Universidade Federal de Pernambuco
Centro de Ciências Exatas e da Natureza
Departamento de Estatística
Ana Beatriz Ramos da Silva

TÓPICOS ESPECIAIS EM ESTATÍSTICA COMPUTACIONAL

REDES NEURAIS PARA CLASSIFICAÇÃO DE RISCO DE CÂNCER

Data: 07 de dezembro de 2025

INTRODUÇÃO

O câncer é um grande problema de saúde pública que ultrapassa limites geográficos e atinge o mundo como um todo. De acordo com Oliveira Santos et al. (2023), na última década observou-se um aumento de 20% na ocorrência de câncer e, para 2030, estima-se que ocorram mais 25 milhões de novos casos registrados. O diagnóstico precoce é um ponto crucial para um tratamento eficaz e maiores chances de controle da doença.

O câncer de pulmão, por sua vez, é a doença maligna mais comum do mundo. Entre os novos casos de câncer, cerca de 13% correspondem a esse tipo. Além disso, segundo o Global Burden of Disease Study 2015, o câncer de pulmão é o mais agressivo e o que registra maior mortalidade. A demora no diagnóstico desse tipo de câncer é um desafio adicional, sobretudo porque o acesso a exames de diagnóstico por imagem é escasso em muitas regiões (ARAÚJO et al., 2018).

Entre as mulheres, o tipo de câncer mais registrado é o de mama. De acordo com o Instituto Nacional do Câncer (2011), fatores como menarca precoce, nuliparidade, idade da primeira gestação, gestação após os 30 anos, menopausa tardia, uso de anticoncepcionais orais, terapia de reposição hormonal e a própria idade da mulher são considerados influentes no risco da doença. Além disso, o atraso no diagnóstico e no início do tratamento pode impedir terapias curativas e reduzir taxas de sobrevivência (SOUZA et al., 2008). Sarris et al. (2018) discutem que o câncer de próstata é o segundo mais comum em homens no Brasil e também o segundo em mortalidade por câncer, representando um grande desafio para a saúde pública.

O câncer de cólon também aparece em destaque e está entre os três mais frequentes no Brasil e no mundo. Ele é responsável por cerca de 1,8 milhão de casos e 862 mil mortes registradas, sendo considerado um dos mais agressivos. O consumo de álcool, o tabagismo, fatores hereditários e hábitos de vida são os principais fatores de risco (INSTITUTO NACIONAL DE CÂNCER, 2023). Para Bomfim, Giotto e Silva (2018), o câncer de pele apresenta dois principais tipos: o não melanoma (CPNM) e o melanoma (CM). O câncer de pele não melanoma é o mais frequente, caracteriza-se por crescimento lento e invasão local, e geralmente apresenta bom prognóstico quando tratado adequadamente. Contudo, o diagnóstico e o tratamento tardios podem levar a ulcerações e deformidades graves.

Diante da magnitude do problema de saúde pública representado pelo câncer e dos desafios gerados pelo diagnóstico tardio, especialmente em cenários onde não há recursos suficientes para um monitoramento completo, o uso de redes neurais pode ser um auxílio importante. Esses métodos permitem analisar dados e antecipar diagnósticos com base em características relevantes diretamente relacionadas ao risco de câncer. O objetivo desse estudo é usar as redes neurais para construir modelos de classificação para risco de câncer e severidade utilizando inteligência artificial para contribuir com o avanço e bem estar da sociedade.

FUNDAMENTAÇÃO TEÓRICA

Redes Neurais Artificiais (RNAs)

Nos últimos anos, com o aumento de informações e o desenvolvimento de processos cada vez mais robustos e complexos que demandam decisões rápidas e contextuais, as Redes Neurais Artificiais (RNAs) têm ganhado destaque, considerando seu desempenho computacional e capacidade de classificação (Fleck et al., 2016). As redes neurais são algoritmos computacionais que utilizam um modelo matemático baseado em estruturas de organismos inteligentes, permitindo simular o funcionamento do cérebro humano computacionalmente. Em suma, as Redes Neurais assemelham-se ao cérebro humano em sua capacidade de aprender e tomar decisões com base no que foi ensinado (Sporl et al., 2011).

