Nome: Gabriel Martins Machado Christo **DRE:** 117217732
Nome: João Vitor de Freitas Barbosa **DRE:** 117055449

Tarefa 7
Árvore de Decisão

link do código no colab: link

O objetivo desta tarefa é realizar um conjunto de experimentos usando o método de árvore de decisão e um dataset escolhido no site kaggle (https://www.kaggle.com/datasets). A base deve ter mais de 500 instâncias.

Deve ser entregue no classroom um relatório contendo as seguintes informações:

1- descrição do dataset selecionado: informe o número de instâncias do dataset, os atributos, assim como seus respectivos tipos e os valores que cada atributo pode assumir;

Nosso dataset é um conjunto de dados do Internet Movie Database (IMDB) com 1000 instâncias, representando os melhores filmes e programas de TV.

Conteúdo do dataset:

Dados:

- Link do poster - URL
- Título do filme - texto
- Ano de lançamento - número inteiro
- Certificado obtido pelo filme - classificação do filme, podendo assumir os seguintes valores: [A, G, GP, PG, PG-13, R, TV-14, TV-MA, TV-PG, U, UA]
- Tempo de duração do filme - número inteiro
- Gênero - texto, sendo a combinação de diversos gêneros de filme: [drama, crime, action, adventure, comedy, history, horror, fantasy …]
- Nota no IMDB - valor de ponto flutuante
- Resumo - texto
- Diretor - texto
- Nota do Metacritc - número inteiro
- Ator principal 1 - texto
- Ator principal 2 - texto
- Ator principal 3 - texto
- Ator principal 4 - texto
- Número de votos - número inteiro
- Valor bruto obtido pelo filme - número inteiro

Vale notar que os atributos numéricos foram convertidos de ponto flutuante para inteiros, devido à discretização requerida pelo framework sklearn.

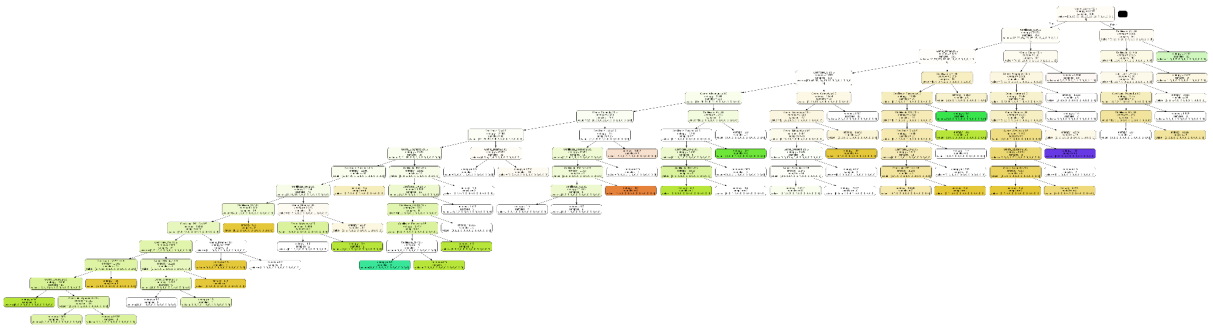
2- defina um experimento base: informe quais atributos do dataset serão considerados (você pode usar todos ou selecionar apenas alguns - neste caso justifique sua decisão); informe os parâmetros usados neste experimento (como o

dataset será dividido em treinamento/teste; no caso de usar k-fold, qual o valor de k foi escolhido, qual função será usada para escolher o atributo durante a construção da árvore - gini ou entropia; como é feita a avaliação da árvore gerada que será, etc).

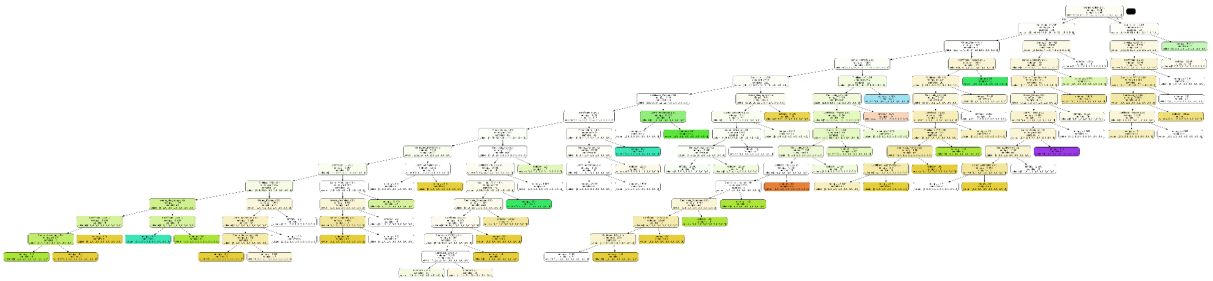
Nosso experimento base consiste em levar em consideração os atributos Gênero e Classificação Etária, ao qual queremos inferir a Nota no IMDB. Utilizamos o critério de entropia para a construção da árvore, pois este apresentou melhor precisão.

