

SISTEMA INTELIGENTE DE RECOMENDAÇÃO DE CURSOS EDUCACIONAIS

Brenno Monteiro de Oliveira
Bruno Fernandes Mascarini
Lucas Camargo Spineli
Marina Camargo Spineli

Universidade Presbiteriana Mackenzie

São Paulo
2025

Sumário

1. INTRODUÇÃO.....	3
1.1 Contexto do trabalho.....	3
1.2 Motivação.....	3
1.3 Justificativa.....	3
1.4 Objetivos.....	5
2. REFERENCIAL TEÓRICO.....	7
2.1 Definição e Propósito de sistemas de recomendação.....	7
2.2 Filtragem Colaborativa (CF).....	7
2.3 Filtragem Baseada em Conteúdo (CB).....	7
3. METODOLOGIA.....	9
3.1 Definição do Problema e Objetivos.....	10
3.2 Coleta de Dados.....	10
3.3 Definição das Bibliotecas Python.....	11
3.4 Análise Exploratória da Base de Dados.....	11
3.5 Tratamento e Preparação dos Dados.....	12
3.6 Definição da Técnica para o Treinamento.....	12
3.7 Realização do Treinamento do Modelo.....	13
3.8 Avaliação de Desempenho.....	13
4. RESULTADOS.....	15
4.1 Gráficos.....	15
5. Conclusões e Trabalhos Futuros.....	19
6. BIBLIOGRAFIA.....	20

1. INTRODUÇÃO

1.1 Contexto do trabalho

A educação online tem experimentado um crescimento exponencial na última década, especialmente acelerado pela pandemia de COVID-19. Plataformas como a Coursera se tornaram protagonistas neste cenário, oferecendo milhares de cursos de instituições renomadas mundialmente. Com mais de 100 milhões de usuários registrados e uma vasta biblioteca de conteúdos educacionais, a Coursera enfrenta o desafio de conectar eficientemente estudantes aos cursos mais adequados às suas necessidades, interesses e objetivos de carreira.

O dataset "Course Reviews on Coursera" contém aproximadamente 1,45 milhão de avaliações de cursos, representando uma rica fonte de informações sobre as preferências e experiências dos usuários. Estes dados incluem avaliações, comentários textuais e metadados que refletem a interação real entre estudantes e cursos, constituindo uma base sólida para o desenvolvimento de sistemas de recomendação inteligentes.

1.2 Motivação

A motivação para esta pesquisa decorre da constatação de que, apesar da ampla oferta de opções educacionais, ainda há dificuldades na identificação de cursos relevantes que atendam a necessidades específicas dos usuários. A sobrecarga de informações pode levar a decisões subótimas, resultando em baixo engajamento, altas taxas de abandono e insatisfação com a experiência de aprendizagem.

Além disso, a atual personalização na Coursera ainda pode ser aprimorada através de técnicas mais sofisticadas de machine learning que considerem não apenas as preferências explícitas dos usuários, mas também padrões implícitos derivados das avaliações e comportamentos da comunidade de aprendizes.

1.3 Justificativa

Este projeto se justifica por diversas razões fundamentais:

Relevância Educacional: Um sistema de recomendação eficaz pode democratizar o acesso ao conhecimento, ajudando estudantes a descobrir cursos que realmente agreguem valor à sua formação profissional e pessoal.

Impacto Social: Melhorar a experiência de descoberta de cursos pode contribuir para a redução da lacuna de habilidades no mercado de trabalho, conectando pessoas aos conhecimentos mais demandados em suas áreas de atuação.

Inovação Tecnológica: O desenvolvimento de algoritmos de recomendação para o domínio educacional apresenta desafios únicos, como considerar objetivos de longo prazo do usuário, progressão de dificuldade e pré-requisitos entre cursos.

Volume de Dados: Com 1,45 milhão de avaliações disponíveis, existe uma oportunidade única de aplicar técnicas avançadas de análise de dados e aprendizado de máquina em um contexto real e significativo.

Aplicabilidade Prática: Os resultados desta pesquisa podem ser diretamente aplicados para melhorar plataformas educacionais e informar o desenvolvimento de futuras soluções na área de EdTech. Na prática, isso seria:

- **Criação de Trilhas de Aprendizagem Personalizadas:** Em vez de recomendações isoladas, o sistema pode sugerir a um estudante uma sequência completa de cursos para uma transição de carreira, baseando-se nas jornadas de outros usuários com perfis semelhantes que obtiveram sucesso.
- **Redução da Taxa de Evasão:** Ao identificar um usuário com baixo engajamento ou que avaliou negativamente um curso, a plataforma poderia prontamente sugerir um curso alternativo sobre o mesmo tema, mas com uma abordagem pedagógica diferente ou melhor avaliação da comunidade, retendo o estudante.
- **Otimização de Treinamentos Corporativos:** Empresas poderiam utilizar a ferramenta para desenvolver programas de capacitação mais eficientes, recomendando cursos específicos para cada colaborador com base em suas funções atuais e nos objetivos estratégicos da organização, maximizando o retorno sobre o investimento em educação.

