

Predição Explicável de Evasão no Ensino Superior baseada em Trajetórias Temporais e Aprendizado de Máquina

Alberto Moura¹, Abílio Nogueira Barros¹, Alexandre Magno Andrade Maciel¹

¹ Universidade de Pernambuco (UPE)

jaomf@ecomp.poli.br, abilionbarros@gmail.com,
alexandre.maciel@upe.br

Abstract. Dropout in higher education compromises institutional efficiency. This paper compares machine learning models such as LSTMs and Random Forests and temporal representations for early prediction of student dropout, using historical academic records. To prevent data leakage, the student's final semester is excluded. Models are trained on sequential or static aggregated data, integrating SHAP for interpretability. For a three-semester window of $N = 3$, the optimized LSTM achieves an F1-Score of 0.785 and 84% accuracy, outperforming the static baseline and matching the sequential Random Forest, which obtained an F1 of 0.784. Additionally, an absences-only model presents 85% precision, with an F1 of 0.67, offering a low-cost warning alternative.

Resumo. A evasão no ensino superior compromete a eficiência institucional. Este trabalho compara modelos de aprendizado de máquina como LSTMs e Random Forests e representações temporais para predição antecipada do abandono acadêmico, usando históricos escolares. Para evitar vazamento de dados, o último semestre do aluno é excluído. Os modelos são avaliados sob representações sequenciais ou agregadas, integrando a técnica SHAP para explicabilidade. Para a janela selecionada de $N = 3$ semestres, o modelo LSTM otimizado apresenta F1-Score de 0,785 e acurácia de 84%, superando o baseline estático condensado e equivalendo-se à RF sequencial, com F1 de 0,784. Ademais, o modelo baseado apenas em faltas apresenta 85% de precisão, viabilizando alertas de baixo custo.

1. Introdução

A evasão no ensino superior compromete as dimensões social, econômica e pedagógica das instituições [Lobo 2012, Tinto 1993]. Definida como a interrupção prematura do ciclo acadêmico, resulta na perda de investimentos e frustra projetos de mobilidade social [Braga et al. 2003]. Taxas persistentes nas carreiras de STEM [MEC/INEP 2024] apontam a relevância de ferramentas de monitoramento precoce para subsidiar políticas de retenção [Silva Filho et al. 2007].

Tradicionalmente, a predição de risco foca em dados demográficos ou socioeconômicos de ingresso [Rovai 2003]. Contudo, por estarem além do controle direto das instituições, direcionar a análise para dados acadêmicos internos apresenta-se como alternativa viável. Teorias clássicas sustentam que a evasão resulta de um processo gradual de desengajamento ao longo do tempo [Tinto 1993]. Assim, a aplicação de MDE (Mineração de Dados Educacionais) ao histórico sequencial discente mostra-se promissora para identificar padrões de risco internamente [Romero and Ventura 2020].

Arquiteturas de redes neurais recorrentes, como LSTM, têm sido investigadas para extrair dependências temporais em trajetórias discentes [Song et al. 2021, Hernandez-Blanco et al. 2019]. Embora eficazes, sua complexidade acarreta menor interpretabilidade das decisões internas [Conati et al. 2018]. Para atenuar essa limitação, técnicas de XAI (Inteligência Artificial Explicável), como SHAP (*Shapley Additive Explanations*), são integradas a sistemas de apoio à decisão pedagógica, fornecendo explicações locais e globais mais transparentes [Lundberg and Lee 2017, Swamy et al. 2022, Barria-Pineda et al. 2021].

Este trabalho compara modelos LSTM e Random Forest sob diferentes representações e volumes de atributos (cenários completo e simplificado) em janelas de N semestres, integrando o SHAP para análise de explicabilidade. Adota-se uma premissa de não-vazamento de dados: o período letivo mais recente do discente é descartado. Isso reduz o viés de retrospectiva no desligamento (comum por excesso de reprovações e faltas), gerando um modelo de alerta precoce baseado no histórico escolar acumulado.

Para guiar o desenvolvimento e a avaliação desta proposta, formulam-se as seguintes QP (Questões de Pesquisa):

- **QP1:** Qual o impacto de utilizar a trajetória temporal do estudante em comparação com resumos estáticos agregados para a previsão de risco de evasão?
- **QP2:** Em janelas temporais curtas, como $N = 3$, modelos de florestas aleatórias com linearização sequencial são competitivos ou equivalentes a redes neurais recorrentes profundas LSTM?
- **QP3:** É viável prever o risco de evasão estudantil com desempenho aceitável utilizando exclusivamente dados de assiduidade, compreendendo faltas e reprovações por falta?
- **QP4:** De que forma a interpretabilidade global e local fornecida pelo SHAP auxilia na identificação dos fatores de risco temporais e na formulação de intervenções pedagógicas direcionadas?

2. Trabalhos Relacionados

A pesquisa em Mineração de Dados Educacionais tem evoluído para modelagens temporais de trajetórias semestrais [Romero and Ventura 2020, Aldowah et al. 2019], visto que o rendimento acadêmico acumulado e a assiduidade correlacionam-se diretamente com a permanência estudantil [Tinto 1993, Gašević et al. 2016].

2.1. Modelagem Baseada em Dados do Histórico Escolar Acadêmico

Enquanto sistemas de alerta precoce recorrem frequentemente a dados demográficos ou de AVA em tempo real [Okubo et al. 2017], para cursos presenciais o histórico acadêmico (notas, reprovações e faltas) constitui um registro representativo da trajetória longitudinal discente. Focar no histórico escolar oferece a vantagem metodológica de utilizar dados já disponíveis nos sistemas acadêmicos internos da instituição, facilitando intervenções direcionadas e viáveis, independentemente de fatores demográficos externos.

