Idade como fator de mortalidade por Neoplasia Maligna de Estômago: Uma análise baseada em *Machine Learning*

Pedro Lemos Mariano

[1]Universidade Federal de Viçosa (UFV) – Rio Paranaíba - Brasil

{pedro.l.mariano}@ufv.br

Abstract. *Among all existing neoplasms, whether benign or malignant, malignant stomach neoplasm stands out due to its high lethality in older populations. This study aims to analyze the influence of age on clinical outcomes in patients with stomach cancer, utilizing Machine Learning techniques to highlight the relevance of predictive and personalized approaches in disease management. Based on data from the Brazilian Mortality Information System (SIM), a predictive model was developed using data from the last 10 years to evaluate the probability of death as a function of the age group of individuals diagnosed with gastric cancer. The results demonstrate a significant correlation between advanced age and increased mortality, providing valuable insights for treatment and prevention strategies, as well as for the formulation of public policies targeted at high-risk age groups.*

Resumo. *Dentre todas as Neoplasias existentes, sejam elas benignas ou malignas, a Neoplasia maligna do estômago destaca-se por sua grande letalidade em populações mais velhas. O objetivo deste estudo é analisar a influência da idade de pacientes com câncer de estômago nos desfechos clínicos, por meio de técnicas de Machine Learning, destacando a relevância da abordagem preditiva e personalizada no manejo da doença. A partir de dados do Sistema de Informações sobre Mortalidade (SIM) do Brasil, um modelo preditivo foi desenvolvido, tendo como base dados dos últimos 10 anos para avaliar a probabilidade de óbito em função da faixa etária dos indivíduos diagnosticados com câncer gástrico. Os resultados demonstram uma correlação significativa entre a idade avançada e o aumento da mortalidade, oferecendo insights valiosos para estratégias de tratamento e prevenção, bem como para a definição de políticas públicas direcionadas a grupos etários de risco.*

1. Contextualização

As neoplasias representam um crescimento anormal e descontrolado de células que pode ocorrer em qualquer tecido do corpo humano. Essas alterações celulares são desencadeadas por modificações genéticas que resultam em proliferação celular independente dos mecanismos regulatórios normais. As neoplasias podem ser classificadas como benignas, quando o crescimento celular é limitado e não invade tecidos adjacentes, ou malignas, caracterizadas pela capacidade de invasão e formação de metástases [Weinberg 2007, Hanahan and Weinberg 2011].

Entre as neoplasias malignas, os cânceres gastrointestinais destacam-se pela sua alta prevalência e mortalidade. Em particular, a neoplasia de estômago, também denominada câncer gástrico, é um dos tipos mais frequentes e letais de neoplasias malignas. Este tipo de câncer frequentemente está associado a fatores como a infecção pela *Helicobacter pylori*, consumo elevado de alimentos salgados ou processados, e predisposição genética [Ferlay et al. 2021, Rawla and Barsouk 2019].

A análise de dados do Sistema de Informações sobre Mortalidade (SIM) do Brasil oferece uma visão detalhada sobre os padrões de mortalidade, permitindo o desenvolvimento de modelos preditivos que podem auxiliar na identificação de grupos de risco. Técnicas de *Machine Learning* surgem como uma ferramenta valiosa para investigar essas correlações e aprimorar o manejo clínico, direcionando estratégias de tratamento mais adequadas a diferentes faixas etárias.

2. Problemática

Apesar dos avanços no diagnóstico e no tratamento, a sobrevida global quando se trata de câncer gástrico, ainda é baixa, especialmente em casos diagnosticados em estágios avançados, o que reforça a importância de estratégias de prevenção e detecção precoce [Cheng 2021], sendo assim, uma das principais causas de morte no Brasil, no qual ocupa a terceira posição como neoplasia mais comum em homens e quinta entre mulheres [Duarte et al. 2020]. A faixa etária tem se mostrado um fator determinante no prognóstico dos pacientes, com uma correlação clara entre o envelhecimento e o aumento do risco de óbito.

