**PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS  
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA**

**Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data**

**Bárbara Gonçalves Oliveira**

**TÍTULO DO PROJETO**

Belo Horizonte

2021

**Bárbara Gonçalves Oliveira**

**TÍTULO DO PROJETO**

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados e Big Data como requisito parcial à obtenção do título de especialista.

Belo Horizonte

2021

**SUMÁRIO**

[1. Introdução 4](#_heading=h.gjdgxs)

[**1.1. Contextualização 4**](#_heading=h.30j0zll)

[**1.2. O problema proposto 4**](#_heading=h.1fob9te)

[**2. Coleta de Dados 4**](#_heading=h.3znysh7)

[**3. Processamento/Tratamento de Dados 5**](#_heading=h.2et92p0)

[**4. Análise e Exploração dos Dados 5**](#_heading=h.17dp8vu)

[**5. Criação de Modelos de Machine Learning 5**](#_heading=h.tyjcwt)

[**6. Apresentação dos Resultados 5**](#_heading=h.3dy6vkm)

[**7. Links 6**](#_heading=h.1t3h5sf)

[**REFERÊNCIAS 7**](#_heading=h.4d34og8)

[**APÊNDICE 8**](#_heading=h.2s8eyo1)

# 1. Introdução

## 1.1. Contextualização

Desde que consigo me recordar, filmes fazem parte da minha vida. Me considero uma fã de filmes densos, difíceis e que dividem a opinião pública. Nesse contexto, sempre que converso com outras pessoas sobre filmes polêmicos, o resultado não poderia ser diferente da classificação desses. Um consenso sempre esteve longe de ser atingido!

Com isso em mente, me veio a ideia de aliar ciência de dados a esse problema a fim de tentar desvendar essas polêmicas por meio de dados e tentar achar um consenso. Para isso, decidi por aliar opiniões de críticos certificados a opiniões de telespectadores em geral.

Os dados de críticos certificados utilizados são provenientes do *Rotten Tomatoes*, um site que de acordo com ele mesmo, é, junto com o *Tomatometer, “*a fonte mais confiável do mundo de recomendações de entretenimento de qualidade. Como o principal agregador online de resenhas de críticos de filmes e programas de TV, oferecem aos fãs um guia completo do que é fresco – e do que é podre – nos cinemas e em casa.”[[1]](#footnote-0) (Rotten Tomatoes, 2021)

A base do *Rotten Tomatoes* foi coletada da plataforma *Kaggle* e é composta por dois conjuntos de dados: dados relacionados aos filmes e dados relacionados aos *reviews*. Esses dados foram raspados da web no final de 2020. Nessa base de dados, como dito acima, além dos *reviews*, há também a classificação dada por estes ao filme, que pode ser *Rotten* ou *Fresh.*

Os dados de telespectadores em geral foram retirados do site IMDB através do IMDB Database, disponibilizado pelo próprio IMDB, que é uma plataforma online que reúne informações sobre filmes, séries, documentários e programas do cinema e televisão em geral. É uma das mais populares em se tratando de filmes e possui um acervo massivo. Lá se encontra reviews deixados por qualquer pessoa.

## 1.2. O problema proposto

O problema a ser resolvido através dos dados aqui, que foi brevemente descrito na contextualização, é tentar responder opiniões polêmicas sobre filmes igualmente polêmicos e chegar a um consenso que alie opiniões de críticos, que podem ser tomados aqui como pessoas com maior entendimento de cinema e artes, e opiniões de telespectadores em geral, que podem enriquecer essa análise e democratizá-la um pouco mais.

Seguindo a linha dos 5-Ws, abaixo são respondidas as perguntas propostas:

* Por que coletar, analisar e comparar informações de filmes?

Toda análise surge por conta de uma dúvida, uma questão a ser resolvida. Nesse caso, essa análise é importante porque ajuda a resolver polêmicas sobre se um filme polêmico é considerado bom ou não, questionamento fruto da minha própria curiosidade. Além disso, é um tópico de meu interesse que se alia ao conhecimento obtido durante as disciplinas desse curso de pós-graduação.

* Quem/quais são esses dados?

Os dados coletados vieram de diversas fontes, porém todos estavam disponibilizados online para esses fins:

1. *Rotten Tomatoes Movies and Critic Review Dataset*:Esse conjunto de dados foi obtido através da plataforma *Kaggle*. Os dados foram previamente raspados do website do *Rotten Tomatoes.* Esses dados são públicos e foram raspados em 31 de outubro de 2020. Possui um conjunto de dados com informações dos filmes em geral e um conjunto de dados com as informações dos reviews, incluindo o comentário do crítico, e o status do review, entre *Fresh* e *Rotten*.
2. *IMDB*

* Qual o objetivo dessa análise?