No cenário nacional, classificadores tradicionais e seleção de atributos têm sido aplicados a históricos escolares, indicando que o rendimento consolidado supera dados de ingresso na predição de evasão [Teodoro and Kappel 2020, Menolli et al. 2025, Chagas et al. 2025]. Apesar do valor dessas abordagens, mapeamentos apontam para

a necessidade de maior exploração de trajetórias temporais [Andrade and Pereira 2024]. Adicionalmente, a maioria das pesquisas correlatas trata os períodos letivos como registros independentes, desconsiderando a dependência temporal ao longo dos semestres.

2.2. Redes Neurais Recorrentes e LSTM na Educação

Para processar trajetórias acadêmicas longitudinais, redes LSTM têm sido empregadas devido à capacidade de reter dependências temporais [Hochreiter and Schmidhuber 1997, Hernandez-Blanco et al. 2019]. Tais modelagens indicam que a evolução sequencial do rendimento antecipa padrões de risco [Hu et al. 2022, Song et al. 2021]. Contudo, a baixa interpretabilidade direta dessas arquiteturas [Conati et al. 2018] motiva a integração de técnicas de XAI, como SHAP e LIME, permitindo quantificar o impacto de atributos escolares e subsidiar intervenções pedagógicas de forma transparente [Lundberg and Lee 2017, Ribeiro et al. 2016, Barredo Arrieta et al. 2021, Swamy et al. 2022, Barria-Pineda et al. 2021]. A interpretabilidade via SHAP também tem sido explorada para analisar a relevância de atributos em indicadores educacionais nacionais [Barros et al. 2024].

No entanto, pesquisas correlatas comumente não tratam o viés comportamental do último período letivo do aluno evasor (que costuma acumular reprovações e faltas após a decisão de desligamento). Incluir esse período gera vazamento de dados, resultando em alertas tardios. Este trabalho propõe prever o risco utilizando o histórico acadêmico sequencial descartando o último semestre ativo, avaliando a relação complexidade-desempenho sob explicações SHAP.

3. Metodologia

Os dados consistem em históricos acadêmicos semestrais (1986-2025) dos cursos de engenharia de uma universidade pública. Foram selecionados discentes com no mínimo dois semestres de registro, totalizando 16.800 estudantes (Tabela 1). Destes, apenas as classes FORMADO e EVASOR foram utilizadas para a modelagem preditiva binária. A rotulação dos estudantes se deu usando o seguinte fluxo, com base nas informações disponibilizadas na base de dados pelo sistema e com auxílio da coordenação:

1. Se o estudante possui alguma situação/status como FORMADO, ele é rotulado como **FORMADO**.
2. Senão, se ele possui a situação/status mais recente como MATRICULADO ou análogos, ele é um estudante **ATIVO**, estudantes que ainda estão cursando a graduação e serão os alvos dos modelos.
3. Senão, se ele possuir a situação/status mais recente como EXCLUÍDO ou DESLIGADO da instituição, ele é rotulado como um estudante **EVASOR**.
4. Senão, ele é rotulado como **OUTROS**.

Para capturar a trajetória do discente, os históricos foram modelados em representações sequenciais e acumuladas.

Após a filtragem para a janela de $N = 3$ semestres letivos (exigindo mínimo de $N + 1 = 4$ períodos devido ao descarte do último semestre), o dataset alinhado final resultou em 8.557 estudantes (5.152 FORMADO e 3.405 EVASOR). Alunos evadidos precocemente (com menos de 4 semestres) são naturalmente excluídos da análise para esta janela.

Tabela 1. Distribuição das situações acadêmicas finais dos discentes com no mínimo dois semestres registrados

Situação Acadêmica	Quantidade de Discentes	Percentual - %
ATIVO	6.299	37,49
FORMADO	5.161	30,72
EVASOR	4.412	26,26
OUTROS	928	5,52
Total	16.800	100,00

3.1. Cuidados Éticos

Os dados utilizados neste estudo são de natureza sensível, tendo sido disponibilizados pela coordenação da instituição sob a assinatura de um NDA (*Non-Disclosure Agreement*). Em conformidade com as resoluções do CNS nº 466/2012 e nº 510/2016, todos os identificadores pessoais diretos, como nomes, registros de identificação civil ou de matrícula, foram desconsiderados e excluídos das etapas de processamento e experimentação. Análises e treinamento dos modelos basearam-se de forma exclusiva em atributos de desempenho acadêmico pareados por meio de identificadores numéricos sintéticos, sem associação com os estudantes.

3.2. Arquitetura da Rede Neural LSTM e Engenharia de Features

Para alimentar a LSTM, os dados tabulares foram reestruturados em tensores tridimensionais no formato (Amostras, Timesteps, Características), possibilitando analisar janelas temporais deslizantes N de 1 a 5 semestres.

Para mitigar o vazamento de dados, o semestre letivo mais recente de cada discente foi descartado. As predições referem-se ao período imediatamente posterior à janela analisada. Sob esta premissa, avaliaram-se duas configurações de atributos:

1. **Modelo Completo**, contendo todas as variáveis: A LSTM é alimentada com as 6 variáveis semestrais originais: SEM_QTD_DISCIPLINAS, SEM_TOT_REPROVACOES_NOTA, SEM_TOT_REPROVACOES_FALTA, SEM_TOT_FALTAS, SEM_MEDIA e SEM_TAXA_APROVACAO.
2. **Modelo Simplificado**, baseado apenas em faltas: Restringe as entradas da LSTM exclusivamente às variáveis associadas à frequência do discente, representadas por SEM_TOT_REPROVACOES_FALTA e SEM_TOT_FALTAS.

