CLASSIFICAÇÃO DE GRUPOS DE RISCO PARA GLIOMAS UTILIZANDO MINERAÇÃO DE DADOS

CLASSIFICATION OF RISK GROUPS FOR GLIOMAS USING DATA MINING

Alef Weslei Moreira dos Santos Cardoso (alef.cardoso@aluno.ifsp.edu.br, Instituto Federal de Educação, Ciência e Tecnologia de São Paulo, São Paulo, Votuporanga)
Ricardo Conde Camillo da Silva (ricardo.conde@ifsp.edu.br, Instituto Federal de Educação, Ciência e Tecnologia de São Paulo, São Paulo, Votuporanga)

RESUMO

Gliomas são os tumores primários mais comuns do cérebro, com classificações distintas baseadas em critérios histológicos, moleculares e clínicos. No entanto, o diagnóstico preciso e o tratamento adequado podem ser desafiadores, e os testes moleculares utilizados para diagnosticá-los são caros. Neste contexto, a mineração de dados surge como uma abordagem promissora para identificar padrões e características que podem melhorar a classificação de gliomas. Neste estudo, foi utilizado um conjunto de dados que inclui 20 genes mutados e 3 características clínicas de pacientes com gliomas. Algoritmos de regressão, *clustering*, *KNN*, bem como diversas ferramentas como IBM Watson e algoritmos no Weka foram empregados para conduzir a mineração, visando identificar o grupo de risco e características clínicas para a classificação de pessoas com maior incidência desses gliomas. Os resultados obtidos foram analisados e discutidos, revelando *insights* importantes para identificação de grupos de indivíduos mais propensos a desenvolver este tipo de tumor. Esta pesquisa destaca a importância da mineração de dados no contexto dos gliomas e seu potencial para melhorar a prevenção e orientar a tomada de decisões.

PALAVRAS-CHAVE: Gliomas, Mineração de Dados, Tumores Cerebrais, Grupo de Risco.

ABSTRACT

Gliomas are the most common primary brain tumors, with distinct classifications based on histological, molecular, and clinical criteria. However, precise diagnosis and appropriate treatment can be challenging, and the molecular tests used to diagnose them can be expensive. In this context, data mining emerges as a promising approach to identify patterns and features that can enhance glioma classification. In this study, a dataset including 20 mutated genes and 3 clinical characteristics of glioma patients was utilized. Regression algorithms, clustering, KNN, as well as various tools such as IBM Watson and algorithms in Weka were employed for mining, aiming to identify the risk group and clinical characteristics for the classification of individuals with a higher incidence of these gliomas. The obtained results were analyzed and discussed, revealing important insights for identifying groups of individuals more prone to developing this type of tumor. This research emphasizes the importance of data mining in the context of gliomas and its potential to improve prevention and guide decision-making.

KEY WORDS: *Gliomas, Data Mining, Brain Tumors, Risk Group.*

INTRODUÇÃO

Os gliomas são tumores primários que se originam no tecido glial do cérebro e são classificados como o tipo mais comum de tumor cerebral (Santos, 2021). Eles podem ser divididos em dois tipos principais: glioma de grau inferior (*lower grade glioma* - LGG) e *glioblastoma multiforme* (GBM). O LGG é um tumor de

crescimento mais lento e menos agressivo, enquanto o GBM é um tumor de alto grau, agressivo e com prognóstico desfavorável (Biterge-sut, B.2020).

O diagnóstico preciso e a classificação adequada dos gliomas são essenciais para o planejamento do tratamento e para determinar as opções terapêuticas mais adequadas para cada paciente (Medeiros Junior, 2021).

Além dos métodos tradicionais de diagnóstico por imagem, a utilização de técnicas de mineração de dados tem se mostrado promissora para auxiliar nesse processo. Existem várias formas de abordar este tema, inclusive através da mineração de dados de gliomas é possível identificar características moleculares, clínicas e genéticas que podem estar relacionadas à classificação dos tumores em LGG ou GBM. Isso permite a descoberta de biomarcadores potenciais e a estratificação de pacientes em grupos de risco.

A Mineração de Dados pode ser definida como um "passo no processo de Descoberta de Conhecimento que consiste na realização da análise dos dados e na aplicação de algoritmos de descoberta que, sob certas limitações computacionais, produzem um conjunto de padrões de certos dados." (Camilo, et al., 2009, apud Fayyad, et al., 1996).

Ela utiliza técnicas estatísticas, algoritmos de aprendizado de máquina e análise exploratória de dados para extrair conhecimento e *insights* valiosos. A aplicação da mineração de dados em oncologia, incluindo o estudo de gliomas, tem se tornado cada vez mais relevante para a compreensão desta doença, para o desenvolvimento de modelos preditivos e para a melhoria de resultados clínicos.

Apesar de ser um tema de extrema relevância, este estudo não teve como objetivo principal prever e classificar indivíduos em grupos propensos a desenvolver gliomas LGG ou GBM com base na combinação de genes, embora tenham sido utilizados estes dados no processamento em *clustering*, IBM Watson e Weka. Em vez disso, focou-se na classificação dos indivíduos em grupos de risco, considerando faixa etária, etnia e sexo. Pois é fundamental compreender os fatores demográficos que podem aumentar a predisposição a esses tipos de tumores cerebrais. Ao analisar esses parâmetros, pode-se obter percepções valiosas para a prevenção e detecção precoce de casos de gliomas.

OBJETIVOS

O objetivo principal deste artigo é explorar a aplicação de técnicas de mineração de dados para a análise de dados de gliomas, buscando identificar características associadas à faixa etária, etnia, sexo e grupos de risco para o desenvolvimento de tumores LGG ou GBM.

O diagnóstico preciso e a classificação adequada dos gliomas são cruciais para o planejamento do tratamento, e este estudo se propõe a investigar como a mineração de dados pode ser uma ferramenta promissora nesse contexto. A análise abrangeu a utilização de algoritmos de descoberta para identificar padrões específicos nos dados sobre gliomas.

REVISÃO DE LITERATURA

A seguir, serão discutidos alguns estudos correlacionados à mineração de dados, aprendizado de máquina e oncologia. "A oncologia é a ciência que estuda os diversos tipos de câncer, com objetivo paliativo ou curativo" (Dos Santos, et al. 2021).

Em seu artigo para o Departamento de Engenharia Biomédica da Universidade de Tianjin, os autores afirmam que o uso do aprendizado de máquina representa uma ferramenta fundamental para moldar o futuro do entendimento sobre gliomas, possibilitando a exploração plena do potencial contido nos extensos conjuntos de dados biomédicos e de pacientes (Wu, Yameng et al., p. 3169, 2021).

Neste mesmo artigo os autores discorrem sobre como o aprendizado de máquina foi aplicado para acelerar o processo de mineração de dados, haja vista a vasta quantidade de informações genéticas e de imagem derivadas das tecnologias computacionais e de publicações biomédicas atuais.

Também abordam a situação atual e delineiam as perspectivas futuras do aprendizado de máquina em estudos relacionados à gliomas. Adicionalmente, comparam as soluções existentes de métodos de aprendizado de máquina, destacando as limitações concernentes à previsão e diagnóstico de gliomas.

