USING MACHINE LEARNING TO PREDICT RECURRENCE OF BRAIN TUMORS

UTILIZAÇÃO DE MACHINE LEARNING PARA PREVER A RECORRÊNCIA DE TUMORES CEREBRAIS

LUPIANEZ FILHO, José Roberto; TAVARES, Lucas Costa; SANTOS, Flávia Aparecida Oliveira; CARVALHO, Jaqueline Corrêa Silva; CARVALHO, Marcos Alberto; RAMOS, Celso de Ávila; BASTOS, Camila; SOUZA, Patrícia Carolina; SILVA, Vinícius Duarte Esteves

José Roberto Lupianez Filho, UNIFENAS, Brasil

Lucas Costa Tavares, UNIFENAS, Brasil

Flávia Aparecida Oliveira Santos, UNIFENAS, Brasil

Jaqueline Corrêa Silva Carvalho, UNIFENAS, Brasil

Marcos Alberto Carvalho, UNIFENAS, Brasil

Celso de Ávila Ramos, UNIFENAS, Brasil

Camila Bastos, UNIFENAS, Brasil

Patrícia Carolina Souza, UNIFENAS, Brasil

Vinícius Duarte Esteves, UNIFENAS, Brasil

Revista Científica da UNIFENAS
Universidade Professor Edson Antônio Velano, Brasil
ISSN: 2596-3481
Publicação: Trimestral
vol. 6, nº. 5, 2024
revista@unifenas.br

Recebido: 08/07/2024
Aceito: 28/08/2024
Publicado: 09/09/2024

URL: https://revistas.unifenas.br/index.php/revistaunifenas/issue/view/52

DOI: 10.29327/2385054.6.5-12

ABSTRACT: Tumors can cause distress in people close to them, and in the process of investigating the subject, discoveries about related topics can be found. This work seeks to use machine learning to predict the recurrence of brain tumors in patients based on some attributes of these patients. This study was developed using Machine Learning techniques, used a Kaggle database, and the development steps were carried out in Google Colab, using the Python language and several specialized libraries. This prediction can help doctors and patients make more informed decisions about the treatment and monitoring of the disease. First, data preprocessing was done, which involved the application of one-hot encoding to the Treatment column, which contains three distinct categories (Radiation, Surgery, and Chemotherapy). This transformation is necessary to convert categorical data into a numerical format that can be interpreted by Machine Learning algorithms. This study demonstrated the effectiveness of using Machine Learning techniques, specifically neural networks, to predict the recurrence of brain tumors. With an accuracy of 98.67%, F1 score of 99.04% and recall of 99.28%, the model proved to be highly accurate, indicating that the use of Machine Learning can be a powerful tool in medicine, helping doctors and patients in making more informed decisions about the treatment and monitoring of brain tumors.

KEYWORDS: Artificial Intelligence, Machine Learning, Supervised, Medicine, Brain Tumors.

RESUMO: Tumores podem causar comoção em pessoas próximas e, no processo de investigação do tema, pode-se encontrar descobertas a respeito de temas próximos. Este trabalho tem como objetivo usar o machine learning para prever a recorrência de tumores cerebrais em pacientes baseando-se em alguns atributos desses pacientes. Este estudo foi desenvolvido utilizando técnicas de Machine Learning, utilizou uma base de dados do Kaggle e as etapas do desenvolvimento foram realizadas no Google Colab, utilizando a linguagem Python e diversas bibliotecas especializadas. Esta previsão pode auxiliar médicos e pacientes na tomada de decisões mais informadas sobre o tratamento e monitoramento da doença. Primeiramente foi feito o Pré-processamento dos dados, que envolveu a aplicação de one-hot encoding para a coluna Treatment, que

contém três categorias distintas (Radiation, Surgery e Chemotherapy). Esta transformação é necessária para converter dados categóricos em um formato numérico que pode ser interpretado pelos algoritmos de Machine Learning. Este estudo demonstrou a eficácia do uso de técnicas de Machine Learning, especificamente redes neurais, para prever a recorrência de tumores cerebrais. Com uma acurácia de 98,67%, F1 score de 99,04% e recall de 99,28%, o modelo mostrou-se altamente preciso, indicando que o uso de Machine Learning pode ser uma ferramenta poderosa na medicina, auxiliando médicos e pacientes na tomada de decisões mais informadas sobre o tratamento e monitoramento de tumores cerebrais.

PALAVRAS-CHAVE Inteligência Artificial, Aprendizado de máquina Supervisionado, Medicina, Tumores cerebrais.

1 INTRODUÇÃO

Os tumores primários do sistema nervoso central representam cerca de 5% das neoplasias e sua incidência é de 6 a 12 casos a cada 100.000 habitantes por ano. Existem vários tipos de tumores cerebrais, alguns com alta porcentagem de cura, outros extremamente agressivos. Comumente, o tratamento envolve cirurgia, quimioterapia e radioterapia, isoladamente ou de maneira combinada [1].