As definições, os tipos de dados e o significado pedagógico de cada uma das características utilizadas para compor o histórico acadêmico sequencial do discente são detalhados na Tabela 2.

A LSTM implementada baseia-se em [Hochreiter and Schmidhuber 1997], contendo uma camada recorrente, uma de *Dropout* para regularização [Srivastava et al. 2014] e uma camada densa com ativação *Sigmoid* para classificação. A otimização dos hiperparâmetros (unidades LSTM entre 16 e 64, taxa de dropout de 0.1 a 0.5, e taxa de aprendizado do otimizador *Adam* [Kingma and Ba 2014]) foi conduzida com o *KerasTuner* por busca aleatória. Adotou-se validação cruzada com 5 dobras (*Stratified K-Fold*) [Kohavi 1995].

Tabela 2. Dicionário de dados das variáveis sequenciais do histórico acadêmico

Variável	Tipo	Descrição Pedagógica
SEM_QTD_DISCIPLINAS	Inteiro	Quantidade de disciplinas matriculadas no semestre.
SEM_TOT_REPROVACOES_NOTA	Inteiro	Total de disciplinas com reprovação por nota.
SEM_TOT_REPROVACOES_FALTA	Inteiro	Total de disciplinas com reprovação por falta.
SEM_TOT_FALTAS	Inteiro	Número absoluto de faltas acumuladas no semestre.
SEM_MEDIA	Real	Média das notas finais obtidas no período, em uma escala de 0 a 10.
SEM_TAXA_APROVACAO	Real	Proporção de disciplinas concluídas com aprovação, variando de 0,0 a 1,0.

Por proporcionalidade ao custo computacional, a LSTM teve seus hiperparâmetros definidos uma única vez sobre uma partição de *holdout* de tuning (80%/20%, semente fixa), isolada das 30 repetições de avaliação. As Random Forests, por serem mais leves, usaram Validação Cruzada Aninhada: hiperparâmetros reotimizados via *GridSearchCV* (3 dobras internas) a cada um dos 30 folds.

3.3. Camada de Explicabilidade via SHAP

Para maior transparência, a metodologia integra o SHAP [Lundberg and Lee 2017], método baseado na teoria dos jogos cooperativos que calcula os Valores de Shapley para atribuir a cada atributo sua respectiva contribuição na predição do modelo. Essa abordagem elucida o impacto das variáveis através das janelas temporais N .

3.4. Modelagem de Retrato Estático Acumulado

Como baseline, implementou-se o **Retrato Estático** para avaliar se a representação temporal ($N = 3$) traz vantagens preditivas frente a uma visão condensada da mesma janela de 3 semestres.

As variáveis do discente em T_0 , T_1 e T_2 foram agregadas por:

- **Média aritmética:** rendimento e matrícula (SEM_QTD_DISCIPLINAS, SEM_MEDIA, SEM_TAXA_APROVACAO).
- **Soma absoluta:** faltas e reprovações (SEM_TOT_FALTAS, SEM_TOT_REPROVACOES_FALTA, SEM_TOT_REPROVACOES_NOTA).

Essa base alimentou dois classificadores Random Forest: o *Modelo Completo Estático*, 6 variáveis e o *Modelo Simplificado Estático*, apenas faltas e reprovações por falta.

4. Resultados e Discussão

A validação da capacidade preditiva da arquitetura LSTM foi conduzida variando a janela temporal N de 1 a 5 semestres. O objetivo desta abordagem foi avaliar o compromisso, conhecido como *trade-off*, entre a quantidade de contexto histórico exigido pelo modelo e a precisão obtida na identificação da classe de interesse EVASOR.

A análise dos relatórios de classificação do Modelo Completo, apresentada na Tabela 3, indica que a acurácia global do modelo tende a crescer conforme o contexto temporal é ampliado, estabilizando-se em 84% para $N = 3$ e atingindo 85% em $N = 5$. O modelo configurado com $N = 3$ semestres consecutivos de histórico prévio apresentou um equilíbrio de desempenho favorável, alcançando acurácia de 84%, precisão de 87%, recall de 71% e F1-Score de 0,78 para a classe de evasores. Esta precisão de 87% pode auxiliar na mitigação de potenciais alarmes falsos, cooperando com a eficiência de programas de intervenção discente.

Tabela 3. Métricas de desempenho do Modelo Completo com todas as variáveis

Janela N	Acurácia	Precisão - Evasor	Recall - Evasor	F1-Score - Evasor	Suporte Total
$N = 1$	0.78 ± 0.01	0.84 ± 0.01	0.65 ± 0.01	0.73 ± 0.01	9573
$N = 2$	0.82 ± 0.01	0.85 ± 0.01	0.70 ± 0.01	0.77 ± 0.01	8975
$N=3$	0.84 ± 0.01	0.87 ± 0.02	0.71 ± 0.01	0.78 ± 0.01	8557
$N = 4$	0.84 ± 0.01	0.82 ± 0.03	0.72 ± 0.02	0.77 ± 0.01	8175
$N = 5$	0.85 ± 0.00	0.82 ± 0.01	0.72 ± 0.01	0.77 ± 0.01	7870

Cabe ressaltar que o suporte total decresce com o aumento de N , uma vez que estudantes com ciclos mais curtos não possuem histórico suficiente para estarem inclusos nas janelas mais longas. A Figura 1 apresenta a Matriz de Confusão do Modelo Completo para $N = 3$, acumulada sobre as partições de teste dos cinco folds da validação cruzada.

