. 111, n. March 2017, p. 45–50, mar. 2018. ISSN 1386-5056.

LIU, C.; ZOWGHI, D.; TALAEI-KHOEI, A. An empirical study of the antecedents of data completeness in electronic medical records. *International Journal of Information Management*, Elsevier, v. 50, n. October 2018, p. 155–170, fev. 2020. ISSN 0268-4012.

MEDHANYIE, A. A. et al. Quality of routine health data collected by health workers using smartphone at primary health care in Ethiopia. *International Journal of Medical Informatics*, Elsevier Ireland Ltd, v. 101, p. 9–14, maio 2017. ISSN 1386-5056.

MIKOLOV, T. et al. Efficient Estimation of Word Representations in Vector Space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, p. 1–12, jan. 2013.

MITCHELL, T. M. *Machine learning*. [S.l.: s.n.], 1997. ISBN 0070428077.

MUTHEE, V. et al. The impact of routine data quality assessments on electronic medical record data quality in Kenya. *PLOS ONE*, v. 13, n. 4, p. e0195362, abr. 2018. ISSN 1932-6203.

NOBLES, A. L. et al. Evaluation of data quality of multisite electronic health record data for secondary analysis. In: *2015 IEEE International Conference on Big Data (Big Data)*. IEEE, 2015. p. 2612–2620. ISBN 978-1-4799-9926-2. Disponível em: ⟨http://ieeexplore.ieee.org/document/7364060/⟩.

NOSELLI, M. et al. MonAT: A Visual Web-based Tool to Profile Health Data Quality. In: *Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies*. SCITEPRESS - Science and Technology Publications, 2017. v. 5, n. Biostec, p. 26–34. ISBN 978-989-758-213-4. Disponível em: ⟨http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0006114200260034⟩.

OSMO, A. A. *Grupos Relacionados em Diagnósticos GRD (DRG – Diagnosis Related Groups): Conceitos, Estudos, Experiência de Implantação, Operação e Análises de Performance*. 2017. 51 p. Disponível em: ⟨http://www.ans.gov.br/images/stories/Particitacao_da_sociedade/2016_gt_remuneracao/4reuniao_2017_apresentacao_andre_osmo.pdf⟩.

PUTTKAMMER, N. et al. An assessment of data quality in a multi-site electronic medical record system in Haiti. *International Journal of Medical Informatics*, Elsevier Ireland Ltd, v. 86, p. 104–116, fev. 2016. ISSN 1386-5056. Disponível em: ⟨https://linkinghub.elsevier.com/retrieve/pii/S1386505615300551⟩.

PUTTKAMMER, N. et al. Identifying priorities for data quality improvement within Haiti's iSanté EMR system: Comparing two methods. *Health Policy and Technology*, Elsevier Ltd, v. 6, n. 1, p. 93–104, mar. 2017. ISSN 2211-8837.

READ, J. et al. Classifier Chains for Multi-label Classification. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. [s.n.], 2009. v. 5782 LNAI, n. PART 2, p. 254–269. ISBN 3642041736. Disponível em: ⟨http://link.springer.com/10.1007/978-3-642-04174-7_17⟩.

REIMER, A. P.; MILINOVICH, A.; MADIGAN, E. A. Data quality assessment framework to assess electronic medical record data for use in research. *International Journal of Medical Informatics*, Elsevier Ireland Ltd, v. 90, p. 40–47, jun. 2016. ISSN 1386-5056.

ROOS, L. L.; SHARP, S. M.; WAJDA, A. Assessing data quality: A computerized approach. *Social Science & Medicine*, v. 28, n. 2, p. 175–182, jan. 1989. ISSN 0277-9536. Disponível em: ⟨https://linkinghub.elsevier.com/retrieve/pii/0277953689901457⟩.

SCHAPIRE, R. E.; SINGER, Y. BoosTexter: A Boosting-based System for Text Categorization. *Machine Learning*, v. 2000, n. 2, p. 135–168, nov. 2000.

SCHOLTE, M. et al. Data extraction from electronic health records (EHRs) for quality measurement of the physical therapy process: comparison between EHR data and survey data. *BMC Medical Informatics and Decision Making*, BMC Medical Informatics and Decision Making, v. 16, n. 1, p. 141, dez. 2016. ISSN 1472-6947.

SINGER, A. et al. Data quality of electronic medical records in Manitoba: do problem lists accurately reflect chronic disease billing diagnoses? *Journal of the American Medical Informatics Association*, v. 23, n. 6, p. 1107–1112, nov. 2016. ISSN 1067-5027.

SKYTTBERG, N. et al. Exploring Vital Sign Data Quality in Electronic Health Records with Focus on Emergency Care Warning Scores. *Applied Clinical Informatics*, v. 08, n. 03, p. 880–892, dez. 2017. ISSN 1869-0327. Disponível em: ⟨http://www.thieme-connect.de/DOI/DOI?10.4338/ACI-2017-05-RA-0075⟩.

SOLLIE, A. et al. Reusability of coded data in the primary care electronic medical record: A dynamic cohort study concerning cancer diagnoses. *International Journal of Medical Informatics*, Elsevier Ireland Ltd, v. 99, p. 45–52, mar. 2017. ISSN 1386-5056.

