Universidade Presbiteriana Mackenzie

Faculdade de Computação e Informática - Ciência de Dados

Projeto Aplicado III – Sistema de recomendação em site de compras

Membros do grupo: Carlos Oliveira, Felipe Ferraz, Erick Isidoro

São Paulo, 01 de maio de 2025

Resumo

Sumário

Introdução

Neste projeto, trabalhamos com sistemas de recomendação para sites de vendas, uma ferramenta que contribui significativamente para o aumento das vendas, podendo, em muitos casos, representar a sobrevivência da empresa. Esses sistemas evitam o acúmulo de produtos, otimizam a alocação de espaço e direcionam esforços para produtos que, de fato, trarão resultados relevantes.

O objetivo principal do sistema é compreender os hábitos de compra dos clientes em sites de comércio eletrônico e sugerir produtos frequentemente adquiridos em conjunto. Por exemplo, ao adicionar farinha ao carrinho, o sistema pode sugerir automaticamente ingredientes como açúcar e ovos, que são comumente comprados juntos para o preparo de uma receita. Além disso, o sistema também pode sugerir receitas completas com base nos ingredientes já selecionados ou recomendar ingredientes adicionais de acordo com a preferência culinária do usuário, como pratos típicos italianos, japoneses ou brasileiros.

Este projeto aplicado considera um objetivo extensionista, atendendo às necessidades dos Objetivos de Desenvolvimento Sustentável (ODS) da ONU, especialmente no que diz respeito ao consumo e produção responsáveis (ODS 12). Através da recomendação de produtos que otimizam o uso de recursos e reduzem o desperdício, o sistema contribui para práticas comerciais mais sustentáveis. Além disso, ao priorizar produtos de pequenos comerciantes e negócios locais, o projeto também apoia o crescimento econômico inclusivo (ODS 8) e promove a inovação na gestão de vendas e estoques (ODS 9).

A integração desses objetivos no sistema de recomendação possibilita não apenas aumentar as vendas e a satisfação do cliente, mas também impulsionar práticas comerciais responsáveis e socialmente conscientes, fortalecendo o impacto positivo do projeto tanto no âmbito econômico quanto social.

Os dados utilizados foram obtidos na plataforma Kaggle, Análise do e-Commerce no Brasil - Olist Dataset, abrangendo o período de 2016 a 2018. Os dados foram disponibilizados pela Olist, uma loja de

departamentos com mais de 100 mil pedidos registrados. Por se tratar de dados reais, foram preservados o anonimato dos indivíduos, parceiros e empresas envolvidas.

Os dados foram organizados em oito conjuntos distintos: Consumidores, Vendedores, Produtos, Pedidos, Artigos dos Pedidos, Pagamento e Geolocalização.

Referencial Teórico

Os sistemas de recomendação têm se tornado ferramentas essenciais em plataformas de comércio eletrônico, redes sociais e serviços de streaming, pois auxiliam os usuários na descoberta de produtos, conteúdos ou informações de interesse. Dentre os algoritmos utilizados nesse contexto, destaca-se o K-Vizinhos Mais Próximos (KNN), um método baseado em instâncias que avalia a similaridade entre itens ou usuários com base em características pré-definidas (RICCI; ROKACH; SHAPIRA, 2015).

O algoritmo KNN foi originalmente proposto por Cover e Hart (1967) e se caracteriza por ser não-paramétrico, realizando classificações ou previsões com base nos exemplos mais próximos em um espaço métrico. Em sistemas de recomendação, o KNN pode ser aplicado tanto na filtragem colaborativa quanto no conteúdo, identificando itens similares com base em atributos ou no comportamento de usuários (SARWAR et al., 2001).

Segundo Hastie, Tibshirani e Friedman (2009), o desempenho do KNN está diretamente relacionado à escolha da métrica de distância e à seleção de características relevantes, sendo comuns o uso das distâncias Euclidiana, Manhattan e Minkowski. Apesar de sua simplicidade, o algoritmo apresenta bons resultados em bases de dados menores ou de média escala, com o benefício adicional de não requerer um treinamento prévio complexo.

Além disso, McKinney (2010) destaca o papel de bibliotecas como o Pandas na manipulação eficiente de dados estruturados, enquanto Géron (2019) explora a aplicação do KNN em projetos práticos utilizando a biblioteca scikit-learn, amplamente adotada na comunidade científica e profissional para implementação de algoritmos de aprendizado de máquina em Python.

