



UNIVERSIDADE PRESBITERIANA MACKENZIE
TECNÓLOGO EM CIÊNCIAS DE DADOS

Caroline Ribeiro Ferreira – 10408052

Lais César Fonseca – 10407066

Leonardo dos Reis Olher – 10407752

Liliane Gonçalves de Brito Ferraz – 10407087

Múcio Emanuel Feitosa Ferraz Filho – 10218691

Otavio Bernardo Scandiuzzi – 10407867

SISTEMA DE RECOMENDAÇÃO DE LIVROS

SÃO PAULO

2024

Sumário

1.	GLOSSÁRIO	3
2.	RESUMO	4
3.	INTRODUÇÃO	5
4.	OBJETIVO.....	7
5.	CRONOGRAMA.....	8
6.	APRESENTAÇÃO DO METADADOS	10
7.	BIBLIOTECAS PYTHON	11
8.	ANÁLISE EXPLORATÓRIA	18
9.	METODOLOGIA	19
10.	MÉTODO DE RECOMENDAÇÃO.....	21
11.	RESULTADOS	24
12.	CONCLUSÃO	25
13.	DIRETÓRIO GITHUB	26
14.	REFERENCIAL TEÓRICO	27

1. GLOSSÁRIO

- **Colaboratory:** Conhecido também como “Colab”, é um produto do Google Research, área de pesquisas científicas do Google. O Colab permite que qualquer pessoa escreva e execute código Python arbitrário pelo navegador e é especialmente adequado para aprendizado de máquina, análise de dados e educação.
- **DataFrame:** É uma estrutura de dados bidimensional com os dados alinhados de forma tabular em linhas e colunas.
- **Datasets:** conjuntos de dados organizados em um formato similar ao das tabelas, com linhas e colunas que contém informações sobre determinado tema.
- **GitHub:** GitHub é uma plataforma de hospedagem de código-fonte e arquivos com controle de versão usando o Git. Ele permite que programadores, utilitários ou qualquer usuário cadastrado na plataforma contribuam em projetos privados e/ou Open Source de qualquer lugar do mundo.
- **Kaggle:** É uma plataforma para aprendizado de ciência de dados. É também uma comunidade, a maior da internet, para assuntos relacionados com Data Science.
- **Machine Learning (Aprendizado de máquina):** É um subcampo da Engenharia e da ciência da computação que evoluiu do estudo de reconhecimento de padrões e da teoria do aprendizado computacional em inteligência artificial.
- **Python:** É uma linguagem de programação de alto nível, interpretada de script, imperativa, orientada a objetos, funcional, de tipagem dinâmica e forte. Foi lançada por Guido van Rossum em 1991.
- **Streaming:** Fluxo contínuo, fluxo de média, fluxo de mídia ou transmissão contínua, é uma forma de distribuição digital, em oposição à descarga de dados.

2. RESUMO

Neste projeto, propomos o desenvolvimento de um sistema de recomendação de livros utilizando técnicas de aprendizado de máquina e análise estatística preditiva. O objetivo principal é melhorar a experiência do usuário ao fornecer recomendações personalizadas individual com base em seu histórico de leituras.

Para o desenvolvimento do projeto utilizaremos um dataset público, que possuem acesso livre no site Kaggle criando um modelo de recomendação de livros, para recomendar livros semelhantes ao leitor com base em seu interesse. Para isto, utilizaremos dos métodos adquiridos nos componentes curriculares de introdução a ciência de dados, como aplicação da linguagem Python com o uso do Colab.

A abordagem do projeto visa aplicarmos o método de filtragem baseada em colaboração, no qual precisamos construir uma máquina preditiva que, com base nas escolhas de leituras de outras pessoas, o livro seja recomendado a outras pessoas com interesses semelhantes.

3. INTRODUÇÃO

No atual cenário literário, a vasta gama de opções de livros disponíveis desafia os leitores a navegarem por um oceano de informações para encontrar obras que verdadeiramente cativem e ressoem com seus interesses individuais. Diante desse desafio, surge a motivação para o desenvolvimento deste projeto: um “Sistema de Recomendações de Livros”. Esta introdução abordará o contexto do trabalho, a motivação, a justificativa e os objetivos, proporcionando uma visão holística do propósito e das metas deste projeto aplicado.

De acordo com o portal de notícias G1, pertencente ao grupo Globo, apenas 16% da população brasileira comprou algum livro no ano de 2023, verifica-se dessa forma a necessidade de impulsionar um mercado tão importante para o desenvolvimento intelectual das pessoas. Um possível reflexo disso foi o baixo desempenho dos jovens brasileiros no Ranking PISA, acendendo um sinal de alerta para esta situação. Desta forma, buscamos encontrar meios que facilitem o acesso à leitura, de forma que seja otimizada a escolha de bons livros que possam impulsionar o desenvolvimento de hábitos saudáveis naqueles que o utilizarem.

No contexto atual, a abundância de opções literárias é tanto um privilégio quanto um desafio. A tecnologia, no entanto, oferece oportunidades significativas para aprimorar a experiência de leitura, simplificando a escolha de livros por meio de sistemas inteligentes de recomendação. Estes, ao analisarem padrões de leitura, hábitos e preferências individuais, podem fornecer sugestões personalizadas, transformando a busca por livros em uma jornada mais enriquecedora.