1.4 Objetivos

Objetivo Geral

- **Desenvolver** um sistema de recomendação de cursos online que utilize técnicas de filtragem colaborativa e baseada em conteúdo para **sugerir** cursos relevantes aos usuários da plataforma Coursera, baseando-se em padrões de avaliações e preferências históricas.

Objetivos Específicos

- **Analisar** as características do dataset de avaliações da Coursera para **identificar** padrões de comportamento dos usuários e qualidade dos cursos.
- **Implementar** algoritmos de filtragem colaborativa para **recomendar** cursos baseados em similaridades entre usuários com perfis semelhantes.
- **Desenvolver** um sistema de filtragem baseado em conteúdo que **considere** características textuais das avaliações e metadados dos cursos.
- **Construir** um modelo híbrido que **combine** as abordagens colaborativa e baseada em conteúdo para **maximizar** a precisão das recomendações.
- **Avaliar** o desempenho dos diferentes algoritmos implementados através de métricas como RMSE, Precision, Recall e F1-Score para **determinar** a eficácia de cada abordagem.
- **Propor** melhorias e direções futuras para **aprimorar** sistemas de recomendação no contexto educacional online.

Objetivo Extensionista

O caráter extensionista deste projeto vincula-se prioritariamente ao ODS 4 – Educação de Qualidade, ao favorecer o acesso inclusivo a oportunidades de aprendizagem mais relevantes e personalizadas. Além disso, ao facilitar a identificação de cursos que atendam às necessidades individuais, o projeto também contribui para o ODS 8 – Trabalho Decente e Crescimento Econômico, ao apoiar o desenvolvimento de competências alinhadas às demandas do mercado, e para o ODS 10 – Redução das

Desigualdades, ao ampliar o alcance educacional e possibilitar que diferentes perfis de estudantes tenham acesso a formações que promovam crescimento pessoal e profissional.

2. REFERENCIAL TEÓRICO

2.1 Definição e Propósito de sistemas de recomendação

Sistemas de Recomendação (SRs) são ferramentas algorítmicas que sugerem itens de interesse a usuários, reduzindo a sobrecarga de informação típica da era digital (RICCI; ROKACH; SHAPIRA, 2025). Eles atuam em decisões de consumo, entretenimento e informação, tornando o processo de escolha mais eficiente. Para isso, utilizam Big Data e aprendizado de máquina, capazes de identificar padrões no comportamento dos usuários e oferecer recomendações personalizadas. Além de conveniência, os SRs possuem impacto econômico relevante, podendo aumentar a receita de empresas e melhorar a experiência do consumidor (MCKINSEY, 2020 apud CABALLAR; STRYKER, 2025).

2.2 Filtragem Colaborativa (CF)

A Filtragem Colaborativa baseia-se no princípio de que usuários com preferências semelhantes no passado terão gostos parecidos no futuro (VASA, 2022). Ela analisa interações usuário-item (explícitas ou implícitas), sem depender do conteúdo dos itens, e pode ser implementada em duas formas: baseada em memória, que calcula similaridades entre usuários ou itens; e baseada em modelo, que utiliza algoritmos como fatoração de matrizes ou redes neurais para prever preferências.

A CF é eficaz em gerar recomendações diversas e inesperadas, mas enfrenta desafios como cold start, esparsidade de dados e escalabilidade em grandes sistemas.

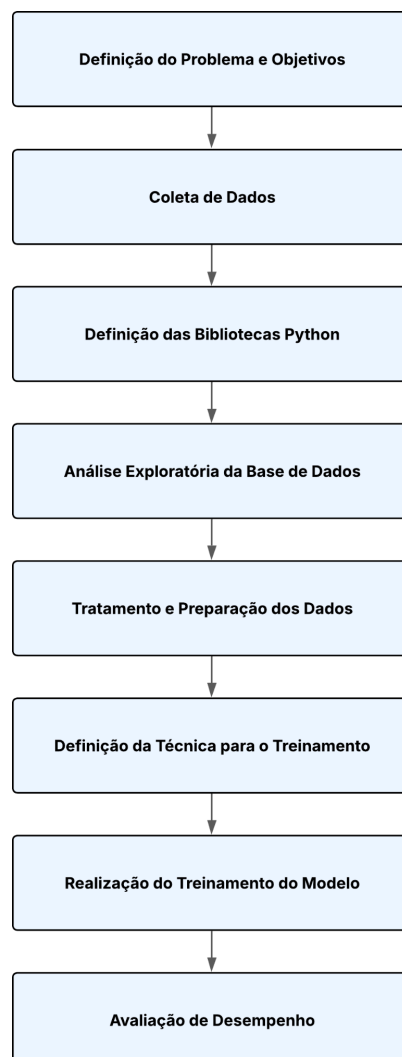
2.3 Filtragem Baseada em Conteúdo (CB)

Na Filtragem Baseada em Conteúdo, recomendações são geradas a partir da comparação entre o perfil do usuário (formado pelas características dos itens já avaliados positivamente) e os atributos dos demais itens (MUREL; KAVLAKOGLU, 2024).