A ausência de políticas públicas abrangentes para o enfrentamento do câncer gástrico representa um desafio significativo para os sistemas de saúde. Em países de baixa e média renda, a carência de programas de prevenção e diagnóstico precoce resulta em diagnósticos tardios, tratamentos menos eficazes e altas taxas de mortalidade. Essas lacunas evidenciam a urgência de iniciativas que ampliem o acesso ao cuidado, promovam rastreamento efetivo e reforcem a atenção primária [Ferlay et al. 2021, Rawla and Barsouk 2019].

3. Revisão de Literatura

3.1. Neoplasia Maligna do Estômago

A idade, predisposição genética, tabagismo, consumo de álcool e infecção pela *Helicobacter pylori* são os principais fatores de risco para o câncer gástrico. Em estágios iniciais, os sintomas incluem indigestão, saciedade precoce, azia, náuseas e perda de apetite, enquanto nos estágios avançados tornam-se mais graves, como dor abdominal, vômitos com sangue, fezes escurecidas, perda de peso e dificuldade para engolir [Ferlay et al. 2021, Cheng 2021]. O rastreamento por endoscopia e medidas preventivas, como a erradicação do *H. pylori*, são essenciais para a detecção precoce e redução da incidência da doença [Huang 2021, Organization 2020].

3.2. Aplicação de *Machine Learning* na Medicina

A análise preditiva tem se mostrado uma ferramenta promissora no contexto oncológico, possibilitando a utilização de técnicas de *Machine Learning* (ML) para prever a ocorrência, progressão e resposta ao tratamento de diversas neoplasias. Algoritmos

de ML, como redes neurais artificiais, árvores de decisão e *random forests*, têm sido empregados para analisar grandes volumes de dados clínicos, genômicos e ambientais, identificando padrões complexos que não seriam detectados por abordagens tradicionais [Esteva 2019, Bibault 2021].

No câncer gástrico, essas ferramentas podem ser aplicadas para estratificar pacientes de acordo com o risco, prever desfechos clínicos e otimizar tratamentos com base em características individuais. Dessa forma, a integração de técnicas de ML à medicina de precisão representa um avanço significativo na luta contra as neoplasias, com potencial para reduzir desigualdades no acesso ao cuidado e melhorar a qualidade de vida dos pacientes [Javed 2022].

3.3. Idade como Fator de Mortalidade

Estudos anteriores apontam a idade como um dos principais fatores associados à mortalidade por câncer, incluindo o câncer gástrico. Com o aumento da idade, há um acúmulo de alterações genéticas e epigenéticas, além de uma redução na capacidade do sistema imunológico em eliminar células tumorais, o que contribui para a maior incidência e gravidade da doença em idosos [Rawla and Barsouk 2019, Cheng 2021].

No caso específico do câncer gástrico, análises demonstram que a mortalidade é significativamente mais alta em indivíduos acima de 60 anos (Figuras 1 e 2), com maior prevalência em estágios avançados devido ao diagnóstico tardio. Esses achados reforçam a necessidade de estratégias de rastreamento específicas para populações idosas, visando reduzir a mortalidade [Huang 2021, Ferlay et al. 2021].

