**Livros em Rede: Desvendando Recomendações Inteligentes**

CURSO: **Tecnologia em Ciência de dados**

POLO DE APOIO PRESENCIAL: **Polo EAD SP - Polo EAD Goiânia**

SEMESTRE: **04**

COMPONENTE CURRICULAR / TEMA: **Projeto aplicado III**

**Gustavo Silva Rios -** RA : **10415824**

**Silas de Souza Ferreira -** RA: **10414793**

**Israel Soares do N. Viana -** RA: **10414894**

**Danilo Brito da Silva -** RA: **10415882**

Sumário

[Introdução 3](#_Toc179055257)

[Dados 4](#_Toc179055258)

[Fonte 4](#_Toc179055259)

[Parâmetros dos Dados 4](#_Toc179055260)

[Apresentação dos Dados 4](#_Toc179055261)

[Metadados 4](#_Toc179055262)

[Link repositório 5](#_Toc179055263)

[Análise Exploratória dos Dados 6](#_Toc179055264)

[1. Limpeza Geral dos Dados 6](#_Toc179055265)

[2. Identificação e Tratamento de Outliers 6](#_Toc179055266)

[3. Engenharia de Features 6](#_Toc179055267)

[4. Visualização de Dados 7](#_Toc179055268)

[5. Insights e Interpretação 8](#_Toc179055269)

[Modelo de recomendação 9](#_Toc179055270)

[Técnica para o Treinamento do Modelo de Recomendação 9](#_Toc179055271)

[Treinamento de um Modelo Inicial como Prova de Conceito 10](#_Toc179055272)

[Avaliação de Desempenho do Modelo 10](#_Toc179055273)

[Referencial Teórico 11](#_Toc179055274)

# Introdução

O presente trabalho visa explorar um conjunto de dados limpo e bem estruturado sobre livros, coletado da API do Goodreads, com o objetivo de desenvolver um sistema de recomendação baseado em filtragem de conteúdo. O conjunto de dados é composto por 6810 livros e inclui informações valiosas, como identificadores, título, subtítulo, autores, categorias, ano de publicação, classificação média e número de avaliações. A relevância de um sistema de recomendação no contexto literário é inegável, pois ele ajuda os leitores a descobrir novos livros de acordo com suas preferências, aumentando assim a satisfação e o engajamento com a leitura. Através da análise exploratória e da engenharia de features, pretende-se entender melhor as características dos livros e a dinâmica das avaliações, possibilitando um modelo de recomendação eficaz e acessível para amantes da literatura.

# Dados

## Fonte

A fonte está no kaggle e foi construída pela API do Goodreads, uma plataforma amplamente reconhecida e utilizada por leitores e escritores para catalogar, avaliar e discutir livros. A API permite acessar informações detalhadas sobre livros e autores, facilitando a coleta de dados de forma organizada e limpa. A escolha dessa fonte foi motivada pela necessidade de obter um conjunto de dados que não apenas fornecesse uma ampla variedade de títulos, mas também garantisse a integridade e a qualidade dos dados, uma vez que muitos conjuntos disponíveis em plataformas como o Kaggle apresentavam informações incompletas ou desatualizadas.

## Parâmetros dos Dados

O conjunto de dados inclui várias colunas com informações essenciais para a análise e recomendação de livros. As principais colunas são: 'isbn13' e 'isbn10', que fornecem identificadores únicos para cada livro; 'title' e 'subtitle', que representam o nome e a subtítulo do livro; 'authors', listando os nomes dos autores, com múltiplos autores separados por uma barra; 'categories', que classifica os livros em diferentes gêneros; 'thumbnail', que oferece uma URL para a imagem da capa; 'description', que contém um resumo do livro; 'published\_year', indicando o ano de publicação; 'average\_rating', que mostra a classificação média recebida; 'num\_pages', que revela o número total de páginas; e 'ratings\_count', que informa quantas avaliações o livro recebeu. Esses parâmetros são fundamentais para compreender as preferências dos leitores e estruturar um sistema de recomendação eficaz.

## Apresentação dos Dados

Os dados foram apresentados em formato CSV delimitado por vírgulas, permitindo fácil manipulação e análise utilizando bibliotecas do Python, como Pandas e NumPy. A estrutura do conjunto de dados é bem organizada, com colunas claramente definidas que facilitam a análise. A apresentação clara e a acessibilidade dos dados são fatores cruciais que contribuem para o desenvolvimento de análises mais complexas e para a criação de um sistema de recomendação robusto.

## Metadados

Os metadados do conjunto de dados oferecem informações adicionais que enriquecem a análise. A inclusão de informações como o número de avaliações, a classificação média e a descrição dos livros permite entender melhor a dinâmica de avaliação e a recepção dos títulos pelos leitores. Além disso, o uso de identificadores exclusivos, como ISBN, garante que cada livro possa ser referenciado de maneira precisa e unívoca. Os metadados também possibilitam a realização de análises mais profundas, como a identificação de padrões de leitura e preferências por gênero, autor ou ano de publicação. Essa camada adicional de informação é fundamental para fundamentar as recomendações que serão geradas pelo sistema proposto.

Uma imagem contendo Texto

Descrição gerada automaticamente

## Link repositório

<https://github.com/gustavosrios/mack-semestre-4>

# Análise Exploratória dos Dados

A análise exploratória dos dados (AED) foi realizada com o intuito de entender as características do conjunto de dados e preparar as informações para a construção do sistema de recomendação. Para isso, foram utilizadas bibliotecas do Python como Pandas, NumPy, Matplotlib e Seaborn.