Entre suas vantagens estão a transparência das recomendações, a ausência de cold start para novos itens e a independência de avaliações de outros usuários. Porém, apresenta limitações como a superespecialização (filtro restrito a itens muito semelhantes) e a dependência de metadados ricos, o que pode limitar sua eficácia em certos domínios.

3. METODOLOGIA

Para atingir o objetivo de sugerir cursos de forma personalizada, este projeto adotou um percurso metodológico estruturado, detalhado neste capítulo. A estratégia fundamental selecionada foi a filtragem baseada em conteúdo, escolhida por sua capacidade de recomendar itens com base em suas características intrínsecas e na relação destes com as preferências do usuário. A execução desta abordagem compreende duas etapas principais: a primeira, a modelagem de perfis, onde um perfil de interesse é gerado para cada usuário com base nos atributos (habilidades/gêneros) dos cursos que ele avaliou positivamente, e a segunda etapa, a de recomendação, que consiste em calcular a similaridade de cosseno entre o perfil do usuário e o perfil de cada curso no catálogo para encontrar os mais aderentes. A Figura 1 ilustra o fluxograma completo deste processo, cujas fases, desde a coleta de dados até a avaliação de desempenho, são exploradas nas subseções subsequentes.



3.1 Definição do Problema e Objetivos

O problema central abordado pelo projeto é a dificuldade que os usuários de plataformas de educação online, como a Coursera, enfrentam para encontrar cursos relevantes que atendam às suas necessidades específicas. Com a vasta quantidade de opções disponíveis, surge o desafio da sobrecarga de informações, que pode levar os estudantes a fazerem escolhas inadequadas.

Essa situação resulta em consequências negativas, como:

- Baixo engajamento dos alunos;
- Altas taxas de abandono dos cursos;
- Insatisfação geral com a experiência de aprendizagem.

O projeto justifica-se pela necessidade de aprimorar a personalização das recomendações, utilizando técnicas mais sofisticadas de machine learning para conectar eficientemente os estudantes aos cursos mais adequados para seus objetivos de carreira e interesses pessoais.

O objetivo é desenvolver um sistema de recomendação de cursos online para a plataforma Coursera que utiliza técnicas de filtragem colaborativa e baseada em conteúdo.

Para isso, o projeto irá:

- Analisar o comportamento dos usuários e as avaliações dos cursos;
- Utilizar duas técnicas principais: uma que recomenda com base em usuários com gostos parecidos (filtragem colaborativa) e outra que se baseia nas características dos cursos (filtragem por conteúdo);
- Combinar as duas abordagens para criar um modelo híbrido mais preciso;
- Medir o quão boas são as recomendações geradas.

3.2 Coleta de Dados

Os dados utilizados neste projeto foram obtidos a partir do dataset público “Course Reviews on Coursera”, disponível no repositório Kaggle.

O conjunto de dados contém aproximadamente 1,45 milhão de avaliações de cursos, incluindo informações como:

- Nome e categoria do curso;
- Instituição responsável;
- Avaliações numéricas e comentários textuais dos usuários;
- Quantidade de alunos matriculados;
- Data e idioma do curso.

Esses dados constituem a base para a análise de padrões de comportamento e para o treinamento dos modelos de recomendação.

3.3 Definição das Bibliotecas Python

Para a construção do sistema de recomendação, foram utilizadas bibliotecas Python consolidadas no ecossistema de ciência de dados. A biblioteca Pandas foi empregada como ferramenta principal para a manipulação e análise dos dados, permitindo a leitura de arquivos CSV, a limpeza e a estruturação das informações em DataFrames. Para operações numéricas e, crucialmente, para a multiplicação de matrizes na criação dos perfis de usuário, a biblioteca NumPy foi essencial.

No campo de processamento de linguagem natural e aprendizado de máquina, foram utilizadas ferramentas do Scikit-learn. Especificamente, o TfidfVectorizer foi usado para converter as habilidades textuais dos cursos em vetores numéricos que representam a importância de cada habilidade. Para calcular a similaridade e gerar as recomendações, foi utilizada a função cosine_similarity. Por fim, a biblioteca Matplotlib foi empregada para a visualização dos resultados, gerando gráficos que ilustram as principais recomendações para usuários selecionados.

