Fatores Preditivos na Avaliação da Pós-Graduação em Computação: Uma Análise com Mineração de Dados Educacionais

Bernardo J. da S. Vieira, Samuel V. de O. Galdino, Sebastião E. A. Filho

Programa de Pós-Graduação em Ciência da Computação – Universidade do Estado do Rio Grande do Norte (UERN) e Universidade Federal Rural do Semi-Árido (UFERSA), Mossoró/RN, Brasil.

{bernardo.vieira,samuel.galdino}@alunos.ufersa.edu.br,
sebastiaoalves@uern.br

Abstract. *This study analyzes the factors that influence the evaluation scores of Computer Science Graduate Programs assessed by CAPES. Using data extracted from the quadrennial evaluation process, synthesized by the Apoema-PG system, and Educational Data Mining techniques with the Extra Trees Regressor algorithm, the most relevant variables for score prediction were identified. Results indicate that the volume of completed doctoral theses, publication output in high-ranked journals, and student participation in conference publications are the main determinants of program performance, offering insights to support better academic management.*

Resumo. *Este trabalho analisa os fatores que influenciam a nota dos Programas de Pós-Graduação em Ciência da Computação avaliados pela CAPES. Utilizando dados extraídos da coleta de informações da avaliação quadrienal, sintetizados pelo sistema Apoema-PG, e técnicas de Mineração de Dados Educacionais com o algoritmo Extra Trees Regressor, identificam-se as variáveis mais relevantes na predição das notas. Os resultados indicam que o volume de teses concluídas, a produção em periódicos de alto estrato e a participação discente em conferências são os principais determinantes do desempenho dos programas, oferecendo subsídios para melhor gestão acadêmica.*

1. Introdução

O Sistema Nacional de Pós-Graduação (SNPG) brasileiro possui um modelo de avaliação consolidado, gerido pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), que classifica os programas com notas de 1 a 7. Essa avaliação quadrienal é crítica, pois determina desde a alocação de recursos até a sobrevivência acadêmica dos cursos. Mais do que um instrumento regulatório, ela exerce influência direta sobre a qualidade da formação de pesquisadores e sobre o desenvolvimento científico da área no país, aspectos centrais dos Grandes Desafios da Educação em Computação identificados pela Sociedade Brasileira de Computação [SBC 2024]. Alcançar notas de excelência ou consolidar programas em níveis intermediários exige que coordenações identifiquem, dentre dezenas de métricas, quais fatores exercem maior influência na composição da nota final.

Na Ciência da Computação, esse desafio é acentuado pela relevância das conferências científicas, que muitas vezes superam os periódicos em impacto e disseminação de conhecimento. Como destacado por Laender *et al.* (2008), métricas genéricas de impacto são insuficientes para a área, exigindo análises que considerem a

dinâmica específica de produção da Computação. Essa particularidade reforça a necessidade de abordagens orientadas a dados capazes de capturar os padrões de produção próprios da área e de relacioná-los ao seu desempenho institucional.

Nesse contexto, a Mineração de Dados Educacionais (EDM) surge como uma ferramenta para extrair conhecimento de grandes volumes de dados acadêmicos. Segundo Romero e Ventura (2020), o uso de EDM permite superar análises puramente manuais, identificando padrões de sucesso por meio de algoritmos de aprendizado de máquina. Recentemente, ferramentas como o *AnyLattes* [Cirilo *et al.* 2025] focaram na automação da extração de dados do Lattes para auxiliar na gestão desses indicadores. Entretanto, ainda há uma lacuna no uso dessas bases para fins preditivos e de ranqueamento de importância de variáveis aplicados especificamente à área de Computação. Diante disso, este trabalho formula a seguinte pergunta de pesquisa: quais fatores quantitativos são os maiores determinantes da nota CAPES em programas de pós-graduação em Ciência da Computação?