Tumores podem causar comoção em pessoas próximas, como membros da família e amigos, e, no processo de investigação do tema, pode-se encontrar, também, descobertas a respeito de temas próximos, como, por exemplo, tumores em outros órgãos e outros problemas cerebrais e/ou neurais.

Segundo a Oracle [2], "O machine learning (ML) é o subconjunto da inteligência artificial (IA) que se concentra na construção de sistemas que aprendem, ou melhoram o desempenho, com base nos dados que consomem. A inteligência artificial é um termo amplo que se refere a sistemas ou máquinas que imitam a inteligência humana".

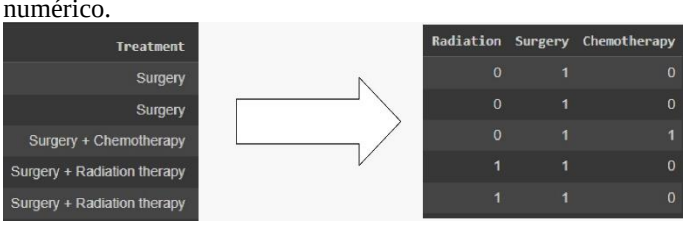
Medicina e computação podem se tornar aliadas de peso no combate ou diminuição de doenças e este trabalho procura usar o machine learning (referido, às vezes, como aprendizado de máquina) para prever a recorrência de tumores cerebrais em pacientes baseando-se na idade, no sexo, no tipo do tumor, no grau do tumor, no local do tumor, no resultado do tratamento e na área de recorrência.

2 METODOLOGIA

Este estudo foi desenvolvido utilizando técnicas de Machine Learning com o objetivo de prever a recorrência de tumores cerebrais. A base de dados utilizada foi obtida do Kaggle (Brain Tumor Stage-Based Recurrence Patterns) [3], um conhecido repositório de bases de dados para análises e competições de Data Science, contendo as seguintes colunas relevantes para o estudo: Idade (Age), Gênero (Gender), Tipo de Tumor (Tumor Type), Grau do Tumor (Tumor Grade), Localização do Tumor (Tumor Location), Tratamento (Treatment), Resultado do Tratamento (Treatment Outcome) e Local de Recorrência (Recurrence Site). As etapas do desenvolvimento e execução dos modelos de Machine Learning foram realizadas no Google Colab, utilizando a linguagem Python e diversas bibliotecas especializadas. Esta previsão pode auxiliar médicos e pacientes na tomada de decisões mais informadas sobre o tratamento e monitoramento da doença.

Primeiramente foi feito o Pré-processamento dos dados, uma etapa crucial em qualquer projeto de Machine Learning. Neste estudo, o pré-processamento envolveu a aplicação de one-hot encoding para a coluna Treatment, que contém três categorias distintas (Radiation, Surgery e Chemotherapy). Esta transformação é necessária para converter dados categóricos em um formato numérico que pode ser interpretado pelos algoritmos de Machine Learning.

Figura 1. Apresentação dos dados categóricos em formato numérico.



Em seguida foi feita a normalização dos dados, uma etapa essencial em Machine Learning para garantir que todas as variáveis numéricas estejam na mesma escala. Isso é importante porque variáveis com magnitudes diferentes podem influenciar de forma desigual o desempenho do modelo. Utilizamos o StandardScaler para normalizar as colunas numéricas, o que ajusta os dados para terem média zero e desvio padrão um, assegurando um tratamento equilibrado de todas as características pelo modelo.

Transformar dados categóricos e numéricos é igualmente crucial, pois algoritmos de Machine Learning requerem entradas numéricas. Utilizamos o LabelEncoder para converter colunas categóricas, como Gender, Tumor Type, Tumor Grade e Tumor Location, em valores numéricos. Isso permite que esses dados categóricos sejam processados de forma eficiente pelos algoritmos, facilitando o aprendizado e a análise.

O dataset foi dividido em 70% para treinamento e 30% para teste usando a função train_test_split da biblioteca scikit-learn, garantindo que o modelo seja avaliado com dados não vistos e tenha boa capacidade de generalização. O modelo de Machine Learning foi construído utilizando a biblioteca Keras. A rede neural sequencial criada incluiu uma camada densa (Dense) com 64 unidades e ativação ReLU, seguida por uma camada Dropout com taxa de 0.4 para prevenir o overfitting. Em seguida, foram adicionadas mais duas camadas densas, uma com 32 unidades e outra com 64 unidades, ambas com ativação ReLU e camadas Dropout com taxa de 0.4. A camada de saída continha uma unidade com ativação sigmoidal para prever a probabilidade de recorrência do tumor. O modelo foi compilado utilizando o otimizador Adam e a função de perda binary_crossentropy,

sendo avaliado pela métrica de acurácia. O treinamento foi realizado em 200 epochs com batch_size de 32.
Após o treinamento, o modelo foi utilizado para fazer previsões nos dados de teste. As previsões retornadas pelo modelo foram probabilidades, que foram então convertidas em valores binários (0 ou 1) utilizando um limiar de 0.5. Este processo permitiu transformar as probabilidades em classificações discretas de recorrência ou não recorrência do tumor.