Nesse caso em específico nossa precisão ficou em torno de 0. 4019, e quando tentamos utilizar o treinamento/teste obtivemos os seguintes resultados:

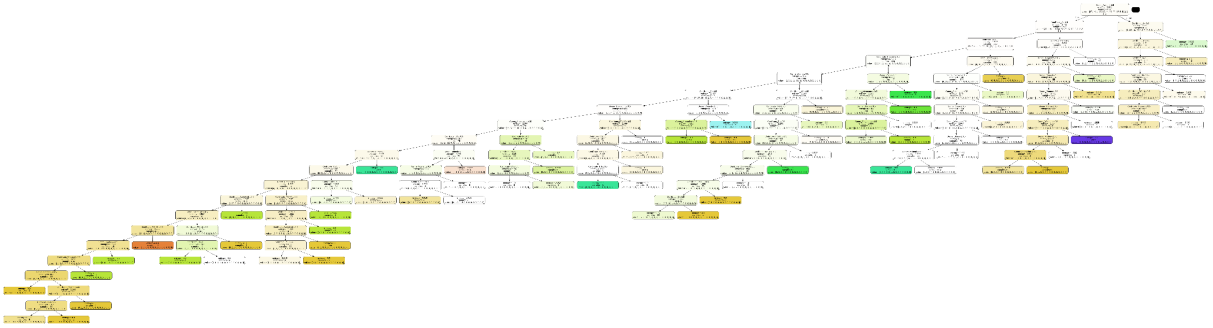
- 0. 6 *treinamento e* 0. 4 *teste*:
    - Precisão do treino: 0,37254901960784315
    - Precisão do teste: 0,3333333333333333



- 0. 8 *treinamento e* 0. 2 *teste*:
    - Precisão do treino: 0.38823529411764707
    - Precisão do teste: 0.28823529411764703



Podemos concluir, observando os dados explicitados acima, que o experimento sem o treinamento/teste possui uma precisão maior:



Com o K-Fold, sendo k = 10, tivemos uma média de acurácia para treino de `0.3874`

3- explore alternativas para melhorar os resultados obtidos: faça isso mudando os parâmetros usados no experimento base e/ou fazendo podas na árvore resultante. Justifique a alteração e o que você esperava que acontecesse quando decidiu fazer a alteração do parâmetro e qual o resultado obtido. Compare seus resultados com o experimento base.

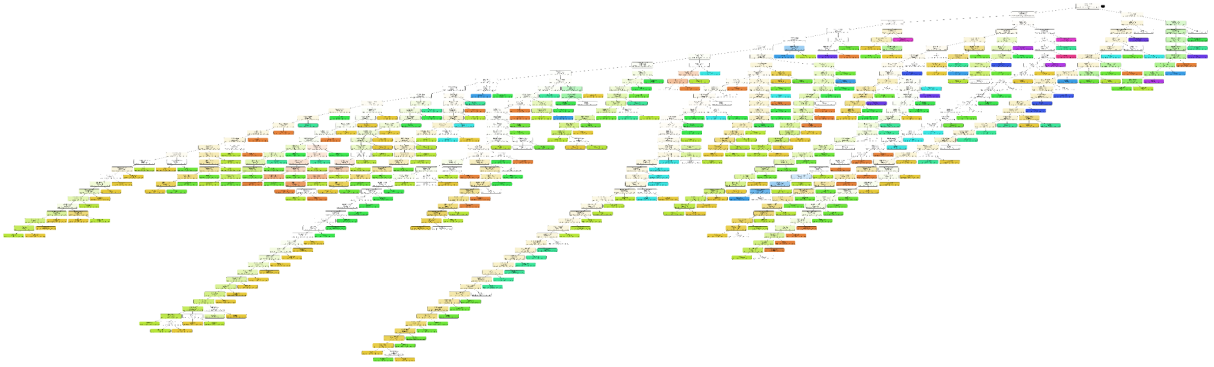
Para melhorar os resultados obtidos, fizemos a troca dos parâmetros usados no experimento base, com a adição do parâmetro de tempo total do filme, esperando obter uma maior acurácia.