Este trabalho busca responder a essa questão por meio de uma análise diagnóstica dos fatores que influenciam a nota dos programas. O objetivo é identificar e rankear as variáveis de maior impacto na determinação da nota, utilizando dados do sistema Apoema-PG[1] e o algoritmo *Extra Trees Regressor*. Os resultados visam oferecer subsídios baseados em evidências para que coordenadores e gestores acadêmicos possam priorizar ações pedagógicas e institucionais que efetivamente contribuam para a melhoria do desempenho de seus programas, com especial atenção ao papel da formação discente na produção científica.

O restante deste artigo está organizado da seguinte forma: a Seção 2 apresenta o referencial teórico; a Seção 3 descreve a metodologia adotada; a Seção 4 discute os resultados; e a Seção 5 apresenta as conclusões e direções futuras.

2. Referencial Teórico

Esta seção apresenta os fundamentos que norteiam a avaliação da pós-graduação no Brasil, as especificidades da área de Ciência da Computação, bem como os conceitos de Mineração de Dados Educacionais e Aprendizado de Máquina (*Ensemble Learning* e Árvores Aleatórias) utilizados neste estudo.

2.1. O sistema de avaliação da Pós-Graduação

A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) realiza, desde 1977, a avaliação dos programas de pós-graduação no Brasil, estabelecendo um ranking de qualidade que subsidia a distribuição de recursos governamentais para pesquisa e educação. O processo de avaliação baseia-se fortemente na revisão por pares e utiliza métricas quantitativas e qualitativas para atribuir notas que variam de 1 a 7 aos programas, sendo 3 a nota mínima para a permanência no SNPG com cursos de mestrado e 4 para cursos de doutorado.

Para a área de Ciência da Computação, a avaliação possui desafios e características únicas. Diferentemente de outras áreas, como Biologia ou Física, onde os

[1] https://sistemas.prpg.usp.br/apoema-pg/

periódicos são o principal veículo de comunicação científica, na Ciência da Computação, as conferências com revisão por pares possuem um peso semelhante na avaliação, pois são fundamentais para o desenvolvimento da área. Laender *et al.* (2008) destacam que métricas importadas de outras áreas, como o fator de impacto do ISI JCR, são insuficientes para a computação, pois cobrem apenas uma fração da produção relevante e ignoram conferências prestigiadas.

O sistema de classificação Qualis é utilizado para estratificar a qualidade dos veículos de publicação (periódicos e conferências), sendo um dos principais indicadores de eficiência dos programas. O Documento de Área da CAPES [CAPES 2025] formaliza esses critérios, orientando que a produção intelectual não deve ser avaliada apenas pela quantidade, mas pelo impacto e qualidade, buscando evitar que a avaliação seja meramente quantitativa.

Além da produção bibliográfica, o modelo de avaliação considera quesitos estruturais como a Proposta do Programa, Corpo Docente, Corpo Discente e Inserção Social. Programas de excelência (notas 6 e 7) devem demonstrar inserção internacional e capacidade de formação de doutores em nível equivalente aos principais centros mundiais. A análise desses múltiplos indicadores gera um cenário complexo, onde técnicas computacionais podem auxiliar na identificação de padrões de sucesso.

2.2. Mineração de Dados Educacionais e Modelos de Ensemble

A Mineração de Dados Educacionais (*Educational Data Mining* - EDM) é uma área de pesquisa interdisciplinar que se ocupa do desenvolvimento de métodos para explorar tipos únicos de dados oriundos de contextos educacionais. O objetivo da EDM é utilizar abordagens computacionais para analisar esses dados e compreender melhor os estudantes e os ambientes onde aprendem.

Segundo Romero e Ventura (2020), o processo de EDM segue etapas similares ao processo geral de descoberta de conhecimento em bancos de dados (KDD), envolvendo pré-processamento, mineração de dados e pós-processamento. Baker e Yacef (2009) observam que, embora a mineração de relacionamentos tenha sido predominante nos primeiros anos da EDM, houve uma mudança significativa de tendência para métodos de predição (classificação e regressão) nos anos mais recentes. Este trabalho se insere na vertente preditiva, aplicando a EDM para gestão institucional.

2.2.1. Métodos de Ensemble e Randomização

Para a tarefa de predição da nota dos programas, este estudo adota o Extra Trees Regressor, um método de ensemble baseado na combinação de múltiplas árvores de decisão. Modelos de ensemble tendem a apresentar maior robustez do que árvores isoladas, pois agregam diferentes preditores e reduzem a instabilidade associada a pequenas variações nos dados [Dietterich 2000].