Estudos recentes continuam explorando variações e aplicações do KNN em sistemas de recomendação, muitas vezes combinando-o com técnicas de pré-processamento, redução de dimensionalidade e análise de agrupamentos para melhorar a performance do sistema (ZHANG et al., 2020). Isso demonstra a relevância contínua desse algoritmo como base para soluções práticas e acessíveis no campo da recomendação de produtos.

Metodologia

O projeto envolveu uma série de decisões metodológicas com o objetivo de garantir a eficácia dos modelos testados. Abaixo, descrevemos as etapas realizadas no pipeline:

1. **Importação do Dataset**

   Foi realizada a importação do *dataframe* disponibilizado na plataforma Kaggle, no repositório `olistbr/brazilian-ecommerce`.

2. **Importação das Bibliotecas**

   Utilizamos as bibliotecas `pandas`, `numpy`, `seaborn`, `matplotlib`, `warnings`, `folium` e `sklearn.cluster` para análise, visualização e modelagem dos dados.

3. **Análise e Cruzamento de Tabelas**

   As tabelas separadas inicialmente — `customers`, `geolocation`, `order_items`, `orders`, `products`, `sellers` e `reviews` — foram gradualmente cruzadas. Em cada etapa, selecionamos apenas as colunas de interesse, resultando em uma tabela consolidada com 113.007 linhas e 28 colunas.

4. **Tratamento de Dados**

   No tratamento da tabela consolidada:

   a. Removemos os registros de 2016 (apenas 375 registros, não cobrindo o ano completo).
   b. Também excluímos setembro de 2018, que continha apenas um registro.
   c. Convertimos as colunas de data para o tipo `datetime`.
   d. Criamos duas novas colunas: uma para o tempo de entrega e outra para a codificação do `product_id` (com Label Encoding).

Ao final do tratamento, o *dataframe* ficou com 112.632 linhas e 30 colunas.

5. **Análises Exploratórias**

   A seguir, realizamos diversas análises sobre os dados tratados:

a. Evolução do volume de compras ao longo do tempo (**Gráfico I**);
b. Gasto médio com frete por estado e volume de compras por estado (**Gráfico II**);
c. Volume de compras por horário, destacando os períodos com maior concentração de pedidos (**Gráfico III**);
d. Volume e quantidade de itens por consumidor;
e. Análise da distância entre consumidores, com base em latitude e longitude de 100 clientes selecionados (**Imagem II**).

6. **Sistema de Recomendação com KNN**

Para o sistema de recomendações, utilizamos as seguintes variáveis:

a. `order_item_id` (identificador único do item no pedido),
b. `product_id` (identificador único do produto),
c. `price` (preço),
d. `product_category_name` (categoria do produto).

Selecionamos os 10 produtos mais vendidos para testar o modelo. Utilizamos *Label Encoding* para transformar os `product_id` em valores numéricos. Em seguida:

e. Dividimos os dados entre treino e teste;
f. Definimos como variáveis preditoras (`X`) o `order_item_id` e o `price`, e como variável-alvo (`y`) o `product_id` codificado;
g. Testamos o modelo KNN com `k = 5` e `k = 10`, mas os melhores resultados foram obtidos com `k = 15`;
h. Também testamos outras composições de variáveis preditoras e alvo, mas todas apresentaram desempenho inferior.

Para a recomendação final, aplicamos o modelo aos 10 produtos mais vendidos. Geramos, para cada um deles, uma lista com os 5 itens mais próximos em termos de similaridade, acompanhados da probabilidade de venda. Os resultados foram consolidados em uma planilha Excel para facilitar a visualização (**Imagem III**).

Resultado

Conclusões e trabalhos Futuros

GitHub

Link para acesso ao projeto no GitHub e Youtube

Abaixo tem-se o link do GitHub onde estão compartilhados, dataset, cronograma, script outros documentos atualizados do projeto:

Demonstração da imagem do cronograma em tabelas de imagem (Imagem 1 – Cronograma de Atividade), arquivo disponibilizado em link Github.

https://github.com/Ferraz0Felipe/Projeto_Aplicado_III.git

Tabela de imagens

Imagem I – Cronograma de Atividade (disponível na integra link Github)