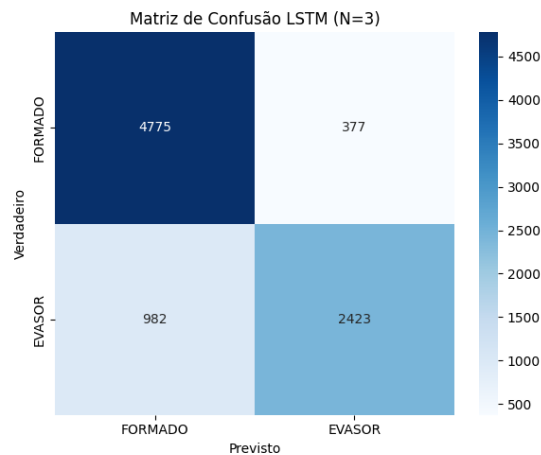


Figura 1. Matriz de Confusão para o Modelo Completo com Janela Temporal $N = 3$

A Figura 2 exibe o SHAP summary plot para $N = 3$, no qual as variáveis do semestre mais recente, T_2 , dominam a predição de risco, com $T_2_SEM_TOT_REPROVACOES_FALTA$ e $T_2_SEM_TAXA_APROVACAO$ nas duas primeiras posições, seguidas pelas médias dos semestres mais antigos da janela, $T_0_SEM_MEDIA$ e $T_1_SEM_MEDIA$. Valores baixos destas médias, em pontos azuis, associam-se a SHAP positivo, sugerindo que um rendimento insatisfatório no início do curso já indica risco de evasão, enquanto valores altos, em pontos vermelhos, atuam como fator de proteção, evidenciando que a trajetória inicial do discente também influencia a decisão do modelo.

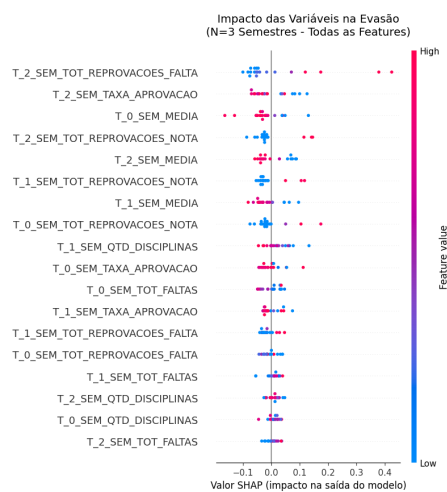


Figura 2. Análise de Impacto Global no Modelo Completo via SHAP Summary Plot para $N = 3$

A título de ilustração prática da explicabilidade local, a Figura 3 apresenta o caso de um discente ATIVO no 8º semestre com probabilidade estimada de evasão de 98,8%, cujo principal fator de risco é a reprovação por falta no semestre mais recente disponível, consistente com o padrão de impacto global discutido anteriormente, enquanto a Figura 4 apresenta um discente no mesmo semestre com probabilidade estimada de 18,5%, cujo perfil de baixo risco é sustentado por ausência de reprovações, taxa de aprovação estável e desempenho acadêmico crescente. Esses exemplos ilustram como os pesos Shapley individuais permitem diagnósticos diferenciados por estudante, para além da leitura agregada do modelo.

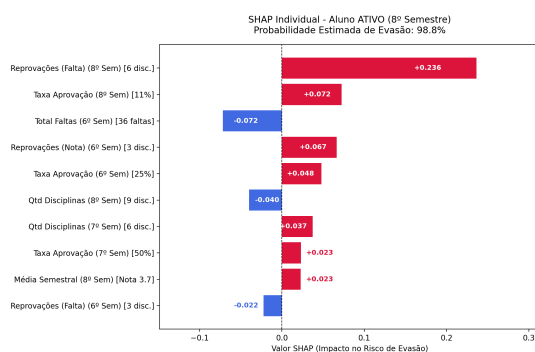


Figura 3. Explicação SHAP individual: discente de alto risco

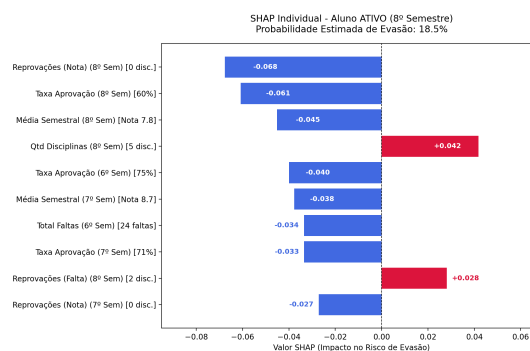


Figura 4. Explicação SHAP individual: discente de baixo risco

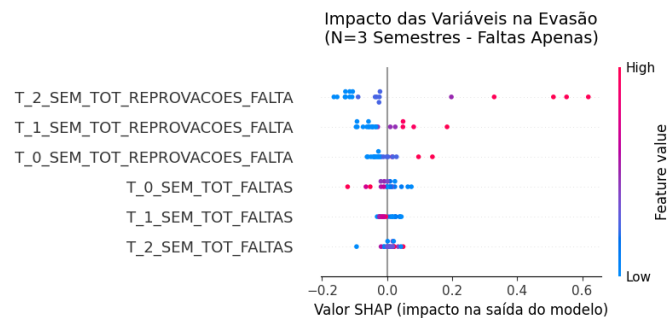
4.1. Modelo Simplificado baseado em Faltas

Para testar a viabilidade de um modelo simplificado e de menor custo operacional, o qual dispensa a coleta de notas semestrais em tempo real, avaliou-se o desempenho da LSTM utilizando apenas as faltas e reprovações por falta, representadas por SEM_TOT_REPROVACOES_FALTA e SEM_TOT_FALTAS, dos semestres anteriores. A Tabela 4 apresenta o desempenho deste Modelo Simplificado.