Em uma outra literatura acerca do aprendizado de máquina no diagnóstico de gliomas, os autores introduzem a aplicação de técnicas de processamento de imagens de tumores cerebrais obtidas por ressonância magnética, aliadas à mineração de séries temporais para a descoberta de padrões (Medeiros e Ferreiro, 2022).

Detalham o método empregado, desde a seleção das imagens no conjunto de dados até a descrição da fase de pré-processamento, abrangendo ainda o agrupamento de séries temporais realizado por meio dos algoritmos de clustering, DTW e HDBSCAN.

Nas literaturas resultantes de pesquisas de artigos científicos no google scholar foram percebidas abordagens teóricas sobre o assunto e técnicas de mineração focada em imagens. Este artigo se faz interessante e útil pois se propõe a descrever e demonstrar resultados da análise de um *dataset* com inúmeros diagnósticos de gliobastoma, e respectivos dados étnicos, genótipos e fenótipos dos pacientes utilizando técnicas de mineração de dados e aprendizado de máquina para tentar obter informações que contribuam para o diagnóstico precoce ou corelações com grupos de riscos ou predispostos a desenvolver a doença. Validando assim um potencial benefício ético para a sociedade.

Nas literaturas mencionadas, e nas provenientes de pesquisas no Google Scholar, foram identificadas abordagens teóricas sobre o tema e técnicas de mineração voltadas para imagens. Este artigo torna-se válido ao propor a análise de um conjunto de dados que contém diversas variações de diagnósticos de glioblastoma, e os respectivos dados étnicos, genotípicos e fenotípicos dos pacientes diagnosticados.

Por meio do uso de técnicas de mineração de dados e aprendizado de máquina, busca-se obter informações que possam contribuir para o diagnóstico precoce ou estabelecer correlações com grupos de riscos ou predispostos a desenvolver a doença. Isso valida um potencial benefício ético para a sociedade.

MATERIAL E MÉTODOS

A análise de dados foi conduzida através do ambiente *Jupyter Notebook*, empregando a linguagem de programação *Python*. Durante o processo, foram empregados três algoritmos de aprendizado de máquina distintos: regressão, *clustering* e KNN (*K-Nearest Neighbors*).

O IBM Watson, que de acordo com SILVA e MATOS (2018) “é uma plataforma tecnológica que utiliza processamento de linguagem natural e aprendizagem de máquina para gerar conhecimento e tomar decisões com grandes quantidades de dados, incluindo também tipos não estruturados”. Foi empregado para realizar a mineração e análise do conjunto de dados, uma vez que incorpora uma variedade de algoritmos para a extração de informações. Ele identificou o algoritmo Classificador XGBoost como algoritmo mais adequado para o conjunto de dados a ser minerado.

A mineração de dados também foi realizada utilizando o software Weka. De acordo com DA SILVA et al. (2021), "O Weka é uma coleção de ferramentas e algoritmos de aprendizado de máquina voltada para tarefas de mineração de dados e áreas específicas de inteligência artificial (IA) dedicadas ao estudo de aprendizagem de máquina"(p.191-197).

Dentro deste contexto foram utilizados os algoritmos *J48*, *NaiveBayes*, *One R* e *Random Forest*. Essa abordagem permitiu explorar diversas técnicas de análise, cada uma com suas particularidades, proporcionando uma análise abrangente do conjunto de dados.

A. Conjunto de Dados

Foi utilizado um conjunto de dados composto por informações demográficas, características clínicas e dados genéticos de pacientes com gliomas. O conjunto de dados foi pré-processado e organizado, garantindo a qualidade e a integridade dos dados para análise. O *Dataset* pode ser encontrado no repositório online da UC Irvine Machine Learning Repository, com o seguinte nome: Glioma Grading Clinical and Mutation Features Dataset. Foi doado para o repositório em 12/13/2022.

B. Mineração

Inicialmente, foi realizada uma análise exploratória dos dados, verificando-se a distribuição das variáveis e identificando possíveis outliers ou dados faltantes. Em seguida, realizou-se o pré-processamento dos dados, removendo instâncias com informações ausentes e convertendo as variáveis categóricas e numéricas, quando necessário. Foram aplicados algoritmos de regressão, *clustering*, *KNN* para realizar a classificação dos indivíduos em grupos de risco. A regressão foi utilizada para analisar a relação entre as variáveis demográficas e clínicas, e as características genéticas foram exploradas por meio do clustering. O algoritmo *KNN* foi aplicado para a classificação propriamente dita, considerando a faixa etária, etnia e sexo como variáveis preditoras.

O algoritmo *KNN* por sua vez é um método de aprendizado de máquina que utiliza a abordagem baseada em vizinhos para classificação e regressão. As predições são realizadas baseando-se nos exemplos mais parecidos com o que deve ser predito (Cambronero e Moreno, 2006).

Enquanto os algoritmos de regressão são específicos para fazer previsões numéricas e os algoritmos de *clustering* para o agrupamento de dados não rotulados.

RESULTADOS E DISCUSSÃO

Os resultados da análise de regressão indicaram associações entre a faixa etária, a etnia e os diversos tipos de gliomas. A seguir, serão apresentados gráficos e figuras pertinentes para aprimorar a compreensão dos resultados.

A Figura 1 exibe o conjunto de dados em sua forma original, sem qualquer forma de manipulação.

Figura 1: *Dataset* Completo

	Grade	Project	Case_ID	Gender	Age_at_diagnosis	Primary_Diagnosis	Race	IDH1	TP53	ATRX	...	FUBP1	
0	LGG	TCGA-LGG	TCGA-DU-8164	Male	51 years 108 days	Oligodendroglioma, NOS	white	MUTATED	NOT_MUTATED	NOT_MUTATED	...	MUTATED	NO
1	LGG	TCGA-LGG	TCGA-QH-A6CY	Male	38 years 261 days	Mixed glioma	white	MUTATED	NOT_MUTATED	NOT_MUTATED	...	NOT_MUTATED	NO
2	LGG	TCGA-LGG	TCGA-HW-A5KM	Male	35 years 62 days	Astrocytoma, NOS	white	MUTATED	MUTATED	MUTATED	...	NOT_MUTATED	NO
3	LGG	TCGA-LGG	TCGA-E1-A7YE	Female	32 years 283 days	Astrocytoma, anaplastic	white	MUTATED	MUTATED	MUTATED	...	NOT_MUTATED	NO
4	LGG	TCGA-LGG	TCGA-S9-A6WG	Male	31 years 187 days	Astrocytoma, anaplastic	white	MUTATED	MUTATED	MUTATED	...	NOT_MUTATED	NO
...	...	...	...	...	...	...	...	...	...	...	...	...	
857	GBM	TCGA-GBM	TCGA-19-5959	Female	77 years 325 days	Glioblastoma	white	NOT_MUTATED	NOT_MUTATED	NOT_MUTATED	...	NOT_MUTATED	NO
858	GBM	TCGA-GBM	TCGA-16-0846	Male	85 years 65 days	Glioblastoma	white	NOT_MUTATED	MUTATED	NOT_MUTATED	...	NOT_MUTATED	NO
859	GBM	TCGA-GBM	TCGA-28-1746	Female	77 years 178 days	Glioblastoma	white	NOT_MUTATED	MUTATED	NOT_MUTATED	...	NOT_MUTATED	NO
860	GBM	TCGA-GBM	TCGA-32-2491	Male	63 years 121 days	Glioblastoma	white	NOT_MUTATED	MUTATED	NOT_MUTATED	...	NOT_MUTATED	
861	GBM	TCGA-GBM	TCGA-06-2557	Male	76 years 221 days	Glioblastoma	black or african	NOT_MUTATED	NOT_MUTATED	NOT_MUTATED	...	NOT_MUTATED	NO

Fonte: Autor, 2023.