Etapa	Titulo	Atividades	Responsável	Inicio	Término	Duração (dias)	Milestones	Status
1	Concepção do Produto	Organizar o grupo	Carlos	2-fev-2025	4-mar-2025	15	6-mar-2024	Concluída
		Escolher Tema	Erick	2-fev-2025	4-mar-2025	3	6-mar-2024	Concluída
		Escolher a base de dados	Felipe	2-fev-2025	4-mar-2025	5	6-mar-2024	Concluída
		definir o cronograma do projeto	Erick	2-fev-2025	4-mar-2025	4	6-mar-2024	Concluída
		Elabora o documento inicial	Carlos	2-fev-2025	4-mar-2025	4	6-mar-2024	Concluída
2	Definição do produto	Analisar a base de dados	Erick	11-mar-2025	10-abr-2025	24	3-abr-2024	Concluída
		limpar e preparar a base de dados	Felipe	11-mar-2025	10-abr-2025	16	3-abr-2024	Concluída
		Escolher a tecnica para treinamento do modelo	Carlos	11-mar-2025	10-abr-2025	4	3-abr-2024	Concluída
		construir uma prova de conceito	Carlos	11-mar-2025	10-abr-2025	10	28-abr-2024	Concluída
		Definir métricas de avaliaçãode desenpenho	Erick	11-mar-2025	10-abr-2025	12	28-abr-2024	Concluída
3	Metodologia	Implementar a técnica proposta	Felipe	3-abr-2025	3-mai-2025	8	28-abr-2024	Concluída
		Ajustar o pipeline de dados	Carlos	3-abr-2025	3-mai-2025	5	28-abr-2024	Concluída
		Documentar os passos implementados	Erick	3-abr-2025	3-mai-2025			
4	Resultados e Clonclusão	Organizar e Documentar os resultados	Carlos	2-mai-2025	1-jun-2025	11	17-mai-2024	Concluída
		Filanilar Documento do projeto	Erick	2-mai-2025	1-jun-2025	10	17-mai-2024	Concluída
		Entregar Materiais produzidos e apresentação	Felipe	2-mai-2025	1-jun-2025	8	17-mai-2024	Concluída

Concluído Programado Atrasado

Fevereiro

16 17 18 19 20 21 22 23 24 25 26 27 28 29 01 02 03 04

Gráfico I - evolução do volume de compras realizadas no decorrer do tempo

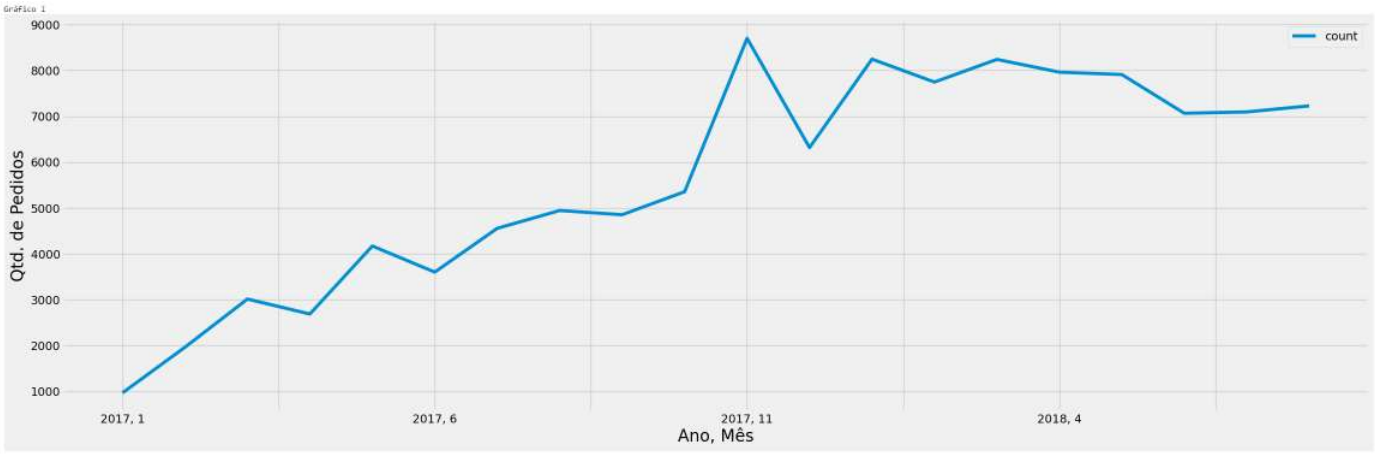

Gráfico II – Volume de vendas por estado e valor médio de gastos com frete por estado.

