|  |
| --- |
| **CURSO:** Tecnologia em Ciência de Dados |
| **POLO DE APOIO PRESENCIAL:** Higienópolis |
| **SEMESTRE:** 4° Semestre - 2/2024 |
| **COMPONENTE CURRICULAR:** PROJETO APLICADO III |
| **NOME COMPLETO DO ALUNO:**  Beatriz de Souza Ferreira – RA: 10414697  Eduardo Nogueira Mota – RA: 10414834  Gustavo Castro Sangali – RA: 23023708  Jéssica Clara Da Silva Santos – RA: 10414974 |
| **TEMA:** Projeto: Sistema de Recomendação de Livros utilizando Dados do Goodreads |
| **NOME DO PROFESSOR:** Carolina Toledo Ferraz |

Sumário

[**Resumo** 3](#_Toc176977996)

[**Introdução** 4](#_Toc176977997)

[**1.** **Impacto** 5](#_Toc176977998)

[**2.** **Desenvolvimento em Python** 5](#_Toc176977999)

[**3.** **Repositório** 6](#_Toc176978000)

[**4.** **Cronograma do projeto:** 6](#_Toc176978001)

[**5.** **Definição da base de dados:** 6](#_Toc176978002)

[**5.1 Principais Campos dos Datasets** 7](#_Toc176978003)

[**Referencial Teórico** 9](#_Toc176978004)

[**Metodologia** 10](#_Toc176978005)

[**Resultados** 11](#_Toc176978006)

[**Referências** 12](#_Toc176978007)

## **Resumo**

Este projeto visa desenvolver um sistema de recomendação de livros utilizando dados do Goodreads. A partir de técnicas de aprendizado de máquina e ciência de dados, o sistema propõe sugestões personalizadas com base em avaliações de usuários e características dos livros. O projeto adota duas abordagens principais: filtragem colaborativa, que identifica padrões de comportamento entre diferentes usuários, e recomendações baseadas em conteúdo, que utilizam a similaridade entre livros para gerar sugestões. A implementação é realizada em Python, utilizando bibliotecas especializadas como Scikit-learn e Surprise, com o objetivo de criar uma solução que possa ser aplicada em ambientes educacionais e bibliotecas comunitárias, incentivando o hábito da leitura.

## **Introdução**

Nos últimos anos, os sistemas de recomendação têm desempenhado um papel crucial em diversos setores, ajudando usuários a encontrar produtos, serviços e informações de interesse de maneira personalizada. Em plataformas de leitura, como o Goodreads, a vasta quantidade de livros disponíveis torna difícil para o leitor identificar quais obras são mais relevantes de acordo com seu gosto e preferências. Por isso, um sistema de recomendação de livros torna-se uma ferramenta importante para guiar os usuários em suas escolhas de leitura, potencializando a descoberta de novos títulos e autores.

Este projeto tem como objetivo desenvolver um sistema de recomendação de livros utilizando dados disponibilizados pelo Goodreads, através do link [Goodreads datasets](https://mengtingwan.github.io/data/goodreads#datasets). O foco será na implementação de um sistema que personalize sugestões de leitura com base no histórico do usuário, similaridade entre livros e classificações realizadas na plataforma.

## **Impacto**

O caráter extensionista do projeto reflete-se no impacto que ele pode ter na comunidade de leitores e na democratização do acesso à literatura. Um sistema de recomendação de livros pode facilitar o acesso a obras que não estão no mainstream, valorizando autores independentes e promovendo a diversidade literária. Além disso, o projeto poderá ser expandido para incluir recomendações baseadas em perfis de diferentes públicos, promovendo o gosto pela leitura em grupos que talvez não tenham fácil acesso a recomendações personalizadas de qualidade.

Este sistema pode ser utilizado por bibliotecas comunitárias, escolas e organizações que promovem a leitura, tornando mais acessível o processo de descoberta de novos livros e incentivando a continuidade no hábito da leitura, em sintonia com o objetivo 4 do desenvolvimento sustentável da ONU (Organização Das Nações Unidas).

O Objetivo de Desenvolvimento Sustentável (ODS) 4 da ONU, Educação de Qualidade, tem como objetivo garantir que todos tenham acesso a uma educação inclusiva, de qualidade e equitativa, e que tenham oportunidades de aprendizagem ao longo da vida.

## **Desenvolvimento em Python**

**2.1. Bibliotecas Utilizadas**

O desenvolvimento do sistema de recomendação envolverá o uso das seguintes bibliotecas:

* Pandas: Manipulação e análise de dados.
* NumPy: Operações matemáticas e processamento de arrays.
* Scikit-Learn: Algoritmos de aprendizado de máquina.
* Surprise: Especializada para sistemas de recomendação.
* TensorFlow/PyTorch: Para modelos mais complexos, como redes neurais profundas.
* Matplotlib/Seaborn: Visualização de dados.
  1. **Modelos de Aprendizado de Máquina**

Filtragem Colaborativa: Com base nas avaliações de livros por diferentes usuários, o sistema identifica padrões de comportamento e sugere livros que outros usuários com gostos semelhantes gostaram. Algoritmos como K-Nearest Neighbors (KNN) ou SVD (Singular Value Decomposition) podem ser utilizados para criar essa recomendação.

## **Repositório**

Será utilizado um repositório no GitHub para organizar os materiais do projeto, como códigos, documentos e planejamentos.

O link do repositório é: <https://github.com/EduNogueiraMota/MACK---projeto-aplicado-III.git>

## **Cronograma do projeto:**



## **Definição da base de dados:**

A base de dados escolhida para o projeto vem da coleção de datasets do **Goodreads**, conforme disponibilizada no repositório [Goodreads datasets](https://mengtingwan.github.io/data/goodreads#datasets). Esses datasets contêm informações detalhadas sobre livros, avaliações e interações de usuários, tornando-se uma excelente fonte para criar um sistema de recomendação.

