# Modelagem Preditiva do Desempenho Acadêmico: Insights e Implicações

1st Kauan Divino Pouso Mariano
*Instituto de Informática*
*Universidade Federal de Goiás*
Goiânia, Goiás
kauan@discente.ufg.br

***Abstract*—Este relatório apresenta uma análise abrangente do desempenho acadêmico de estudantes da fictícia "University of Exampleville". Utilizando um conjunto de dados disponível, foi conduzida uma Análise Exploratória de Dados para identificar tendências e padrões chave no desempenho estudantil. Posteriormente, modelos preditivos foram desenvolvidos, com ênfase na Floresta Aleatória, que se mostrou particularmente eficaz com um coeficiente de determinação $R^2$ de 0.998. O estudo revela implicações práticas para a educação, como a possibilidade de intervenções proativas baseadas nas previsões do modelo e sugestões para refinamentos curriculares. Limitações e recomendações para pesquisas futuras também são discutidas.**

***Index Terms*—Desempenho acadêmico, Modelagem Preditiva, Floresta Aleatória, Aprendizado de Máquina, Hiperparâmetros.**

## I. Introdução

O projeto em questão teve como objetivo principal a análise e previsão do desempenho acadêmico de estudantes, utilizando um conjunto de dados oriundo de uma instituição educacional fictícia chamada "University of Exampleville". Os dados foram adquiridos a partir de avaliações acadêmicas aplicadas em variados cursos e níveis acadêmicos, priorizando a avaliação do desempenho dos estudantes em tópicos de gestão geral e áreas específicas.

A relevância desta análise decorre da demanda crescente das instituições educacionais em entender os elementos que afetam o desempenho acadêmico dos estudantes. Ao compreender esses elementos de maneira aprofundada, é possível para as instituições formular e aplicar estratégias mais efetivas para otimizar o processo de ensino e aprendizado, identificar pontos críticos e prover o suporte adequado aos estudantes, quando necessário.

Neste relatório, detalhar-se-ão as etapas conduzidas ao longo do projeto, englobando desde a obtenção e tratamento dos dados até a elaboração de modelos preditivos. As metodologias empregadas, assim como os resultados e percepções adquiridos, serão expostos de maneira objetiva, proporcionando uma visão integral do projeto.

## II. Coleta e Descrição dos Dados

Os dados empregados neste projeto provêm de um conjunto disponibilizado na plataforma Kaggle, concebido para retratar os resultados de avaliações de estudantes da fictícia "University of Exampleville". Este conjunto é estruturado em 12 colunas que delineiam diversos atributos e indicadores de desempenho dos estudantes. Entre esses atributos, incluem-se: a identificação anonimizada do estudante (por questões de privacidade); a universidade de matrícula; o programa acadêmico em que o estudante está inscrito (BBA ou MBA); a área de especialização ou major escolhido; o semestre ou período acadêmico de realização do exame; e o domínio da avaliação, que pode ser dividido em gestão geral e domínio específico. Além disso, as colunas registram as pontuações obtidas em gestão geral (com um máximo de 50 pontos), no domínio específico (também até 50 pontos) e a soma total dessas pontuações, alcançando até 100 pontos.

Quanto ao volume e qualidade dos dados, o conjunto totaliza 500 registros, representando individualmente os estudantes da "University of Exampleville". Uma análise exploratória preliminar revelou a consistência e qualidade dos dados. Eventuais ausências de valores foram identificadas e tratadas de forma adequada, garantindo a integridade das análises que se seguiriam.

## III. Análise Exploratória de Dados (EDA)

A Análise Exploratória de Dados foi conduzida com o intuito de entender a distribuição, tendência e interações entre os atributos contidos no conjunto de dados. Tal fase possibilitou a identificação de percepções preliminares sobre o desempenho estudantil, assim como possíveis anomalias ou padrões notáveis nos registros.

Ao longo da EDA, empregaram-se variadas técnicas de visualização, juntamente com estatísticas descritivas, para investigar o conjunto de dados. Dentre as análises fundamentais, destacam-se: a distribuição das pontuações em gestão geral e domínio específico; a comparação de desempenho acadêmico entre os programas BBA e MBA; a análise do desempenho baseado nas especializações selecionadas pelos estudantes; e a correlação entre diferentes atributos para discernir os fatores de maior impacto no desempenho acadêmico.

Os insights centrais derivados da análise exploratória são:

- Estudantes matriculados no programa MBA, em média, apresentaram desempenho levemente superior em relação aos inscritos no programa BBA.
- A escolha de especialização demonstrou exercer impacto significativo nas pontuações, notadamente no domínio específico.

- Identificou-se uma correlação robusta entre as pontuações de gestão geral e domínio específico, sugerindo que estudantes com bom desempenho em um domínio frequentemente performam bem no outro.
- A distribuição das pontuações indicou uma concentração majoritária de estudantes na faixa de desempenho intermediário, com uma quantidade reduzida nas extremidades de alto e baixo desempenho.