3 RESULTADOS E DISCUSSÃO

Os resultados do estudo foram obtidos após a realização do treinamento do modelo de rede neural com os dados de treino e a subsequente avaliação com os dados de teste. A seguir, foram apresentados os principais resultados obtidos, incluindo as métricas de desempenho do modelo e as análises das previsões realizadas.
As métricas de desempenho foram calculadas para avaliar a eficácia do modelo em prever a recorrência de tumores cerebrais. As principais métricas incluem a acurácia, a pontuação F1 e a sensibilidade (recall). A Tabela 1 apresenta os valores obtidos para cada uma dessas métricas.

Tabela 1. Apresentação dos valores obtidos para as métricas de desempenho

Métrica	Valor
Acurácia	987
Pontuação F1	990
Sensibilidade	993

A matriz de confusão fornece uma visão detalhada do desempenho do modelo, mostrando o número de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos conforme apresentado na tabela 2.

Tabela 2. Visão detalhada do desempenho do modelo

Coluna 1	Predito Negativo	Predito Positivo
Real Negativo	181	5
Real Positivo	3	411

Os resultados obtidos indicam que o modelo de rede neural desenvolvido é altamente eficaz na previsão da recorrência de tumores cerebrais, com uma acurácia de 98.7%, que é significativamente alta. A alta sensibilidade (99.3%) também indica que o modelo é extremamente eficaz em identificar pacientes que terão recorrência, o que é crucial para intervenções precoces. A pontuação F1 de 99.0% reforça a qualidade do modelo, demonstrando um equilíbrio entre precisão e recall.

Os resultados corroboram a importância do pré-processamento dos dados. A normalização das variáveis numéricas e a transformação dos dados categóricos em numéricos foram essenciais para garantir que o modelo tratasse todas as características de maneira equilibrada.
Apesar dos resultados promissores, este estudo apresenta algumas limitações. A base de dados utilizada pode conter vieses que não foram totalmente mitigados, o que pode influenciar os resultados. Além disso, a falta de dados demográficos detalhados limita a generalização dos resultados para populações diferentes. A utilização de um único tipo de modelo (rede neural) sem comparação com outros algoritmos de Machine Learning também representa uma limitação.

CONCLUSÃO

Este estudo demonstrou a eficácia do uso de técnicas de Machine Learning, especificamente redes neurais, para prever a recorrência de tumores cerebrais. Com uma acurácia de 98,67%, F1 score de 99,04% e recall de 99,28%, o modelo mostrou-se altamente preciso, indicando que o uso de Machine Learning pode ser uma ferramenta poderosa na medicina, auxiliando médicos e pacientes na tomada de decisões mais informadas sobre o tratamento e monitoramento de tumores cerebrais.
Os resultados confirmam a importância do pré-processamento de dados, incluindo a normalização das variáveis numéricas e a transformação de variáveis categóricas, essenciais para o desempenho do modelo. No entanto, o estudo apresenta algumas limitações, como a possibilidade de vieses nos dados e a falta de comparação com outros algoritmos de Machine Learning.
Futuros estudos poderiam explorar a inclusão de mais variáveis clínicas e demográficas para melhorar a acurácia e a generalização do modelo. Além disso, a utilização de outras técnicas de Machine Learning, como Random Forests ou SVM, poderia ser investigada para comparar o desempenho com o modelo de rede neural utilizado neste estudo. A validação do modelo em diferentes bases de dados também seria útil para avaliar sua robustez e aplicabilidade em contextos variados.

REFERÊNCIAS

[1] Friedman R, Sherman Jr CD. Câncer de pele. In: Blaquiere RM, Bosch FX, Boyd NF, Brada M, Brennan MF, Bruera E, organizadores. Manual de oncologia clínica. São Paulo: Fundação Oncocentro de São Paulo;1999. p.245-253

[2] Oracle. O que é Machine Learning? c2024 [Acesso em18 de jun. de 2024] Disponível em: https://www.oracle.com/br/artificial-intelligence/machine-learning/what-is-machine-learning/#:~:text=O%20machine%20learning%20(ML)%20%C3%A9,que%20imitam%20a%20intelig%C3%AAncia%20humana.

[3] Prathamesh Pradeep Dessai. BRAIN TUMOR STAGE-BASED RECURRENCE PATTERNS [base de dados na Internet]. Plataforma Kaggle; 2024 [Acesso em 18 jun

2024]. Disponível em https://www.kaggle.com/datasets/thegoanpanda/brain-tumor-stage-based-recurrence-patterns