A motivação para este projeto emerge da percepção das dificuldades enfrentadas pelos leitores contemporâneos ao selecionar livros entre uma ampla variedade de opções. A escolha deste tema é impulsionada pelo desejo de simplificar e aprimorar a experiência de leitura, tornando-a mais personalizada e acessível. Além disso, a paixão pela literatura e o reconhecimento do impacto positivo que uma recomendação bem-sucedida pode ter na vida do leitor constituem fortes motivadores.

Justificamos o desenvolvimento deste sistema pela sua capacidade potencial de democratizar o acesso à literatura, ampliando o alcance de obras significativas e diversificando o repertório literário dos usuários. Acreditamos que o projeto pode não apenas facilitar a escolha de livros, mas também promover a descoberta de novos gêneros e autores, enriquecendo assim a experiência de leitura de forma significativa.

Os sistemas de recomendação de livros desempenham um papel crucial não só na facilitação da descoberta de novas obras para os leitores, mas também na promoção da diversidade

cultural, no estímulo ao diálogo intercultural e na ampliação do conhecimento sobre uma vasta gama de culturas e perspectivas, com isto atendemos as questões relacionadas aos ODS (Objetivos de Desenvolvimento Sustentável no Brasil).

Ao sugerir livros que se relacionam com interesses específicos ou áreas de estudo, esses sistemas auxiliam os usuários a enriquecer seu entendimento sobre diferentes culturas, períodos históricos e pontos de vista. Além disso, ao possibilitar a descoberta de livros de uma variedade de gêneros, autores e temas, os sistemas de recomendação contribuem para enriquecer a diversidade cultural presente na leitura dos usuários.

A abordagem híbrida foi a escolhida para compor este projeto, aliando a filtragem colaborativa, baseada em interações de usuários, e a filtragem baseada em conteúdo, que foca nas características dos livros presentes nos arquivos. O uso da aprendizagem supervisionada auxiliará na obtenção de informações de múltiplos tipos de dados, como dados textuais, avaliações de usuários e demais metadados editoriais, todos esses divididos entre três datasets escolhidos para compor este projeto. Isso resultará em um modelo final de machine learning que pode prever diretamente os interesses dos usuários com a adição de itens específicos.

4. OBJETIVO

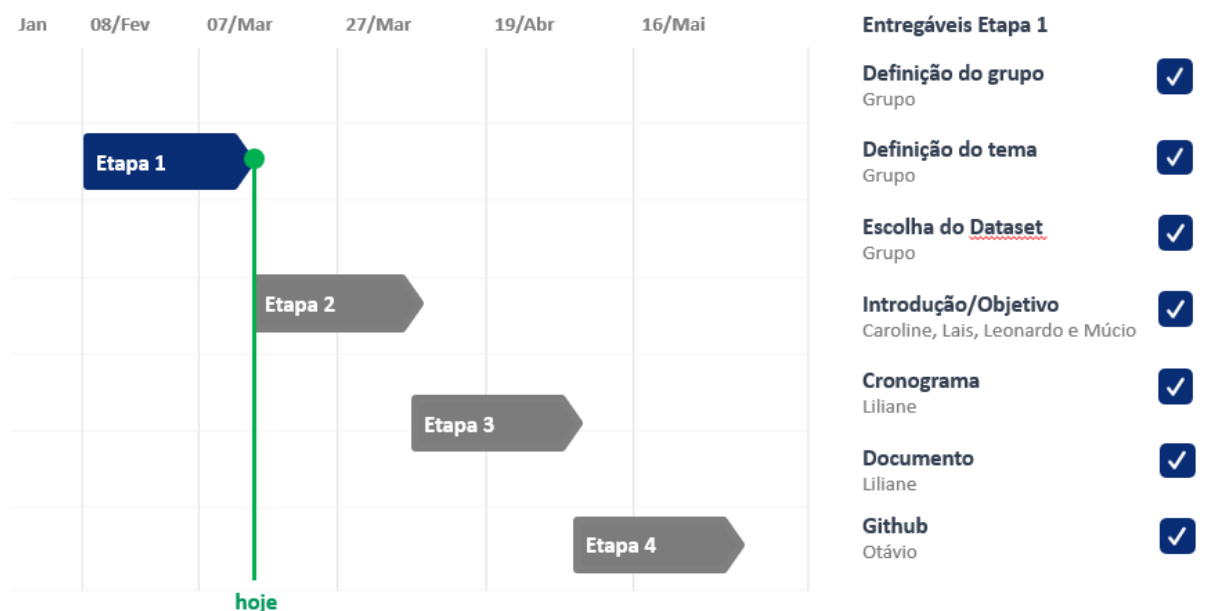
Este projeto tem como principal objetivo desenvolver um Sistema de Recomendações de Livros eficiente e inovador, aplicando técnicas avançadas de recomendação para proporcionar sugestões personalizadas aos usuários. Além disso, busca-se:

- Construir uma plataforma interativa que permita aos usuários interagirem entre si, compartilhando recomendações, resenhas e insights literários.
- Analisar e avaliar criticamente diferentes abordagens de recomendação, propondo melhorias e ajustes conforme necessário.
- Aumentar o engajamento dos usuários e estimular o hábito da leitura, com o intuito de contribuir para a formação de uma comunidade de leitores mais diversificada, tornando a leitura mais inclusiva e abrangente, contribuir para uma sociedade mais informada, empática e culturalmente enriquecida.

Através desses objetivos, almejamos não apenas criar um sistema eficaz de recomendação de livros, mas também fomentar uma comunidade engajada e apaixonada pela leitura. Esses elementos fundamentais formam a base do nosso projeto, delineando a trajetória que percorreremos para alcançar nossos objetivos e oferecer uma solução valiosa no universo literário contemporâneo.