Para Ferreira et al. (2016), as RNAs são modelos não paramétricos não lineares que, além de simular sistemas complexos, conseguem generalizar os resultados obtidos para situações desconhecidas. Com o aprendizado do treinamento, respostas coerentes e apropriadas para os padrões observados são geradas para os casos não vistos. Em geral, essa metodologia pode ser aplicada em diversas áreas do conhecimento para a solução de variados problemas, possuindo diferentes arquiteturas e componentes.

Rede Neural Perceptron Multicamadas (MLP)

O Perceptron, proposto por Rosenblatt (1958), é uma forma simples de rede neural focada na classificação de padrões. A principal desvantagem desse método é que ele só consegue identificar e classificar padrões lineares, o que frequentemente não representa a realidade do problema a ser solucionado. Dessa forma, o uso de um Perceptron de multicamadas torna-se indispensável (Ambrósio, 2002). Dentre as arquiteturas de redes neurais, as do tipo MLP são os modelos mais utilizados e conhecidos. Essa arquitetura é composta por camadas: camada de entrada, camadas ocultas e camada de saída (Nied, 2007). Segundo Fleck et al. (2016), a rede MLP é comumente aplicada em problemas de classificação, aproximação, previsão e modelagem temporal nas mais diversas áreas.

Algoritmo de Treinamento MLP

As redes neurais são capazes de aprender e generalizar seu aprendizado. Os procedimentos usados para construir o aprendizado de uma RNA a partir de uma função são chamados de algoritmo de aprendizado (Igus, 1996). O aprendizado por retropropagação consiste em propagar e retropropagar a informação. Propagação: Um padrão de ativação é aplicado aos nós da camada de entrada da rede, e seu efeito se propaga pelas camadas. Na última camada, a saída é produzida, tornando-se a resposta da rede. Retropropagação: Os pesos sinápticos sofrem ajuste de acordo com uma regra de correção de erro. O sinal de erro é propagado para trás, na direção contrária das conexões sinápticas, ajustando os pesos para que a resposta real da rede seja estatisticamente próxima da desejada (Nied, 2007).

Funções de Ativação e Função de Perda

A função de ativação representa o efeito que a entrada interna e o estado atual de ativação exercem na definição do próximo estado de ativação da unidade (Fleck et al., 2016, p. 50), sendo a responsável por introduzir a não linearidade do modelo. De acordo com Haykin (2001), diversas funções de ativação podem ser usadas. Neste estudo, foram abordadas as seguintes: ReLU, Função tanh, ELU e Sigmoid. Para avaliar a perda em classificação multiclasse, a função cross-entropy é amplamente utilizada.

Regularização da Rede Neural e Otimização

O overfitting ocorre quando o modelo se ajusta excessivamente aos dados de treino, aprendendo padrões específicos em vez de capturar relações mais gerais que se aplicariam a novos dados (Bejani; Ghatee, 2021). Para evitar esse fenômeno, a técnica de Dropout é utilizada, baseando-se em "desligar" alguns neurônios, fazendo com que a rede não dependa de pontos específicos. De acordo com Srivastava et al. (2014), níveis mais elevados de dropout são mais benéficos ao se trabalhar com variáveis fortemente correlacionadas. Para otimização, foram utilizados os otimizadores Adam, que adapta a taxa de aprendizado por parâmetro; RMSprop, que trabalha bem na presença de ruídos; e SGD que, proporciona melhor generalização. A otimização também incluiu o uso de pesos de classe para tratar o desequilíbrio da variável alvo.

Métricas de Avaliação

As métricas de avaliação utilizadas foram:

- Acurácia: Avalia a proporção de acertos totais.
- Precisão: Mede a taxa de acerto entre as predições positivas.
- Recall: Avalia a proporção de acertos de uma classe (sensibilidade).
- F1-Score: Realiza uma média harmônica entre a Precisão e o Recall.
- Matriz de Confusão: Detalha os acertos e erros do modelo por classe.