É perceptível a presença de várias grandezas que estão atuando como cabeçalhos das colunas neste conjunto de dados. Para viabilizar a aplicação do algoritmo de regressão, foi imprescindível normalizar o *dataset* e remover as colunas desnecessárias. O resultado dessa operação pode ser visualizado na Figura 2:

Figura 2: *Dataset* Normalizado

	Age_at_diagnosis	Gender	Race
0	51.0	0	0
1	38.0	0	0
2	35.0	0	0
3	32.0	1	0
4	31.0	0	0

Fonte: Autor, 2023.

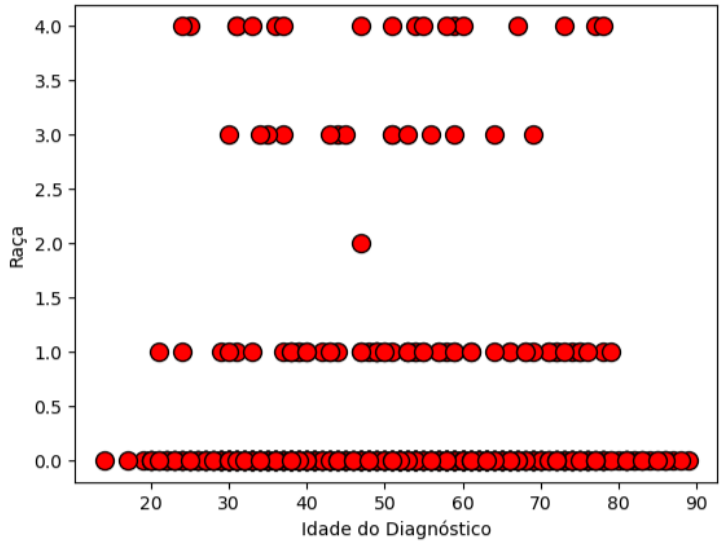
Além das colunas excluídas, todo o conteúdo do conjunto foi transformado em números, possibilitando a manipulação dos dados pelo algoritmo de regressão. Este processo é repetido em todas as outras técnicas de mineração utilizadas neste estudo com poucas variações, portanto não será exibido novamente imagens do d*ataset*.

Algoritmo de Regressão

A Figura 3 foi obtida importando a biblioteca *matplotlib.pyplot* considerando as colunas "*Age at diagnosis*" e "*Race*". Para facilitar o entendimento do gráfico, segue legendas do eixo y (Raça):

•0 - Pessoas Brancas
•1 - Pessoas Afroamericanas e Afrodescendente
•2 - Nativo do Alaska ou índio americano
•3 - Asiáticos
•4 - Não Reportado

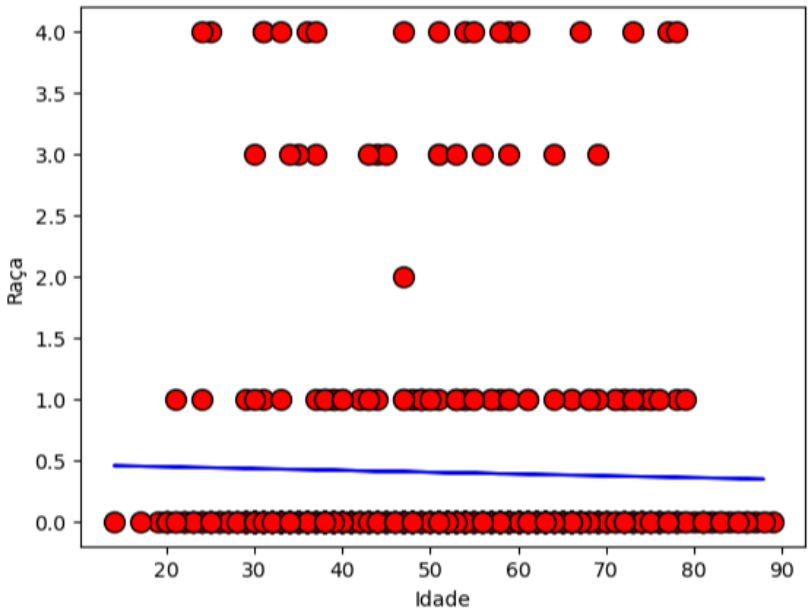
Figura 3: *Dataset* Normalizado



Fonte: Autor, 2023.

Por fim, a Figura 4 mostra o resultado do gráfico de regressão linear. A linha azul foi gerada pelo algoritmo de regressão, que realizou uma predição a partir dos dados informados de Raça e Idade. Ele assumiu uma relação linear entre as variáveis dependentes e independentes, ou seja, procurou encontrar uma linha reta que melhor se ajustasse aos dados.

Figura 4: Dataset Normalizado



Fonte: Autor, 2023.

Após a aplicação do algoritmo de regressão, procedeu-se à apresentação do erro quadrático médio e do erro médio absoluto como métricas de avaliação de desempenho. Embora o algoritmo não tenha alcançado um desempenho excepcional, ele atendeu às expectativas estabelecidas pelo autor.

Figura 5: Desempenho do algoritmo

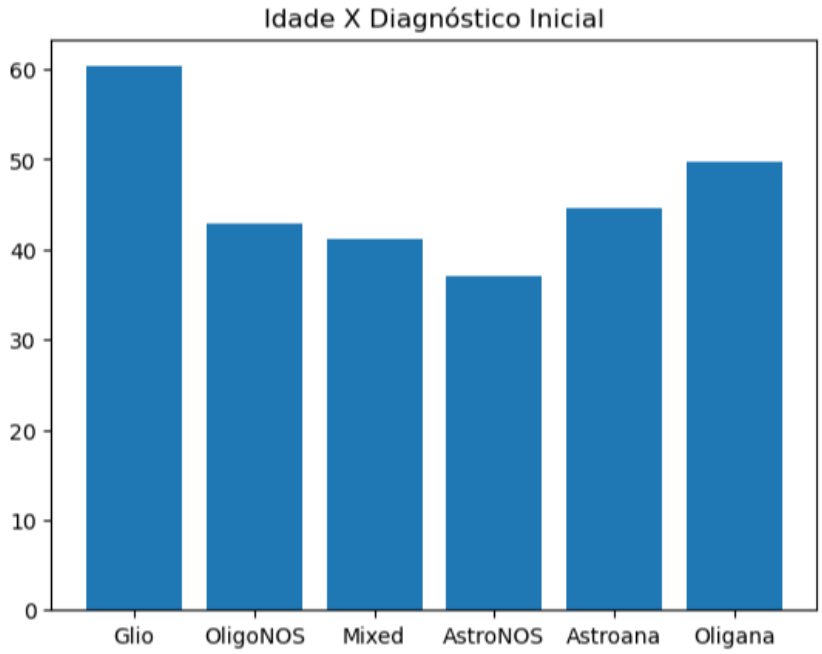
```
In [25]: y_testPred = metodoRegressao.predict(x_test.reshape(-1, 1))
         import numpy as np
         #print(y_testPred)

         MSE = np.mean(np.square(y_testPred - y_test))
         print('Erro quadrático médio:', MSE)

         EMA = np.mean(np.absolute(y_testPred - y_test))
         print('Erro médio absoluto:', EMA)

         Erro quadrático médio: 0.2517238863293587
         Erro médio absoluto: 0.49169086581469545
```

Fonte: Autor, 2023.