5. CRONOGRAMA

Reportar o desenvolvimento das etapas propostas e o percentual de evolução de entregas referente ao componente curricular de Projeto Aplicado III do curso de Tecnologia em Ciências de Dados.





6. APRESENTAÇÃO DO METADADOS

Os metadados escolhidos para o desenvolvimento do nosso projeto aplicado de “Sistema de Recomendação de Livros”, tem as características apresentadas a seguir:

6.1. Tipo de arquivo

As bases de dados adquirida é de extensão csv. Denominadas de “BX_Books.csv”, “BX-Users.csv” e “Book-Ratings.csv”.

6.2. Origem dos dados

Os dados são de domínio público/aberto, do site da Kaggle.

<https://www.kaggle.com/code/isaienkov/book-review-ratings-analysis-and-visualization/input>

6.3. Sensibilidade / LGPD

Esta base de dados não possui dados sensíveis e está de acordo com a Lei Geral de Proteção de Dados Pessoas – LGPD.

6.4. Proprietário do dado

KOSTIANTYN ISAIENKOV

6.5. Descrição e atributos dos dados

O conjunto de dados BX_Books.csv é o metadados que apresenta informações dos livros, no qual contém 271.379 linhas e 8 colunas, e dispões dos seguintes atributos:

- ISBN: International Standard Book Number (Padrão Internacional de Numeração de Livro) é um padrão numérico criado com o objetivo de fornecer uma espécie de "RG" para publicações de livros.
- Book-Title: Apresenta o Título do livro no qual constam 242.154 livros distintos registrados.
- Book-Author: Apresenta o Nome do Autor do livro o qual constam livros de 102.028 autores distintos registrados.
- Year-Of-Publication: Apresenta o ano de publicação do livro no qual constam publicações de 116 anos cadastrados distintos, buscando os tops 20 anos com mais livros publicados observamos que consta livros de 1985 a 2004.
- Publisher: Apresenta o nome da Editora do livro, no qual apresenta 16.806 nomes distintos de editoras registrada.
- Image-URL: Apresenta a url da imagem/foto da capa do livro.

O conjunto de dados BX-Users.csv é o metadados que apresenta informações dos usuários, no qual contém 278.858 linhas e 3 colunas, e dispõe dos seguintes atributos:

- User-ID: Código do usuário no qual contém 278.858 usuários distintos cadastrado.
- Location: Localização do usuário, apresentado por cidade, estado e país. Apresenta 57.339 localizações distintas cadastrada.
- Age: Apresenta a idade do usuário, no qual constam 165 idades distintas cadastrada.

O conjunto de dados BX-Book-Ratings.csv é o metadados que apresenta informações das avaliações dos livros realizadas pelos leitores, no qual contém 1.149.780 linhas e 3 colunas, e dispõe dos seguintes atributos:

- User-ID: Código do usuário no qual contém 105.283 usuários distintos cadastrado.
- ISBN: Código de registro do livro, no qual contém 340.556 livros distintos cadastro.
- Book-Rating: Apresenta a nota de avaliação do usuário para cada livro, no qual contém 11 tipos de notas distinta cadastrada.

7. BIBLIOTECAS PYTHON

Para a implementação do sistema de recomendação, após estudos e análises realizadas sobre o tema, avaliamos que as bibliotecas da linguagem *Python* que serão utilizadas neste projeto para coletar os dados, processamento e tratamento dos dados, modelagem e avaliação. Selecionamos as bibliotecas a seguir:

1. Pandas:

A biblioteca Pandas é uma poderosa ferramenta de linguagem de programação *Python*, de código aberto e gratuito, que desempenha um papel fundamental na análise, limpeza e manipulação de dados. Além disso, ela permite a criação de gráficos e a manipulação de tabelas, tornando-a uma escolha essencial para programadores e cientistas de dados. *Python* é amplamente utilizado em diversas áreas, incluindo aprendizado de máquina, cibersegurança, mineração de dados, ciência de dados, programação web e muitas outras. A biblioteca Pandas é uma das razões pelas quais *Python* é tão popular para lidar com grandes estruturas de dados.

A biblioteca Pandas oferece uma ampla gama de recursos e funcionalidades para programadores e analistas de dados e suas funcionalidades são:

- **Manipulação de Dados:** O Pandas permite importar, manipular e processar dados de diversas fontes, como arquivos CSV, TSV ou bancos de dados SQL. Ele transforma esses dados em objetos *Python* chamados DataFrames, que se assemelham a tabelas, facilitando a análise e a manipulação.
- **Análise de Dados:** Com o Pandas, é possível realizar análises elaboradas dos dados. Ele oferece funções para agregar, agrupar, filtrar e calcular estatísticas, tornando a análise de dados eficiente e poderosa.
- **Limpeza de Dados:** A biblioteca simplifica a tarefa de limpar dados, permitindo a detecção e remoção de valores ausentes, duplicados e inconsistentes. Isso resulta em conjuntos de dados mais confiáveis e prontos para análise.
- **Visualização de Dados:** O Pandas se integra à biblioteca Matplotlib, facilitando a criação de gráficos e visualizações de dados. Isso torna a comunicação dos insights obtidos a partir dos dados mais eficazes.
- **Manipulação de Séries Temporais:** O Pandas oferece suporte robusto para lidar com dados de séries temporais, permitindo análises avançadas de dados ao longo do tempo.
- **Combinação de DataFrames:** É possível combinar DataFrames horizontal ou verticalmente, o que é útil quando se lida com grandes conjuntos de dados fragmentados.
- **Trabalho com Dados Categóricos:** O Pandas facilita a categorização de dados, simplificando a criação de modelos de aprendizado de máquina e a visualização de dados categóricos.