METODOLOGIA

A base dados utilizada foi a Cancer Risk Factors Dataset, disponível em https://www.kaggle.com/code/tarekmasryo/cancer-risk-factors-prediction que possui informações de cada individuo que estão relacionadas ao risco de câncer. A variável resposta é o nível de risco de câncer, sendo estes: baixo, médio ou alto. Na etapa de pré-processamento variáveis irrelevantes foram excluídas, as variáveis numéricas foram padronizadas e nas variáveis categóricas a codificação foi One-Hot Enconding e a variável alvo com LabelEnoconder. A variável resposta mostrou desbalanceamento entre suas categorias e foi contornado com class weights. Os dados foram divididos em 80% para treino e 20% para teste. A arquitetura de Rede Neural escolhida foi a MLP que lida bem com dados tabulares. O treinamento dos dados foi feito com função de perda categorical cross-entropy, otimizador Adam, 50 épocas e no teste 20%. Para otimizar os Hiperparâmetros utilizou-se o GridSearchCV com Scikeras para combinar funções de ativação distintas, quantidade de neurônios, número de épocas, otimizadores e taxa de dropout. Por fim, para avaliar o melhor modelo a validação foi realizada conforme desempenho no conjunto de teste com as métricas: acurácia, precisão, recall, f1-score e a matriz de confusão.

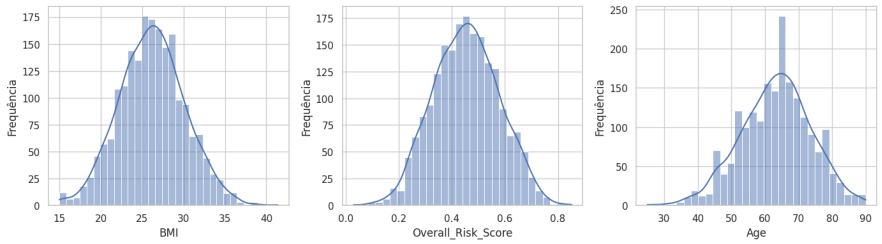
APLICAÇÃO

Inicialmente a análise exploratória foi realizada, considerando a natureza de cada variável. As variáveis Idade, IMC e pontuação de risco geral são numéricas e estatísticas básicas podem ser calculadas. Os resultados obtidos mostram que a idade média dos indivíduos é de 63 anos, o mínimo 25 anos e o máximo de 70, mostrando uma variabilidade que alcança diferentes grupos etários. A média de IMC é de 26,18 e o máximo de 41,40 indicando um sobrepeso médio leve no grupo avaliado e, para pontuação de risco geral, revela-se uma média de 0,45, um mínimo de 0,02 e o máximo de 0,85, mostrando que alguns indivíduos têm muito mais propensão a desenvolver câncer que outros.

Considerando as variáveis categóricas, os tipos de câncer mapeados são: Pulmão com 527 registros, Mama com 460 registros, Cólon com 418, Próstata com 305 e Pele com 290 registros. Para

o gênero, têm-se 0 e 1 para representar o gênero feminino e masculino, respectivamente, e dentre os indivíduos a maior classe é do gênero feminino, correspondendo a 1022 registros. Outras variáveis funcionam como pontuação de 0 até 10 para demonstrar a intensidade da característica observada para cada indivíduo mapeamendo seu perfil clínico e comportamental.

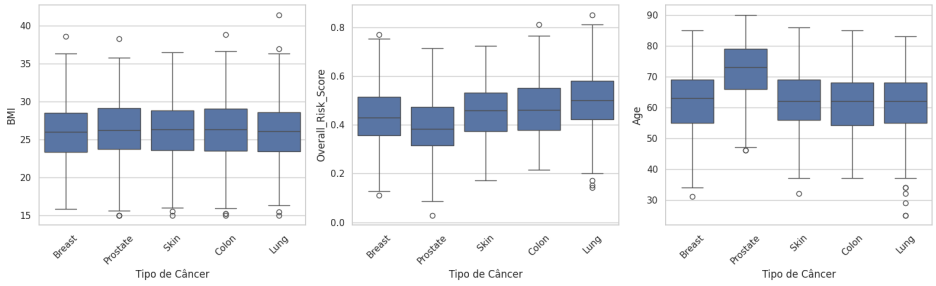
Figura 1 – Distribuição das variáveis númericas (IMC, Pontuação de risco total e idade)



Fonte: Elaborada pela autora (2025).

Na figura 1, a distribuição das variáveis numéricas pode ser observada, em que, o comportamento destas se assemelha a distribuição normal. Além disso, a relação entre essas mesmas variáveis e o tipo de câncer foi observada e os resultados podem ser observados na figura 2.