KNN

O algoritmo *KNN* (*k-nearest neighbors*, ou k-vizinhos mais próximos) é um algoritmo de aprendizado supervisionado utilizado para classificação e regressão. Ele é baseado no princípio de que objetos similares tendem a estar próximos uns dos outros. O objetivo do *KNN* é encontrar os K pontos de um conjunto de referência que estão mais próximos de cada ponto de um conjunto de consulta.
Foi utilizado o *KNN* neste conjunto de dados. A seguir serão apresentados alguns resultados. Na imagem abaixo podemos ver o resultado de um gráfico relacionando Idade x Diagnóstico Inicial:

Figura 6: Idade x Diagnóstico Inicial Knn



Fonte: Autor, 2023.

1) Glio - Glioblastoma
2) OligoNOS - Oligodendroglioma, NOS
3) Mixed - Mixed glioma
4) AstroNOS - Astrocytoma, NOS
5) Astroana - Astrocytoma, anaplastic

6) Oligana - Oligodendroglioma, anaplastic

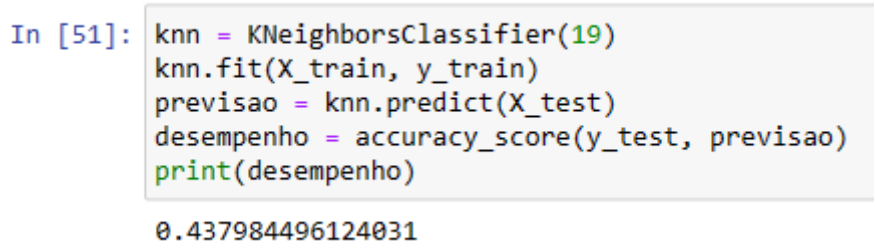
Analisando o gráfico é perceptível que tumores do tipo Glioblastoma (Glio) ocorrem em todas as faixas etárias, enquanto os gliomas do tipo Astrocytoma, NOS (AstroNOS) aparentemente não são diagnosticados em indivíduos acima de quarenta anos. A seguir é mostrado a capacidade de predição do algoritmo.
O algoritmo foi treinado e em seguida foi solicitado que ele realizasse a predição a partir dos dados inseridos manualmente. O primeiro atributo é referente a idade, o segundo ao sexo e o terceiro à raça. Foi inserido um valor de sessenta e sete anos para a idade, do sexo masculino (0) e da raça branca.

Figura 7: Predição Knn

Fonte: Autor, 2023.

O resultado obtido no *array* é referente ao diagnóstico inicial, que no caso da imagem é Gliobastoma (0). Ou seja, o algoritmo previu que o diagnóstico inicial de glioma para uma pessoa com estas características será Gliobastoma.

Figura 8: Desempenho Knn



Fonte: Autor, 2023.

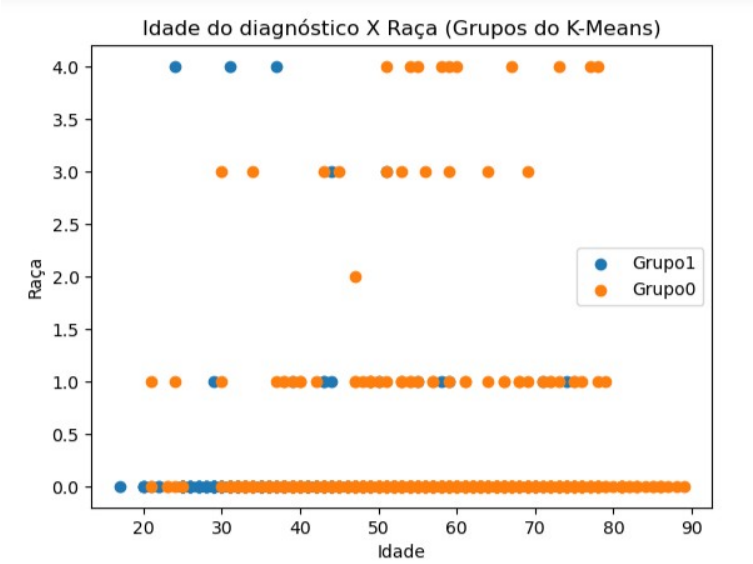
Na imagem acima podemos observar o desempenho obtido pelo algoritmo *Knn.*

Clustering KMeans

O algoritmo de clustering é uma técnica de aprendizado não supervisionado que agrupa dados semelhantes em grupos ou clusters. Com o intuito de encontrar padrões, reúne amostras em clusters onde elas são mais semelhantes entre si do que com amostra de outros clusters (Medeiros e Ferreiro, 2022).

Esta técnica também foi utilizada no conjunto de dados deste artigo com o intuito de extrair informações relevantes. Na figura 9, nota-se a relação entre Idade do Diagnóstico Inicial x Raça. Segue legendas para simplificar a interpretação:

Figura 9: *Clustering*



Fonte: Autor, 2023.

• 0 - Pessoas Brancas

• 1 - Pessoas Afroamericanas e Afrodescendente

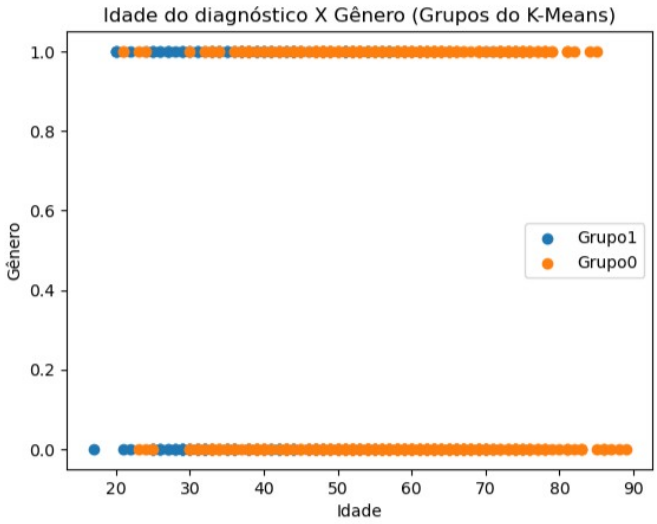
• 2 - Nativo do Alaska ou índio americano

• 3 - Asiáticos

• 4 - Não Reportado

A figura 10 mostra a relação entre o sexo da pessoa e a idade de diagnóstico, revelando que não há correlação de gênero para gliomas:

Figura 10: *Clustering*



Fonte: Autor, 2023.