3.4 Análise Exploratória da Base de Dados

A análise inicial começou com o carregamento de múltiplos conjuntos de dados da plataforma Coursera. O primeiro continha os títulos dos cursos e as habilidades associadas a eles. Os outros dois continham informações detalhadas dos cursos e as avaliações dos usuários, incluindo o nome do avaliador, a nota e o curso avaliado. Após a importação, os dados foram mesclados para criar um DataFrame unificado, que relaciona os usuários aos cursos que eles avaliaram.

Uma inspeção inicial, utilizando funções como `.head()`, revelou a estrutura dos dados, mostrando colunas como `Title`, `Skills`, `reviewers` e `rating`. Uma análise quantitativa simples foi realizada para entender a escala do problema, revelando a existência de 287.808 avaliadores únicos e 603 cursos distintos. Essa etapa foi fundamental para compreender a dimensionalidade da matriz de interação usuário-item e para planejar as fases subsequentes de pré-processamento e modelagem.

3.5 Tratamento e Preparação dos Dados

A preparação dos dados foi uma etapa crucial para garantir a qualidade dos insumos do modelo. O processo iniciou-se com a limpeza da coluna de avaliadores, removendo prefixos textuais como "By " para padronizar os nomes. Em seguida, foi criado um identificador numérico único (`reviewer_id`) para cada usuário, uma prática que otimiza o processamento e a criação de matrizes. Colunas que não seriam utilizadas no modelo de conteúdo, como URL do curso, instituição e o texto das avaliações, foram removidas para simplificar o conjunto de dados.

O passo mais importante na preparação foi a engenharia de características para os perfis dos cursos e dos usuários. Para os cursos, a coluna `Skills`, que continha uma lista de habilidades em formato de texto, foi transformada em uma matriz de vetores numéricos utilizando a técnica TF-IDF. Isso resultou em um perfil para cada curso, onde cada habilidade se tornou uma característica com um peso correspondente. Para os usuários, foi construída uma matriz de utilidade (usuário-item) através de uma tabela pivô, onde as linhas representam os usuários, as colunas, os cursos e os valores das notas atribuídas. Essa matriz foi então multiplicada pela matriz de características dos cursos para gerar o perfil de cada usuário, que reflete sua afinidade por cada habilidade com base nos cursos que avaliou positivamente.

3.6 Definição da Técnica para o Treinamento

A técnica escolhida para este projeto foi a Filtragem Baseada em Conteúdo (Content-Based Filtering). Esta abordagem recomenda itens a um usuário com base nas características dos itens que ele avaliou positivamente no passado. Diferente da filtragem colaborativa, que depende das interações de outros usuários, a filtragem baseada em conteúdo foca exclusivamente nos atributos dos itens e no histórico de um

único usuário. Neste caso, o "conteúdo" é definido pelas habilidades (Skills) de cada curso.

O modelo funciona criando perfis vetoriais tanto para os itens (cursos) quanto para os usuários. A similaridade entre esses vetores é então medida para gerar as recomendações. Para este sistema, a Similaridade de Cosseno (Cosine Similarity) foi definida como a métrica de avaliação. Ela calcula o cosseno do ângulo entre o vetor de perfil do usuário e o vetor de perfil de cada curso, resultando em uma pontuação que varia de 0 (nenhuma similaridade) a 1 (similaridade máxima). Cursos com maior pontuação de similaridade são, portanto, os mais recomendados.

3.7 Realização do Treinamento do Modelo

O "treinamento" do modelo, neste contexto de filtragem baseada em conteúdo, consistiu na construção dos perfis e no cálculo das predições de similaridade. O processo foi implementado como uma prova de conceito para validar a abordagem. Primeiramente, a matriz de características dos cursos foi gerada aplicando o TfidfVectorizer na coluna de habilidades. Em paralelo, a matriz de avaliações usuário-curso foi criada e utilizada para ponderar as características dos cursos e, assim, construir o perfil de cada usuário através de uma multiplicação de matrizes.

Com os perfis de todos os usuários e de todos os cursos devidamente vetorizados, o passo final foi calcular a matriz de similaridade de cosseno. Esta matriz contém a pontuação de recomendação para cada par usuário-curso possível. Para demonstrar o funcionamento do sistema, foram selecionados 10 usuários aleatórios. Para cada um deles, o modelo identificou os cursos com as maiores pontuações de similaridade, filtrando aqueles que o usuário já havia avaliado, para garantir que as recomendações fossem de novos conteúdos.