O Extra Trees Regressor introduz aleatoriedade tanto na seleção de subconjuntos de atributos quanto nos pontos de corte utilizados na construção das árvores [Geurts et al. 2006]. Essa estratégia reduz a variância do modelo e o custo computacional, preservando capacidade preditiva adequada para bases com múltiplos indicadores acadêmicos.

2.3. Trabalhos Relacionados

A literatura referente à análise de dados da pós-graduação brasileira e à aplicação de EDM pode ser categorizada em três vertentes principais: estudos bibliométricos descritivos, predição de desempenho discente e análise institucional via mineração de dados.

Primeiramente, há uma vasta produção focada na análise bibliométrica. Trabalhos clássicos na Ciência da Computação [Laender *et al.* 2008; Mena-Chalco & Cesar Jr. 2009] utilizam dados da Plataforma Lattes para mapear redes de coautoria e produtividade. Embora fundamentais para compreender a dinâmica de colaboração na área, esses estudos geralmente possuem caráter descritivo e focam na contagem de publicações, sem correlacioná-las com a nota final atribuída pela CAPES.

Em uma segunda vertente, a Mineração de Dados Educacionais tem sido aplicada na predição de desempenho e evasão discente. Trabalhos como os de Manhães et al. (2011) e Eashwar et al. (2017) demonstram o uso de algoritmos para identificar estudantes com risco de evasão, baixa produtividade ou baixo desempenho acadêmico. No entanto, o foco dessas abordagens é o aluno individual, e não o Programa de Pós-Graduação como entidade estratégica.

A terceira vertente, na qual este trabalho se insere, é a Mineração de Dados para Gestão de Programas. O trabalho de Gualhano et al. (2018) é o correlato mais próximo desta proposta. Os autores aplicaram técnicas de mineração de dados, por meio do algoritmo J48, nas fichas de avaliação da Quadrienal 2013-2016 da área Interdisciplinar, buscando diagnosticar desfechos como nota final e evolução do programa. Na área de Computação, Nunes da Silva et al. (2022) analisaram redes de coautoria de programas de pós-graduação e sua relação com as notas CAPES, porém com foco em medidas topológicas e caracterização estrutural das redes, sem modelagem supervisionada da nota.

O presente trabalho avança em relação aos demais ao focar especificamente a predição da nota CAPES em Programas de Pós-Graduação em Ciência da Computação a partir de indicadores institucionais granulares de produção científica e formação discente. Até onde foi possível identificar na literatura, não foram encontrados trabalhos que apliquem regressão supervisionada para predizer a nota CAPES especificamente na área de Computação a partir desse tipo de dado. Diferentemente de abordagens descritivas ou baseadas em árvores de decisão únicas, este estudo utiliza o Extra Trees Regressor, um método de ensemble menos suscetível ao overfitting, oferecendo não apenas uma estimativa de desempenho, mas também um ranqueamento empírico dos fatores de maior peso preditivo na avaliação da área.

3. Metodologia

Do ponto de vista metodológico, esta pesquisa caracteriza-se como aplicada, quantitativa, exploratória, descritiva e documental [Gil 2002]. Essa classificação se justifica pelo uso de dados numéricos secundários extraídos do sistema Apoema-PG, analisados por técnicas de aprendizado de máquina para identificar fatores preditivos ainda pouco estudados na avaliação da pós-graduação em Computação. Operacionalmente, o estudo divide-se em quatro etapas: coleta de dados,

pré-processamento, modelagem preditiva automatizada e análise exploratória de fatores.

3.1. Coleta e descrição de dados

Os dados utilizados neste estudo foram extraídos do sistema Apoema-PG, que integra informações sobre a pós-graduação brasileira e disponibiliza indicadores acadêmicos. O universo da pesquisa compreendeu 85 Programas de Pós-Graduação Stricto Sensu em Ciência da Computação, considerando registros disponíveis no período de 2013 a 2024. As informações foram obtidas a partir de diferentes planilhas, contendo indicadores de produção científica, formação discente e características estruturais dos programas, posteriormente integradas em uma base consolidada.