Figura 2 – Relação entre as variáveis IMC, Pontuação de risco total e idade e o tipo de câncer



Fonte: Elaborada pela autora (2025).

Observando a figura 2, a distribuição do IMC por tipo de câncer observa-se que a mediana está entre 24 e 27 para todas as classes. Existem alguns outliers, mas, não há grandes diferenças dentre as categorias. Assim, sugere-se que o IMC não varia significativamente dentre os tipos de câncer. Observando a pontuação de risco geral por tipo de câncer é visto que a mediana é menor na classe de câncer de próstata, podendo evidenciar um menor risco em média. O grupo de câncer de pulmão demostra maior variabilidade, maiores valores máximos e maior mediana. Por fim, considerando a idade por tipo de câncer as constatações são de que para o câncer de próstata os pacientes são aparentemente mais velhos, com uma mediana próxima aos 70 anos e um máximo de 90. Para as classes de câncer de mama, pele e colón a idade mediana está entre 55 e 65 anos. Para o câncer de pulmão tem-se a maior variabilidade e essa atinge extremos.

Após uma análise descritiva seguiu-se para a preparação dos dados, sendo esta preparação a normalização das características numéricas e codificação da variável reposta. O banco de dados foi separado em treino e teste, com separação 80,20 respectivamente. O primeiro modelo testado foi um MLP (Multilayer Perceptron) com camadas densas (totalmente conectadas), em que na primeira camada oculta foram usados 64 neurônios, com função de ativação ReLU e dropout de 30%, na segunda camada oculta 32 neurônios, função ReLU e 30% de dropout. Para a camada de saída tem se um neurônio para categoria da variável resposta. O otimizador usado foi o Adam, para a perda a Categorial Cross Entropy e métrica de acurácia resultados expostos nas tabelas 1 e 2.

Tabela 1 — Desempenho por Classe

Classe	Precisão	Sensibilidade	F1-score	Support
High	0.75	0.45	0.56	20
Low	0.92	0.88	0.90	65

Classe	Precisão	Sensibilidade	F1-score	Support
Medium	0.94	0.97	0.96	315

Elaborada pela autora (2025).

Tabela 2 — Métricas Gerais

Métrica	Precisão	Sensibilidade	F1-score
Macro Avg	0.87	0.77	0.81
Weighted Avg	0.93	0.93	0.93
Acurácia	—	—	**0.935**

Fonte: Elaborada pela autora (2025).

Observa-se uma boa acurácia, de mais de 93%. Porém, a acurácia não é uma boa métrica quando temos classes desbalanceadas que é o caso. Ao avaliar o desempenho por classe observa-se que a classe High tem um baixo desempenho, certamente provocado pelo desbalanceamento das classes.

Para contornar o desbalanceamento de classes utilizou-se pesos de classe para que a categoria desfavorecida receba um peso maior e o modelo possa ter um melhor aprendizado. Com a otimização do Grid Search várias combinações de hiperparâmetros foram testadas. Em suma, os melhores hiperparâmetros encontrados foram 128 neurônios, um dropout de 40%, otimizador Adam e 75 épocas que monstrou bons resultados um ligeiro ganho na acurácia chegando até 0,9350. Houve aumento na precisão e aumento nas métricas de recall e f1-score para a classe minoritaria que busca-se melhorar. Seguindo para o modelo otimizado 2, otimizou-se a função de ativação, quantidade de neurônios das duas camadas, taxa de dropout e o otimizador. O modelo otimizado 2 contou com 50 épocas, 64 neurônios na primeira camada, função de ativação tangente hiperbólica que lida bem com dados padronizados, dropout de 20%, segunda camada oculta com 32 neurônios e a mesma função de ativação da primeira, para a camada de saída função de ativação softmax e 3 neurônios para representar as 3 classes e os resultados podem ser observados nas tabela 3 e 4.

Tabela 3 — Desempenho por Classe (Modelo Otimizado 2)

Classe	Precisão	Sensibilidade	F1-score	Support
High	0.86	0.95	0.90	20
Low	0.94	0.97	0.95	65
Medium	0.99	0.98	0.98	315

Fonte:Elaborada pela autora (2025).