Observa-se a acurácia do algoritmo *kmeans* de clusterização na figura 11:

Figura 11: *Clustering*

```
: from sklearn.metrics import accuracy_score

: desempenho = accuracy_score(mutacoes['Primary_Diagnosis'], mutacoes['KMeans'])
  desempenho

: 0.46867749419953597
```

Os resultados das análises de regressão, *clustering* e KNN revelaram que faixa etária e etnia estão associadas a diferentes tipos de gliomas. Alguns *insights* obtidos:

1) As etnias indígenas ou nativas do alasca possuem baixíssimas incidências de gliomas.

2) A etnia de cor branca possui maior probabilidade de ocorrência de gliomas.

3) Poucas ocorrências de gliomas abaixo dos vinte anos.

4) Poucas ocorrências de gliomas após os 90 anos.

5) Pessoas de etnia afrodescendente são o segundo maior grupo atingido por gliomas.

6) Indivíduos afrodescendentes têm maior probabilidades de serem diagnosticados com gliomas após a idade de trinta anos.

7) Indivíduos asiáticos possuem poucas ocorrências de gliomas.

8) O sexo do indivíduo não possui influência relevante para o diagnóstico dos gliomas

IBM Watson

A plataforma do IBM Watson foi utilizada para analisar o *dataset* pré-processado. Foi necessário realizar novas etapas de tratamento no conjunto de dados para otimizar o desempenho do algoritmo. Foram removidas algumas linhas e colunas que não contribuíam para a classificação, com a finalidade de diminuir a confusão do classificador.

Especificamente foi utilizado a ferramenta *watsonx.ai* dentro do ambiente *cloud* da IBM, para realizar a classificação foi utilizada a versão gratuita para teste. (Disponível em: https://dataplatform.cloud.ibm.com/. Acesso em: 20 de outubro de 2023)

De acordo com VENTURA-FERNÁNDEZ et al. (2021 apud Nagwanshi e Dubey, 2017), o IBM Watson Analytics é um sistema inteligente que faz uso de processamento de linguagem natural, aprendizado dinâmico, e geração e avaliação de hipóteses. Após o carregamento do conjunto de dados na nuvem, esse sistema possui a capacidade de antecipar consultas, relacionar dados e reunir informações, visando proporcionar pontuações de qualidade de dados, análises detalhadas e associações de campos.

Na figura 12 pode-se visualizar os resultados obtidos após o processo de mineração e aprendizado de máquina realizado de forma automatizada pelo IBM Watson.

Figura 12: IBM Watson

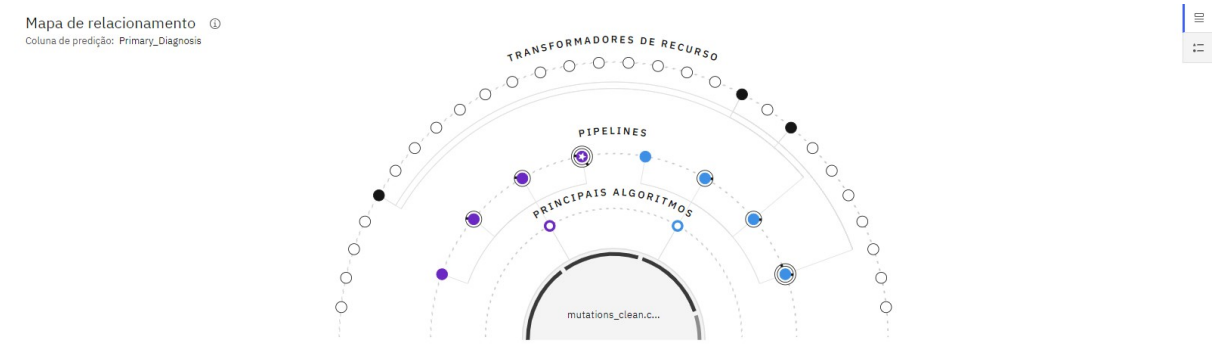


Tabela de classificação de pipeline

	Classificação	Nome	Algoritmo	Precisão (Otimizado) Validação Cruzada	Melhorias	Tempo de construção
★	1	Pipeline 4	Classificador do XGB	0.645	HPO-1 FE HPO-2	00:01:35

Fonte: Autor, 2023.

No experimento acima pode-se observar que o algoritmo que obteve a melhor precisão foi o Classificador XGB. Entretanto os resultados não foram suficientemente satisfatórios pois a precisão foi relativamente baixa, cerca de 60% de assertividade.

Esse fenômeno pode ser explicado pela necessidade de uma amostragem mais extensa, a fim de permitir que o algoritmo realize previsões mais precisas para esse tipo de problema. Além disso, contribui para essa situação o fato de que o conjunto de dados não foi adequadamente equilibrado.

Weka: J48

De acordo com DA SILVA et al. (2021) é “baseado em árvores de decisão que reimplementa na suíte Weka o algoritmo C4.5 desenvolvido por Quinlan”. O algoritmo C4.5 por sua vez:

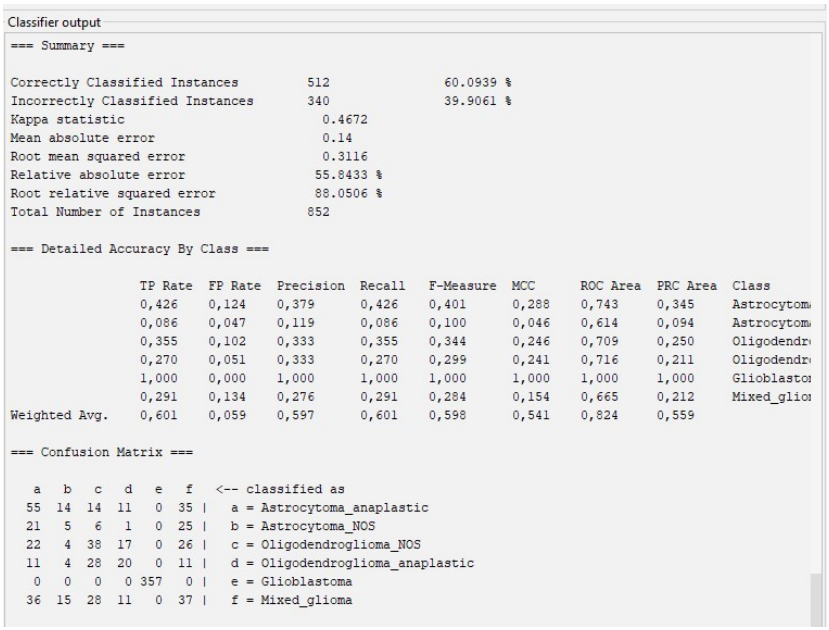
> “tem como objetivo gerar um modelo classificador na forma de uma árvore de decisão, apresentando dois estados durante o processo, os quais são: folha que indica um ponto no final da classificação, sendo atribuída a uma classe e nó de decisão, onde baseando-se no atributo em análise, poderá conter uma ramificação seguida de uma folha ou uma sub-árvore para cada possível valor encontrado na base” (Palhano, Maicon Bastos et al., 2012, Anais SULCOMP, v. 6).

