UNIVERSIDADE FEDERAL DO CEARÁ

# TECHNICAL REPORT

Aluno: Erick Ramos Coutinho

## 1. Introdução

*Este projeto em Python tem como objetivo construir um modelo de aprendizado de máquina capaz de identificar se uma pessoa tem um tumor cerebral ou não, utilizando os dados de um extenso dataset. O câncer cerebral é uma doença grave que pode ter consequências devastadoras se não for detectado e tratado precocemente.*

*Usando técnicas, o modelo será treinado com um conjunto de dados que contém os dados de pacientes com e sem tumores cerebrais. O objetivo é desenvolver um modelo preciso e confiável que possa ser utilizado como uma ferramenta auxiliar pelos profissionais de saúde no diagnóstico e tratamento de pacientes com tumores cerebrais.*

## 2. Observações

*O relatório sobre o projeto com o dataset de estudo sobre tumores cerebrais destaca a tarefa de descobrir se uma pessoa tem ou não um tumor cerebral a partir de dados de um dataset. Isso se deve ao fato de que as características presentes no dataset dos tumores cerebrais nem sempre são facilmente identificáveis ou descritíveis. Além disso, o dataset é extenso e composto por uma grande quantidade de dados de difícil compreensão, o que pode dificultar a análise e entendimento dos dados de modo geral.*

*Como os dados são apenas numéricos, pode ser difícil entender o que cada coluna significa e qual é a sua importância para a identificação do tumor. Por esse motivo, é importante realizar uma análise dos dados, selecionando algumas colunas para serem analisadas, além da classe e assim tirarmos as conclusões.*

## 3. Resultados e discussão

***Questão 1:***

*Na primeira questão foi feito o pré-processamento de dados, como por exemplo: verificar a existência de NA no dataset, que não havia no caso.*

```
 Verificando a existência de NA
Unnamed: 0    0
X53416        0
M83670        0
X90908        0
M97496        0
             ..
M13699.1      0
X54489        0
T55008        0
M10065.2      0
y             0
Length: 7466, dtype: int64
```

*Além disso, foi feita a transformação da classe de string para inteiro, e assim foi feita a distribuição da classe, tendo 18 com "tumor" e 18 "normal".*

```
DISTRIBUIÇÃO DE CLASSE, QUANTIDADE DE 1 E QUANTIDADE DE 0:
0    18
1    18
Name: classe, dtype: int64
```

***Questão 2:***

*Na questão 2 foi feita a divisão dos dados em treino em teste, sem a normalização. logo após foi implementado o knn, e obtive o seguinte resultado:*

```
[36 rows x 7467 columns]
Acurácia sem normalizar:
0.875
```

***Questão 3:***

*Na questão 3, o dataset foi normalizado de duas formas: normalização logarítmica e normalização de média 0 e variância unitária e das duas formas foi implementado o knn.*

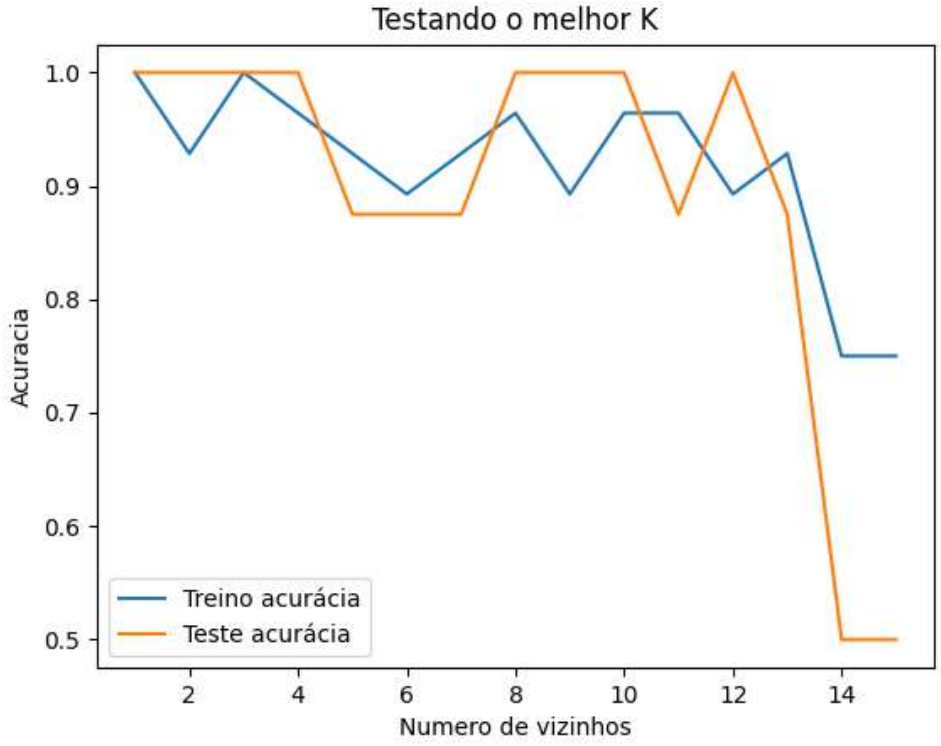
*Após a implementação foi feita a comparação das acurácias:*

```
Acurácia Normalizada com log e com scaler:
0.875 0.9642857142857143
```

***Questão 4:***

*Na questão 4, usei a melhor normalização para fazer o gráfico mais preciso possível, que no caso foi a normalização "Scaler", que foi 96,4%.*

*Após isso foi plotado o gráfico para a identificação do melhor "K":*




*Por meio do gráfico percebemos que o melhor K está entre os intervalos de 1-4 e 8-10. e 12.*

```
Acuracia no teste:  {1: 1.0, 2: 1.0, 3: 1.0, 4: 1.0, 5: 0.875, 6: 0.875, 7: 0.875, 8: 1.0, 9: 1.0, 10: 1.0, 11: 0.875, 12: 1.0, 13: 0.875, 14: 0.5, 15: 0.5}
```

## 4. Conclusões

Após a análise e processamento do dataset, podemos dizer que sim, os resultados esperados foram satisfatórios. Apesar das dificuldades encontradas ao longo do processo, mesmo com as tabelas sem identificação e a grande quantidade de dados, foi feito o pré-processamento e a tabela ficou totalmente utilizável.

Com a análise dos dados, podemos perceber que o dataset ele é bem completo contando com todas as informações necessárias, mesmo que elas sejam de difícil compreensão, pois a acurácia foi ótima, tendo um alto índice de assertividade, usando

o algoritmo de machine learning **Knn** (K-nearest neighbors, ou "K-vizinhos mais próximos). Somente após a normalização do dataset, conseguimos os melhores "K" possíveis, satisfazendo o objetivo.

## 5. Próximos passos

*Após apresentar os resultados e as conclusões do projeto de identificação de tumores cerebrais por meio de uma base de dados, alguns dos próximos passos para o relatório podem incluir:*

*Discussão das limitações do estudo: É importante destacar as limitações do estudo, como a qualidade dos dados, o tamanho da amostra, ou outras limitações metodológicas, e discutir possíveis estratégias para superá-las.*

*Sugestões para trabalhos futuros: O relatório pode apresentar sugestões para trabalhos futuros, como a utilização de outras técnicas de aprendizado de máquina, a inclusão de outras variáveis na análise ou o uso de outros algoritmos para lidar com o grande volume de dados.*

*Discussão sobre a aplicabilidade do modelo: É importante discutir sobre a aplicabilidade do modelo de identificação de tumores cerebrais em situações clínicas reais e como ele pode ser incorporado na prática médica.*