STAFF, M.; ROBERTS, C.; MARCH, L. The completeness of electronic medical record data for patients with Type 2 Diabetes in primary care and its implications for computer modelling of predicted clinical outcomes. *Primary Care Diabetes*, Primary Care Diabetes Europe, v. 10, n. 5, p. 352–359, out. 2016. ISSN 1751-9918.

SUN, W. et al. Data processing and text mining technologies on electronic medical records: A review. *Journal of Healthcare Engineering*, v. 2018, p. 1–9, 2018. ISSN 2040-2295. Disponível em: ⟨https://www.hindawi.com/journals/jhe/2018/4302425/⟩.

TAGGART, J.; LIAW, S.-T.; YU, H. Structured data quality reports to improve EHR data quality. *International Journal of Medical Informatics*, Elsevier Ireland Ltd, v. 84, n. 12, p. 1094–1098, dez. 2015. ISSN 1386-5056. Disponível em: ⟨https://linkinghub.elsevier.com/retrieve/pii/S1386505615300435⟩.

TERRY, A. L. et al. A basic model for assessing primary health care electronic medical record data quality. *BMC Medical Informatics and Decision Making*, BMC Medical Informatics and Decision Making, v. 19, n. 1, p. 30, dez. 2019. ISSN 1472-6947.

TOFTDAHL, A. K. S. et al. Collect once - Use many times: The research potential of low back pain patients' municipal electronic healthcare records. *Studies in Health Technology and Informatics*, v. 247, p. 211–215, 2018. ISSN 1879-8365.

TSOUMAKAS, G.; KATAKIS, I. Multi-Label Classification. *International Journal of Data Warehousing and Mining*, v. 3, n. 3, p. 1–13, jul. 2007. ISSN 1548-3924. Disponível em: ⟨http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/jdwm.2007070101⟩.

TSOUMAKAS, G.; KATAKIS, I.; VLAHAVAS, I. Effective and efficient multilabel classification in domains with large number of labels. *Proc. ECML/PKDD 2008 Workshop on Mining Multidimensional Data (MMD'08)*, p. 30–44, 2008. Disponível em: ⟨http://lpis.csd.auth.gr/publications/tsoumakas-mmd08.pdf⟩.

TU, K. et al. Are family physicians comprehensively using electronic medical records such that the data can be used for secondary purposes? A Canadian perspective. *BMC Medical Informatics and Decision Making*, BMC Medical Informatics and Decision Making, v. 15, n. 1, p. 67, dez. 2015. ISSN 1472-6947.

VERAS, C. M. T.; MARTINS, M. S. A confiabilidade dos dados nos formulários de Autorização de Internação Hospitalar (AIH), Rio de Janeiro, Brasil. *Cadernos de Saúde Pública*, v. 10, n. 3, p. 339–355, set. 1994. ISSN 0102-311X. Disponível em: ⟨http://www.scielo.br/scielo.php?script=sci{_}arttext{\&}pid=S0102-311X1994000300014{\&}lng=p⟩.

WEISKOPF, N. G. et al. A Data Quality Assessment Guideline for Electronic Health Record Data Reuse. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, v. 5, n. 1, p. 14, set. 2017. ISSN 2327-9214.

WENNBERG, S. et al. Providing quality data in health care - almost perfect inter-rater agreement in the Norwegian tonsil surgery register. *BMC Medical Research Methodology*, BMC Medical Research Methodology, v. 19, n. 1, p. 6, dez. 2019. ISSN 1471-2288.

WHANG, S. E.; LEE, J.-G. Data collection and quality challenges for deep learning. *Proceedings of the VLDB Endowment*, v. 13, p. 3429–3432, 8 2020. ISSN 2150-8097. Future tutorial for ML pipelines, includes data cleaning and data validation. Disponível em: ⟨https://dl.acm.org/doi/10.14778/3415478.3415562⟩.

XIE, P.; XING, E. A neural architecture for automated ICD coding. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, 2018. p. 1066–1076. Disponível em: ⟨https://aclanthology.org/P18-1098⟩.

XU, K. et al. Multimodal machine learning for automated ICD coding. *CoRR*, abs/1810.13348, 2018. Disponível em: ⟨http://arxiv.org/abs/1810.13348⟩.

YOUNG, T. et al. Recent Trends in Deep Learning Based Natural Language Processing [Review Article]. *IEEE Computational Intelligence Magazine*, IEEE, v. 13, n. 3, p. 55–75, ago. 2018. ISSN 1556-603X. Disponível em: ⟨https://ieeexplore.ieee.org/document/8416973/⟩.

ZHANG, M.-L.; ZHOU, Z.-H. A Review on Multi-Label Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 26, n. 8, p. 1819–1837, ago. 2014. ISSN 1041-4347. Disponível em: ⟨http://ieeexplore.ieee.org/document/6471714/⟩.

ZHONG, J.; GAO, C.; YI, X. Categorization of patient disease into icd-10 with nlp and svm for chinese electronic health record analysis. In: *Proceedings of the 2018 International Conference on Artificial Intelligence and Pattern Recognition*. New York, NY, USA: Association for Computing Machinery, 2018. (AIPR 2018), p. 101–106. ISBN 9781450365246. Disponível em: ⟨https://doi.org/10.1145/3268866.3268877⟩.