Recomendação Personalizada de Recursos Educacionais com Aprendizado de Máquina: Comparação Experimental entre Baselines Clássicos e SAKT

Ana Clara Fabres[1], **Carlos Santos**[1], **Cristhiano Vasconcellos**[1]

[1] Instituto Federal de Educação, Ciência e Tecnologia Farroupilha (IFFar)
Alegrete – RS – Brasil

ana.55013@aluno.iffar.edu.br
carlos.santos@iffarroupilha.edu.br
cristhiano.vasconcellos@iffarroupilha.edu.br

Abstract. *This paper presents an experimental study on personalized recommendation of educational resources using ASSISTments interaction data. We compare classical recommenders (Popularity, Item-kNN and TruncatedSVD) with a sequential knowledge tracing model (SAKT) in a temporal next-item prediction protocol. The evaluation uses HitRate@10, NDCG@10, Precision@10 and Recall@10 with 200 sampled negatives per test user. Classical baselines achieved the best ranking results, while SAKT showed overfitting after the first epochs in the 20-epoch run. The paper discusses methodological implications, threats to validity, and practical next steps for educational recommendation.*

Resumo. *Este artigo apresenta um estudo experimental sobre recomendação personalizada de recursos educacionais com dados de interações do ASSISTments. Comparamos recomendadores clássicos (Popularidade, Item-kNN e TruncatedSVD) com um modelo sequencial de knowledge tracing (SAKT) em um protocolo temporal de predição do próximo item. A avaliação utiliza HitRate@10, NDCG@10, Precision@10 e Recall@10 com 200 negativos amostrados por usuário de teste. Os baselines clássicos obtiveram os melhores resultados, enquanto o SAKT apresentou sinais de sobreajuste na execução com 20 épocas. O artigo discute implicações metodológicas, ameaças à validade e próximos passos para recomendação educacional.*

1. Introdução

A personalização em ambientes educacionais digitais depende da capacidade de modelar interações entre estudantes e recursos de aprendizagem. Em plataformas com grande volume de exercícios e trilhas formativas, recomendar o próximo recurso pode reduzir sobrecarga cognitiva, aumentar engajamento e apoiar a progressão em habilidades específicas. No contexto brasileiro, isso é particularmente relevante no ensino médio, etapa marcada por grande escala de atendimento, heterogeneidade de perfis e desigualdades no acesso e uso pedagógico de tecnologias digitais [Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) 2024, Comitê Gestor da Internet no Brasil (CGI.br) 2024].

Este trabalho integra um projeto institucional de pesquisa aplicada voltado ao desenvolvimento de um sistema de recomendação personalizada de recursos educacionais

com aprendizado de máquina. A motivação prática reside na carência de *pipelines* reprodutíveis e de baixo custo computacional que possam ser testados, auditados e adaptados por instituições de ensino, especialmente na educação básica.

Na literatura, coexistem duas linhas principais para esse problema: técnicas clássicas de recomendação, como popularidade, vizinhança e fatoração matricial; e modelos sequenciais de *knowledge tracing* (KT), que modelam a evolução temporal do estudante e têm avançado com arquiteturas baseadas em atenção, grafos e mecanismos temporais, como SAKT, GKT, AKT e HawkesKT [Piech et al. 2015, Pandey and Karypis 2019, Shen et al. 2024, Wang et al. 2021]. Revisões recentes sobre sistemas de recomendação educacional destacam desafios relacionados à explicabilidade, adaptação contextual, escalabilidade, diversidade pedagógica e validação em cenários reais de ensino [Li et al. 2024, Mhagama and Garg 2025, Raza et al. 2026].

Neste trabalho, investigamos a comparação entre essas duas linhas em um cenário de *next-item prediction* com o dataset ASSISTments [ASSISTments Data Mining 2024]. Comparamos Popularidade, Item-kNN, TruncatedSVD e SAKT em um protocolo temporal por usuário, avaliando desempenho com métricas de ranking. Como contribuições, apresentamos: (i) um protocolo offline reprodutível em notebook; (ii) uma comparação objetiva entre baselines clássicos e SAKT; (iii) análise do comportamento de treinamento do SAKT em 20 épocas; e (iv) discussão aplicada sobre adoção incremental de modelos de recomendação em contextos educacionais com restrições práticas.

Diferentemente de avaliações puramente orientadas a benchmark, este estudo busca também produzir evidências úteis para decisões iniciais de implantação em contextos educacionais reais. Assim, o foco não está apenas em identificar o modelo com maior desempenho numérico, mas em compreender a relação entre desempenho, custo computacional, simplicidade de implementação, reprodutibilidade e potencial de adaptação institucional.