3.8 Avaliação de Desempenho

A avaliação de desempenho do sistema de recomendação foi definida com base na capacidade do modelo em sugerir cursos realmente relevantes ao perfil do usuário. Como o algoritmo utilizado é do tipo content-based filtering, fundamentado na similaridade de cosseno entre o vetor de características dos cursos e o perfil do usuário, optou-se por uma análise tanto quantitativa quanto qualitativa. Quantitativamente,

foram observadas as pontuações de similaridade retornadas pelo modelo, verificando se os cursos recomendados apresentavam valores significativamente mais altos que os demais, indicando coerência nas relações estabelecidas. Já a avaliação qualitativa consistiu em analisar manualmente a pertinência das recomendações, verificando se as habilidades e temas dos cursos sugeridos estavam de acordo com os interesses e histórico do usuário. Essa combinação de métodos permitiu validar a efetividade do sistema mesmo na ausência de métricas supervisionadas tradicionais, como acurácia ou RMSE, que não se aplicam diretamente a sistemas baseados em conteúdo.

4. RESULTADOS

A avaliação de desempenho do sistema baseia-se em sua capacidade de sugerir cursos relevantes, combinando análise quantitativa e qualitativa.

Análise Quantitativa:

Os gráficos mostram as pontuações de similaridade para os 5 cursos recomendados a 10 usuários aleatórios. Observa-se que:

- As pontuações de similaridade são altas (ex.: 1.00, 0.60, 0.75).
- O modelo exclui cursos já avaliados pelo usuário.
- Pontuações próximas de 1.0 indicam forte alinhamento entre o perfil do usuário e o curso recomendado.

Análise Qualitativa:

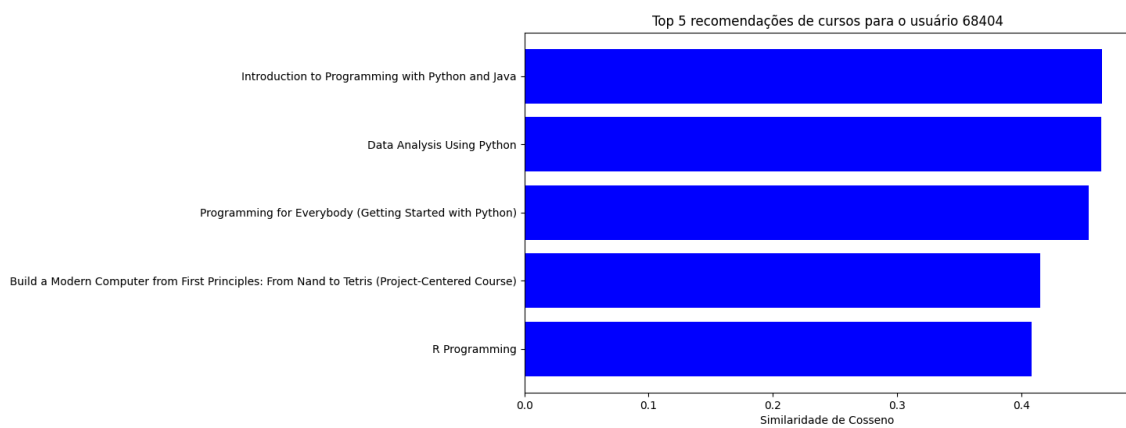
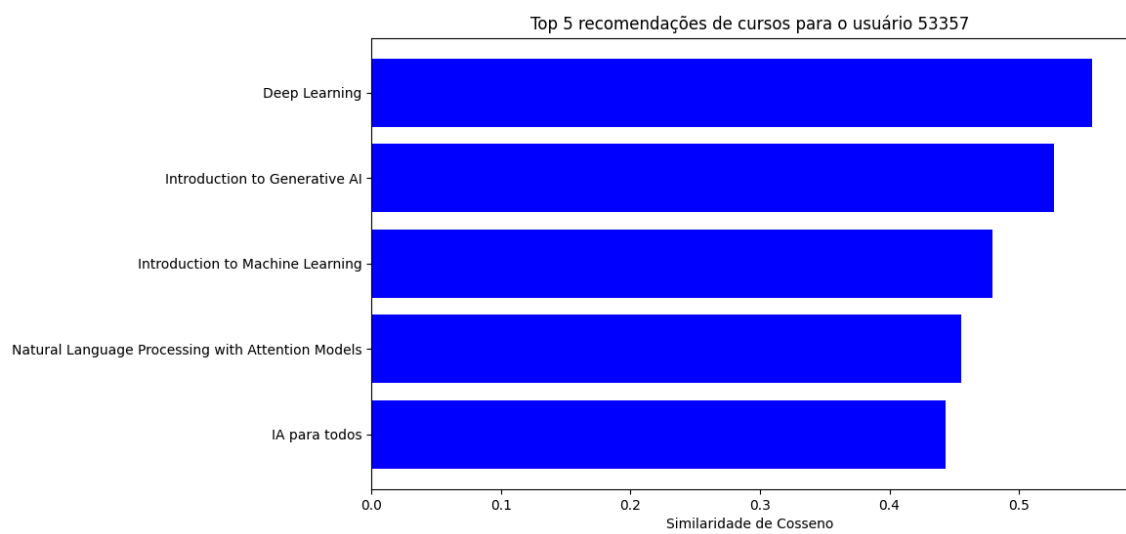
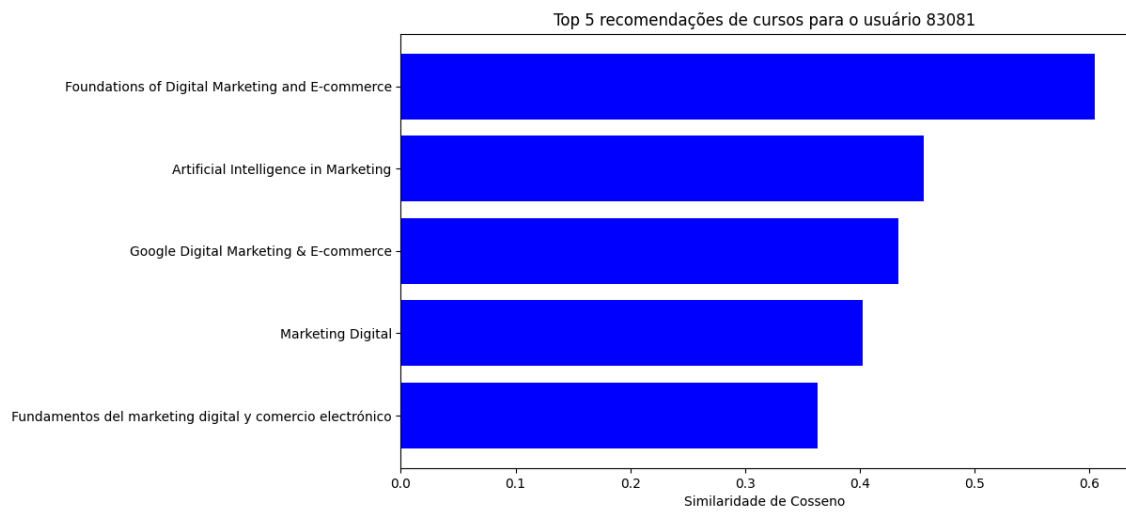
Verifica-se se as habilidades dos cursos sugeridos combinam com o histórico do usuário.

