# Etapa 1 do Projeto Final de Ciência De Dados

Danilo Balman Garcia
Rafael Machado Wanner
Ra: 2482088
Ra: 2021013
[1]Departamento Acadêmico de Computação (DACOM)
Universidade Tecnológica Federal do Paraná (UTFPR)

**Abstract**

This paper is a study of the data analysis process, from problem definition to the communication of results. This first stage will present the chosen dataset, meeting the minimum requirements, the problem formulation, presenting its justification and context (stating its relevance and who would benefit), and the presentation of testable hypotheses (prioritizing those related to predictive modeling).

**Resumo**

Este artigo é um estudo do processo de análise de dados, desde a definição do problema até a comunicação dos resultados. Esta primeira etapa apresentará o dataset escolhido, atendendo aos requisitos mínimos, a formulação do problema, apresentando justificativa e contexto (explicitando sua relevância e quem seria beneficiado), e a apresentação de hipóteses testáveis (priorizando aquelas relacionadas à modelagem preditiva).

## 1 Introdução

O conjunto de dados selecionado para este projeto é uma compilação de três fontes distintas, que reúnem informações sobre filmes e programas de TV disponíveis nas plataformas de streaming Netflix, Amazon Prime e Disney+. Coletivamente, esses conjuntos de dados totalizam aproximadamente 20.000 registros, cada um descrito por 12 atributos iniciais. As variáveis abrangem diversas naturezas, incluindo dados textuais, temporais e categóricos. Prevê-se a necessidade de tratamento de dados, especialmente no atributo ”gênero”, e planeja-se o enriquecimento do dataset por meio da integração com fontes externas, como as bases de dados do IMDb ou Rotten Tomatoes, para ampliar o potencial analítico.

Os dados selecionados foram publicados no Kaggle pelo usuário Shivam Bansal, que disponibilizou conjuntos de dados para cada uma das principais plataformas de streaming Bansal (2021a,b,c). Os datasets contêm os seguintes atributos:

| Atributo | Descrição |
|---|---|
| show_id (textual) | Identificador único do filme/série. |
| type (textual) | Indica se a obra é um filme ou uma série. |
| title (textual) | Título da obra. |
| director (textual, topológico) | Diretores da obra. |
| cast (textual, topológico) | Atores que compõem o elenco. |
| country (textual, espacial) | País em que a obra foi produzida. |
| date_added (data, temporal) | Data em que a obra foi adicionada na plataforma. |
| release_year (inteiro, temporal) | Ano em que a obra foi lançada. |
| rating (textual) | Classificação indicativa. |
| duration (textual) | Duração da obra (minutos para filmes, temporadas para séries). |
| listed_in (textual) | Gêneros em que a obra é listada. |
| description (textual) | Descrição breve da obra. |

O código-fonte completo da implementação, incluindo todos os scripts utilizados para a análise de dados, foi disponibilizado publicamente. O projeto pode ser acompanhado através do seu repositório oficial no GitHub:

nosso repositório online

## 2 Formulação do Problema

Este projeto tem como objetivo conduzir uma análise aprofundada do catálogo de filmes e séries disponíveis nas principais plataformas de streaming (Netflix, Amazon Prime e Disney+). A análise busca descobrir padrões nos dados e desenvolver modelos preditivos que respondam a questões centrais sobre a estratégia de conteúdo e a recepção do público. Para guiar esta investigação, foram formuladas as seguintes perguntas de pesquisa:

1. **Análise de Desempenho por Gênero:**

   Gêneros com maior volume de produção em um determinado período (anualmente, por exemplo) tendem a receber avaliações médias (como as notas do IMDb) mais altas ou mais baixas em comparação com gêneros de nicho com menor volume de produção?

   - *Relevância:* A análise busca investigar a existência de um possível viés onde a popularidade de um gênero, medida pelo seu volume de produção, pode influenciar as avaliações médias, o que tem implicações diretas na percepção de qualidade do conteúdo.
   - *Impacto Prático:* Os resultados podem fornecer insights estratégicos para as plataformas, auxiliando na decisão entre investir em conteúdo de nicho, potencialmente mais aclamado pela crítica, ou em gêneros de alta popularidade para atrair o público geral.
   - *Beneficiários:* As empresas de streaming seriam as principais beneficiárias, utilizando os insights para aprimorar a tomada de decisão, otimizar a alocação de orçamento e mitigar os riscos financeiros associados à aquisição e produção de novos títulos.

2. **Classificação de Plataforma de Origem:**

   É possível prever em qual plataforma de streaming um título foi originalmente lançado (Netflix, Amazon Prime ou Disney+) com base em suas características, tais como gênero, país de produção, classificação indicativa e duração?

- *Relevância:* A capacidade de classificar um título permite decodificar a identidade de marca e a estratégia de curadoria de conteúdo de cada plataforma, mapeando o "DNA"de seus catálogos e revelando seus focos de investimento.

- *Impacto Prático:* A análise tem um impacto direto para produtoras e estúdios, que podem utilizar os resultados para direcionar suas propostas de projetos para a plataforma com maior afinidade, otimizando seus esforços de venda.

- *Beneficiários:* Além das produtoras, as próprias equipes de marketing das plataformas se beneficiam ao poder desenvolver campanhas mais assertivas, que reforcem os atributos que alinham um novo lançamento à identidade da marca.

## 3 Hipóteses Testáveis

- *H1: Filmes de terror costumam ter uma duração menor*

- *H2: Filmes no Disney Plus chegam mais rápido ao streaming*

- *H3: Atores dedicados ao gênero de terror participam de menos produções*

## Referências

Bansal, Shivam. 2021a. Amazon Prime Movies and TV Shows. Acessado em: 14 de outubro de 2025. https://www.kaggle.com/datasets/shivamb/amazon-prime-movies-and-tv-shows.

Bansal, Shivam. 2021b. Disney+ Movies and TV Shows. Acessado em: 14 de outubro de 2025. https://www.kaggle.com/datasets/shivamb/disney-movies-and-tv-shows.

Bansal, Shivam. 2021c. Netflix Movies and TV Shows. Acessado em: 14 de outubro de 2025. https://www.kaggle.com/datasets/shivamb/netflix-shows.