O objetivo dessa análise é extrair o texto e encontrar padrões nos comentários dos críticos e classificá-los de acordo com os rótulos *Fresh* ou *Rotten* utilizando de métodos supervisionados de aprendizado de máquina para obter um modelo treinado que posteriormentepoderá ser utilizado para interpretar outros comentários.

* Quais os aspectos geográficos dessa análise?

A base utilizada para treino, apesar de conter diversos filmes de diversos países, é americana e todo o conteúdo está em inglês, portanto, é esperado que se tenha uma visão mais ocidental nas análises.

* Qual o período que está sendo analisado?

Os comentários analisados são dos últimos 20 anos, mais precisamente do ano 2000 até o ano de 2020, quando esses dados foram raspados. Já os filmes revisados aqui, são de um período muito mais amplo – de 1914 até 2020.

# 2. Coleta de Dados

Os dados utilizados nesse trabalho são provenientes de duas fontes distintas, abaixo são detalhados os métodos de coleta e as variáveis presentes em cada um deles.

## 2.1. *Rotten Tomatoes Movies and Reviews Dataset*

A base de dados do *Rotten Tomatoes* está disponibilizada em .csv no site do *Kaggle[[2]](#footnote-1)* e foi obtida em 26/06/2021. Contém dois conjuntos de dados: *movies* e *critic\_reviews*.

1. Conjunto de dados *Movies*

Esse conjunto possui dados de filmes com data de lançamento de 1914 a 2020, é composto previamente por 17.712 registros e 22 colunas.

| **Nome da coluna/campo** | **Descrição** | **Tipo** |
| --- | --- | --- |
| Rotten\_tomatoes\_link | URL do filme no Rotten Tomatoes website. | Object |
| Movie\_title | Título do filme em inglês. | Object |
| Movie\_info | Breve sinopse do filme. | Object |
| Critics\_consensus | Consenso da avaliação dos críticos. | Object |
| Content\_rating | Classificação indicativa de idade do filme. | Object |
| Genres | Gêneros do filme. | Object |
| Directors | Diretores do filme. | Object |
| Authors | Autores do filme. | Object |
| Actors | Atores que atuaram no filme. | Object |
| Original\_release\_date | Data original de estreia do filme nos Estados Unidos. | Object |
| Streaming\_release\_date | Data de início do streaming do filme. | Object |
| Runtime | Duração do filme em minutos. | Object |
| Production\_company | Produtora do filme. | Object |
| Tomatometer\_status | Avaliação qualitativa do filme entre *Rotten* e *Fresh.* | Object |
| Tomatometer\_rating | Percentual de avaliações positivas do filme. | Float64 |
| Tomatometer\_count | Número de críticas totais utilizadas para o cálculo da variável tomatometer\_status. | Float64 |
| Audience\_status | Classificação qualitativa do filme pelos telespectadores usuários do website. | Object. |
| Audience\_rating | Percentual de avaliações positivas dos usuários. | Float64 |
| Audience\_count | Número de avaliações totais dos usuários utilizadas para o cálculo da variável audience\_status. | Float64 |
| Tomatometer\_top\_critics\_count | Número de avaliações de top críticos. | Int64 |
| Tomatometer\_fresh\_critics\_count | Número de avaliações do tipo *Fresh.* | Int64 |
| Tomatometer\_rotten\_critics\_count | Número de avaliações do tipo *Rotten.* | Int64 |

1. Conjunto de dados *Critic\_reviews*

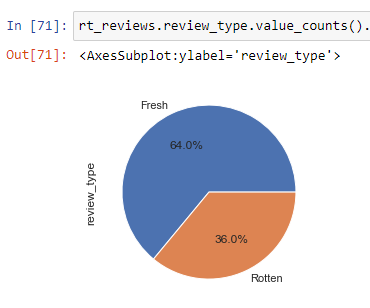
Esse conjunto possui dados das revisões dos filmes disponíveis no *Rotten Tomatoes,* os mesmos filmes do conjunto de dados Movies. É composto previamente por 1.130.017 registros e 8 colunas.

| **Nome da coluna/campo** | **Descrição** | **Tipo** |
| --- | --- | --- |
| Rotten\_tomatoes\_link | URL do filme no Rotten Tomatoes website. | Object |
| Critic\_name | Nome do crítico que escreveu a revisão. | Object |
| Top\_critic | Valor booleano que indica se esse crítico é um top crítico ou não. | Bool |
| Publisher\_name | Nome da publicadora na qual o crítico trabalha. | Object |
| Review\_type | Classificação do filme de acordo com o crítico, entre *Fresh* e *Rotten.* | Object |
| Review\_score | Nota avaliativa dada pelo crítico ao filme. Não possui uma consistência. | Object |
| Review\_date | Data da revisão. | Object |
| Review\_content | Conteúdo da revisão em si. | Object. |