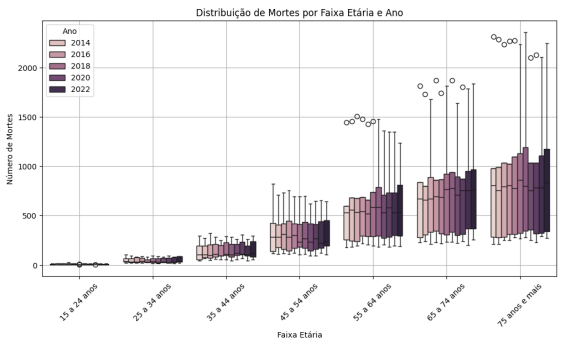

Figure 1. ***Distribuição de Mortes por Faixa Etária e Ano***

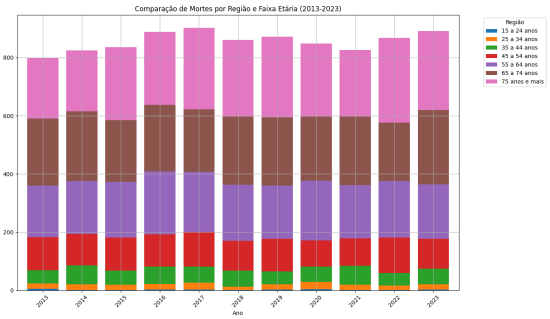

Figure 2. ***Comparação de Mortes por Região e Faixa Etária entre 2013 e 2023***

Figure 3. Mortes por Faixa Etária

3.4. Mortalidade nas Regiões do Brasil

O câncer gástrico é uma das principais causas de mortalidade por neoplasias no Brasil, apresentando variações regionais significativas que refletem as desigualdades socioeconômicas e o acesso aos serviços de saúde. Segundo dados recentes, as taxas de mortalidade são mais altas nas regiões Sudeste e Nordeste, possivelmente devido à menor disponibilidade de programas de rastreamento, diagnóstico precoce e tratamento adequado, além da dimensão dessas regiões [Organization 2020, Rawla and Barsouk 2019].

Na Região Sul, embora a mortalidade também seja elevada, observa-se maior detecção precoce da doença, atribuída a melhores indicadores de desenvolvimento humano e maior acesso a serviços especializados [Cheng 2021]. Por outro lado, nas região Nordeste, fatores como baixa cobertura de exames endoscópicos, elevada prevalência de infecção por *Helicobacter pylori* e condições de vida precárias contribuem para diagnósticos tardios e altas taxas de letalidade [Huang 2021].

A Região Sudeste, que concentra os maiores centros urbanos e uma infraestrutura de saúde mais robusta, apresenta taxas de mortalidade mais altas, pois enfrenta desafios relacionados à desigualdade no acesso ao tratamento entre populações de baixa renda. Por fim, na Região Centro-Oeste, o cenário é intermediário, com menor densidade populacional e serviços de saúde concentrados em áreas urbanas [Ferlay et al. 2021].

Essas disparidades destacam a necessidade de políticas públicas específicas para cada região, focadas em estratégias como a ampliação do rastreamento, o acesso universal a tecnologias de diagnóstico e o fortalecimento da atenção primária à saúde.

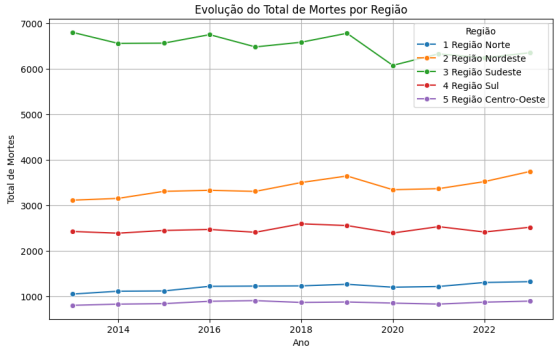

Figure 4. ***Evolução do Total de Mortes por Região***

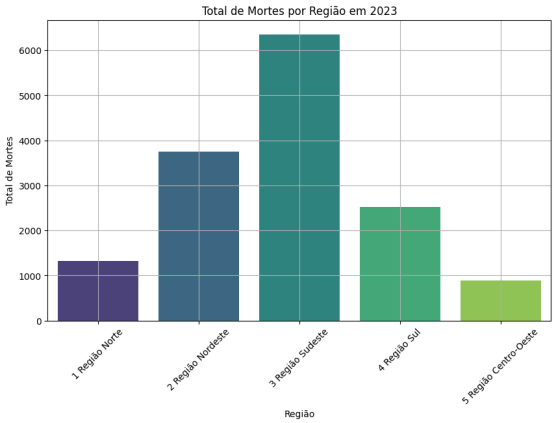

Figure 5. ***Total de Mortes por Região em 2023***

Figure 6. Mortes por Região

4. Metodologia

O modelo preditiviro realizado, usa como base os últimos 10 anos, em função da idade por cada região do Brasil. Os dados foram normalizados e preenchidos em casos de valores ausentes. Foram aplicadas técnicas como oversampling para lidar com desbalanceamento nas classes. O modelo foi ajustado utilizando validação cruzada e hiperparâmetros otimizados com busca em grade.