A biblioteca Pandas oferece várias vantagens distintas para os programadores e analistas de dados, são estas as principais vantagens:

- **Produtividade Elevada:** O Pandas é altamente produtivo e eficiente, economizando tempo na análise e manipulação de dados.
- **Facilidade de Acesso:** A biblioteca é conhecida por sua facilidade de uso e acessibilidade, tornando-a adequada para iniciantes e especialistas.
- **Versatilidade:** O Pandas é extremamente versátil e pode ser aplicado em diversas áreas, desde análise de dados até aprendizado de máquina.
- **Comunidade Ativa:** Com uma comunidade de colaboradores ativos, o Pandas está sempre em constante desenvolvimento e melhoria.
- **Integração com Outras Bibliotecas:** A integração com bibliotecas como Matplotlib e NumPy amplia ainda mais suas capacidades.
- **Manipulação de Grandes Dados:** Mesmo em grandes conjuntos de dados, o Pandas mantém seu desempenho e eficiência, tornando-o uma escolha sólida para projetos de qualquer escala.

2. Seaborn:

O Seaborn é uma biblioteca de visualização de dados em *Python* que se baseia no popular Matplotlib. Ela foi projetada para criar gráficos estatísticos elegantes e informativos com facilidade, exigindo apenas algumas linhas de código. O Seaborn é particularmente útil ao lidar com dados complexos, fornecendo diversas ferramentas para simplificar o processo de visualização e apresentação de resultados. Suas principais funções são:

- **Gráficos de Barras:** Os gráficos de barras são ideais para visualizar dados categóricos. O Seaborn oferece diversos tipos de gráficos de barras, incluindo gráficos de barras agrupadas, empilhadas e horizontais. Esses gráficos ajudam a representar informações de forma clara e eficaz.
- **Gráficos de Dispersão:** Os gráficos de dispersão são usados para visualizar a relação entre duas variáveis. O Seaborn oferece vários tipos de gráficos de dispersão, como aqueles com linhas de regressão e gráficos de dispersão com hexágonos, que ajudam a identificar tendências e padrões nos dados.
- **Gráficos de Caixa:** Os gráficos de caixa são úteis para representar a distribuição de uma variável numérica. O Seaborn oferece diferentes tipos de gráficos de caixa, incluindo aqueles com distribuição e pontos, permitindo a análise da dispersão e dos valores atípicos nos dados.
- **Gráficos de Densidade:** Os gráficos de densidade ajudam a visualizar a distribuição de uma variável numérica. O Seaborn oferece gráficos de densidade uni variada e bivariada, fornecendo insights sobre a distribuição conjunta de duas variáveis numéricas.

Suas vantagens de utilização são:

- **Facilidade de Uso:** O Seaborn é reconhecido por sua facilidade de uso, permitindo que os usuários criem visualizações complexas com código conciso.
- **Estilo Elegante:** A biblioteca oferece uma ampla variedade de estilos e paletas de cores elegantes, tornando as visualizações atraentes e informativas.
- **Integração com o Matplotlib:** O Seaborn é baseado no Matplotlib, o que significa que você pode combinar as funcionalidades dessas duas bibliotecas, aproveitando o poder do Matplotlib com a simplicidade do Seaborn.
- **Visualização Estatística:** O foco do Seaborn está na visualização estatística, o que o torna uma escolha qualitativa para análises exploratórias de dados e apresentação de resultados em um formato informativo.
- **Flexibilidade:** Apesar de sua simplicidade, a Seaborn oferece opções avançadas de personalização para atender às necessidades específicas de visualização.

3. Numpy:

NumPy, abreviatura de "*Numeric Python*", é uma biblioteca poderosa da linguagem de programação Python que se destaca por suas estruturas de dados multidimensionais, conhecidas como arrays. Além disso, o NumPy oferece uma extensa coleção de rotinas e funções que facilitam o processamento de arrays,

O NumPy é extremamente reconhecido por fornecer um conjunto abrangente de recursos e operações que simplifica o desenvolvimento de cálculos numéricos. Esses cálculos desempenham um papel fundamental em diversas áreas, incluindo:

- **Modelos de Machine Learning:** Em algoritmos de *Machine Learning*, é comum realizar uma variedade de cálculos numéricos, como multiplicação de matrizes, transposição e adição. O NumPy oferece uma biblioteca eficiente para executar esses cálculos de maneira fácil e rápida. Os arrays do NumPy são frequentemente usados para armazenar dados de treinamento e intervalos de modelos de *Machine Learning*.
- **Processamento de Imagem e Computação Gráfica:** Para manipular imagens de forma eficiente, o NumPy fornece funções que simplificam tarefas como espelhamento e rotação de imagens, entre outras operações de processamento de imagem.
- **Tarefas Matemáticas:** O NumPy é uma ferramenta útil para executar diversas tarefas matemáticas, incluindo integração numérica, diferenciação, interpolação e extrapolação. Além disso, a biblioteca oferece funções internas para álgebra linear e geração de números aleatórios. O NumPy é frequentemente combinado com outras bibliotecas, como SciPy e Matplotlib, para realizar tarefas complexas de análise e visualização de dados. Ele também é considerado uma alternativa ao MATLAB para aplicações matemáticas.