Tabela 4. Métricas de desempenho do Modelo Simplificado baseado em faltas

Janela N	Acurácia	Precisão - Evasor	Recall - Evasor	F1-Score - Evasor	Suporte Total
$N = 1$	0.72 ± 0.00	0.85 ± 0.01	0.48 ± 0.01	0.61 ± 0.01	9573
$N = 2$	0.76 ± 0.01	0.82 ± 0.02	0.55 ± 0.02	0.66 ± 0.02	8975
$N=3$	0.78 ± 0.01	0.85 ± 0.02	0.56 ± 0.02	0.67 ± 0.02	8557
$N = 4$	0.79 ± 0.00	0.80 ± 0.02	0.59 ± 0.01	0.68 ± 0.00	8175
$N = 5$	0.81 ± 0.01	0.84 ± 0.02	0.57 ± 0.03	0.68 ± 0.02	7870

A exclusão das notas reduziu o F1-Score da classe evasor para 0,67 em $N = 3$, uma diminuição de 11 pontos percentuais em relação ao Modelo Completo, e o *Recall* recuou para 0,56, o que significa que aproximadamente 44% dos estudantes que efetivamente evadem não são sinalizados por este modelo, indicando que a ausência de notas reduz a sensibilidade do modelo em detectar trajetórias de risco sutis. Ainda assim, a precisão de 85% obtida com apenas duas variáveis de frequência, SEM_TOT_REPROVACOES_FALTA e SEM_TOT_FALTAS, indica viabilidade para sistemas de alerta precoce de baixo custo. A análise do SHAP na Figura 5 mostra que SEM_TOT_REPROVACOES_FALTA em T_2 é o principal vetor de risco, seguido pelas faltas no mesmo período, o que sugere que a frequência pode ser monitorada em tempo real para identificar perfis de risco sem a necessidade de aguardar o fechamento das notas.

**Figura 5. Impacto Global das Características via SHAP no Modelo Simplificado com Janela $N = 3$**

4.2. Comparação de Modelos: Random Forest vs. LSTM

Para avaliar se o uso de redes neurais recorrentes LSTM se justifica frente a classificadores tradicionais que recebem a informação temporal linearizada de forma sequencial plana ou condensada, realizou-se uma comparação de desempenho sistemática.

No cenário do **Modelo Completo**, a LSTM Completo obteve F1-Score de 0,785, com acurácia de 84%, precisão de 87% e recall de 71%, e a RF Sequencial obteve F1-Score de 0,784, com acurácia de 84%, precisão de 86% e recall de 72%, apresentando desempenho praticamente equivalente entre si. A LSTM, com capacidade reduzida a 32 unidades e dropout de 0,2, mostrou-se capaz de capturar dependências temporais mesmo sob janelas curtas de tamanho $N = 3$, enquanto a RF Sequencial também se mostrou competitiva, tratando as posições temporais como variáveis tabulares independentes, com menor custo computacional.

No cenário do **Modelo Simplificado**, a RF Seq Faltas obteve F1-Score de 0,68, com precisão de 84% e recall de 57%, superando ligeiramente a LSTM correspondente, com F1-Score de 0,67, precisão de 85% e recall de 56%, e o baseline RF Estático, com F1-Score de 0,65, precisão de 81% e recall de 55%. Dado o custo computacional e a complexidade de parametrização da rede recorrente, a RF Sequencial apresenta-se como alternativa viável e operacionalmente mais simples para alertas de evasão.

Como comparação adicional a um classificador linear tradicional, avaliou-se a Regressão Logística sobre o cenário Completo, com hiperparâmetro de regularização $C = 0,1$ definido sob o mesmo protocolo de tuning aplicado à LSTM. A RL obteve F1-Score médio de $75,11\% \pm 1,13\%$, contra $78,56\% \pm 0,82\%$ da LSTM Completa, diferença confirmada como estatisticamente significativa pelo teste pareado de Wilcoxon, com $p = 1,86 \times 10^{-9}$. Essa comparação foi mantida isolada do teste multimodelo da Seção 4.3, uma vez que a correção estatística do teste de Nemenyi escala com o número de classificadores comparados simultaneamente, e a inclusão de mais dois modelos ampliaria a Diferença Crítica exigida para significância entre todos os pares.