O algoritmo J48 gera árvores de decisão, com estruturas hierárquicas representando decisões e suas possíveis consequências. Cada nó interno da árvore representa uma decisão com base em um atributo específico, e as folhas representam as classes ou valores de saída. Ele também pode incluir uma fase chamada de poda que ajuda a evitar o sobreajuste (*overfitting*). Da Silva et al. (2021)

Para este artigo, a execução do algoritmo ocorreu no software Weka, para realizar esta operação, foi preciso converter o conjunto de dados de um formato “.csv” para um formato “. arff”.

Os demais algoritmos do conjunto de ferramentas do Weka abordados neste trabalho a seguir, também utilizaram o mesmo conjunto de dados em formato do tipo “. arff”. Visto que o Weka não permite a mineração de arquivos do tipo “.csv”.

Figura 13: J48



Classifier output

```
=== Summary ===

Correctly Classified Instances         512               60.0939 %
Incorrectly Classified Instances       340               39.9061 %
Kappa statistic                          0.4672
Mean absolute error                      0.14
Root mean squared error                  0.3116
Relative absolute error                 55.8433 %
Root relative squared error             88.0506 %
Total Number of Instances              852

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,426    0,124    0,379      0,426    0,401      0,288    0,743     0,345     Astrocytom
                 0,086    0,047    0,119      0,086    0,100      0,046    0,614     0,094     Astrocytom
                 0,355    0,102    0,333      0,355    0,344      0,246    0,709     0,250     Oligodendr
                 0,270    0,051    0,333      0,270    0,299      0,241    0,716     0,211     Oligodendr
                 1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     Glioblasto
                 0,291    0,134    0,276      0,291    0,284      0,154    0,665     0,212     Mixed_glio
Weighted Avg.    0,601    0,059    0,597      0,601    0,598      0,541    0,824     0,559

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
  55  14  14  11   0  35 |   a = Astrocytoma_anaplastic
  21   5   6   1   0  25 |   b = Astrocytoma_NOS
  22   4  38  17   0  26 |   c = Oligodendroglioma_NOS
  11   4  28  20   0  11 |   d = Oligodendroglioma_anaplastic
   0   0   0   0 357   0 |   e = Glioblastoma
  36  15  28  11   0  37 |   f = Mixed_glioma
```

Fonte: Autor 2023

É possível visualizar na imagem que o algoritmo classificou corretamente cerca de 60% das instâncias, resultado próximo ao obtido no IBM Watson.

NaiveBayes

O *NaiveBayes* é um algoritmo de *machine learning*, onde o classificador aprende por meio de um algoritmo de classificação de documentos (Buzic e Dobsa, 2018).

De acordo com Da Silva et al. (2021), ele é um classificador probabilístico que se baseia no Teorema de Bayes para realizar a classificação de instâncias em categorias. Ele é chamado de "naive" (ingênuo) porque assume independência condicional entre os atributos, o que significa que a presença de um atributo em uma classe não está relacionada à presença de outros atributos.

Isso significa que a presença ou ausência de um atributo não afeta a presença ou ausência de outros atributos, o que é uma suposição forte, mas pode ser útil na resolução de alguns problemas.

No conjunto de dados deste trabalho, o algoritmo *NaiveBayes* obteve um desempenho de aproximadamente 61 % de classificações corretas.

Na Figura 14, observa-se que 61,85% das instâncias foram corretamente classificadas, enquanto 38,14% foram erroneamente categorizadas. A análise da matriz de confusão revela que o algoritmo demonstrou melhor precisão ao diagnosticar o atributo Glioblastoma: de 357 instâncias, classificou corretamente 355 e falhou apenas em duas classificações.

O desempenho do *NaiveBayes* foi menos eficaz na classificação dos atributos do tipo Oligodendroglioma_NOS, acertando somente 13 instâncias, em contraste com o total de 107 instâncias que deveriam ter sido classificadas corretamente.

Figura 14: *NaiveBayes*

Classifier output

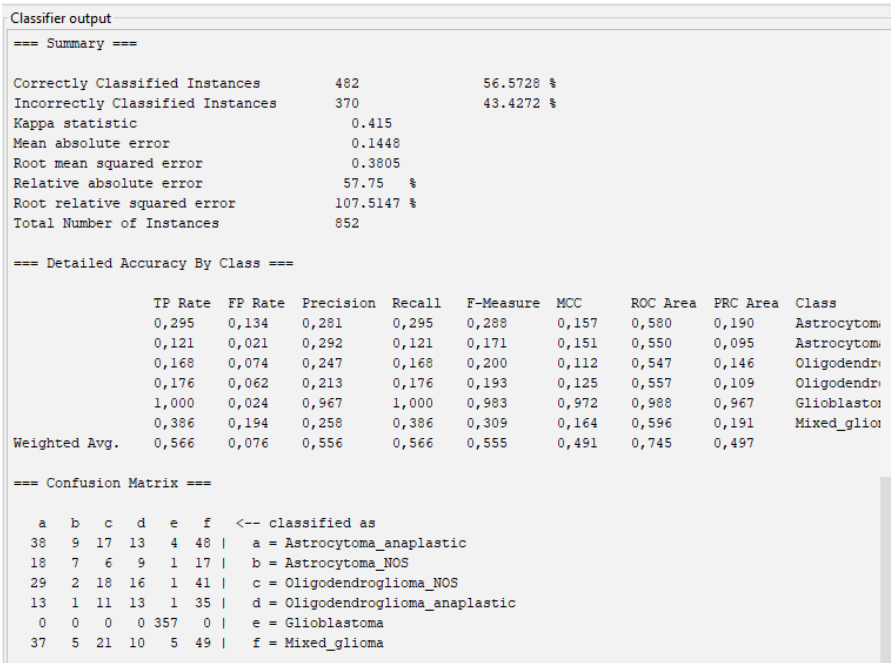
```
=== Summary ===

Correctly Classified Instances         527               61.8545 %
Incorrectly Classified Instances       325               38.1455 %
Kappa statistic                          0.4949
Mean absolute error                      0.1344
Root mean squared error                  0.2725
Relative absolute error                 53.6218 %
Root relative squared error             77.0177 %
Total Number of Instances              852

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,419    0,068    0,524      0,419    0,466      0,386    0,876     0,473     Astrocytom
                 0,276    0,108    0,157      0,276    0,200      0,130    0,827     0,171     Astrocytom
                 0,439    0,087    0,420      0,439    0,429      0,345    0,841     0,380     Oligodendr
                 0,257    0,050    0,328      0,257    0,288      0,231    0,856     0,333     Oligodendr
                 0,994    0,000    1,000      0,994    0,997      0,995    1,000     1,000     Glioblasto
                 0,283    0,119    0,295      0,283    0,289      0,168    0,778     0,294     Mixed_glio
Weighted Avg.    0,619    0,051    0,634      0,619    0,624      0,573    0,904     0,623

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
  54  34   1   7   0  33 |   a = Astrocytoma_anaplastic
   9  16   2   2   0  29 |   b = Astrocytoma_NOS
   9  13  47  22   0  16 |   c = Oligodendroglioma_NOS
  11   4  33  19   0   7 |   d = Oligodendroglioma_anaplastic
   0   1   0   0 355   1 |   e = Glioblastoma
  20  34  29   8   0  36 |   f = Mixed_glioma
```

Fonte: Autor 2023

One R

One Rule ou *One R* é um algoritmo de classificação simples, porém eficaz. De acordo com DE CAMPOS JR et al. (2022) o algoritmo gera uma regra para cada atributo e seleciona aquela com o menor erro total como a única regra a ser utilizada. Esse algoritmo, emprega um método de classificação de custo reduzido, mantendo uma alta acurácia.