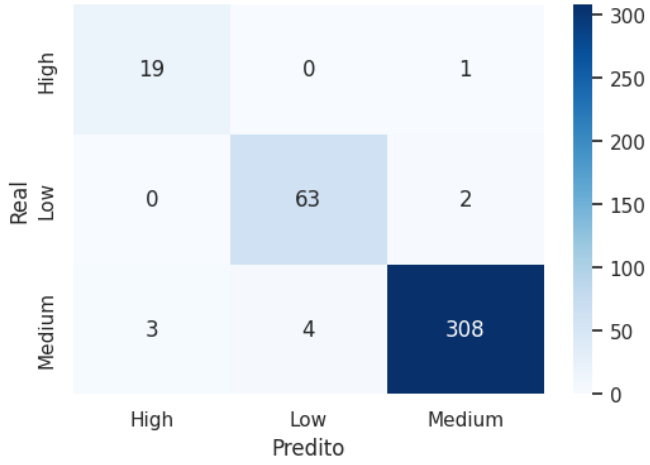
Tabela 4 — Métricas Gerais (Modelo Otimizado 2)

Métrica	Precisão	Sensibilidade	F1-score
Macro Avg	0.93	0.97	0.95
Weighted Avg	0.98	0.97	0.98
Acurácia	—	—	**0.975**

Fonte: Elaborada pela autora (2025).

Além destes, outros modelos foram testados mantendo a otimização e o balanceamento de classes. Estes, tiveram como diferença as taxa de dropout e a função de ativação alterada para sigmoide respectivamente. Em suma, o ganho não foi superior ao modelo otimizado já testado. Dessa forma, o modelo eleito como melhor foi o modelo otimizado 2.

Figura 3 – Matriz de confusão modelo otimizado 2.



Fonte: Elaborada pela autora (2025).

Observando a figura 3, que corresponde a matriz de confusão é visto que o modelo erra apenas uma vez para a classe minóritária e quando as classes menores são classificadas erradas caem sempre na classe majorirária.

CONCLUSÃO

Em geral, esse estudo refere-se a uma aplicação de Redes Neurais Artificiais (RNAs) para classificar o risco de câncer de um indivíduo a partir de seus fatores clínicos, comportamentais e hereditários. Sabendo das severidades trazidas pelo câncer e seu alcance global, técnicas que mapeiam o perfil do indivíduo para que, em caso de risco alarmante, seja realizado um maior acompanhamento que posteriormente pode se tornar um diagnóstico precoce, o que aumenta as chances de cura e sobrevivência, são de grande valia, especialmente em situações que o acesso a tecnologias de ponta é escasso.

O primeiro modelo construído mostrou um bom desempenho, com acurácia do modelo acima de 90%. Contudo, em casos de classes desbalanceadas, a acurácia pode não corresponder ao desempenho efetivo do modelo, que é este caso. Para a classe minoritária “High”, o desempenho é inferior. Para contornar esse problema, a técnica “class weight” foi utilizada para melhoria de sensibilidade.

Ademais, utilizando a otimização GridSearchCV, realizou-se a combinação de hiperparâmetros que tornassem o modelo mais eficaz e robusto. Nessa otimização, foram determinados o número de neurônios, a taxa de dropout, a quantidade de épocas e a otimização. O modelo com otimização trouxe melhorias, a começar pela acurácia, que atingiu 93,50%. O aumento não é tão expressivo, pois o modelo já fazia boas previsões para a classe majoritária. A melhoria expressiva foi na classe minoritária, trazendo mais equilíbrio entre as métricas de precisão, sensibilidade e F1-score. O modelo otimizado 2 foca em otimizar toda a rede, avaliando funções de ativação, número de neurônios das duas camadas, taxa de dropout e otimizador e este, apresentou os melhores resultados de forma geral, com acurácia superior a 97%. Outros modelos foram avaliados com diferentes configuração e não mostraram resultados superiores.

Logo, os resultados obtidos mostram que as redes neurais artificiais de arquitetura MLP podem ser ferramentas eficientes para avaliação da severidade do câncer, trazendo assim informações úteis do ponto de vista de saúde pública. Embora essa classificação não substitua os diagnósticos propostos pela medicina, ela pode ser um auxílio para controle de perfil clínico, priorização e acompanhamentos mais efetivos para aqueles que demonstrarem maiores riscos, buscando um diagnóstico precoce. Para trabalhos futuros, outras arquiteturas de redes neurais podem ser avaliadas e outras técnicas de balanceamento, visando um melhor desempenho e confiabilidade do modelo.