## IV. Modelagem Preditiva

A fase de modelagem preditiva representou um marco central deste projeto, com ênfase na elaboração e aperfeiçoamento de modelos voltados à previsão do desempenho acadêmico dos estudantes. A importância desta etapa é indiscutível, visto que a capacidade de antever tendências e padrões no rendimento estudantil pode servir como guia para instituições e educadores. Adicionalmente, um modelo preditivo eficaz pode se traduzir em uma ferramenta de intervenção valiosa, permitindo ajustes focados e personalizados no processo de ensino.

No tocante à metodologia e à abordagem utilizadas, o conjunto de dados foi estrategicamente segmentado em treinamento e teste. Enquanto o primeiro conjunto auxiliou no treinamento e ajuste dos modelos, o segundo serviu para testar a habilidade dos modelos de realizar previsões em dados inéditos, avaliando assim sua eficácia e precisão.

Diversos modelos foram explorados ao longo deste processo:

- Regressão Linear: Optou-se por este devido à sua abordagem direta e capacidade de identificar relações lineares, sendo utilizado como um modelo base de referência.
- Árvore de Decisão: Este modelo, notável pela sua clareza e aptidão em detectar relações não lineares, segmenta os dados em diferentes regiões para atribuir previsões.
- Floresta Aleatória: Esta técnica, que agrega previsões de múltiplas árvores de decisão, demonstra resiliência a anomalias e ruídos, favorecendo a robustez.
- Gradient Boosting e XGBoost: Ambos são modelos de conjunto que constroem árvores sequencialmente, com cada nova árvore buscando corrigir as falhas da anterior. Esses modelos são frequentemente reconhecidos por sua precisão e adaptabilidade.

A otimização de hiperparâmetros, que são parâmetros intrínsecos que influenciam o comportamento dos modelos, foi rigorosamente conduzida. Utilizou-se o método Grid Search para sistematicamente testar combinações de hiperparâmetros, objetivando a configuração mais adequada para cada modelo.

Em relação aos resultados:

- A Regressão Linear, mesmo sendo um modelo básico, revelou uma performance notável com um coeficiente de determinação $R^2$ de 0.996.
- A Árvore de Decisão apresentou um $R^2$ de 0.995.
- A Floresta Aleatória, após otimizações, alcançou um $R^2$ de 0.998.
- Tanto o Gradient Boosting quanto o XGBoost mostraram-se eficientes, com $R^2$ de 0.998 e 0.997, respectivamente.

Ao final, considerando as métricas, complexidade dos modelos, e sua capacidade de generalização, a Floresta Aleatória foi determinada como o modelo ótimo para este projeto, graças à sua precisão, robustez e relativa simplicidade.

## V. Conclusão

Este projeto envolveu uma análise meticulosa do desempenho acadêmico dos estudantes da "University of Exampleville". A Análise Exploratória de Dados possibilitou a identificação de padrões notáveis, como a predominância de desempenho intermediário entre os estudantes, o impacto considerável da especialização na performance dos exames específicos e a existência de uma correlação expressiva entre as pontuações de gestão geral e domínio específico. Estas observações sugerem uma consistência no rendimento dos estudantes em distintas matérias. No segmento de modelagem preditiva, os modelos de aprendizado de máquina demonstraram competência em antever o rendimento acadêmico, com destaque para a Floresta Aleatória, que, após refinamentos, alcançou um coeficiente de determinação $R^2$ de 0.998.

Os resultados obtidos carregam implicações relevantes para a prática educacional:

- Intervenções Proativas: O modelo preditivo habilita a identificação precoce de estudantes que possivelmente necessitem de suporte, facilitando ações proativas.
- Desenvolvimento Curricular: As descobertas acerca das áreas de especialização podem influenciar decisões relacionadas ao currículo e distribuição de recursos.
- Benchmarking: O modelo estabelecido pode funcionar como parâmetro comparativo do desempenho entre diferentes cursos ou instituições.

No entanto, é essencial considerar algumas restrições:

- Origem dos Dados: Os registros provêm de uma única instituição fictícia, limitando a representatividade em relação a instituições reais.
- Foco em Avaliações Finais: O conjunto de dados se atém majoritariamente às pontuações finais, desconsiderando aspectos como envolvimento em sala de aula e avaliações contínuas.

Recomenda-se para investigações futuras:

- Ampliação do Conjunto de Dados: Inclusão de variáveis adicionais, como histórico acadêmico anterior e participação em atividades extracurriculares.
- Modelos para Diferentes Resultados: Além da pontuação acumulada, seria relevante prever outras métricas acadêmicas, como taxas de retenção ou sucesso pós-graduação.
- Validação com Dados Reais: A eficácia do modelo poderia ser testada em dados oriundos de instituições educacionais autênticas.

## References

[1] Desconhecido. (2023). Examination results from "University of Exampleville". Kaggle. Disponível em: https://www.kaggle.com/datasets/atharvbharaskar/students-test-data. Acesso em: 23 set 2023.

[2] ames, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning (Vol. 112). New York: springer.
[3] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.
[4] Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(Feb), 281-305.
[5] Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. International journal of forecasting, 22(4), 679-688.
[6] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).