2. Trabalhos Relacionados

Recomendação em educação combina elementos de sistemas de recomendação tradicionais com particularidades pedagógicas, como dependência temporal, domínio por habilidade, sequência de aprendizagem e feedback binário de acerto ou erro. Baselines clássicos seguem relevantes por sua robustez, interpretabilidade e baixo custo computacional.

Modelos por popularidade fornecem uma referência simples em cenários nos quais determinados itens são amplamente recorrentes. Métodos Item-kNN exploram coocorrência e similaridade entre itens a partir do histórico de interação dos usuários. Já modelos de fatoração matricial, como os baseados em SVD, capturam fatores latentes entre usuários e itens e frequentemente apresentam bom desempenho em tarefas de ranking com dados esparsos [Ricci et al. 2022, Sarwar et al. 2001, Koren et al. 2009, Halko et al. 2011].

Em *knowledge tracing*, trabalhos como DKT [Piech et al. 2015] e SAKT [Pandey and Karypis 2019] consolidaram o uso de modelos neurais para estimar a probabilidade de acerto do estudante em uma próxima interação. Abordagens posteriores passaram a explorar mecanismos de atenção, grafos de conhecimento, esquecimento, relações entre habilidades e efeitos temporais cruzados [Shen et al. 2024, Wang et al. 2021]. Apesar

desse avanço, o desempenho de modelos de KT depende fortemente do alinhamento entre tarefa, representação dos dados, protocolo de avaliação e estratégia de amostragem. Por isso, comparações com baselines fortes são fundamentais em estudos aplicados, especialmente quando o objetivo é orientar uma implementação institucional inicial.

3. Metodologia

3.1. Dados e pré-processamento

O experimento foi conduzido com o conjunto ASSISTments, amplamente utilizado em mineração de dados educacionais e em estudos de *knowledge tracing* [Heffernan and Heffernan 2014]. Os dados foram carregados a partir de arquivos CSV em um notebook reprodutível, com leitura robusta e auto-detecção das colunas principais. Após padronização, o conjunto passou a utilizar os campos `user_id`, `item_id`, `rating` (acerto binário), `skill_id` e `timestamp`.

Foram removidos registros sem identificador de usuário ou item, e os identificadores foram reindexados para inteiros consecutivos. Após esse processamento, a base resultante ficou composta por 942.816 interações, 1.709 usuários, 3.162 itens e 102 habilidades. A distribuição de interações por usuário apresentou perfil de cauda longa, com mínimo de 2 interações, mediana de 441 e máximo de 3.057, padrão compatível com bases educacionais reais, nas quais há forte heterogeneidade de uso entre estudantes.

3.2. Protocolo experimental

A avaliação foi estruturada como um problema de *next-item prediction*, com divisão temporal por usuário. Para cada estudante, as interações foram ordenadas por tempo e a última interação foi reservada para teste, enquanto todas as anteriores foram utilizadas para treino. Esse procedimento preserva a ordem temporal dos eventos e evita vazamento de informação futura.

Com esse protocolo, o conjunto final foi dividido em 941.107 interações de treino e 1.709 interações de teste, correspondendo a uma interação de teste por usuário. Para o cálculo das métricas de ranking, foi adotado um esquema com um item relevante por usuário e 200 itens negativos amostrados aleatoriamente (`N_NEG`=200). A partir desse conjunto de candidatos, os modelos geram um ranking e o desempenho é medido em Top-K, com $K = 10$.

As métricas reportadas foram *HitRate@10*, *NDCG@10*, *Precision@10* e *Recall@10*. O *HitRate@10* indica a proporção de usuários para os quais o item verdadeiro aparece entre as dez primeiras posições; o *NDCG@10* avalia a qualidade da posição ocupada por esse item; e *Precision@10* e *Recall@10* complementam a análise no Top-10 sob o protocolo adotado.

3.3. Modelos avaliados

Foram comparados quatro modelos com diferentes níveis de complexidade e requisitos computacionais. Como referência inicial, utilizou-se o baseline de Popularidade, que recomenda os itens mais frequentes no conjunto de treino. Embora simples, esse baseline continua sendo amplamente empregado em estudos comparativos por sua robustez e por servir como parâmetro mínimo de desempenho em tarefas de recomendação [Ricci et al. 2022, Li et al. 2024].