# 3. Processamento/Tratamento de Dados

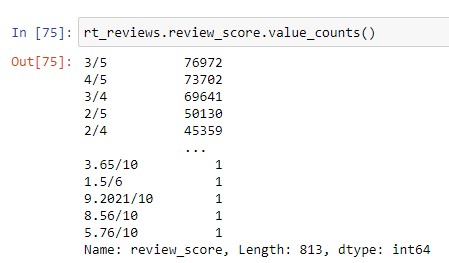
1. Conjunto de dados *Critic\_reviews*

Para treinamento do modelo foram utilizados dados do *dataset* de reviews apenas. As colunas utilizadas foram *review\_content* e *review\_type* que, correspondem, respectivamente ao texto da avaliação do crítico e ao tipo de classificação dada por este ao filme entre *Fresh* e *Rotten.* Esse dataset, composto por 1.130.017 registros possuía 65.806 textos de *reviews* nulos, que por sua vez, foram removidos dos dados. Em relação ao rótulo dos dados, o dataset se apresenta levemente desbalanceado, com 64% dos rótulos sendo *Fresh* e os outros 36% sendo *Rotten*, conforme figura 1.



*Figura 1: Distribuição dos dados de rótulo*

A coluna *review\_score* não foi utilizada por apresentar dados de classificações que utilizaram parâmetros diferentes. Conforme exemplificado na figura 2.



*Figura 2: parâmetros diferenciados utilizados na variável review\_score*

Foi criada uma coluna para fins de análise exploratória, chamada *review\_year,* baseada da data do review (*review\_date*).

O restante dos dados do *dataset* de review não foi utilizado por não serem relevantes à modelagem.

1. Conjunto de dados *Movies*

# 4. Análise e Exploração dos Dados

Nessa seção você deve mostrar como foi realizada a análise e exploração dos seus. Mostre as hipóteses levantadas durante essa etapa e os padrões e *insights* identificados.

# 5. Criação de Modelos de Machine Learning

Conforme o documento de instruções para o TCC, essa etapa é obrigatória. Nessa seção você irá descrever as ferramentas e algoritmos utilizados. Se você utilizou ferramentas visuais como Knime e Rapid Miner, coloque aqui um print do seu modelo e a descrição detalhada do workflow do seu modelo. Caso você tenha escrito scripts em Python, por exemplo, coloque aqui o seu script. Explique as *features* utilizadas, faça a comparação entre diferentes algoritmos/modelos, justifique a escolha por determinado modelo, os parâmetros utilizados, etc. Por fim, salienta-se que embora você possa utilizar ferramentas como KNIME e RapidMiner para testar protótipos do seu modelo de ML, encorajamos você a fazer seus modelos em Python ou R.

# 6. Apresentação dos Resultados

Nessa seção você deve apresentar os resultados obtidos. Apresente gráficos, *dashboards*, conte a sua história de forma bastante criativa. Aqui você pode utilizar os modelos de Canvas propostos por Dourard (clique [aqui](https://www.louisdorard.com/machine-learning-canvas)) ou por Vasandani (clique [aqui](https://towardsdatascience.com/a-data-science-workflow-canvas-to-kickstart-your-projects-db62556be4d0)).



# 7. Links

Aqui você deve disponibilizar os links para o vídeo com sua apresentação de 5 minutos e para o repositório contendo os dados utilizados no projeto, scripts criados, etc.

Link para o vídeo: youtube.com/...

Link para o repositório: github.com/...

# REFERÊNCIAS

Um projeto/relatório técnico de Ciência de Dados não requer revisão bibliográfica. Portanto, a inclusão das referências não é obrigatória. No entanto, caso você deseje incluir referências relacionadas às tecnologias ou às metodologias usadas em seu trabalho, relacione-as de acordo com o modelo a seguir.

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.

# APÊNDICE

**Programação/Scripts**

Cole aqui seus scripts em Python e/ou R.

**Gráficos**

Cole aqui workflows (KNIME, RapidMiner, etc), gráficos e figuras que você tenha gerado e não colocou no texto principal.

**Tabelas**

Cole aqui tabelas de dados que você tenha gerado e não colocou no texto principal.

1. Traduzido do original: “Rotten Tomatoes and the Tomatometer score are the world’s most trusted recommendation resources for quality entertainment. As the leading online aggregator of movie and TV show reviews from critics, we provide fans with a comprehensive guide to what’s Fresh – and what’s Rotten – in theaters and at home.”. Disponível em: https://www.rottentomatoes.com/about, acesso em 26/06/2021. [↑](#footnote-ref-0)
2. https://www.kaggle.com/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset [↑](#footnote-ref-1)