Para representar cada programa por um único registro, os dados das planilhas foram agregados a uma base principal de identificação dos programas, e, nos casos com múltiplos registros para um mesmo indicador, utilizou-se a média dos valores correspondentes. A base final contém 31 atributos, incluindo indicadores de produção intelectual, formação discente e características do corpo docente, tendo como variável alvo a nota atribuída aos programas na avaliação da pós-graduação. A distribuição das notas compreendeu 26 programas com nota 3, 35 com nota 4, 12 com nota 5, 4 com nota 6 e 8 com nota 7. Eventuais ausências de informação sobre a variável alvo foram verificadas manualmente por consulta à Plataforma Sucupira.

3.2. Pré-processamento e tratamento de dados

Na etapa de preparação dos dados, foi aplicado um *pipeline* de pré-processamento com o objetivo de garantir a consistência e a adequação das variáveis para a modelagem preditiva. Inicialmente, realizou-se a padronização dos tipos de dados, convertendo atributos numéricos originalmente formatados com vírgula decimal para o formato de ponto flutuante, compatível com as bibliotecas utilizadas no ambiente de análise.

Após a padronização dos tipos de dados, foi conduzida uma filtragem qualitativa das instâncias. A base bruta iniciou com 91 programas, porém três programas foram removidos por não possuírem nota numérica atribuída pela CAPES, dois programas foram removidos por estarem em processo de desativação e um programa foi removido por apresentar nota 2, resultando na base final de 85 instâncias utilizadas na modelagem. A base final apresenta desbalanceamento entre as faixas de nota, especialmente nos níveis mais elevados, nos quais há quatro programas com nota 6 e oito programas com nota 7. Essa limitação é discutida na Seção 5.

O *pipeline* de preparação também incluiu a normalização das variáveis numéricas, executada pelo ambiente de configuração da biblioteca *PyCaret*[2]. Esse procedimento foi adotado para reduzir possíveis efeitos de escala entre os atributos, por exemplo, variáveis que representam contagem de publicações e variáveis associadas a tempo de titulação garantindo que diferenças de magnitude entre os indicadores não influenciassem de forma indevida o processo de aprendizado do modelo.

3.3. Seleção e treinamento do modelo

Para a etapa de modelagem preditiva, foi utilizada a biblioteca *PyCaret* em seu módulo

[2] Documentação da biblioteca *PyCaret*: https://pycaret.readthedocs.io/en/latest/

de regressão, que automatiza diferentes etapas do fluxo de aprendizado de máquina, incluindo preparação dos dados, treinamento e comparação de modelos. Essa abordagem permite avaliar múltiplos algoritmos de regressão de forma padronizada, facilitando a seleção do modelo mais adequado ao conjunto de dados analisado.

Embora as notas atribuídas pela CAPES constituam uma escala discreta e ordinal, variando de 3 a 7 na base analisada, este trabalho adota intencionalmente uma formulação de regressão em vez de classificação. Essa escolha se justifica pela maior granularidade informacional da saída contínua, uma vez que a regressão permite estimar não apenas a faixa de desempenho associada a um programa, mas também sua proximidade em relação às faixas adjacentes. Por exemplo, um programa com nota 3 e valor predito de 3,7 pode ser interpretado como mais próximo da transição para a nota 4 do que de uma permanência estável na nota 3. Essa informação é particularmente útil para gestores acadêmicos, pois permite identificar programas em transição e orientar ações de melhoria com maior precisão do que uma predição puramente categórica.

O conjunto de dados foi particionado automaticamente em 59 instâncias para treinamento (≈70%) e 26 para teste (≈30%), com validação cruzada de 10 folds (k=10, *KFold*) aplicada ao conjunto de treinamento. Esse procedimento permite obter estimativas mais robustas do desempenho dos modelos, reduzindo o risco de sobreajuste e aumentando a confiabilidade dos resultados.