O segundo modelo é o Item-kNN, um método de filtragem colaborativa baseada em vizinhança entre itens, no qual os itens candidatos são ranqueados a partir da similaridade com o histórico do usuário. Esse tipo de abordagem permanece relevante pela boa relação entre desempenho, interpretabilidade e custo de implementação, especialmente em cenários com infraestrutura limitada [Sarwar et al. 2001].

O terceiro modelo é o TruncatedSVD, empregado como uma forma de fatoração matricial para aprendizado de fatores latentes de usuários e itens. Em recomendação, modelos de fatores latentes consolidaram-se como uma base importante para capturar padrões de preferência em dados esparsos [Koren et al. 2009]. Do ponto de vista computacional, a decomposição truncada também é suportada por métodos numéricos eficientes para aproximações de baixa-rank [Halko et al. 2011].

O quarto modelo é o SAKT (*Self-Attentive Knowledge Tracing*), uma abordagem sequencial baseada em atenção proposta para modelar dependências temporais em interações educacionais e selecionar, via autoatenção, eventos passados mais relevantes para a predição [Pandey and Karypis 2019]. No notebook utilizado neste trabalho, o SAKT foi configurado com embeddings de item e resposta, atenção multi-cabeças e camada de saída para pontuação dos candidatos. Os principais hiperparâmetros adotados foram `MAX_SEQ_LEN`=100, `BATCH_SIZE`=256, taxa de aprendizado `LR`=0,001, `EPOCHS`=20, dimensão latente 64, 4 cabeças de atenção e *dropout* de 0,1. A execução foi realizada em GPU, com `DEVICE=cuda`.

4. Resultados e Discussão

A Tabela 1 resume os resultados. O TruncatedSVD foi o melhor modelo em todas as métricas, seguido de Item-kNN. O SAKT apresentou desempenho inferior inclusive ao baseline de Popularidade.

Tabela 1. Resultados comparativos (Top-10) com 200 negativos por usuário.

Modelo	HitRate@10	NDCG@10	Precision@10	Recall@10
Popularidade	0,119953	0,060410	0,011995	0,119953
Item-kNN	0,316559	0,169322	0,031656	0,316559
TruncatedSVD	**0,370392**	**0,197892**	**0,037039**	**0,370392**
SAKT	0,032768	0,013497	0,003277	0,032768

O TruncatedSVD recuperou o item correto no Top-10 para cerca de 37,0% dos usuários, enquanto o SAKT atingiu aproximadamente 3,3%. Isso corresponde a uma redução relativa de 91,2% no HitRate do SAKT em relação ao TruncatedSVD. Em comparação ao baseline de Popularidade, o SAKT também ficou abaixo, com redução relativa de 72,7%.

Os resultados indicam que os métodos clássicos, especialmente TruncatedSVD e Item-kNN, foram mais adequados ao protocolo adotado. Uma possível explicação é que, nesse protocolo, cada usuário possui apenas um item relevante no teste e o conjunto de candidatos é formado por amostragem negativa. Como a mediana de interações por usuário é de 441, métodos baseados em padrões globais de coocorrência e fatores latentes conseguem explorar informação histórica consolidada.

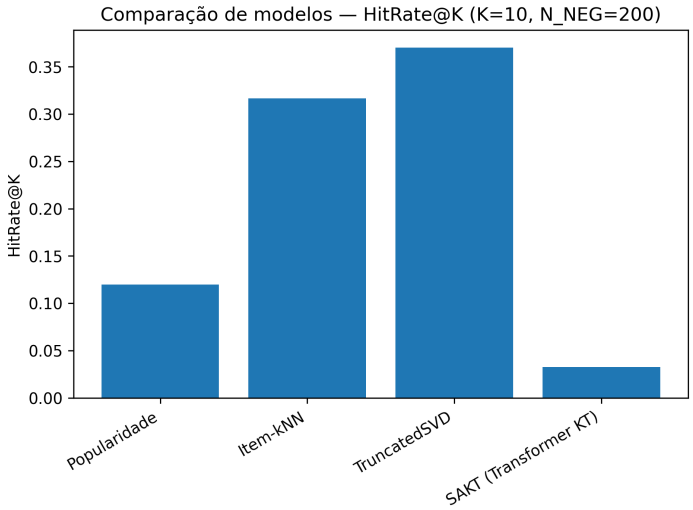

Figura 1. HitRate@10 dos modelos avaliados.

Uma leitura complementar é que o protocolo adotado favorece a reconstrução de padrões históricos consolidados. Como cada usuário contribui com apenas uma interação de teste, métodos capazes de capturar regularidades globais de coocorrência tendem a apresentar vantagem. Isso não significa que modelos sequenciais sejam inadequados para educação, mas que sua avaliação deve estar fortemente alinhada ao objetivo pedagógico e à forma de geração dos candidatos.