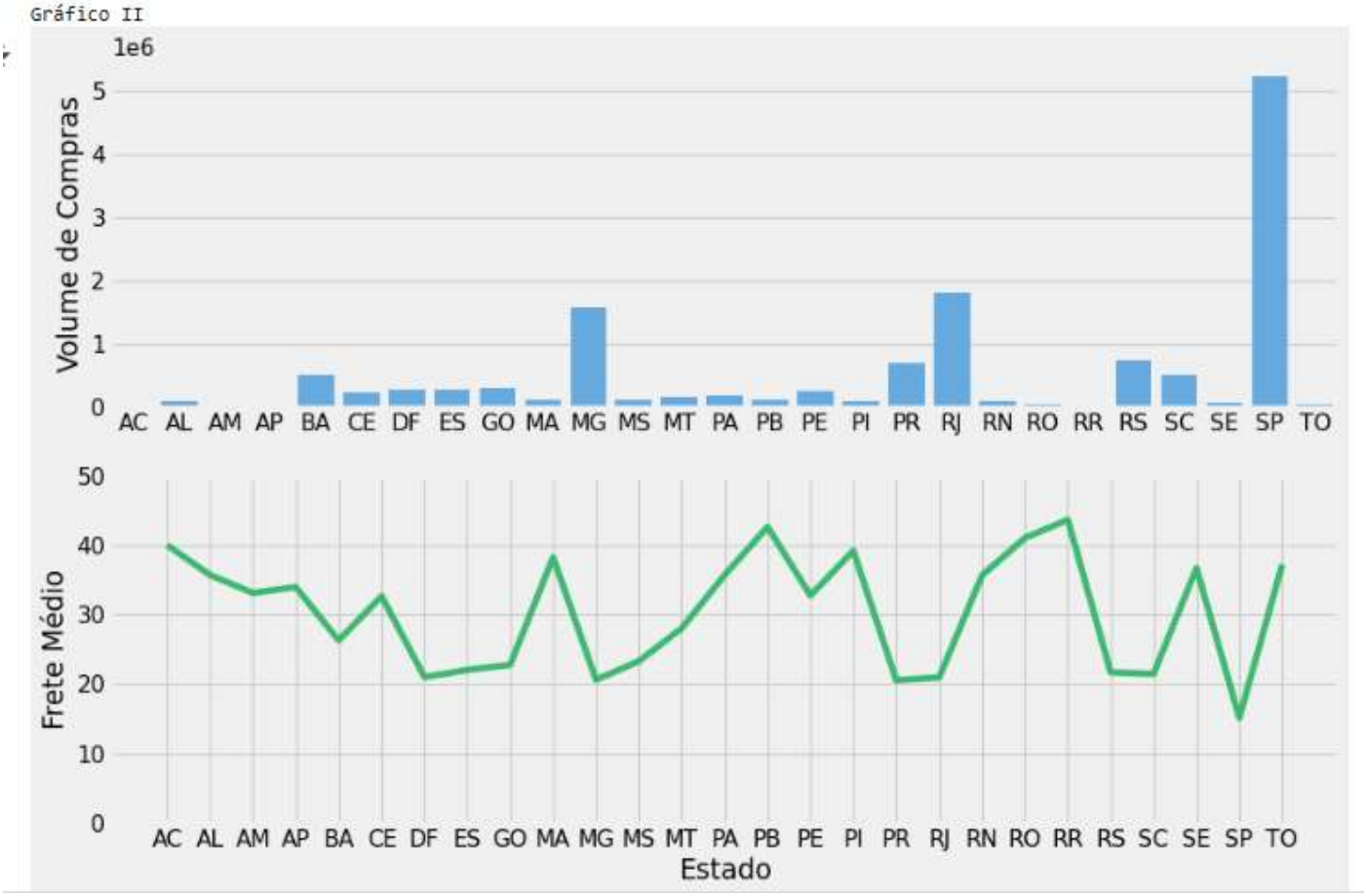

Gráfico III – Volume de vendas por horas.