Durante a fase de comparação de modelos, diferentes algoritmos de regressão foram avaliados utilizando métricas de desempenho apropriadas para problemas de predição contínua, especialmente o Coeficiente de Determinação (R^2) e o Erro Médio Absoluto (MAE). A partir dessa análise comparativa, o algoritmo *Extra Trees Regressor* apresentou o melhor desempenho geral entre os modelos avaliados, sendo selecionado para a etapa final de análise e interpretação dos resultados.

Ressalta-se que as variáveis preditoras utilizadas, como publicações em conferências com participação discente e teses concluídas, são também critérios explícitos do processo de avaliação da CAPES. Embora esse alinhamento reforce a validade de construto do modelo, ele implica que os coeficientes de importância refletem a estrutura formal da avaliação vigente, e não necessariamente relações causais independentes, distinção discutida na Seção 4.

3.4. Extração de Conhecimento e Análise

Após o treinamento, utilizou-se a funcionalidade de interpretabilidade do modelo para gerar o gráfico de Importância de Atributos, permitindo identificar quais variáveis exerceram maior peso matemático na predição da nota.

Para corroborar os achados do modelo com a realidade dos programas, realizou-se uma análise descritiva complementar. Os programas foram agrupados por faixas de nota (Clusters 3, 4, 5, 6 e 7), calculando-se as médias de produção e formação para cada grupo. Essa abordagem híbrida (Modelo Preditivo + Estatística Descritiva) permitiu traçar os perfis quantitativos dos programas de excelência e daqueles em consolidação.

4. Resultados e discussão

A avaliação dos modelos de regressão foi realizada utilizando um ambiente automatizado de aprendizado de máquina, que comparou diferentes algoritmos com base em métricas de desempenho preditivo. A Tabela 1 apresenta os algoritmos avaliados, ordenados pelo coeficiente de determinação R^2, adotado como métrica principal de seleção por indicar a proporção da variância da variável alvo explicada pelo modelo. As demais colunas apresentam o Erro Médio Absoluto (MAE) e o Erro Quadrático Médio (MSE), métricas nas quais valores menores indicam melhor desempenho preditivo.

Tabela 1. Comparação de desempenho entre algoritmos avaliados

Model	MAE	MSE	R^2
Extra Trees Regressor	0.4459	0.3113	0.4813
Random Forest Regressor	0.4753	0.3596	0.3421
K Neighbors Regressor	0.5073	0.3841	0.3361
AdaBoost Regressor	0.4994	0.4840	0.1132
Decision Tree Regressor	0.6400	0.8067	-0.4215

Em termos práticos, o MAE de 0,4459 obtido pelo *Extra Trees Regressor* indica que, em média, as predições do modelo se desviam menos de meio ponto da nota real dos programas, o que sugere utilidade para estimar o desempenho relativo dos programas dentro da escala analisada. Destaca-se que a árvore de decisão simples Decision Tree obteve R^2 negativo de -0.4215, indicando que o modelo não generaliza bem para além dos dados de treinamento, evidência empírica do problema de overfitting característico de árvores isoladas, que o Extra Trees mitiga por meio de ensemble e injeção de aleatoriedade.

Como apresentado na Figura 1, o modelo selecionado obteve coeficiente de determinação (R^2) de 0,787 no conjunto de teste, indicando que aproximadamente 78,7% da variabilidade das notas dos programas pode ser explicada pelos indicadores quantitativos utilizados no estudo. Esse resultado sugere que os atributos considerados possuem capacidade explicativa relevante em relação ao processo de avaliação dos programas de pós-graduação, embora parte da variabilidade ainda possa estar associada a fatores qualitativos não representados nos dados analisados.

