Análise de Risco de Câncer com Regressão Logística e Linear: Um Estudo com Fatores Ambientais e Clínicos

Felipe Perrella dos Santos - 210298
Gabriel Pereira de Camargo - 210039
Luis Felipe dos Santoa Gianoni - 210206

lfelipe2305@hotmail.com.br | perrella.felipe@hotmail.com | gabriel.pc2001@gmail.com

Resumo

Este artigo apresenta uma abordagem baseada em aprendizado de máquina para análise de dados clínicos e ambientais relacionados ao risco de câncer. Foram aplicadas técnicas de regressão logística e linear sobre um dataset contendo 1.000 registros de pacientes, com variáveis como idade, exposição à poluição, tabagismo, dieta e sintomas clínicos. A regressão logística foi utilizada para classificar os pacientes quanto ao nível de risco de câncer (baixo, médio e alto), enquanto a regressão linear previu a perda de peso, um dos sintomas associados. O modelo de classificação obteve acurácia de 100%, e o modelo de regressão linear alcançou um R^2 de 0,72, indicando boa performance preditiva.

Palavras-chave: Regressão logística; Regressão linear; Câncer; Poluição; Classificação.

Abstract

This paper presents a machine learning-based approach to analyze clinical and environmental data related to cancer risk. Logistic and linear regression techniques were applied to a dataset with 1,000 patient records, including variables such as age, pollution exposure, smoking, diet, and clinical symptoms. Logistic regression was used to classify cancer risk levels (low, medium, high), while linear regression predicted weight loss, one of the associated symptoms. The classification model achieved 100% accuracy, and the linear regression model reached an R^2 of 0.72, indicating strong predictive performance.

Keywords: Logistic regression; Linear regression; Cancer; Pollution; Classification.

1. Introdução

O câncer é uma das principais causas de mortalidade no mundo, e diversos fatores ambientais e comportamentais contribuem para seu desenvolvimento. Entre esses fatores, destacam-se a poluição do ar, o consumo de álcool, o tabagismo e a obesidade. A análise preditiva desses elementos pode auxiliar na detecção precoce de grupos de risco, otimizando políticas públicas e ações preventivas. Este trabalho tem como foco o uso de técnicas de regressão para classificar e prever informações de saúde associadas ao risco de câncer. O dataset utilizado contém dados de pacientes com diferentes níveis de exposição a fatores de risco e sintomas clínicos. A análise baseou-se em duas abordagens principais: regressão logística para classificação do risco de câncer e regressão linear para prever o nível de perda de peso.

2. Objetivos

Objetivo Geral:

Aplicar técnicas de regressão logística e linear para classificar o nível de risco de câncer e prever sintomas associados, com base em dados ambientais, clínicos e comportamentais.

Objetivos Específicos:

• Realizar análise exploratória para identificar padrões relevantes entre variáveis.

• Classificar os níveis de risco de câncer utilizando regressão logística.

• Prever perda de peso com regressão linear.

• Avaliar os modelos com métricas como accuracy, F1-score, R^2 e MSE.

3. Fundamentação Teórica

A regressão logística é uma técnica estatística amplamente utilizada para problemas de classificação binária ou multiclasse. Já a regressão linear é aplicada em tarefas de predição contínua, como a estimativa de valores quantitativos. Ambas são técnicas interpretáveis e eficientes para modelos com variáveis independentes numéricas ou categóricas. Em aplicações de saúde, essas técnicas têm sido empregadas para prever diagnósticos, identificar fatores de risco e analisar sintomas clínicos.
Os modelos de regressão logística e linear foram escolhidos por sua simplicidade, boa capacidade preditiva e elevada interpretabilidade. No contexto médico, onde a compreensão dos fatores de risco é tão importante quanto a previsão em si, modelos transparentes e explicáveis são mais apropriados do que arquiteturas complexas de aprendizado profundo.

4. Trabalhos Relacionados

Estudos anteriores exploram a relação entre fatores ambientais e o desenvolvimento de doenças respiratórias e oncológicas. Modelos supervisionados como regressão logística, random forest e redes neurais têm sido aplicados com sucesso na identificação de padrões em dados médicos. Embora modelos avançados possam ter melhor desempenho, métodos mais simples como regressões ainda se destacam pela fácil interpretação e aplicabilidade prática em ambientes clínicos.

5. Materiais e Métodos

O dataset utilizado contém 1.000 amostras de pacientes, cada uma com 26 atributos, incluindo idade, gênero, nível de poluição, uso de álcool, tabagismo, sintomas clínicos (fadiga, tosse, dor no peito) e risco de câncer. Os dados foram limpos e pré-processados utilizando as bibliotecas Pandas, NumPy e Seaborn, e os modelos foram implementados em Python com scikit-learn.