4.3. Comparação Estatística Multimodelo via Teste de Friedman e Nemenyi

Para analisar a diferença de desempenho entre as abordagens e avaliar o valor da informação sequencial em relação à acumulada, realizou-se um teste comparativo multimodelo envolvendo seis cenários sob a mesma população de 8.557 discentes:

1. **LSTM Completo** para $N = 3$: Modelo sequencial baseado nas 6 características dos últimos 3 semestres.
2. **RF Sequencial Completo** para $N = 3$: Classificador Random Forest treinado com as características dos últimos 3 semestres dispostas de forma linear plana.
3. **RF Estático Completo** para $N = 3$: Classificador Random Forest treinado com o resumo estático condensado por meio da média e soma das características do mesmo intervalo de 3 semestres.
4. **LSTM Simplificado** para $N = 3$: Modelo sequencial simplificado, contendo apenas reprovações por falta e faltas, dos últimos 3 semestres.
5. **RF Sequencial Simplificado** para $N = 3$: Classificador Random Forest treinado com as características dos últimos 3 semestres, contendo apenas faltas e reprovações por falta, dispostas de forma linear plana.
6. **RF Estático Simplificado** para $N = 3$: Classificador Random Forest treinado com o resumo estático acumulado por meio da soma de reprovações por falta e faltas do mesmo intervalo de 3 semestres.

Seguindo recomendações metodológicas clássicas [Demšar 2006], coletaram-se 30 amostras pareadas de F1-Score para cada classificador por meio de validação cruzada estratificada repetida 6 vezes com 5 dobras, utilizando sementes de inicialização controladas para buscar a reprodutibilidade dos pesos e das partições.

O teste global de Friedman resultou em uma estatística de 140,36, com p -value = $1,50 \times 10^{-28}$, inferior ao nível de significância de 0,05, rejeitando-se a hipótese nula de equivalência de desempenho. Os rankings médios obtidos, nos quais valores menores indicam melhor posicionamento médio no teste, foram: LSTM Completo com rank de 1,37, RF Sequencial Completo com rank de 1,67, RF Estático Completo com rank de 2,97, RF Sequencial Simplificado com rank de 4,40, LSTM Simplificado com rank de

4,63 e RF Estático Simplificado com rank de 5,97. O teste pós-hoc de Nemenyi, com nível de confiança de 95% e $\alpha = 0,05$, resultou em uma Diferença Crítica, referida como CD, de 1,377, cujo diagrama correspondente é apresentado na Figura 6.

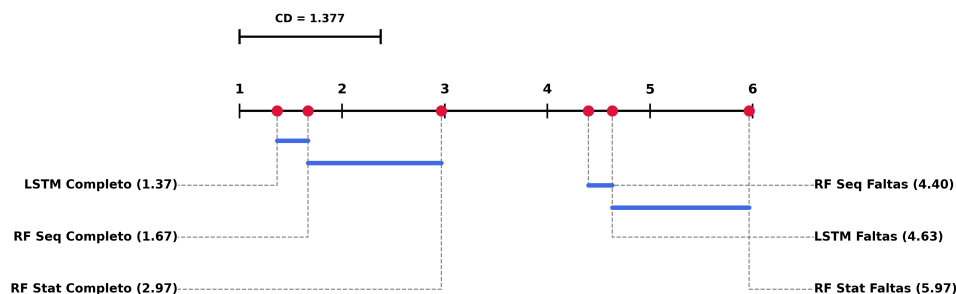


Figura 6. Diagrama de Diferença Crítica de Nemenyi comparando os Rankings Médios dos 6 Modelos sob a janela de $N = 3$

A análise das comparações pareadas traz elementos para discussão:

- O modelo **LSTM Completo** obteve ranking médio de 1,37, apresentando diferença estatisticamente significativa favorável em relação a seu homólogo condensado tradicional, representado pela RF Estático Completo com ranking de 2,97, dado que a diferença de rankings (1,60) supera a Diferença Crítica de 1,377. Esse resultado traz indícios favoráveis de que o uso de redes neurais recorrentes, sob parametrização adequada, pode apresentar vantagens de desempenho frente à condensação estática agregada tradicional das características estudantis.
- O modelo **LSTM Completo** é estatisticamente equivalente ao **RF Sequencial Completo**, que possui ranking de 1,67, visto que a diferença de ranking entre eles, igual a 0,30, é muito inferior à Diferença Crítica de 1,377.
- O modelo **RF Sequencial Completo** obteve ranking médio de 1,67, mas sua diferença em relação ao **RF Estático Completo**, de 1,30, ficou ligeiramente abaixo da Diferença Crítica de 1,377, tornando-os estatisticamente equivalentes sob este teste pós-hoc, embora a RF Sequencial apresente F1-Score médio ligeiramente maior.
- No cenário de atributos simplificados de frequência, ou seja, as faltas, a RF Sequencial Simplificada, com ranking de 4,40, apresentou diferença estatisticamente significativa favorável em relação à RF Estática Simplificada, que possui ranking de 5,97, uma vez que a diferença, de 1,57, supera a Diferença Crítica de 1,377.
- Não foi detectada diferença estatisticamente significativa entre a LSTM Simplificada, de ranking 4,63, e a RF Estática Simplificada, de ranking 5,97, cuja diferença, de 1,34, foi inferior à Diferença Crítica, sugerindo que a modelagem recorrente profunda com restrição de atributos a dados de faltas apresenta-se similar às florestas ordenadas.

Esse conjunto de resultados apoia a relevância de modelar as trajetórias sequenciais discentes. Ao contrário da comparação estática agregada tradicional, ao parear os modelos em uma janela de $N = 3$ semestres, os dados sugerem que a preservação da evolução

temporal das características acadêmicas está associada a ganhos de desempenho preditivo.