O conjunto de dados é dividido em várias partes, incluindo:

* **Books (Livros)**: Contém metadados sobre os livros disponíveis na plataforma.
* **Ratings (Avaliações)**: Contém as avaliações feitas pelos usuários em livros.
* **Users (Usuários)**: Contém informações gerais sobre os usuários (anônimos para proteger a privacidade).
* **Reviews (Resenhas)**: Contém textos de resenhas escritos pelos usuários.

### **5.1 Principais Campos dos Datasets**

#### **1. Dataset de Livros (books.csv)**

Este dataset contém informações sobre os livros cadastrados no Goodreads. Os principais campos incluem:

* **book\_id**: Identificador único do livro.
* **title**: Título do livro.
* **authors**: Nome(s) do(s) autor(es) do livro.
* **average\_rating**: Avaliação média do livro (de 0 a 5) com base nas classificações dos usuários.
* **isbn**: Código ISBN do livro (identificação internacional de obras publicadas).
* **isbn13**: Versão de 13 dígitos do ISBN.
* **language\_code**: Código do idioma em que o livro está publicado.
* **num\_pages**: Número de páginas do livro.
* **ratings\_count**: Quantidade total de avaliações recebidas pelo livro.
* **text\_reviews\_count**: Quantidade de resenhas textuais sobre o livro.
* **publication\_date**: Data de publicação.
* **publisher**: Editora do livro.

#### **2. Dataset de Avaliações (ratings.csv)**

Esse dataset contém as classificações numéricas feitas por usuários nos livros disponíveis. Os principais campos são:

* **user\_id**: Identificador único do usuário que fez a avaliação.
* **book\_id**: Identificador do livro avaliado.
* **rating**: Classificação dada pelo usuário, geralmente de 1 a 5 estrelas.

#### **3. Dataset de Resenhas (reviews.csv)**

Esse dataset contém resenhas textuais escritas pelos usuários. Os principais campos incluem:

* **review\_id**: Identificador único da resenha.
* **user\_id**: Identificador único do usuário que escreveu a resenha.
* **book\_id**: Identificador do livro ao qual a resenha se refere.
* **review\_text**: Texto da resenha escrita pelo usuário.
* **rating**: Classificação dada pelo usuário no contexto da resenha (em alguns casos, pode diferir do rating principal).

#### **4. Dataset de Usuários (users.csv)**

Este dataset contém informações sobre os usuários que interagem com o Goodreads. Ele geralmente inclui:

* **user\_id**: Identificador único do usuário.
* **location**: Localização do usuário (geralmente país ou cidade, dependendo do que for permitido no cadastro).
* **age**: Idade do usuário.

## **Referencial Teórico**

O desenvolvimento de sistemas de recomendação é amplamente estudado na área de ciência de dados e inteligência artificial. Existem três abordagens principais:

* Filtragem Colaborativa: Essa técnica identifica padrões de comportamento entre usuários com preferências similares. A filtragem colaborativa pode ser dividida em duas categorias: baseada em usuários e baseada em itens. No primeiro caso, o sistema recomenda livros que foram bem avaliados por outros usuários com gostos semelhantes. No segundo, livros similares àqueles que o usuário já leu e avaliou positivamente são sugeridos.
* Recomendações Baseadas em Conteúdo: Nessa abordagem, o foco está nas características dos livros, como gênero, autor e descrições. O sistema analisa as preferências do usuário com base em livros que ele já avaliou e sugere novos títulos com características semelhantes.
* Modelos Híbridos: Combina as abordagens de filtragem colaborativa e recomendação baseada em conteúdo para maximizar a precisão das sugestões.

Além dessas abordagens, a análise de dados e a implementação de técnicas de aprendizado de máquina, como a decomposição em valores singulares (SVD) e algoritmos de vizinhos mais próximos (KNN), são amplamente utilizadas para melhorar a performance dos sistemas de recomendação.

## **Metodologia**

O desenvolvimento deste sistema de recomendação segue uma abordagem estruturada em quatro etapas:

**1. Coleta de Dados**

Utilizaremos os datasets do Goodreads, que contêm informações sobre livros, avaliações de usuários e resenhas. A coleta será feita por meio de downloads dos arquivos CSV contendo os principais campos, como título, autor, avaliação média e avaliações de usuários.

**2. Pré-processamento**

Os dados coletados passarão por um processo de limpeza, onde serão removidos valores nulos, duplicados ou inconsistentes.

**3. Modelos de Aprendizado de Máquina**

Filtragem Colaborativa: Com base nas avaliações de livros por diferentes usuários, o sistema identifica padrões de comportamento e sugere livros que outros usuários com gostos semelhantes gostaram. Algoritmos como K-Nearest Neighbors (KNN) ou SVD (Singular Value Decomposition) podem ser utilizados para criar essa recomendação.

**4. Implementação**

A implementação será realizada em Python, utilizando bibliotecas como Pandas, Scikit-learn e Surprise. O sistema será implementado como uma API utilizando Flask, possibilitando a integração com outras aplicações.

## **Resultados**

Os resultados esperados incluem a capacidade do sistema de fornecer recomendações de livros personalizadas, com alto grau de precisão, e sua aplicabilidade em bibliotecas comunitárias e plataformas educacionais.

## **Referências**

1. MCAULEY, Julian. **Dados do Steam**. Disponível em: https://cseweb.ucsd.edu/~jmcauley/datasets.html#steam\_data. Acesso em: 06 set. 2024.
2. WAN, Mengting. **Dados Goodreads**. Disponível em: https://mengtingwan.github.io/data/goodreads . Acesso em: 06 set. 2024.