Embora o dataset seja amplamente utilizado em plataformas como o Kaggle, não foi possível localizar a fonte original dos dados para verificação da sua autenticidade. Essa limitação deve ser considerada, especialmente em estudos que envolvam decisões clínicas ou generalizações populacionais.

A regressão logística foi aplicada à variável categórica 'Level', com três classes: Low, Medium e High. A acurácia do modelo foi de 100%, indicando excelente performance. Já a regressão linear foi utilizada para prever o valor da variável contínua 'Weight Loss', com um R^2 de 0,72 e erro quadrático médio de 1.41.

6. Resultados

A regressão logística apresentou performance perfeita, com precisão, revocação e F1-score iguais a 1.00 para todas as classes. A matriz de confusão indicou nenhuma classificação incorreta. Isso pode ser explicado pelo forte padrão presente nos dados. Embora os resultados da regressão logística tenham sido perfeitos, com acurácia de 100%, isso pode indicar sobreajuste (overfitting), uma vez que o modelo pode ter aprendido padrões específicos do conjunto de treino. O dataset relativamente pequeno (1.000 amostras) também limita a generalização dos resultados para populações maiores e mais diversas.

Na regressão linear, o modelo obteve um erro médio quadrático de 1.41 e um coeficiente de determinação (R^2) de 0.72, o que indica boa capacidade preditiva. O gráfico de resíduos mostrou uma distribuição aceitável, com poucos desvios sistemáticos.

Todo o desenvolvimento, tratativas dos dados, e os modelos utilizados podem ser checados no código pelo github no link: https://github.com/rageagainst10/AF-IA.git.

7. Conclusão e Trabalhos Futuros

Este estudo demonstrou o potencial de técnicas simples de regressão para análise de dados médicos relacionados ao risco de câncer. Os resultados obtidos mostram que, mesmo com abordagens clássicas, é possível alcançar alto desempenho em tarefas preditivas quando os dados são bem estruturados. Como continuidade, recomenda-se testar modelos mais complexos, aplicar validação cruzada e aumentar a base de dados com amostras de diferentes regiões geográficas.

Como contribuição prática, este estudo pode auxiliar em iniciativas de saúde pública voltadas à prevenção de câncer em áreas urbanas com alta poluição. A identificação de padrões associados ao risco permite que autoridades possam desenvolver campanhas de triagem preventiva com base em dados ambientais e comportamentais.

8. Referências

SANTOS, Felipe Perrella dos; CAMARGO, Gabriel Pereira de; GIANONI, Luis Felipe dos Santos. Análise de Risco de Câncer com Regressão Logística e Linear: Código-fonte. 2025. Disponível em: https://github.com/rageagainst10/AF-IA. Acesso em: 13 jun. 2025

SOUZA, F. C. d. BERTimbau: modelos BERT pré-treinados para português brasileiro. Dissertação (Mestrado) – UNICAMP, 2020.

LIMA, D. C. M. de. Gerenciamento e tratamento de feedbacks com RPA. FACENS, 2023.

THE DEVASTATOR. Lung Cancer Prediction: Air Pollution, Alcohol, Smoking & Risk of Lung Cancer. Disponível em: https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link. Acesso em: 13 jun. 2025.

9. Discussão Complementar

A escolha pelos modelos de regressão adotados neste estudo se justifica tanto pela simplicidade quanto pela capacidade de interpretação dos coeficientes, o que é fundamental em cenários clínicos. Além disso, eles exigem menos dados para treinamento em comparação com redes neurais, sendo mais apropriados em contextos com volume limitado de amostras, como neste caso (1.000 registros).

A análise exploratória revelou que variáveis como poluição do ar, obesidade, consumo de álcool e fadiga têm distribuições que sugerem relevância na separação entre os níveis de risco. Esses achados foram confirmados nos modelos aplicados, reforçando a importância dessas variáveis para estratégias de triagem clínica e preventiva.

Além disso, o comportamento bimodal da variável 'Weight Loss' — com concentração nos níveis 2 e 7 — aponta para dois grupos de pacientes: os que apresentam perda de peso leve e os que apresentam sintomas mais agressivos. Esse padrão sugere a possibilidade de estratificação clínica com base nesse indicador.

Para estudos futuros, é recomendável testar modelos mais robustos como Random Forest, SVM e XGBoost, e realizar validação cruzada para garantir maior generalização dos resultados. Também é relevante aplicar análise de importância de variáveis (feature importance) para interpretar o impacto relativo de cada fator no modelo preditivo.