A Figura 2 apresenta o gráfico de importância relativa das variáveis gerado pelo modelo Extra Trees Regressor. Nesse gráfico, cada barra representa uma variável preditora, e seu comprimento indica sua contribuição relativa para a predição da nota; quanto maior a barra, maior a importância atribuída pelo modelo à variável. Os resultados indicam que Teses_concluidas apresentou a maior relevância, seguida de Pub_peridico_A2 e Publicacoes_Conferencias_Part_Dis. Esse achado sugere que a formação discente, especialmente no doutorado, é central para o desempenho institucional, pois a conclusão de teses reflete permanência no programa, envolvimento em pesquisa e contribuição para a produção acadêmica. Embora a relação entre teses concluídas e ingressantes pudesse indicar retenção ou evasão no doutorado, essa análise extrapola o escopo deste trabalho e é indicada como possibilidade para estudos futuros. Variáveis como Duracao_Doutorado_Med e Matricula_Doutorado_Ano_Med reforçam

a importância da consolidação da formação doutoral para níveis mais elevados de avaliação.

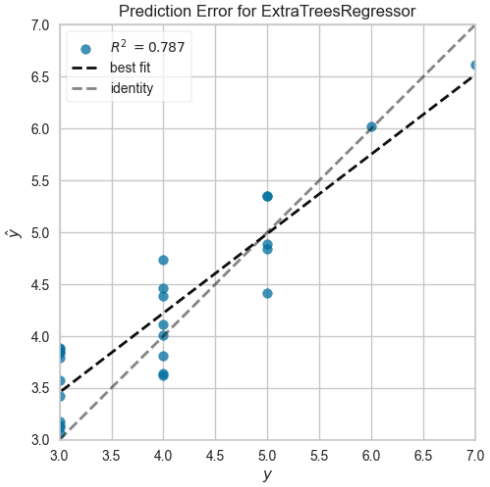

Figura 1. Erro de predição do Extra Trees Regressor

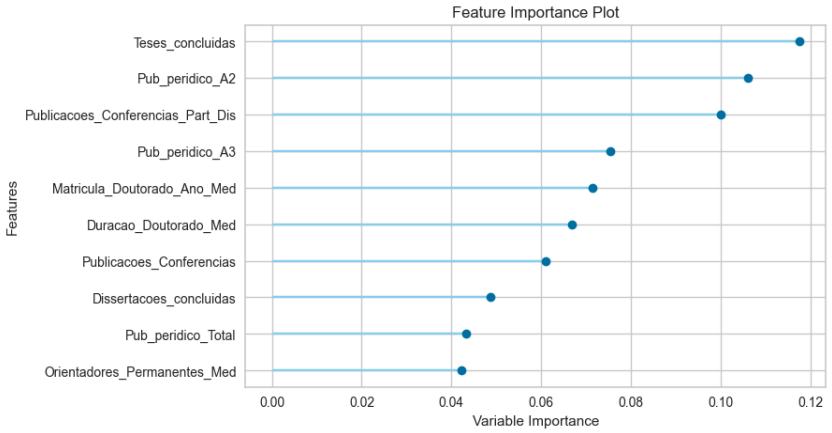

Figura 2. Importância relativa das variáveis segundo modelo Extra Trees Regressor

Esses achados dialogam com Gualhano et al. (2018), que também identificaram Produção Intelectual e Corpo Discente, Teses e Dissertações como dimensões relevantes na avaliação de programas da área Interdisciplinar. A diferença é que, ao utilizar indicadores quantitativos mais granulares, este estudo evidencia fatores específicos da Computação, como teses concluídas, publicações qualificadas e participação discente em conferências. Esta última variável aproxima os resultados da educação em computação, pois indica inserção dos estudantes na comunidade científica, contato com o estado da arte e desenvolvimento de competências de pesquisa em uma área de rápida evolução.

A relevância das variáveis associadas à conclusão de teses, duração do doutorado e produção científica converge com as diretrizes do Documento de Área da CAPES para Ciência da Computação, segundo o qual programas com nota 4 devem possuir doutorado em andamento ou funcionamento, enquanto programas com nota 5 ou superior devem demonstrar consolidação na formação de doutores. Essa aderência reforça a validade de construto do modelo, pois indica que os fatores de maior peso

preditivo correspondem a dimensões efetivamente consideradas no processo avaliativo oficial. No entanto, tal resultado deve ser interpretado com cautela, uma vez que as importâncias estimadas pelo modelo indicam utilidade preditiva dentro do sistema de avaliação vigente, mas não permitem inferir relações causais independentes entre produção científica, formação discente e qualidade do programa, já que esses próprios indicadores compõem a lógica avaliativa da CAPES.