Figura 15: One R



```
Classifier output
=== Summary ===

Correctly Classified Instances         482               56.5728 %
Incorrectly Classified Instances       370               43.4272 %
Kappa statistic                          0.415
Mean absolute error                      0.1448
Root mean squared error                  0.3805
Relative absolute error                 57.75   %
Root relative squared error            107.5147 %
Total Number of Instances              852

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,295    0,134    0,281      0,295    0,288      0,157    0,580     0,190     Astrocytom
                 0,121    0,021    0,292      0,121    0,171      0,151    0,550     0,095     Astrocytom
                 0,168    0,074    0,247      0,168    0,200      0,112    0,547     0,146     Oligodendr
                 0,176    0,062    0,213      0,176    0,193      0,125    0,557     0,109     Oligodendr
                 1,000    0,024    0,967      1,000    0,983      0,972    0,988     0,967     Glioblasto
                 0,386    0,194    0,258      0,386    0,309      0,164    0,596     0,191     Mixed_glio
Weighted Avg.    0,566    0,076    0,556      0,566    0,555      0,491    0,745     0,497

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
  38   9  17  13   4  48 |   a = Astrocytoma_anaplastic
  18   7   6   9   1  17 |   b = Astrocytoma_NOS
  29   2  18  16   1  41 |   c = Oligodendroglioma_NOS
  13   1  11  13   1  35 |   d = Oligodendroglioma_anaplastic
   0   0   0   0 357   0 |   e = Glioblastoma
  37   5  21  10   5  49 |   f = Mixed_glioma
```

Fonte: Autor 2023

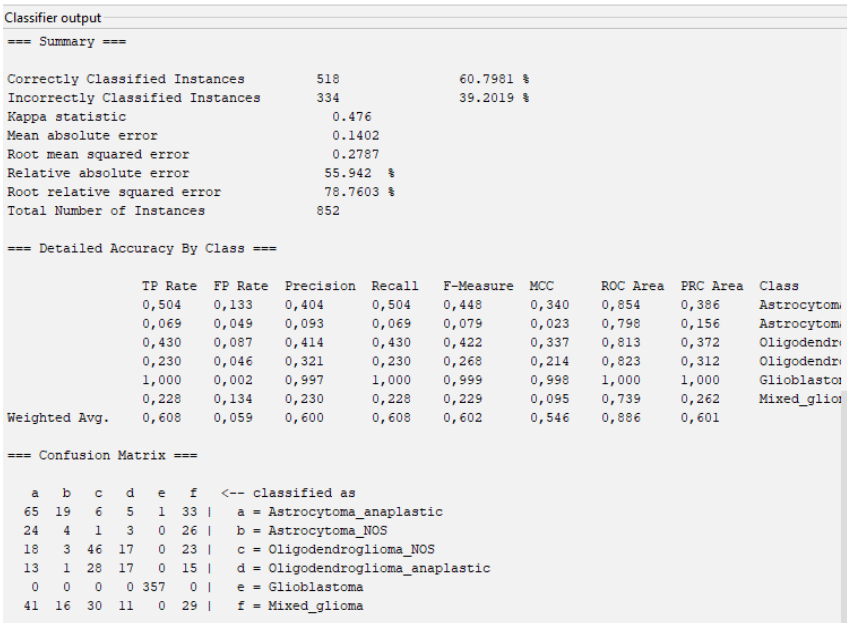
No algoritmo *One R*, a porcentagem de classificação das instâncias foi ainda menor com 56% de instâncias classificads corretamente.

RandomForest

RandomForest é um algoritmo de aprendizado de máquina, que combina as previsões de vários modelos para melhorar a precisão e o desempenho geral.

Segundo TELOKEN, et al. (2016), o *RandomForest* envolve a agregação de classificadores do tipo árvore de decisão, cuja estrutura é construída de maneira aleatória. Para determinar a classe de uma instância, o método combina os resultados de várias árvores de decisão através de um mecanismo de votação. Cada árvore contribui com uma classificação ou voto para uma classe, e a classificação final é determinada pela classe que recebeu a maioria dos votos entre todas as árvores da floresta.

Figura 16: *RandomForest*



Classifier output

```
=== Summary ===

Correctly Classified Instances         518               60.7981 %
Incorrectly Classified Instances       334               39.2019 %
Kappa statistic                          0.476
Mean absolute error                      0.1402
Root mean squared error                  0.2787
Relative absolute error                 55.942  %
Root relative squared error             78.7603 %
Total Number of Instances              852

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,504    0,133    0,404      0,504    0,448      0,340    0,854     0,386     Astrocytoma
                 0,069    0,049    0,093      0,069    0,079      0,023    0,798     0,156     Astrocytoma
                 0,430    0,087    0,414      0,430    0,422      0,337    0,813     0,372     Oligodendr
                 0,230    0,046    0,321      0,230    0,268      0,214    0,823     0,312     Oligodendr
                 1,000    0,002    0,997      1,000    0,999      0,998    1,000     1,000     Glioblasto
                 0,228    0,134    0,230      0,228    0,229      0,095    0,739     0,262     Mixed_glio
Weighted Avg.    0,608    0,059    0,600      0,608    0,602      0,546    0,886     0,601

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
  65  19   6   5   1  33 |   a = Astrocytoma_anaplastic
  24   4   1   3   0  26 |   b = Astrocytoma_NOS
  18   3  46  17   0  23 |   c = Oligodendroglioma_NOS
  13   1  28  17   0  15 |   d = Oligodendroglioma_anaplastic
   0   0   0   0 357   0 |   e = Glioblastoma
  41  16  30  11   0  29 |   f = Mixed_glioma
```

Fonte: Autor 2023

De acordo com a Figura 16. O algoritmo *RandomForest* teve um desempenho de 60% de classificações corretas.

Os resultados obtidos por meio das análises utilizando o IBM Watson e o conjunto de ferramentas do Weka ficaram abaixo das expectativas, foram identificados alguns fatores que contribuíram para esse desempenho aquém do esperado. Alguns dos problemas causais são destacados a seguir.