4. Matplotlib:

A biblioteca Matplotlib é uma ferramenta poderosa na linguagem de programação *Python*, voltada para a plotagem de gráficos 2D. Ela foi lançada em 2003 e seu desenvolvimento foi liderado pelo neurologista americano John D. Hunter. A origem do Matplotlib está ligada à pesquisa de pós-doutorado de Hunter, onde ele visualiza dados de eletrocorticografia em pacientes com epilepsia.

O Matplotlib oferece uma ampla gama de funcionalidades e recursos para criar gráficos 2D de alta qualidade e suas funcionalidades são:

- **Visualização de Dados:** A principal função do Matplotlib é criar gráficos e visualizações de dados de maneira eficaz. Ela suporta uma variedade de tipos de gráficos, incluindo gráficos de dispersão, barras, linhas, histogramas, entre outros.
- **Personalização:** A biblioteca permite personalizar todos os aspectos dos gráficos, incluindo núcleos, tamanhos, fontes e estilos. Isso possibilita a criação de visualizações únicas e informativas.
- **Suporte a Diferentes Backends:** A Matplotlib oferece suporte a uma ampla variedade de backends e saídas. Isso significa que os gráficos criados podem ser salvos em diferentes formatos de arquivo e exibidos em várias plataformas, tornando-os altamente portáteis.
- **Integração com outras bibliotecas:** O Matplotlib é frequentemente usado em conjunto com outras bibliotecas de análise de dados, como Pandas e NumPy, facilitando a criação de visualizações a partir de dados processados por essas ferramentas.

Existem várias vantagens em escolher a Matplotlib para criar gráficos e visualizações de dados, sendo elas:

- **Facilidade de Uso:** O Matplotlib é conhecido por sua facilidade de uso, tornando-a acessível tanto para iniciantes quanto para profissionais experientes.
- **Ampla Comunidade:** A biblioteca possui uma comunidade ativa de desenvolvedores e usuários.

5. SKLearn ou Scikit-Learn:

O *Scikit-learn*, originalmente chamado de *scikits.learn*, é uma biblioteca de aprendizado de máquina de código aberto para *Python*. Oferece uma ampla variedade de algoritmos para classificação, regressão e agrupamento, como máquinas de vetores de suporte, florestas aleatórias, gradient boosting, k-means e DBSCAN. Essa biblioteca é projetada para integrar-se perfeitamente com as bibliotecas *Python* numéricas e científicas, como NumPy e SciPy. Além disso, o Scikit-learn é uma ferramenta gratuita e versátil para modelagem estatística, análise de dados e aprendizado supervisionado e não supervisionado, tornando-o uma escolha popular para machine learning em *Python*. Suas principais aplicações são:

- **Algoritmos de Classificação:** Identificam categorias associadas aos dados, úteis para tarefas como classificar e-mails como spam ou não.
- **Algoritmos de Regressão:** Criam modelos para compreender a relação entre dados de entrada e saída, usados, por exemplo, para prever o comportamento dos preços das ações.

- **Algoritmos de Agrupamento (*Clustering*):** Agrupam automaticamente dados com características semelhantes, como segmentar clientes por idade ou localização.
- **Redução de Dimensionalidade:** Diminuem o número de variáveis para análise, aprimorando eficiência na visualização e processamento de dados.
- **Seleção de Modelo:** Oferecem ferramentas para comparar, validar e selecionar os melhores modelos e parâmetros para projetos de ciência de dados.
- **Pré-processamento:** Extrai e normaliza recursos nos dados, sendo útil para transformar dados de entrada, como texto, durante a análise.

Seus principais recursos são:

- **Pré-processamento:** Realiza transformações e manipulações nos dados brutos, incluindo tratamento de valores ausentes, conversão de valores categóricos em formatos numéricos e seleção de recursos.
- **Estimadores:** Oferece uma variedade de algoritmos predefinidos para aprendizado supervisionado e não supervisionado, como classificadores, regressões, SVM, árvores de decisão e algoritmos de *clustering*.
- **Avaliação do Modelo:** Fornece métricas estatísticas para avaliar o desempenho dos modelos, incluindo validação cruzada e funções de métricas individuais.
- **Otimização do Modelo:** Permite a otimização de hiper parâmetros, incluindo aprendizado conjunto, pesquisa em grade e pesquisa aleatória para melhorar o desempenho dos modelos de aprendizado de máquina

6. Torch:

É uma biblioteca de aprendizado de máquina de código aberto para Python. Oferece uma ampla variedade de algoritmos para redes neurais, otimização, processamento de tensores e muito mais. Projetado para integração perfeita com bibliotecas Python numéricas, como NumPy, e científicas, como SciPy, Torch é amplamente utilizada em pesquisas e aplicações de aprendizado profundo. Suas principais aplicações incluem:

- **Redes Neurais:** Implementação de redes neurais artificiais, incluindo modelos convolucionais, recorrentes e completamente conectados.
- **Processamento de Tensores:** Manipulação eficiente de estruturas de dados multidimensionais, essenciais para operações de álgebra linear e aprendizado profundo.
- **Otimização:** Fornecimento de otimizadores para ajustar os parâmetros dos modelos durante o treinamento, utilizando técnicas como descida de gradiente estocástica e algoritmos de segunda ordem.

- **Visualização:** Ferramentas para visualizar e diagnosticar modelos durante o treinamento, permitindo análise detalhada do comportamento do modelo e do processo de aprendizado.