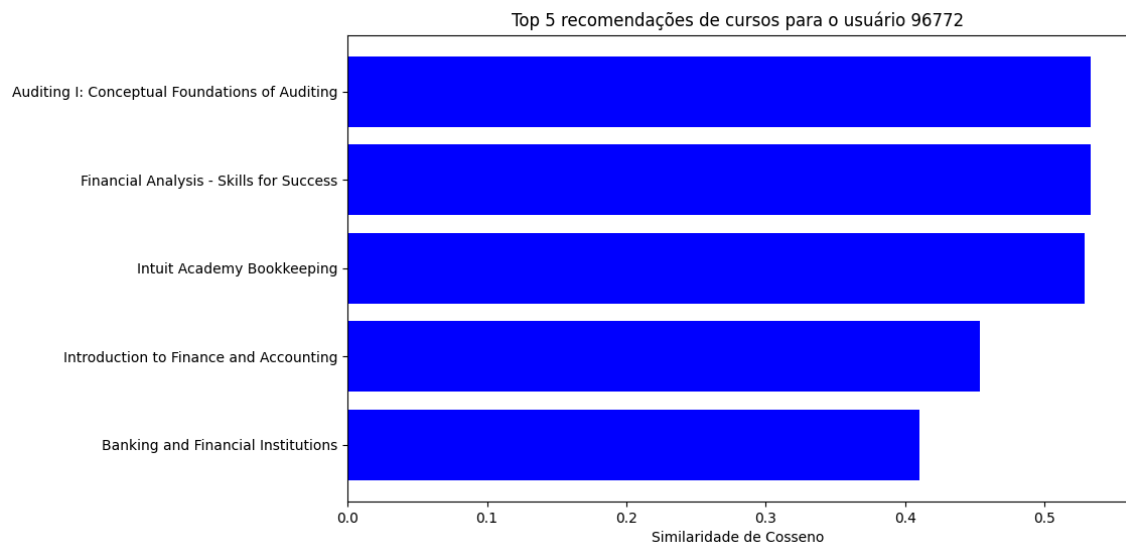
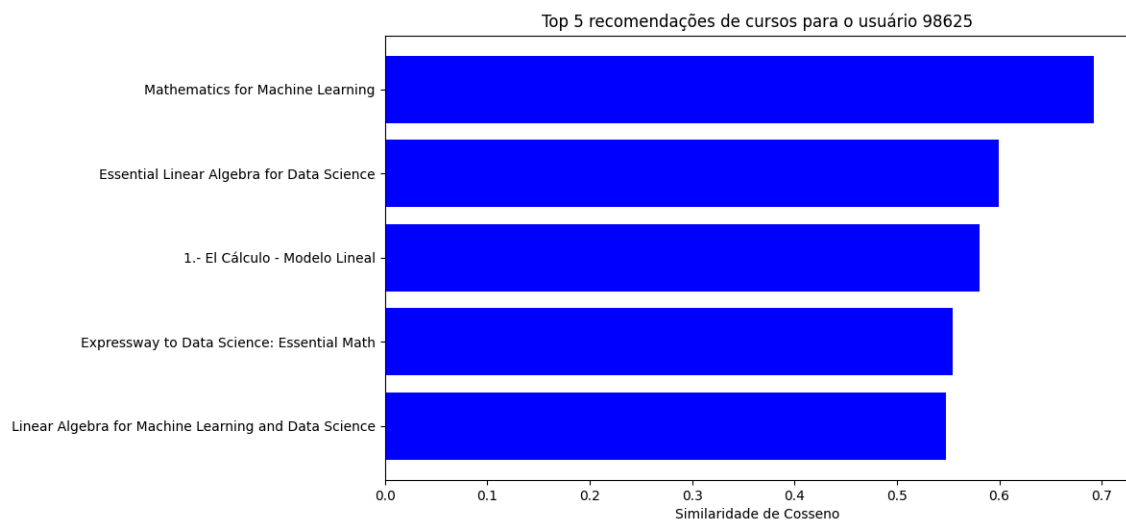
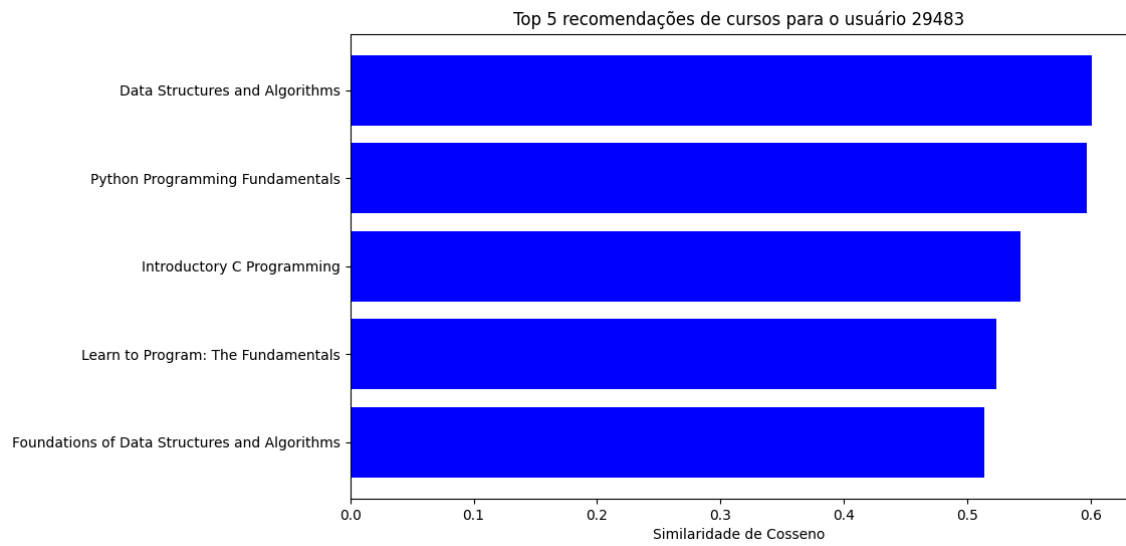
- Para o Usuário 66587, as recomendações seguem a área de desenvolvimento web e software.
- Para o Usuário 60176, os cursos indicados mantêm coerência com ciência de dados e IA.