Para complementar a interpretação dos resultados obtidos, realizou-se uma análise descritiva dos programas agrupados por nota. A Tabela 2 apresenta a média dos principais indicadores de produção científica e formação discente para cada faixa de avaliação. Observa-se uma tendência consistente de aumento na produção científica e no número de teses concluídas à medida que a nota dos programas se eleva, evidenciando diferenças estruturais entre programas em consolidação e programas de excelência.

Tabela 2. Médias dos principais indicadores por nota CAPES

Nota CAPES	Pub. Conf. Part. discente	Teses Concluídas	Duração média doutorado	Pub. em conferências
7	1100,3	282	5,04	1652
6	653,8	146,3	4,96	1135
5	473,8	72,5	4,77	849,2
4	204,7	17,6	4,45	541,2
3	65,8	0	0	281,2

Observa-se que programas com nota 7 apresentam, em média, maior volume de publicações em conferências com participação discente e um número significativamente superior de teses concluídas quando comparados aos programas com notas inferiores. Esse padrão reforça os resultados identificados pelo modelo de aprendizado de máquina, indicando que a produção científica envolvendo estudantes e a formação consistente de doutores estão fortemente associadas aos níveis mais elevados de avaliação.

Com o objetivo de analisar o desempenho do modelo por faixa de nota, foi calculado o MAE por nível de avaliação, conforme apresentado na Tabela 3. O erro é menor nas faixas intermediárias, com MAE inferior a 0,21 para as notas 3, 4 e 5, e aumenta nas notas 6 e 7, o que está associado ao número reduzido de instâncias nesses estratos. Esse padrão reforça a limitação discutida na Seção 5 quanto à generalização dos resultados para programas de excelência.

Tabela 3. Erro médio absoluto do modelo por faixa de nota CAPES

Nota CAPES	n	MAE
7	8	0,34
6	4	0,30
5	12	0,19
4	35	0,20
3	26	0,20

5. Conclusão

Este trabalho apresentou um estudo sobre os fatores preditivos do desempenho dos Programas de Pós-Graduação em Ciência da Computação na avaliação quadrienal da CAPES, com base em dados do sistema Apoema-PG referentes ao período de 2013 a 2024. Utilizando técnicas de Mineração de Dados Educacionais com o algoritmo *Extra Trees Regressor*, foram identificadas e ranqueadas as variáveis de maior importância preditiva sobre as notas atribuídas a 85 programas.

Os resultados indicam que a conclusão de teses, a produção qualificada em periódicos de alto estrato *Pub_peridico_A2* e a participação discente em publicações de conferências são os principais fatores preditivos do desempenho institucional, com R^2 de 0,787 no conjunto de teste. Além disso, variáveis relacionadas à eficiência do doutoramento, como *Duracao_Doutorado_Med* e *Matricula_Doutorado_Ano_Med*, também figuram entre os fatores mais relevantes, convergindo com as diretrizes estabelecidas no Documento de Área da CAPES [CAPES 2025]. As principais contribuições deste trabalho são: (i) um ranqueamento empírico das variáveis de maior peso preditivo na avaliação da pós-graduação em Computação; (ii) a validação comparativa do *Extra Trees Regressor* frente a outros algoritmos, incluindo a demonstração do overfitting da árvore de decisão simples; e (iii) um perfil quantitativo por faixa de nota que pode subsidiar decisões pedagógicas e administrativas de coordenadores de programas, especialmente aqueles em consolidação (notas 3 e 4).

Como limitações, destacam-se o tamanho reduzido da amostra nos estratos superiores (nota 6: n=4; nota 7: n=8), que limita a generalização dos resultados para programas de excelência. Esse desbalanceamento é estrutural, reflexo da própria distribuição do SNPG, e não pode ser contornado pela inclusão de dados de outras áreas, uma vez que os critérios de avaliação CAPES diferem substancialmente entre elas, o que inviabiliza comparações diretas sem ajustes metodológicos específicos.