Já o SAKT foi originalmente proposto para estimar desempenho em *knowledge tracing*, isto é, a probabilidade de acerto em uma próxima interação. Sua adaptação para ranqueamento entre um item positivo e 200 negativos pode exigir calibração mais cuidadosa, especialmente na forma de pontuar candidatos, selecionar negativos e ajustar hiperparâmetros. Assim, a amostragem negativa, a ausência de validação temporal para ajuste fino e o uso de uma configuração inicial do modelo podem ter contribuído para o desempenho inferior do modelo de atenção.

A Figura 2 apresenta as curvas de perda de treino e teste do SAKT. A perda de treino caiu de 0,6073 para 0,4833, indicando aprendizado no conjunto de treino. Já a perda de teste caiu inicialmente, atingindo seu melhor valor na época 6, e depois aumentou, terminando em 0,7185.

Esse comportamento caracteriza sobreajuste, em que o modelo continua reduzindo o erro de treino, mas perde generalização após poucas épocas. Na prática, isso ajuda a explicar por que aumentar o treinamento para 20 épocas não trouxe ganho nas métricas de ranking. O resultado não invalida modelos sequenciais em recomendação educacional, mas indica que sua adoção requer validação explícita, *early stopping*, ajuste de hiperparâmetros e possivelmente um protocolo mais alinhado à predição de conhecimento por habilidade.

Do ponto de vista de implantação, os achados favorecem uma estratégia incremental. Em uma fase inicial, modelos clássicos como TruncatedSVD e Item-kNN combinam melhor desempenho, menor custo computacional, maior simplicidade operacional e maior facilidade de auditoria. Por outro lado, modelos sequenciais continuam relevantes para pesquisas futuras, especialmente quando o objetivo é modelar evolução de conhecimento,

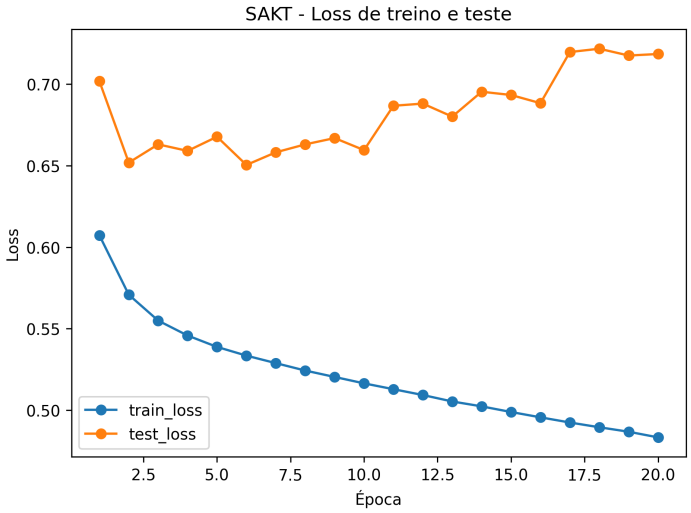

Figura 2. Curvas de perda do SAKT: evidência de sobreajuste após melhora inicial.

dependência temporal entre habilidades e probabilidade de acerto em interações específicas. Além das métricas de ranking, uma implantação educacional deve considerar indicadores pedagógicos, como diversidade de habilidades recomendadas, cobertura de conteúdos, coerência curricular, engajamento e impacto na aprendizagem.

Do ponto de vista institucional, os resultados reforçam a importância de uma estratégia incremental de adoção tecnológica. Em vez de iniciar diretamente por modelos neurais mais complexos, uma instituição pode começar com abordagens clássicas, avaliar sua utilidade pedagógica com professores e estudantes, ajustar critérios curriculares e somente depois incorporar modelos sequenciais mais sofisticados. Essa trajetória reduz riscos de implantação, facilita auditoria e permite que decisões técnicas sejam discutidas com equipes pedagógicas.

5. Ameaças à Validade e Trabalhos Futuros

Este estudo possui limitações. Primeiro, os experimentos foram conduzidos em um único conjunto de dados, o que limita a generalização para ambientes com diferentes padrões de densidade, sequência e granularidade de habilidades. Segundo, a avaliação foi exclusivamente offline e baseada em métricas de ranking, o que não permite concluir, por si só, que as recomendações produziriam benefícios pedagógicos em sala de aula ou em uma plataforma real. Além disso, métricas offline não capturam aspectos como aceitabilidade pelos professores, clareza das recomendações, adequação ao planejamento didático ou impacto real no percurso de aprendizagem dos estudantes.