4.1. Ambiente de Testes

- **Ambiente:** Google Colaboratory.
- **Software:** Python, utilizando as bibliotecas: Pandas, Numpy, Scikit Learn, Seaborn, Matplotlib, Pytz, Tensorflow, Keras.
- **Dados:** A avaliação utiliza como entrada o Conjunto de Dados do DATASUS [da Saúde 2023], tendo como parâmetros analisádos entre os anos de 2013 a 2023.

[1]

4.2. Algoritmo

1. **LSTM:** O algoritmo é uma arquitetura poderosa para trabalhar com dados sequenciais, especialmente quando há necessidade de capturar dependências de longo prazo. Ele é amplamente utilizado em diversas áreas, como NLP, séries temporais e reconhecimento de padrões. A implementação no modelo é mostrada no Apêndice A 9.

4.3. Avaliação de Desempenho

O treinamento do modelo ocorreu em 100 épocas, e a perda no conjunto de validação é calculada a cada 10 épocas. Sendo assim, o modelo foi capaz de fazer uma previsão dos próximos 10 anos.

- **Val Loss:** O modelo obteve constante em todas as épocas (0.5266).

5. Resultados e Discussão

5.1. Análise Exploratória dos Dados:

A análise dos dados mostrou 2019 como o ápice dos niveis de mortalidade por Neoplasia Maligna de Estômago, como mostrado na Figura 7, já 2020 obteve os menores números do uúltimos 10 ano.

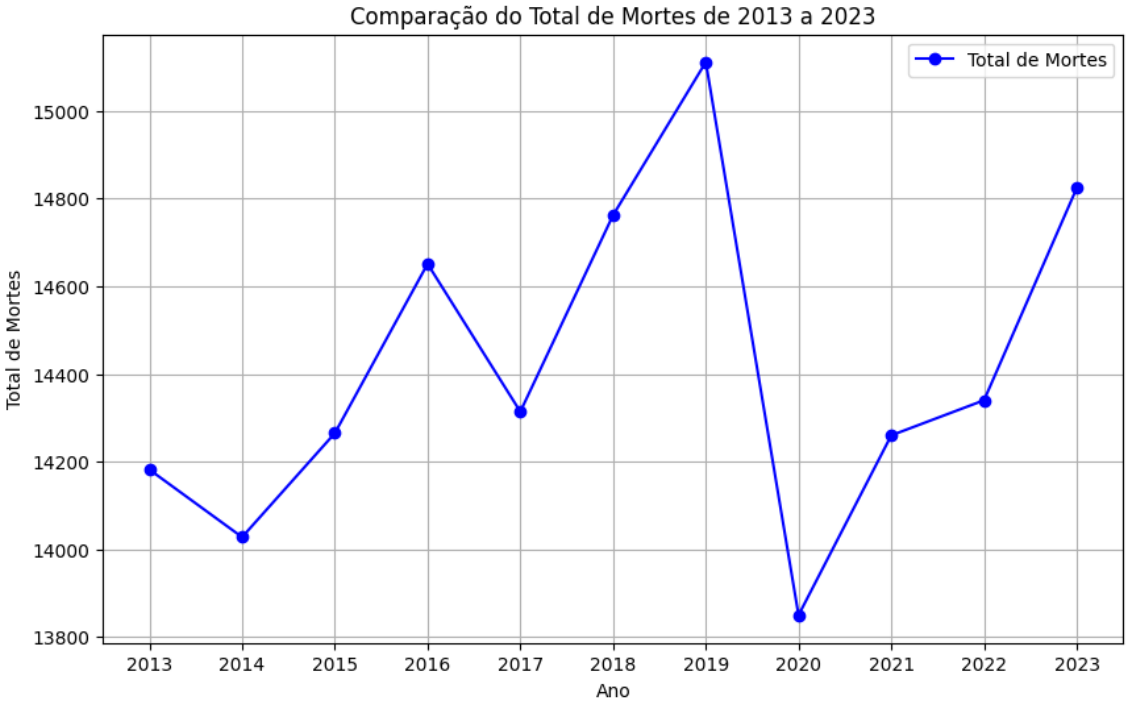

Figure 7. Comparação do Total de Mortes de 2013 a 2023

[1] https://github.com/PedroLemosMariano/CID-Neoplasia-Maligna-do-Estomago.git

Baseado no modelo gerado, é possiver ver na Figura 8, o sudeste se apresenta como o primeiro em comparação com as demais regiões do Brasil, seguido pela região Nordeste.