Como trabalhos futuros, pretende-se expandir a base de dados para incluir ciclos avaliativos anteriores, viabilizando uma análise longitudinal do sistema, e aplicar técnicas de Processamento de Linguagem Natural sobre os textos das propostas dos programas, com o objetivo de identificar fatores subjetivos que complementem a compreensão de maneira mais qualitativa do sucesso institucional na pós-graduação brasileira.

6. Declaração sobre uso de Inteligência Artificial

Durante a preparação deste trabalho, os autores utilizaram as ferramentas de IA generativa ChatGPT e Claude com a finalidade de apoiar a revisão textual. Após a utilização dessas ferramentas, os autores revisaram e editaram integralmente o conteúdo gerado, assumindo plena responsabilidade pelo conteúdo da publicação.

Referências

Baker, R. S. and Yacef, K. (2009) "The state of educational data mining in 2009: A review and future visions", Journal of Educational Data Mining, v. 1, n. 1, p. 3--17.

Breiman, L. (1996) "Bagging predictors", Machine Learning, v. 24, n. 2, p. 123-140.

CAPES (2025). Documento de Área: Ciência da Computação. Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brasília, DF. Disponível em: https://www.gov.br/capes/pt-br/acesso-a-informacao/acoes-e-programas/avaliacao/sobre-a-avaliacao/areas-avaliacao/sobre-as-areas-de-avaliacao/colegio-de-ciencias-exatas-tecnologicas-e-multidisciplinar/ciencias-exatas-e-da-terra/COMPUTACAO_DOCAREA_2025_2028.pdf.

Cirilo, A. C. S., Santos, I. M. and Mota, M. P. (2025) "AnyLattes: An Application for Continuous Assessment of Lattes Curriculum Information", In: Anais do XXI Simpósio Brasileiro de Sistemas de Informação (SBSI 2025), p. 439-448.

Dietterich, T. G. (2000) "Ensemble Methods in Machine Learning", In: Kittler, J.; Roli, F. (Eds.). Proceedings of the First International Workshop on Multiple Classifier Systems. London: Springer-Verlag, p. 1-15.

Eashwar, K. B., Venkatesan, R. and Ganesh, D. (2017) "Student performance prediction using SVM", International Journal of Mechanical Engineering and Technology, v. 8, n. 11, p. 649–662.

Geurts, P., Ernst, D. and Wehenkel, L. (2006) "Extremely randomized trees", Machine Learning, v. 63, n. 1, p. 3-42.

Gil, A. C. (2002) “Como elaborar projetos de pesquisa”, 4. ed., São Paulo: Atlas.

Gualhano, M. A., Salles, S. A. F. and Hora, H. R. M. da (2018) “Mineração de dados das fichas da Avaliação Quadrienal da Capes dos Programas da área Interdisciplinar: Engenharia, Tecnologia e Gestão”, Meta: Avaliação, v. 10, n. 29, p. 417–442.

Laender, A. H. F. et al. (2008) "Assessing the research and education quality of the top Brazilian Computer Science graduate programs", ACM SIGCSE Bulletin, v. 40, n. 2, p. 135-139.

Manhães, L. M. B., Cruz, S. M. S., Costa, R. J. M., Zavaleta, J. and Zimbrão, G. (2011) "Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados", In: Anais do XXII Simpósio Brasileiro de Informática na Educação (SBIE), Aracaju, SE.

Mena-Chalco, J. P. and Cesar Jr., R. M. (2009) "ScriptLattes: An open-source knowledge extraction system from the Lattes platform", Journal of the Brazilian Computer Society, v. 15, n. 4, p. 31-39.

Nunes da Silva, A. J., Breve, M. M., Mena-Chalco, J. P. and Lopes, F. M. (2022) "Analysis of co-authorship networks among Brazilian graduate programs in computer science", PLOS ONE, v. 17, n. 1, e0261200. DOI: 10.1371/journal.pone.0261200.

Romero, C. and Ventura, S. (2020) "Educational Data Mining and Learning Analytics: An Updated Survey", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, v. 10, n. 3, e1355.

SBC (2024). Grandes Desafios da Computação no Brasil (2024–2034). Sociedade Brasileira de Computação, Porto Alegre, RS.