Continuamos usando o critério de entropia para a construção da árvore, pois este critério apresentou melhor precisão.

Com isso, fizemos diversos testes para entender melhor como o programa reagia, e assim, levamos em consideração os seguintes parâmetros:

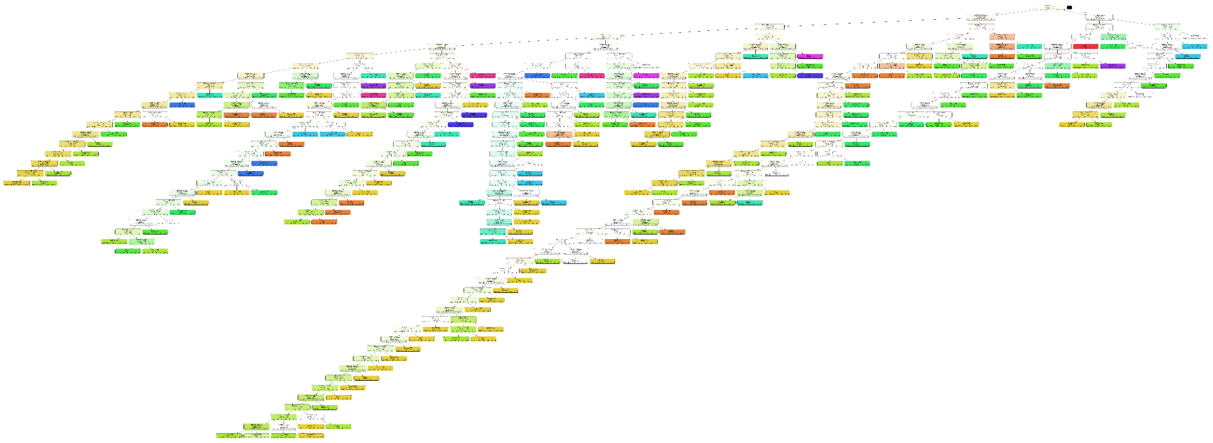
- Gênero, Classificação Etária e Tempo de duração

Com a predição sendo realizada com todas as entradas do dataset tivemos uma acurácia de 0.9411:



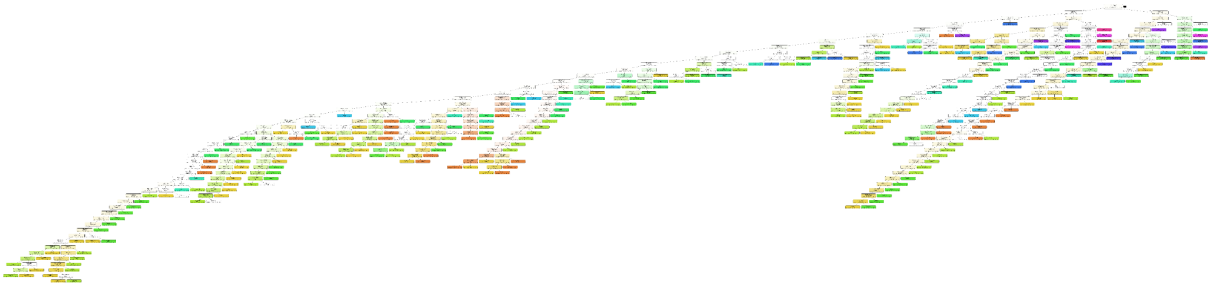
Já com a divisão de treinamento/teste do dataset, tivemos os seguintes resultados:

- 0.6 *treinamento e* 0.4 *teste*:
  - Precisão do treino: 0,6529
  - Precisão do teste: 0,6169



- 0.8 *treinamento e* 0.2 *teste*:

- Precisão do treino: 0.8215
- Precisão do teste: 0.7932



Com o K-Fold, sendo k = 10, tivemos uma média de acurácia para treino de `0.8619`

4- compare os resultados que você obteve na base que você escolheu com os obtidos por outros métodos e que estejam disponíveis no kaggle.

O único notebook de predição disponível no kaggle estava inferindo o lucro por filme, porém como não havia a correção monetária de acordo com o ano de lançamento do filme, houve um certo viés nos dados. Por conta disso resolvemos inferir a nota qualitativa dos filmes.