Um dos desafios enfrentados é relacionado ao tamanho do conjunto de dados. Para que os algoritmos consigam realizar previsões e identificar padrões com eficácia a partir da coluna especificada, é crucial considerar a complexidade do *dataset*. Neste caso, seria benéfico contar com um conjunto de dados mais amplo para melhorar o desempenho do algoritmo.

Além disso, enfrentamos a questão do desbalanceamento do *dataset*. Para otimizar o desempenho dos algoritmos, é fundamental que o conjunto de dados seja equilibrado. O desbalanceamento pode introduzir viés nos resultados, comprometendo a capacidade dos algoritmos de generalizar efetivamente.

É importante ressaltar que, apesar dos desafios encontrados, os dados analisados contêm informações valiosas que ainda podem ser extraídas. Para os objetivos específicos deste projeto, os resultados alcançados podem ser considerados satisfatórios. No entanto, para futuras análises ou projetos semelhantes, recomenda-se considerar estratégias para lidar com o tamanho do conjunto de dados e o desbalanceamento, visando aprimorar a precisão e a confiabilidade dos resultados.

CONCLUSÕES

Embora este estudo não tenha como objetivo prever a ocorrência de gliomas com base na combinação de genes, os resultados destacam a importância dos fatores demográficos na predisposição a esses tumores cerebrais. A faixa etária e a etnia foram identificadas como variáveis significativas na classificação dos indivíduos em grupos de risco de gliomas LGG ou GBM.

Compreender esses fatores demográficos pode ter implicações importantes para a prevenção e detecção precoce de gliomas. Estratégias de triagem e monitoramento mais direcionadas podem ser desenvolvidas com base nessas associações, permitindo uma abordagem mais personalizada para a detecção de gliomas em grupos de risco específicos.

No entanto, é importante destacar que este estudo tem algumas limitações. A amostra de dados utilizada pode não ser representativa da população em geral, e a inclusão de um maior número de características moleculares e clínicas pode expandir as conclusões. Além disso, não foram considerados todos os genes disponíveis no conjunto de dados e o conjunto de dados minerado não é ideal para o tipo de mineração realizada. Todavia, para

o propósito deste estudo a mineração de dados aplicada à classificação de grupos de risco para gliomas mostrou-se uma abordagem eficaz para auxiliar na identificação de indivíduos com maior propensão ao desenvolvimento desses tumores e apesar de a precisão dos algoritmos ter ficado aquém do esperado, ainda assim foi possível identificar direções para análises futuras mais assertivas.

A análise das variáveis demográficas e clínicas permitiu a identificação de grupos de risco distintos, contribuindo para estratégias de prevenção e detecção precoce mais direcionadas. Esses resultados destacam a importância da mineração de dados como uma ferramenta poderosa no campo da oncologia, possibilitando avanços significativos no diagnóstico e tratamento de gliomas.

REFERÊNCIAS BIBLIOGRÁFICAS

BITERGE-SUT, B. (2020). A comprehensive analysis of the angiogenesis-related genes in glioblastoma multiforme vs. brain lower grade glioma. *Arquivos De Neuro-psiquiatria*, *78*(1), 34–38.

BUŽIĆ, D.; DOBŠA, J.Lyrics Classification using Naive Bayes. International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO),p. 21-25,2018.

CAMILO, Cássio Oliveira; SILVA, João Carlos da. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. **Universidade Federal de Goiás (UFC)**, v. 1, n. 1, p. 1-29, 2009.

C. G. CAMBRONERO and I. G. MORENO, “Algoritmos de aprendizaje: knn & kmeans,” Intelgencia en Redes de Comunicación, Universidad Carlos III de Madrid, vol. 23, 2006.

CORDEIRO, Michel B.; MEYER, Bruno H.; ZOLA, Wagner M. Nunan. KNN exato em GPU. In: Anais da XXIII Escola Regional de Alto Desempenho da Região Sul. SBC, 2023. p. 17-20.
DOS SANTOS, Ana Suelen Alves et al. ATUAÇÃO DO FARMACÊUTICO NA ONCOLOGIA. Encontro de Extensão, Docência e Iniciação Científica (EEDIC), v. 8, 2021.

DA SILVA, Ricardo Conde Camillo et al. An Intrusion Detection System for Web-Based Attacks Using IBM Watson. **IEEE Latin America Transactions**, v. 20, n. 2, p. 191-197, 2021.

DE CAMPOS JR, Arion; HUBIE, Lucas Ferreira; GONÇALVES, Mateus Bueno. A MINERAÇÃO DE DADOS COMO FERRAMENTA PARA AVALIAÇÃO DE DADOS RELACIONADOS À PANDEMIA DO COVID-19. REVISTA DE ENGENHARIA E TECNOLOGIA, v. 14, n. 4, 2022.

DE MEDEIROS JÚNIOR, José Gilberto B.; FERRERO, Carlos Andrés. Agrupamento de Imagens Tumorais de MRI utilizando Extração de Descritores baseados em Séries Temporais. In: **Anais do XVII Escola Regional de Banco de Dados**. SBC, 2022. p. 109-118.

MEDEIROS JÚNIOR, Jose Gilberto B. de. Transformando imagens de segmentos de tumores cerebrais em séries temporais para mineração de dados. 2021.

Santos, A. L. dos. (2021). GLIOMAS, TUMORES MALIGNOS QUE SURGEM NO SISTEMA NERVOSO. *Revista Ibero-Americana De Humanidades, Ciências E Educação*, *7*(2), 12.

SILVA, André Marcos; MATTOS, Rogério. IBM Watson como Ambiente para Desenvolvimento e Execução de um Chatbot–Um Estudo de Caso Aplicado ao Processo de Atendimento ao Usuário. São Paulo:[sn], p. 1-9, 2018.

VENTURA-FERNÁNDEZ, Tania; VIDALÓN-SOLDEVILLA, Ethel; VENTURA-FERNÁNDEZ, Freddy. Predictibilidad en el diagnóstico utilizando Watson de IBM. Vive Revista de Salud, v. 4, n. 10, p. 86-96, 2021.

WU, Yameng et al. Research progress of gliomas in machine learning. Cells, v. 10, n. 11, p. 3169, 2021.

Documento Digitalizado Restrito

TCC - Pós-graduação em Gestão da Tecnologia da Informação e Comunicação - Alef Cardoso

Assunto: TCC - Pós-graduação em Gestão da Tecnologia da Informação e Comunicação - Alef Cardoso
Assinado por: Marcelo Murari
Tipo do Documento: Anexo
Situação: Finalizado
Nível de Acesso: Restrito
Hipótese Legal: Direito Autoral - conservar a obra inédita (Art. 24, III, da Lei nº 9.610/1998)
Tipo do Conferência: Documento Digital

Documento assinado eletronicamente por:

- **Marcelo Luis Murari**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 20/12/2023 20:01:12.

Este documento foi armazenado no SUAP em 20/12/2023. Para comprovar sua integridade, faça a leitura do QRCode ao lado ou acesse https://suap.ifsp.edu.br/verificar-documento-externo/ e forneça os dados abaixo:

Código Verificador: 1532050
Código de Autenticação: 011bd6ce92