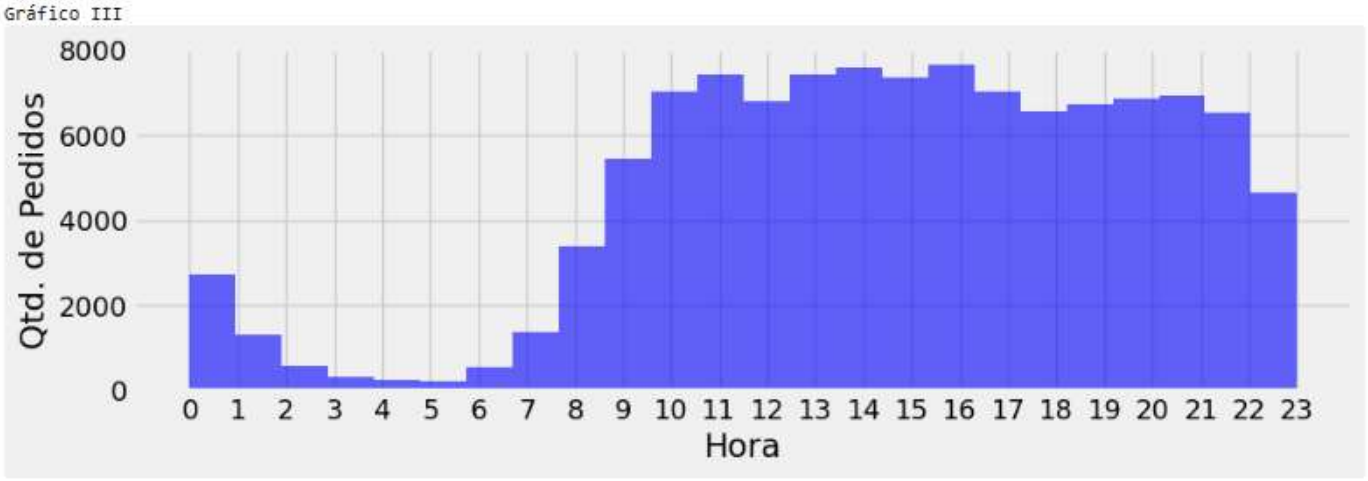

Imagem II - Análise de distância entre consumidores.

```
Cluster -1: Ciano
Cluster  0: Rosa
Cluster  1: Vermelho
Cluster  2: Azul
Cluster  3: Laranja
Cluster  4: Marrom
```

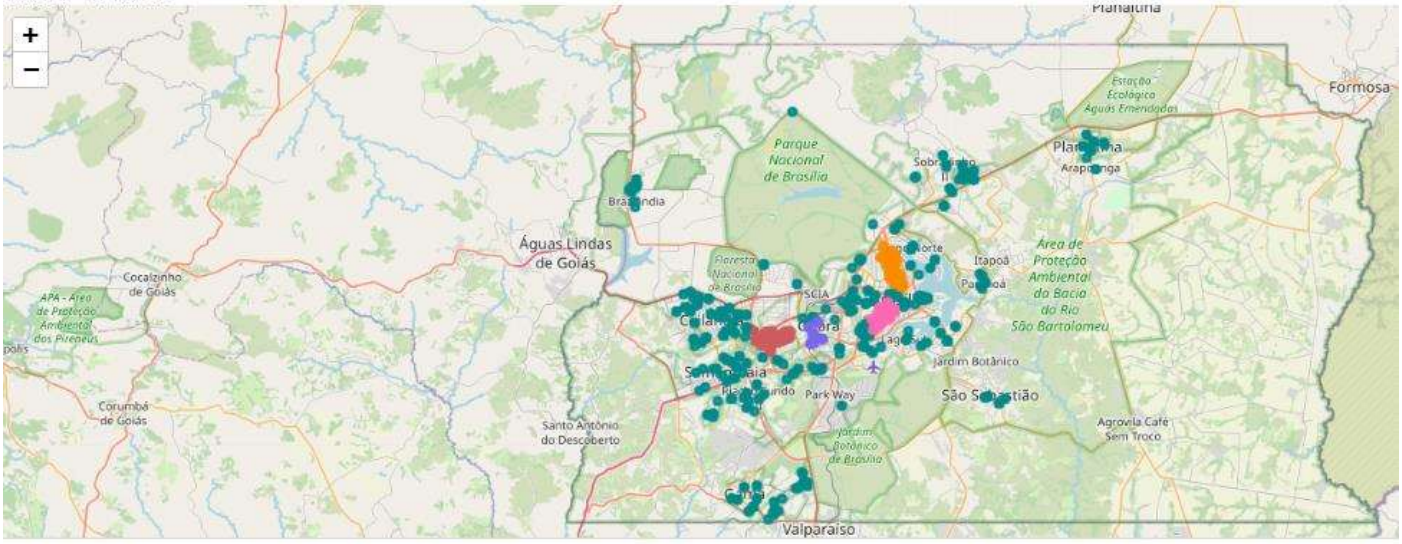

Imagem III - demonstração de Excel (disponível na integra link Github).