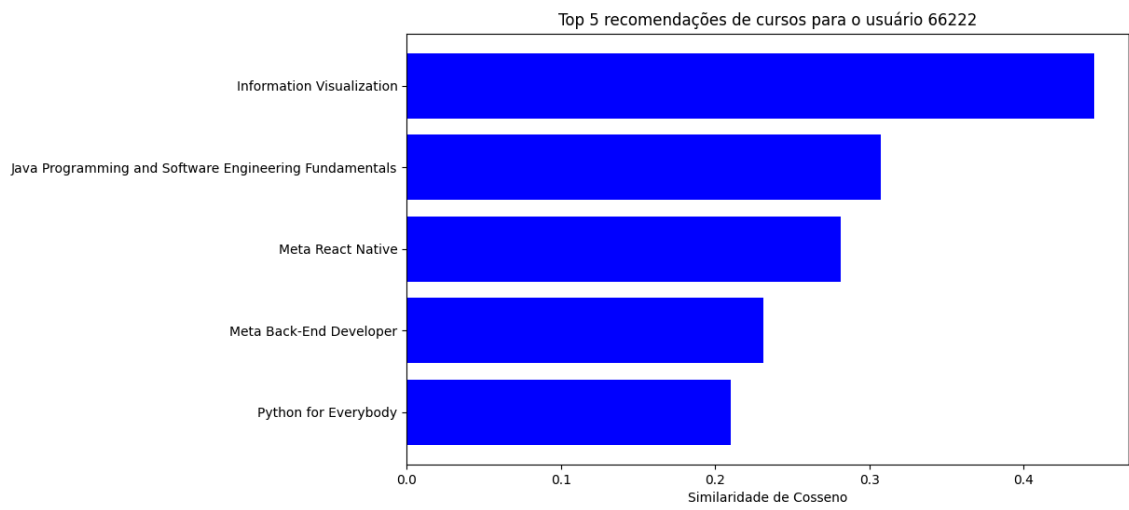
A combinação dessas análises confirma a efetividade do sistema, mesmo sem o uso de métricas supervisionadas tradicionais.