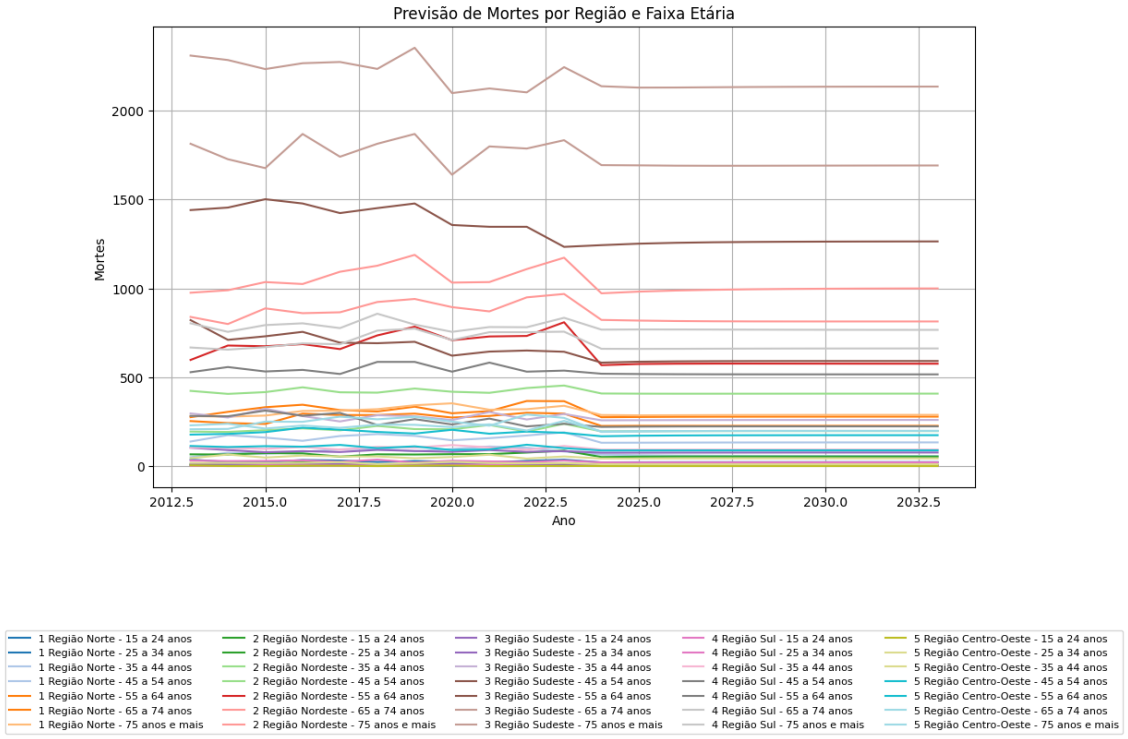

Figure 8. Previsão de Mortes por Região e Faixa Etária

5.2. Interpretação dos Resultados:

A partir de 2020, conforme ilustrado na Figura 7, observa-se um aumento contínuo nos números, evidenciando a necessidade urgente de políticas públicas voltadas para o cuidado desse grupo de risco.

Embora o algoritmo utilizado tenha gerado previsões coerentes, os resultados indicam um comportamento linear constante após o ano de 2024, sugerindo uma estabilização nos números de mortalidade. Em algumas regiões, como o Sudeste, o modelo aponta uma possível redução desses índices no período de 2023 para 2024.