## 1. Limpeza Geral dos Dados

Asseguramos a integridade dos dados no Python através de uma série de processos de limpeza. Isso incluiu a remoção de valores nulos, o tratamento de dados ausentes e a exclusão de entradas duplicadas. Todas as inconsistências identificadas durante essas etapas foram devidamente tratadas, utilizando o heatmap da biblioteca seaborn para identificar rapidamente.

2. Identificação e Tratamento de Outliers

* Identificar valores atípicos que podem distorcer a análise;
* Utilização de métodos estatísticos (como IQR ou z-score) para detectar outliers;
* Análise do impacto desses outliers na distribuição das variáveis e na modelagem;
* Decisão sobre a remoção ou ajuste dos outliers, com base na análise contextual.

Gráfico

Descrição gerada automaticamente

## 3. Engenharia de Features

* Criar novas variáveis que enriqueçam a análise e melhoram o modelo de recomendação;
* Criação de categorias adicionais a partir de colunas existentes (por exemplo, agrupamento de livros em quantis com base na classificação média);
* Transformação de dados textuais em variáveis numéricas, se aplicável;
* Seleção de características relevantes para o modelo a ser desenvolvido.

Gráfico, Gráfico de barras

Descrição gerada automaticamente

## 4. Visualização de Dados

* Explorar as distribuições e relações entre variáveis;
* Geração de histogramas para visualizar a distribuição das classificações médias, número de páginas, entre outros;
* Criação de gráficos de dispersão e correlação para investigar relações entre diferentes variáveis, como a relação entre avaliação e número de páginas dos livros;
* Utilização de bibliotecas como Matplotlib e Seaborn para facilitar a visualização e interpretação dos dados.

|  |  |  |  |
| --- | --- | --- | --- |
| Aplicativo  Descrição gerada automaticamente com confiança baixa | Gráfico, Gráfico de dispersão  Descrição gerada automaticamente | Gráfico, Gráfico de barras  Descrição gerada automaticamente |  |
|  |  |  |  |

Gráfico

Descrição gerada automaticamente

Uma imagem contendo Esquemático

Descrição gerada automaticamente

## 5. Insights e Interpretação

Extrair informações valiosas a partir das visualizações e análises realizadas, como: Média de número de páginas por livros, desempenho dos autores, gênero literário e livros com mais e menos páginas.

# Modelo de recomendação

## Técnica para o Treinamento do Modelo de Recomendação

A técnica escolhida para o treinamento do modelo de recomendação por filtragem baseada em conteúdo se fundamenta na análise das características dos itens disponíveis e nas preferências do usuário. Esse tipo de abordagem utiliza informações dos próprios itens, como descrição, gênero, autor e outras características relevantes, para gerar recomendações personalizadas. Um dos métodos comuns é a representação dos itens em um espaço vetorial, onde as características são transformadas em vetores. Utiliza-se, então, a similaridade entre esses vetores para recomendar itens que são mais semelhantes aos que o usuário já apreciou. A similaridade pode ser calculada através de métrica do cosseno, permitindo que o modelo forneça recomendações que alinhem as preferências do usuário com as propriedades dos itens.

Texto

Descrição gerada automaticamente

## Treinamento de um Modelo Inicial como Prova de Conceito

O treinamento de um modelo inicial, como prova de conceito, envolve a utilização de um subconjunto dos dados disponíveis para validar a eficácia da abordagem escolhida. Neste estágio, um pequeno conjunto de livros com suas respectivas características e interações dos usuários é selecionado para construir o modelo. O objetivo é garantir que a técnica escolhida funcione conforme o esperado, permitindo ajustes antes da implementação completa. Durante este treinamento, é importante registrar as interações e feedbacks iniciais para aprimorar o modelo. Uma vez que o modelo é treinado, ele pode ser testado com um conjunto de dados separado para observar como ele se comporta em termos de precisão e relevância das recomendações.

Uma imagem contendo Linha do tempo

Descrição gerada automaticamente

## Avaliação de Desempenho do Modelo

A avaliação de desempenho do modelo é um passo crucial para determinar sua eficácia e qualidade nas recomendações. Uma abordagem comum é dividir os dados em conjuntos de treinamento e teste, onde o modelo é treinado em um conjunto e avaliado em outro. Métricas como precisão, recall, F1-score e a média de precisão em K podem ser empregadas para quantificar o desempenho do modelo. Além disso, a validação cruzada pode ser utilizada para garantir que o modelo não esteja superajustado aos dados de treinamento.

# Referencial Teórico

O referencial teórico que embasa a modelagem do algoritmo de recomendação é fundamental para entender as bases e a evolução das técnicas de recomendação. A filtragem baseada em conteúdo, em particular, tem suas raízes em teorias de recuperação da informação, onde a similaridade entre documentos ou itens é avaliada. Estudos prévios, como os trabalhos de Sarwar et al. (2001) e Pazzani & Billsus (2007), são cruciais para compreender as práticas recomendadas na construção de sistemas de recomendação. Além disso, a literatura sobre aprendizado de máquina e técnicas de pré-processamento de dados, como a limpeza e normalização de texto, oferece insights valiosos para garantir a qualidade dos dados utilizados no treinamento do modelo. Esse embasamento teórico não apenas valida a escolha das técnicas aplicadas, mas também orienta a implementação e avaliação do sistema de recomendação.