4.4. Análise Estatística do Impacto das Janelas Temporais

Para analisar o impacto do tamanho da janela temporal, com $N \in \{1, 2, 3, 4, 5\}$, no desempenho preditivo das LSTMs, aplicou-se o teste não-paramétrico de Friedman, que é a metodologia padrão recomendada para comparar classificadores de aprendizado de máquina sobre partições repetidas de dados [Demšar 2006]. O teste foi executado sobre os F1-Scores obtidos nas 5 dobras da validação cruzada.

Para o Modelo Completo, o teste de Friedman resultou em uma estatística de 11,04 com $p\text{-value} = 0,0261$. Para o Modelo Simplificado, baseado apenas em faltas, obteve-se a estatística de 13,60 com $p\text{-value} = 0,0087$. Como ambos os $p\text{-values}$ foram inferiores ao nível de significância padrão, com $p < 0,05$, rejeita-se a hipótese nula de que o tamanho da janela não afeta as previsões.

Os resultados sugerem que janelas curtas, com $N = 1$, tendem a limitar a sensibilidade da rede neural, ao passo que expandir o histórico além de três períodos ($N > 3$) associa-se a uma estabilização ou leve declínio na acurácia e no F1-Score. Esse comportamento indica que a janela de três semestres ($N = 3$) pode representar um equilíbrio adequado entre o desempenho do modelo e a quantidade de discentes de suportados.

4.5. Custo Computacional e Complexidade dos Modelos

A Tabela 5 apresenta o custo computacional dos classificadores avaliados, medido sobre as mesmas 30 partições de validação cruzada.

Tabela 5. Custo computacional e tamanho dos modelos (janela $N = 3$)

Modelo	Tempo Treino (s)	Tempo Predição		F1-Score
		p/ Partição (ms)	Tamanho	
LSTM Completo	12,98 ± 0,97	204,58 ± 19,92	84,7 KB	0,7856 ± 0,0082
RF Seq Completo	3,04 ± 0,51	28,63 ± 7,67	10.800,9 KB	0,7841 ± 0,0096
RF Estático Completo	2,12 ± 0,20	21,55 ± 7,23	22.822,0 KB	0,7641 ± 0,0109
RL Completo	0,009 ± 0,002	0,17 ± 0,04	1,0 KB	0,7511 ± 0,0113
LSTM Simplificado	7,46 ± 1,10	230,69 ± 55,68	227,5 KB	0,6723 ± 0,0142
RF Seq Simplificado	1,83 ± 0,17	23,37 ± 8,74	4.332,6 KB	0,6751 ± 0,0154
RF Estático Simplificado	1,42 ± 0,11	19,15 ± 6,47	10.511,1 KB	0,6537 ± 0,0123
RL Simplificado	0,005 ± 0,001	0,12 ± 0,03	0,9 KB	0,6751 ± 0,0136

Observa-se uma inversão ao argumento de simplicidade computacional da Random Forest: seu treinamento é mais rápido, mas seu tamanho em disco chega a ser 270 vezes maior que o da LSTM, por armazenar a estrutura completa de diversas árvores. Adicionalmente, a LSTM Completa resultou em um tamanho menor do que a Simplificada, apesar de mais variáveis de entrada, já que o *KerasTuner* selecionou 64 unidades recorrentes para o cenário Simplificado contra 32 do Completo, possivelmente compensando a menor quantidade de informação disponível. A Regressão Logística apresenta o menor custo em todas as métricas, por ordens de magnitude.

4.6. Ameaças à Validade

Seguindo a taxonomia de [Wohlin et al. 2012], discutem-se as ameaças à validade do estudo em quatro dimensões: interna, de construto, externa e de conclusão.

Validade Interna. O histórico acadêmico abrange quase quatro décadas (1986–2025), período no qual a instituição pode ter passado por reformas curriculares e alterações nos critérios administrativos de desligamento discente, constituindo possíveis variáveis de confusão não controladas.

Validade de Construto. A variável `SEM_TOT_FALTAS` agrega faltas justificadas e ausências associadas a desengajamento em um único valor numérico, sem distinção no sistema institucional. Esse construto medido não corresponde de forma unívoca ao construto teórico de interesse (desengajamento acadêmico), o que pode ter restringido a granularidade dos padrões comportamentais capturados pelos modelos.

Validade Externa. Os dados são oriundos de uma única instituição pública, restritos aos cursos de Engenharia, com taxa de evasão de 26,26% (Tabela 1). Áreas de conhecimento e instituições distintas podem apresentar dinâmicas de evasão diferentes, não representadas nesta amostra.

Validade de Conclusão. A Tabela 5 reporta o desvio padrão sobre 30 medições, 6 sementes $\times$ 5 dobras, obtidas na validação cruzada robusta para $N = 3$, mitigando a dependência de um único particionamento. Para as Tabelas 3 e 4, a variabilidade é reportada sobre as 5 dobras de uma única execução da validação cruzada devido ao custo computacional de replicar as 30 repetições para todas as janelas N . Contudo, a comparação estatística multimodelo da Seção 4.3 permanece restrita à janela $N = 3$, e o ranking entre os seis classificadores não foi replicado para as demais janelas exploradas.

5. Conclusões

A exclusão do último período letivo mitiga o viés de vazamento de dados, permitindo que as predições ocorram com antecedência útil à intervenção. Os testes estatísticos confirmam as vantagens da representação sequencial frente ao retrato estático condensado, uma vez que a LSTM Completa obteve o melhor posicionamento médio no teste de Nemenyi, superando os resumos estáticos tradicionais e equivalendo-se às florestas aleatórias sequenciais linearizadas. A frequência escolar, expressa por faltas e reprovações por falta, apresenta-se por sua vez como um indicador de fácil coleta, uma vez que o modelo simplificado manteve precisão de 85% na janela de $N = 3$, viabilizando alertas em tempo real sem depender da consolidação tardia de notas. Por fim, a análise via SHAP confere transparência às previsões e subsidia a compreensão de fatores de risco individuais, sustentando a viabilidade de sistemas de alerta de baixo custo para a retenção discente.