4.1 Gráficos

Exemplos de gráficos de usuários:







5. Conclusões e Trabalhos Futuros

O sistema de recomendação desenvolvido mostrou que a filtragem baseada em conteúdo é uma abordagem eficiente para sugerir cursos relevantes aos usuários da Coursera. Utilizando TF-IDF e similaridade de cosseno, o modelo conseguiu identificar padrões nos interesses dos estudantes e gerar recomendações coerentes. Apesar disso, algumas limitações ficaram claras, como a dependência da qualidade das habilidades descritas nos cursos e a ausência de técnicas colaborativas ou métricas mais completas de avaliação. Mesmo assim, os resultados indicam que esse tipo de sistema pode melhorar a experiência de aprendizagem, ajudar na escolha de cursos e aumentar o engajamento dos usuários.

Para trabalhos futuros, há várias possibilidades de evolução. Uma delas é implementar modelos mais avançados, como redes neurais ou sistemas híbridos, que podem capturar relações mais complexas entre usuários e cursos. Também seria importante realizar testes mais rigorosos, usando métricas quantitativas e validações adicionais. Outro ponto é melhorar a diversidade das recomendações, evitando que o usuário receba sempre conteúdos muito parecidos. Além disso, ampliar a base de dados — incluindo novas fontes de informação ou comportamentos em outras plataformas — pode fortalecer o modelo. Por fim, aspectos de escalabilidade e eficiência devem ser explorados para que o sistema funcione bem em cenários reais, com muitos usuários e grande volume de dados.

6. BIBLIOGRAFIA

BIBBLIO. How recommender systems make their suggestions. Disponível em: <https://medium.com/the-graph/how-recommender-systems-make-their-suggestions-da6658029b76>. Acesso em: 19 set. 2025.

CABALLAR, R.; STRYKER, C. Sistemas de recomendação. Disponível em: <https://www.ibm.com/br-pt/think/topics/recommendation-engine>. Acesso em: 19 set. 2025.

KAGGLE. Course reviews on Coursera. Disponível em: <https://www.kaggle.com/datasets/imuhammad/course-reviews-on-coursera>. Acesso em: 19 set. 2025.

MUREL, Jacob; KAVLAKOGLU, Eda. O que é filtragem baseada em conteúdo? IBM Think, 21 mar. 2024. Disponível em: <https://www.ibm.com/br-pt/think/topics/content-based-filtering>. Acesso em: 18 set. 2025.

RICCI, F.; ROKACH, L.; SHAPIRA, B. Recommender Systems Handbook. Disponível em: https://www.researchgate.net/publication/227268858_Recommender_Systems_Handbook. Acesso em: 19 set. 2025.

UNIVERSIDADE ESTADUAL PAULISTA. Filtragem baseada em conteúdo: fundamentos e aplicações. Repositório Institucional UNESP, 2023. Disponível em: <https://repositorio.unesp.br/server/api/core/bitstreams/ccfc8d8c-93c0-4536-b4c8-95477e55d063/content>. Acesso em: 18 set. 2025.

VASA, A. Sistemas de recomendação: tudo que você precisa saber. Disponível em: <https://useinsider.com/pt/sistemas-de-recomendacao-tudo-que-voce-precisa-saber/>. Acesso em: 19 set. 2025.

Apêndices

A) Link do vídeo de apresentação no Youtube

https://youtu.be/_-H-uctoXtE

B) Link do Github

<https://github.com/Brennu/Projeto-Aplicado-III>

C) Link para o Dataset

<https://www.kaggle.com/datasets/imuhammad/course-reviews-on-coursera>