Os principais recursos do Torch incluem:

- **Módulos e Camadas:** Fornece uma variedade de módulos e camadas predefinidos para construir redes neurais, incluindo convolucionais, lineares, recorrentes e de pooling.
- **Autograd:** Capacidade de calcular automaticamente gradientes para tensores, permitindo o treinamento eficiente de modelos por meio de retro propagação.
- **CUDA e Computação na GPU:** Suporte para computação na GPU por meio do CUDA, permitindo treinamento acelerado de modelos em hardware compatível.
- **API Flexível:** Interface de programação de aplicativos flexível que permite construir e personalizar modelos de aprendizado profundo de acordo com as necessidades específicas do projeto.
- **Integração com Ferramentas Existentes:** Compatibilidade com várias bibliotecas Python, incluindo NumPy e SciPy, para facilitar a interoperabilidade com outras ferramentas de análise de dados e aprendizado de máquina.

8. ANÁLISE EXPLORATÓRIA

Aplicando os métodos de análise exploratória dos dados podemos obter o entendimento de todo o conteúdo dos dados presente no dataset, para uma avaliação de como será aplicado o tratamento necessário para a melhor adequação ao modelo proposto e assim obter as medidas de acurácia mais precisas, na implementação do algoritmo de recomendação.

Neste processo de exploração passamos por algumas etapas:

- Importar as bibliotecas de necessário para realização de análise exploratória, gráficos e tratamento, sendo elas: Pandas, Numpy, Seaborn, Matplotlib.pyplot e Matplotlib.ticker.
- Importar os três conjuntos de dados: informações sobre livros, usuários e avaliações.
- Os conjuntos de dados foram unificados, com base nos identificadores únicos de usuários e livros.
- Realizou-se a análise das dimensões dos dataframes, identificação dos atributos e tipos de dados.
- Foram verificadas as distribuições de dados, identificando a quantidade de usuários, livros e avaliações disponíveis.
- Foi realizado o tratamento de valores ausentes, removendo as observações com informações faltantes.
- Calculou-se a distribuição dos usuários por faixa etária e a distribuição da quantidade de leitores por país.
- Exploramos a contagem de livros por ano de publicação, editora e autor, identificando os mais populares.
- Apresentamos as avaliações dos usuários por livro, identificando os mais avaliados.

Os dados foram tratados e preparados para o treinamento do modelo de recomendação. Isso incluiu a união dos conjuntos de dados relevantes, limpeza de dados para remover duplicatas e valores ausentes, bem como a transformação de variáveis categóricas em formato adequado para o treinamento do modelo.

- Realizou-se o tratamento dos dados removendo as observações com valores nulos.
- Selecionou-se apenas os usuários que fizeram mais de 50 avaliações para garantir uma base de dados mais robusta.
- Após o tratamento, o conjunto de dados final possui 765.672 registros e 8 atributos.

O código de análise exploratória consta disponibilizado em nosso diretório do Github:

https://github.com/OtavioBer/ProjAplicadoIII/blob/OtavioBer-patch-1/Projeto_Aplicado_III_An%C3%A1lise_Explorat%C3%B3ria.ipynb

9. METODOLOGIA

Para o desenvolvimento do sistema de recomendação de livros abordaremos a técnica de “recomendação” através de algoritmos de aprendizado de máquina a DNNs (Deep Neural Networks) e o método de avaliação MSE (Mean Squared Error), para oferecer sugestões personalizadas com base nas preferências do usuário.

Neste processo de implementação seguir um processo bem definido é muito importante para um desenvolvimento bem estruturado e eficaz, e para isto seguimos o modelo tradicional, se apresenta-se na figura 1.

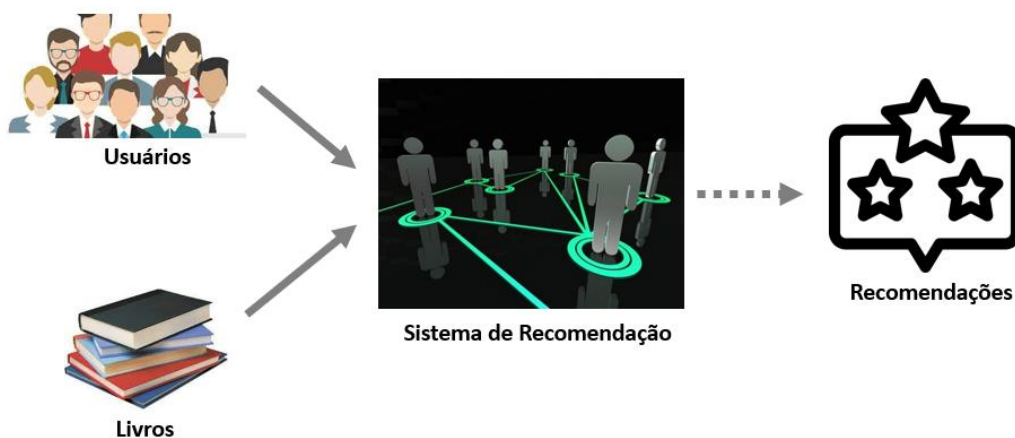
Figura 1: Processo de Recomendação



Elaborado pelo autor, 2024.

Sistemas de recomendação são algoritmos e técnicas que oferecem sugestões personalizadas de itens ou informações para usuários com base em suas preferências, comportamentos passados ou características similares de outros usuários, para exemplificar como esta tecnologia de recomendação funciona, ilustramos na figura 2, este sistema funciona a partir de dois tipos de informações: Características (sobre itens, categorias etc.) e interações entre usuários e item (avaliações, números de compras, etc.). Com base nisso, temos as recomendações baseada em conteúdo e a filtragem colaborativa.