5.3. Idade como Fator Preditor:

No modelo gerado na Figura 8, a região Sudeste se destaca como a principal em comparação com as demais regiões do Brasil. A faixa etária com maior destaque é a de pessoas com mais de 75 anos. No entanto, o modelo sugere que, a partir dos 65 anos, já são observados números significativos nas previsões de mortalidade.

A análise realizada por meio de técnicas de *Machine Learning* demonstrou que a idade dos pacientes é um dos fatores mais influentes nos desfechos clínicos, superando, em alguns casos, a relevância de variáveis como sexo ou localização geográfica. Os modelos preditivos utilizados identificaram padrões claros, indicando que indivíduos em faixas

etárias mais avançadas apresentam uma vulnerabilidade significativamente maior, o que pode ser atribuído à presença de comorbidades, fragilidade imunológica e limitações na resposta aos tratamentos disponíveis.

Esses achados não apenas confirmam a relevância da idade como um marcador prognóstico, mas também destacam a necessidade de abordagens personalizadas no manejo do câncer gástrico. Além disso, os resultados reforçam a importância de desenvolver políticas públicas que priorizem intervenções preventivas e estratégias terapêuticas voltadas para grupos etários mais vulneráveis, de forma a reduzir as taxas de mortalidade e melhorar a qualidade de vida dos pacientes diagnosticados.

5.4. Limitações do Estudo:

Este estudo apresenta algumas limitações que devem ser consideradas ao interpretar seus resultados. Primeiramente, a dependência de dados históricos pode restringir a capacidade do modelo de prever mudanças abruptas ou não previstas nos padrões de mortalidade. Além disso, a análise se baseia em um conjunto de dados específicos, o que pode não refletir a totalidade das variações regionais e sociais que influenciam a mortalidade no Brasil. A qualidade e a precisão dos dados coletados também podem afetar os resultados, uma vez que possíveis erros de registro ou inconsistências nos dados podem comprometer a acurácia do modelo. Por fim, o modelo desenvolvido não considera variáveis externas, como fatores econômicos e políticos, que podem influenciar diretamente as taxas de mortalidade.

5.5. Sugestões para Trabalhos Futuros

Com base nas limitações identificadas neste estudo, algumas direções para trabalhos futuros podem ser consideradas. Primeiramente, seria interessante ampliar a base de dados utilizada, incorporando variáveis adicionais, como fatores socioeconômicos e comportamentais, que podem influenciar mais diretamente as taxas de mortalidade. Além disso, a inclusão de modelos mais sofisticados, como redes neurais, pode melhorar a precisão das previsões, principalmente em cenários de mudanças abruptas. Outro aspecto relevante seria a validação do modelo em outras regiões do Brasil, a fim de verificar sua aplicabilidade e generalização. Por fim, investigações futuras podem explorar o impacto de fatores externos, como crises econômicas e políticas, que não foram considerados neste estudo, mas que podem desempenhar um papel significativo nas previsões de mortalidade.

6. Conclusão

No caso do câncer gástrico, a ausência de iniciativas como a triagem populacional para a identificação precoce de lesões pré-cancerígenas, a implementação de campanhas de conscientização e a redução dos fatores de risco, como o consumo excessivo de alimentos processados, agrava significativamente a situação. Além disso, o subfinanciamento da saúde pública impede a disseminação de tecnologias avançadas de diagnóstico e limita o acesso aos tratamentos mais recentes e eficazes [Cheng 2021, Huang 2021].

Políticas públicas eficazes poderiam incluir a ampliação de programas de vacinação contra o *Helicobacter pylori*, o fortalecimento da pesquisa científica no campo oncológico e a criação de centros de referência para diagnóstico e tratamento. Sem essas medidas, as desigualdades no acesso ao cuidado oncológico continuarão, perpetuando os altos índices de mortalidade [Organization 2020].

Além disso, fatores sociodemográficos, como idade, sexo, localização geográfica, nível socioeconômico e comportamentos de saúde, desempenham um papel significativo no aumento da mortalidade por doenças específicas. A falta de acesso a cuidados de saúde adequados em algumas regiões do país contribui para desigualdades que perpetuam altos índices de mortalidade. É fundamental que políticas públicas regionais considerem esses fatores para reduzir as disparidades e melhorar a qualidade de vida da população.