5.1. Trabalhos Futuros

Como perspectivas de trabalhos futuros, pretende-se validar os modelos propostos em uma base de dados mais ampla e diversificada, contemplando cursos de outras áreas, com diferentes padrões, além das engenharias. Essa validação externa é necessária para avaliar a capacidade de generalização das redes neurais LSTM e mitigar potenciais vieses comportamentais e acadêmicos específicos das áreas de Ciências Exatas e Tecnológicas.

Declaração de Uso de Inteligência Artificial Generativa

Durante a preparação deste trabalho, foi utilizado o agente do Gemini na IDE Antigravity no desenvolvimento, refinamento do código-fonte dos experimentos e formatação do manuscrito. Todo o conteúdo científico, análises estatísticas, modelagem experimental e redação final foram revisados pelos autores, que assumem total responsabilidade pelo conteúdo final e integridade intelectual deste trabalho.

Referências

- Aldowah, H., Al-Samarraie, H., and Faupuin, W. (2019). Educational data mining and learning analytics in higher education: A systematic review. *IEEE Access*, 7:43721–43737.
- Andrade, F. and Pereira, R. (2024). Investigação da evasão estudantil por meio de mineração de dados e aprendizado de máquina: um mapeamento sistemático. *Revista Brasileira de Informática na Educação*, 32:1–25.
- Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2021). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.
- Barria-Pineda, J. et al. (2021). Explainable student models: What, how, and why? In *International Conference on Artificial Intelligence in Education*, pages 312–323. Springer.
- Barros, A. N., Alves, G., and Mello, R. F. (2024). Entendendo os fatores de previsão do tdi nas escolas públicas brasileiras: Uma abordagem usando a técnica shap. In *Anais do XXXV Simpósio Brasileiro de Informática na Educação*, pages 1529–1542. SBC.
- Braga, M. M., Peixoto, M. C. L., and Bogutchi, T. F. (2003). A evasão no ensino superior brasileiro: novos dados, velhos problemas. *Avaliação: Revista da Avaliação da Educação Superior*, 8(1):161–179.
- Chagas, F. R., Gomide, A., and Santana, A. (2025). Predição de risco de evasão universitária em cursos de engenharia utilizando mineração de dados. *Revista Brasileira de Informática na Educação*, 33:1–22.
- Conati, C. et al. (2018). Add explainable ai to intelligent tutoring systems. In *International Conference on Intelligent Tutoring Systems*, pages 5–15. Springer.
- Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.
- Gašević, D. et al. (2016). Learning analytics should not fit all: Measuring the influence of context on predictor variables in learning analytics models. *International Review of Research in Open and Distributed Learning*, 17(1):67–84.
- Hernandez-Blanco, A. et al. (2019). Predicting student dropout in higher education using deep learning and time series. *IEEE Latin America Transactions*, 17(10):1588–1597.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

- Hu, Q. et al. (2022). Improving student dropout prediction with lstm and attention mechanism. *Interactive Learning Environments*, pages 1–15.
- Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 1137–1143.
- Lobo, M. B. (2012). Evasão no ensino superior brasileiro: uma análise sob a ótica dos dados do censo da educação superior. *Cadernos de Pesquisa*, 42(145):232–247.
- Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30:4765–4774.
- MEC/INEP (2024). Censo da educação superior 2024: Notas estatísticas. Ministério da Educação/Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Disponível em: <https://www.gov.br/inep>. Acesso em: 12 mar. 2026.
- Menolli, A., Alves, G., and Brandão, W. (2025). Processo baseado em seleção de características para criação e avaliação de modelos de aprendizado de máquina aplicados à predição de evasão no ensino superior. *Revista Brasileira de Informática na Educação*, 33:1–24.
- Okubo, F. et al. (2017). Predicting graduate-level students’ performance using deep learning. In *Proceedings of the 7th International Learning Analytics and Knowledge Conference*, pages 15–24.
- Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). ”why should i trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144.
- Romero, C. and Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):e1355.
- Rovai, A. P. (2003). In search of higher education college student persistence models. *Internet and Higher Education*, 6(1):1–16.
- Silva Filho, R. L. L. et al. (2007). A evasão no ensino superior brasileiro: novos e velhos temas. *Scientia Paedagogica*, 13(25):203–220.
- Song, X. et al. (2021). Time-series performance prediction based on long short-term memory (lstm) neural network model. *Proceedings of the IEEE*, 109(3):247–278.
- Srivastava, N. et al. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.
- Swamy, V. et al. (2022). Evaluating explainable ai for student performance prediction. In *Proceedings of the 15th International Conference on Educational Data Mining*.
- Teodoro, L. H. and Kappel, M. A. A. (2020). Aplicação de técnicas de aprendizado de máquina para predição de evasão escolar no ensino superior público brasileiro. *Revista Brasileira de Informática na Educação*, 28:1003–1025.

Tinto, V. (1993). *Leaving college: Rethinking the causes and cures of student attrition*. University of Chicago Press.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (2012). *Experimentation in Software Engineering*. Springer Science & Business Media, Berlin, Heidelberg.