Figura 2: Sistema de Recomendação



Elaborado pelo autor, 2024.

Um mecanismo de recomendação é uma classe de aprendizado de máquina que oferece sugestões relevantes ao cliente. Antes do sistema de recomendação, a grande tendência para comprar era aceitar sugestões de amigos. Mas agora o Google sabe quais notícias você vai ler, o Youtube sabe que tipo de vídeos você vai assistir com base em seu histórico de pesquisa, histórico de exibição ou histórico de compra.

Um sistema de recomendação ajuda uma organização a criar clientes fiéis e construir a confiança deles nos produtos e serviços desejados para os quais vieram em seu site. Os sistemas de recomendação de hoje são tão poderosos que também podem lidar com o novo cliente que visitou o site pela primeira vez. Eles recomendam os produtos que estão em alta ou com alta classificação e também podem recomendar os produtos que trazem o máximo de lucro para a empresa.

Existem três tipos principais de sistemas de recomendação:

1. Filtragem Colaborativa:

- A filtragem colaborativa é uma abordagem popular em sistemas de recomendação, que faz recomendações com base nas preferências de usuários semelhantes.
- Existem duas principais técnicas de filtragem colaborativa: baseada em usuário e baseada em item.
- Na filtragem colaborativa baseada em usuário, recomenda-se itens com base nas preferências de usuários semelhantes.
- Já na filtragem colaborativa baseada em item, os itens são recomendados com base em sua similaridade com outros itens preferidos pelo usuário.

2. Filtragem Baseada em Conteúdo:

- A filtragem baseada em conteúdo recomenda itens com base em características ou atributos dos itens e preferências do usuário.
- Essa técnica considera tanto as características dos itens quanto o perfil do usuário para fazer recomendações personalizadas.
- Por exemplo, se um usuário gosta de livros de mistério, a filtragem baseada em conteúdo recomendará outros livros de mistério com características semelhantes.

3. Sistemas Híbridos:

- Os sistemas híbridos combinam diferentes abordagens, como filtragem colaborativa e filtragem baseada em conteúdo, para melhorar a qualidade das recomendações.
- Essa abordagem pode mitigar as limitações de cada técnica individualmente e fornece recomendações mais precisas e personalizadas.

Os sistemas de recomendação são fundamentais para melhorar a experiência do usuário, aumentar o engajamento e auxiliar os usuários na descoberta de novos produtos, serviços ou conteúdos que possam ser de seu interesse. Eles são aplicados em uma variedade de setores.

Para o nosso modelo de recomendação de livros, optamos por implementar uma abordagem híbrida, combinando elementos de filtragem colaborativa e filtragem baseada em conteúdo. Utilizamos a filtragem colaborativa para identificar padrões de preferências de usuários semelhantes e recomendar livros com base nas avaliações de usuários com perfis semelhantes. Além disso, incorporamos a filtragem baseada em conteúdo para considerar características dos livros, como gênero, autor e editora, e personalizar ainda mais as recomendações com base nas preferências individuais dos usuários. Essa abordagem híbrida permite que nosso modelo aproveite as vantagens de ambas as técnicas, fornecendo recomendações mais precisas e relevantes aos usuários.

10.MÉTODO DE RECOMENDAÇÃO

Para o treinamento do modelo de recomendação, optamos por utilizar redes neurais profundas (DNN) com Embeddings, uma abordagem que se destaca pela sua capacidade de capturar relações complexas entre os dados e aprender representações eficazes para a recomendação de livros.

O que são Redes Neurais Profundas (DNN):

As redes neurais profundas é um tipo de modelo de aprendizado de máquina que consiste em múltiplas camadas de neurônios artificiais. Essas camadas são organizadas em uma arquitetura profunda, permitindo que o modelo aprenda representações hierárquicas dos dados, capturando características complexas e abstratas.

Em nosso modelo, utilizamos embeddings para representar usuários e livros de forma vetorizada em um espaço de características latentes.

Os embeddings são vetores de números reais de dimensões reduzidas, que são aprendidos durante o treinamento do modelo.

Essas representações densas e de baixa dimensão capturam informações importantes sobre os usuários e os livros, como preferências, padrões de comportamento e características intrínsecas dos itens.

A utilização de redes neurais profundas com embeddings oferece várias vantagens para o nosso modelo de recomendação de livros:

- **Captura de Relações Complexas:** As DNNs são capazes de capturar relações não lineares e complexas entre os usuários e os livros, considerando múltiplos fatores e interações entre eles.
- **Representações Eficientes:** Os embeddings aprendidos fornecem representações eficazes e compactas dos usuários e dos livros, permitindo uma modelagem mais precisa das preferências e características individuais.
- **Generalização e Personalização:** O modelo é capaz de generalizar padrões de comportamento para fazer recomendações para novos usuários e, ao mesmo tempo, personalizar as recomendações com base nas preferências individuais de cada usuário.

Avaliação de Desempenho do Modelo

O desempenho do modelo foi avaliado utilizando a métrica de erro quadrático médio (MSE), calculada sobre um conjunto de dados de teste separado. Quanto menor o MSE, melhor é o desempenho do modelo em prever as avaliações dos usuários.

O MSE (Mean Squared Error), ou Erro Quadrático Médio, é uma métrica comumente utilizada para avaliar a precisão de um modelo de regressão. Ele mede a média dos quadrados dos erros entre os valores previstos pelo modelo e os valores reais observados nos dados de teste. Matematicamente, o MSE é calculado pela média dos quadrados das diferenças entre as previsões do modelo (\hat{y}) e os valores reais (y) para cada amostra nos dados de teste.