7. Apêndice A

```python
# Modelo LSTM
class LSTMModel(nn.Module):
    def __init__(self, input_size, hidden_size, output_size):
        super(LSTMModel, self).__init__()
        self.lstm = nn.LSTM(input_size, hidden_size, batch_first=True)
        self.fc = nn.Linear(hidden_size, output_size)

    def forward(self, x):
        _, (hidden, _) = self.lstm(x)
        out = self.fc(hidden[-1])
        return out

# Configurações do modelo
input_size = X.shape[2]
hidden_size = 50
output_size = y.shape[1]
model = LSTMModel(input_size, hidden_size, output_size)

criterion = nn.MSELoss()
optimizer = torch.optim.Adam(model.parameters(), lr=0.01)

# Treinar o modelo
epochs = 100
for epoch in range(epochs):
    model.train()
    optimizer.zero_grad()
    output = model(X_train)
    loss = criterion(output, y_train)
    loss.backward()
    optimizer.step()

    if (epoch + 1) % 10 == 0:
        model.eval()
        val_output = model(X_val)
        val_loss = criterion(val_output, y_val)
        print(f"Epoch {epoch + 1}/{epochs}, Loss: {loss.item():.4f}, Val Loss: {val_loss.item():.4f}")

# Previsão para os próximos 10 anos
model.eval()
future_years = 10
predictions = []

last_sequence = X[-1].detach().numpy()
for _ in range(future_years):
    input_seq = torch.tensor(last_sequence[np.newaxis, :, :], dtype=torch.float32)
    pred = model(input_seq).detach().numpy()
    predictions.append(pred[0])
    last_sequence = np.vstack([last_sequence[1:], pred])

predictions = scaler.inverse_transform(np.array(predictions))

future_df = pd.DataFrame(predictions, columns=pivot_data.columns, index=range(2024, 2024 + future_years))
```

Figure 9. Código de modelo preditivo utilizado (LSTM)

References

[Bibault 2021] Bibault, J.-E. e. a. (2021). Artificial intelligence and machine learning in oncology: Toward personalized medicine. *Nature Reviews Clinical Oncology*, 18:540–550.

[Cheng 2021] Cheng, N. e. a. (2021). Gastric cancer: epidemiology, genomics, and treatment implications. *Nature Reviews Gastroenterology & Hepatology*, 18:543–558.

[da Saúde 2023] da Saúde, M. (2023). Datasus. tabnet. Brasília, DF: Ministério da Saúde.

[Duarte et al. 2020] Duarte, A. C. d. S. F., Wanderley, R. L., Silva, G. J. T. d., Silva, Z. C. d., Souza, A. A., Torres, V. C., Silva, E. B. d., and Rocha, T. J. M. (2020). Perfil

epidemiológico das internações por neoplasia maligna de estômago durante a última década no brasil. *Brazilian Journal of Development*, 6(10):78528–78539. ISSN 2525-8761.

[Esteva 2019] Esteva, A. e. a. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25:24–29.

[Ferlay et al. 2021] Ferlay, J. et al. (2021). Global cancer observatory: Cancer today. *International Agency for Research on Cancer*. Accessed January 2025.

[Hanahan and Weinberg 2011] Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674.

[Huang 2021] Huang, W. e. a. (2021). Healthcare access and gastric cancer outcomes: A global perspective. *Lancet Gastroenterology & Hepatology*, 6:421–430.

[Javed 2022] Javed, A. e. a. (2022). Predictive analytics and machine learning in gastrointestinal cancer: Opportunities and challenges. *World Journal of Gastroenterology*, 28(10):1204–1216.

[Organization 2020] Organization, W. H. (2020). Health equity and cancer prevention. *World Cancer Report 2020*. Accessed January 2025.

[Rawla and Barsouk 2019] Rawla, P. and Barsouk, A. (2019). Epidemiology of gastric cancer: global trends, risk factors and prevention. *Przeglad gastroenterologiczny*, 14(1):26.

[Weinberg 2007] Weinberg, R. A. (2007). *The Biology of Cancer*. Garland Science.