Terceiro, o desempenho do SAKT pode ter sido afetado pela adaptação do modelo para ranqueamento com amostragem negativa, pela ausência de validação temporal para seleção de hiperparâmetros e pela falta de *early stopping*. Quarto, não foram avaliadas variantes recentes de KT, como modelos baseados em grafos, atenção contextual, mecanismos de esquecimento ou efeitos temporais cruzados. Essas comparações não foram realizadas nesta etapa por exigirem novos experimentos, maior tempo de execução e revisão adicional do protocolo experimental.

Como trabalhos futuros, pretende-se: (i) executar múltiplas sementes e validação temporal; (ii) testar diferentes estratégias de amostragem negativa; (iii) comparar variantes recentes de *knowledge tracing*; (iv) avaliar outros conjuntos de dados educacionais; e (v) incorporar métricas pedagógicas, como diversidade de habilidades, cobertura de conteúdos, coerência curricular, engajamento e impacto em aprendizagem. Também se pretende avançar para uma validação aplicada, com professores e estudantes, de modo a verificar se as recomendações são compreensíveis, úteis e pedagogicamente adequadas.

6. Conclusão

Este artigo apresentou um estudo experimental de recomendação personalizada de recursos educacionais comparando baselines clássicos e SAKT em um protocolo temporal de predição do próximo item. O TruncatedSVD apresentou o melhor desempenho, seguido de Item-kNN, enquanto o SAKT teve desempenho inferior no cenário analisado.

Os resultados indicam que modelos clássicos podem ser uma escolha inicial robusta para implantação institucional, especialmente quando há restrições de infraestrutura, necessidade de reprodutibilidade e demanda por soluções de menor complexidade operacional. Ao mesmo tempo, a análise do SAKT mostra que modelos sequenciais exigem calibração cuidadosa para tarefas de ranqueamento e não devem ser adotados apenas por sua maior sofisticação arquitetural.

Como contribuição aplicada, o estudo oferece um pipeline simples, reproduzível e auditável em notebook para apoiar decisões iniciais de implantação de recomendação educacional em instituições, priorizando evidência empírica antes de aumentar a complexidade do modelo. Os próximos passos incluem reavaliar modelos sequenciais com validação temporal, parada antecipada, múltiplas bases de dados e métricas mais diretamente associadas à utilidade pedagógica.

Referências

ASSISTments Data Mining (2024). Assistments dataset. `https://sites.google.com/view/assistmentsdatamining/dataset`. Acesso em: 21 fev. 2026.

Comitê Gestor da Internet no Brasil (CGI.br) (2024). Sete em cada dez alunos do ensino médio usam ia generativa em pesquisas escolares, revela tic educação. `https://bit.ly/4kNPl9H`. Acesso em: 21 fev. 2026.

Halko, N., Martinsson, P.-G., and Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288.

Heffernan, N. T. and Heffernan, C. L. (2014). Assistments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, 24(4):470–497.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) (2024). Mec e inep contextualizam resultados do censo escolar 2024. `https://www.gov.br/inep/pt-br/centrais-de-conteudo/noticias/censo-escolar/mec-e-inep-contextualizam-resultados-do-censo-escolar-2024`. Acesso em: 21 fev. 2026.

Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Li, Y. et al. (2024). Recent developments in recommender systems: A survey. *IEEE Computational Intelligence Magazine*.

Mhagama, J. T. and Garg, K. (2025). A systematic review of educational recommender systems: Techniques, target users, and emerging trends in personalized learning. *International Journal of Technology in Education Science*, 2(1):79–98.

Pandey, S. and Karypis, G. (2019). A self-attentive model for knowledge tracing. In *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)*, pages 384–389.

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L., and Sohl-Dickstein, J. (2015). Deep knowledge tracing. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28.

Raza, S. et al. (2026). A comprehensive review of recommender systems. *Journal of Network and Computer Applications*.

Ricci, F., Rokach, L., and Shapira, B. (2022). *Recommender Systems Handbook*. Springer, 3 edition.

Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web (WWW '01)*, pages 285–295. ACM.

Shen, S., Liu, Q., et al. (2024). A survey of knowledge tracing: Models, variants, and applications. *IEEE Transactions on Learning Technologies*.

Wang, C., Ma, W., Zhang, M., Chen, C., Liu, Y., and Ma, S. (2021). Temporal cross-effects in knowledge tracing. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. ACM.