1. Interpretação do MSE:

- Quanto menor o valor do MSE, melhor é o desempenho do modelo. Isso significa que as previsões do modelo estão mais próximas dos valores reais.
- Um MSE igual a zero indicaria um modelo perfeito, onde as previsões do modelo são exatamente iguais aos valores reais para todas as amostras de teste.
- No entanto, um MSE muito baixo nos dados de treinamento pode indicar superajustamento (overfitting), onde o modelo se ajusta muito bem aos dados de treinamento, mas não generaliza bem para novos dados.

2. Avaliação no Contexto do Modelo de Recomendação:

No contexto do modelo de recomendação apresentado, o MSE é calculado para avaliar a precisão das previsões do modelo em relação às avaliações reais dos usuários para os livros. Uma vez que o modelo é treinado e otimizado para minimizar o MSE durante o treinamento, a avaliação do MSE nos dados de teste fornece uma medida objetiva do quão bem o modelo

generaliza para novos dados não vistos. Um MSE baixo indica que o modelo é capaz de fazer previsões precisas das preferências dos usuários para os livros, enquanto um MSE alto indica que o modelo tem dificuldade em fazer previsões precisas e pode precisar de ajustes ou melhorias.

3. Utilidade do MSE:

O MSE é uma métrica robusta e amplamente utilizada para avaliar a precisão de modelos de regressão, incluindo modelos de recomendação. Ele fornece uma medida quantitativa do desempenho do modelo, facilitando a comparação entre diferentes modelos e ajustes.

No contexto de um sistema de recomendação, um baixo MSE indica que o sistema é capaz de fazer recomendações precisas e relevantes para os usuários, contribuindo para uma melhor experiência do usuário e possivelmente aumentando a satisfação e a retenção do cliente.

A implementação do algoritmo de recomendação, a nossa prova de conceito, está disponível em nosso diretório do GitHub, no link:

https://github.com/OtavioBer/ProjAplicadoIII/blob/OtavioBer-patch-1/Modelo_Recomendacao_Livros.ipynb

11.RESULTADOS

12. CONCLUSÃO

13. DIRETÓRIO GITHUB

Todo o conteúdo do projeto estará disponível no site da GitHub, que poderá ser acessado pelo link:

<https://github.com/OtavioBer/ProjAplicadoIII>

O diretório está organizado por pastas:

Pasta “Códigos” será disponibilizado os códigos em Python que foram utilizados para realizar a análise exploratória, tratamento dos dados e o sistema de recomendação.

Pasta “Dados” temos os arquivos utilizados para o estudo.

Pasta “Documentos” temos o cronograma de entrega do projeto e as versões de entrega deste documento.

Temos também o arquivo README.md com informações relevantes do projeto.

14. REFERENCIAL TEÓRICO

14.1. Sites de pesquisa:

<https://medium.com/tech-grupozap/sistemas-de-recomenda%C3%A7%C3%A3o-5bd1626326fe>
<https://www.sciencedirect.com/topics/chemical-engineering/deep-neural-network>
<https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-recommendation-engine-python/>
<https://learn.microsoft.com/pt-br/azure/architecture/solution-ideas/articles/build-content-based-recommendation-system-using-recommender>
<https://ieducacao.ceie-br.org/sistemas-recomendacao/>
<https://www.voitto.com.br/blog/artigo/biblioteca-pandas>
<https://medium.com/ensina-ai/entendendo-a-biblioteca-numpy-4858fde63355>
<https://www.crawly.com.br/blog/python-e-big-data-fique-por-dentro-de-3-bibliotecas-essenciais>
<https://dadosaocubo.com/analise-de-dados-com-seaborn-python/>
<https://surpriselib.com/>
<https://awari.com.br/scikit-learn/>
<https://pt.wikipedia.org/wiki/Scikit-learn>
<https://aws.amazon.com/pt/what-is/neural-network/#:~:text=As%20redes%20neurais%20profundas%2C%20ou,um%20n%C3%B3%20reprimir%20o%20outro.>
https://en.wikipedia.org/wiki/Mean_squared_error
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html

14.2. Livros da minha biblioteca

FERREIRA, Rogério; Aprendizado Profundo. Editora Saraiva, 2021.
NETTO, Amílcar; NETO, Francisco. Python Para Data Science e Machine Learning Descomplicado. Rio de Janeiro: Alta Books, 2021.
GRUS, Joel; Data Science do Zero. Editora Alta Books, 2021.
RICCI, F.; ROKACH, L.; SHAPIRA, B. Introdução ao Manual de Sistemas de Recomendação. Springer, 2015.
KOREN, Y.; BELL, R. Avanços na filtragem colaborativa. Springer, 2015.
Su, X.; KHOSHGOFTAAR, TM. Uma pesquisa de técnicas de filtragem colaborativa. Avanços em Inteligência Artificial, 2009.
HERLOCKER, JL; KONSTAN, JA; TERVEEN, LG; RIEDL, JT. Avaliando Sistemas de Recomendação de Filtragem Colaborativa. Transações ACM em Sistemas de Informação, 2004.

RESNICK, P.; VARIAN, HR. Sistemas de recomendação. Comunicações da ACM, 1997.

Material de apoio dos componentes curriculares: Aprendizado de Máquina, Análise Estatística Preditiva, Aquisição e Preparação de Dados