D1 | Categoria do Recomendado

	A	B	C	D	E	F
1	**Top 10 Produto**	**Categoria do Produto**	**Produto Recomendado**	**Categoria do Recomendado**	**Chance (%)**	
2	aca2eb7d00ea1a7b8eb	moveis_decoracao	30d1842be9c62546d3fb853119602ac2	eletrodomesticos_2	87,5	
3	aca2eb7d00ea1a7b8eb	moveis_decoracao	fe5827d110a9ba59d4ab407e86ccd6eb	construcao_ferramentas_jardim	65,52	
4	aca2eb7d00ea1a7b8eb	moveis_decoracao	ee8b16196604ed47b34b094351ce68e2	bebes	73,45	
5	aca2eb7d00ea1a7b8eb	moveis_decoracao	cc9c93a7dc6ba4b5913c8f3fba62612e	esporte_lazer	72,78	
6	aca2eb7d00ea1a7b8eb	moveis_decoracao	43423cdffde7fda63d0414ed38c11a73	relogios_presentes	55,4	
7	99a4788cb24856965c3	cama_mesa_banho	8ac47b3ab13c68f49f10dde899674149	dvds_blu_ray	58,48	
8	99a4788cb24856965c3	cama_mesa_banho	437c05a395e9e47f9762e677a7068ce7	beleza_saude	78,04	
9	99a4788cb24856965c3	cama_mesa_banho	f30de5fdde000c5debf03cc49d782249	casa_conforto	94,07	
10	99a4788cb24856965c3	cama_mesa_banho	7c1e2b3fa0233e46fb3bcdcb9919a72f	papelaria	51,91	
11	99a4788cb24856965c3	cama_mesa_banho	a6ad77b15e566298a4e8ee2011ab1255	moveis_decoracao	70,13	
12	422879e10f46682990d	ferramentas_jardim	b9142260cefbbd5688748061179bb7fe	cama_mesa_banho	88,46	
13	422879e10f46682990d	ferramentas_jardim	167b19e93baccb17916b9a6dd03264e7	eletronicos	57,93	
14	422879e10f46682990d	ferramentas_jardim	6413f7a28e149a324c4a914000399fb2	cool_stuff	62,25	
15	422879e10f46682990d	ferramentas_jardim	17606c7d7254ed1f0351fd48a28be932	cama_mesa_banho	82,92	
16	422879e10f46682990d	ferramentas_jardim	54d9ac713e253fa1fae9c8003b011c2a	cool_stuff	62,06	
17	389d119b48cf3043d31	ferramentas_jardim	b7605b5b483063d12bd90a772bff9d21	cama_mesa_banho	87,84	
18	389d119b48cf3043d31	ferramentas_jardim	362b773250263786dd58670d2df42c3b	esporte_lazer	57,39	
19	389d119b48cf3043d31	ferramentas_jardim	ea546947e2412f88a20ac74103592b42	beleza_saude	67,77	
20	389d119b48cf3043d31	ferramentas_jardim	777d2e438a1b645f3aec9bd57e92672c	cama_mesa_banho	61,59	
21	389d119b48cf3043d31	ferramentas_jardim	9007d9a8a0d332c61d9dd611fa341f4b	papelaria	56,97	
22	368c6c730842d78016ac	ferramentas_jardim	b9142260cefbbd5688748061179bb7fe	cama_mesa_banho	82,49	
23	368c6c730842d78016ac	ferramentas_jardim	167b19e93baccb17916b9a6dd03264e7	eletronicos	81,91	
24	368c6c730842d78016ac	ferramentas_jardim	6413f7a28e149a324c4a914000399fb2	cool_stuff	84,62	
25	368c6c730842d78016ac	ferramentas_jardim	17606c7d7254ed1f0351fd48a28be932	cama_mesa_banho	88,68	

Bibliografia

COVER, T.; HART, P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, v. 13, n. 1, p. 21-27, 1967.

GÉRON, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. 2. ed. Sebastopol: O'Reilly Media, 2019.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2. ed. New York: Springer, 2009.

MCKINNEY, W. Data Structures for Statistical Computing in Python. In: *Proceedings of the 9th Python in Science Conference*. 2010.

RICCI, F.; ROKACH, L.; SHAPIRA, B. *Recommender Systems Handbook*. 2. ed. Springer, 2015.