REFERÊNCIAS

AMBRÓSIO, P. E. Redes neurais artificiais no apoio ao diagnóstico diferencial de lesões intersticiais pulmonares. 2002. Dissertação (Mestrado) – Faculdade de Filosofia, Universidade de São Paulo, Ribeirão Preto, SP.

ARAUJO, L. H.; BALDOTTO, C.; CASTRO JR., G. D.; KATZ, A.; FERREIRA, C. G.; MATHIAS, C.; BARBIOS, C. H. Câncer de pulmão no Brasil. *Jornal Brasileiro de Pneumologia*, v. 44, p. 55-64, 2018.

BEJANI, M. M.; GHATEE, M. A systematic review on overfitting control in shallow and deep neural networks. Artificial Intelligence Review, p. 1–48, 2021.

BOMFIM, Simara Silva; GIOTTO, Ani Cátia; SILVA, Anna Gabriella e. CÂNCER DE PELE: CONHECENDO E PREVENINDO A POPULAÇÃO. **REVISA**, *[S. l.]*, v. 7, n. 3, p. 255–259, 2018.

OLIVEIRA SANTOS, M.; LIMA, F. C. D. S.; MARTINS, L. F. L.; OLIVEIRA, J. F. P.; ALMEIDA, L. M.; CAMARGO CANCELA, M. Estimativa de incidência de câncer no Brasil, 2023-2025. *Revista Brasileira de Cancerologia*, v. 69, n. 1, 2023.

FERREIRA, A.; FERREIRA, R. P.; SILVA, A. M. da; FERREIRA, A.; SASSI, R. J. Um estudo sobre previsão da demanda de encomendas utilizando uma rede neural artificial. Blucher Marine Engineering Proceedings, v. 2, n. 1, p. 353–364, 2016.

FERNANDES, M. G. S.; REIS, E. B. B.; PERINI, H. F.; MIGUEL NETO, J.; MIGUEL, C. B.; FELIPE, A. G. B. Microbioma intestinal versus câncer colorretal.

FLECK, L.; TAVARES, M. H. F.; EYNG, E.; HELMANN, A. C.; ANDRADE, M. D. M. Redes neurais artificiais: princípios básicos. Revista Eletrônica Científica Inovação e Tecnologia, v. 1, n. 13, p. 47–57, 2016.

HAYKIN, S. Redes neurais: princípios e práticas. 2. ed. São Paulo: Bookman, 2001. 900 p.

IGUS, J. P. Data mining with neural network: solving business problems from applications development to decision support. New York: McGraw-Hill, 1996.

INSTITUTO NACIONAL DE CÂNCER. Estimativa 2023: Incidência de Câncer no Brasil. Rio de Janeiro: INCA, 2023.

NIED, A. Treinamento de redes neurais artificiais baseado em sistemas de estrutura variável com taxa de aprendizado adaptativa. 2007. Tese (Doutorado) – Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Minas Gerais, Belo Horizonte, MG.

SARRIS, A. B.; CANDIDO, F. J. L. F.; PUCCI FILHO, C. R.; STAICHAK, R. L.; TORRANI, A. C. K.; SOBREIRO, B. P. Câncer de próstata: uma breve revisão atualizada. Visão Acadêmica, v. 19, n. 1, p. 137–151, 2018.

SRIVASTAVA, N.; HINTON, G.; KRIZHEVSKY, A.; SUTSKEVER, I.; SALAKHUTDINOV, R. Dropout: uma maneira simples de evitar o sobreajuste em redes neurais. Journal of Machine Learning Research, v. 15, n. 1, p. 1929–1958, 2014.

SOUZA, V.O.; GRANDO, J.P.S.; FILHO, J.O.; Tempo decorrido entre o diagnóstico de câncer de mama e o início do tratamento, em pacientes atendidas no Instituto de Câncer de Londrina (ICL). RBM Rev Bras Med, 2008.

SPÖRL, C.; CASTRO, E. G.; LUCHIARI, A. Aplicação de redes neurais artificiais na construção de modelos de fragilidade ambiental. Revista do Departamento de Geografia, v. 21, n. 1, p. 113–135, 2011.