PREDIÇÃO DA EVASÃO ESCOLAR COM ALGORITMOS DE MACHINE LEARNING: UM ESTUDO DE CASO NO IFCE - CAMPUS ARACATI

PREDICTION OF SCHOOL DROPOUT USING MACHINE LEARNING ALGORITHMS: A CASE STUDY AT IFCE ARACATI CAMPUS

Davi Moreira Cardoso*
Felipe Bastos Nunes**

RESUMO

Este artigo tem como objetivo utilizar *machine learning* para prever a evasão no Instituto Federal do Ceará (IFCE), campus Aracati, utilizando dados públicos do site IFCE em Números. A metodologia utilizada foi baseada no processo de KDD (*Knowledge Discovery in Databases*), composto pelas etapas de seleção, pré-processamento, transformação, mineração e interpretação dos dados. Foram testados 14 algoritmos de diferentes famílias, com destaque para o desempenho do *Random Forest* e do *Gradient Boosting*. Os resultados demonstram a eficácia dos modelos na identificação de padrões relacionados à evasão, possibilitando o desenvolvimento de soluções estratégicas com foco na permanência dos alunos.

Palavras-chave: Evasão escolar. *Machine Learning*. Educação superior. Previsão. IFCE.

ABSTRACT

This study aims to apply machine learning techniques to predict student dropout at the Federal Institute of Ceará (IFCE), Aracati campus, using public data from the IFCE em Números portal. The methodology was based on the KDD (Knowledge Discovery in Databases) process, which includes the steps of selection, preprocessing, transformation, mining, and data interpretation. Fourteen algorithms from different families were tested, with the performance of Random Forest and Gradient Boosting standing out. The results demonstrate the effectiveness of the models in identifying patterns related to dropout, enabling the development of strategic solutions focused on student retention.

Keywords: School dropout. Machine Learning. Higher education. Prediction. IFCE.

* Graduando em Ciência da Computação, Instituto Federal de Educação, Ciência e Tecnologia do Ceará, Aracati. CE, Brasil. E-mail: davimoreirax2021@gmail.com
** Especialista em Docência no Ensino Técnico, docente do Instituto Federal de Educação. Ciência e Tecnologia do Ceará, Aracati, CE, Brasil. E-mail: felipebastos@ifce.edu.br

1 INTRODUÇÃO

A evasão escolar no ensino superior é um dos principais desafios enfrentados pela educação brasileira. Abandonar o curso antes da conclusão compromete não apenas os projetos de vida dos estudantes, mas também representa um desperdício de recursos e esforços das instituições de ensino, principalmente nas redes públicas. Segundo o Mapa do Ensino Superior no Brasil, produzido pelo Instituto Semesp (2024), a taxa de evasão no ensino superior chega a 57,2% nas redes públicas e privadas, considerando tanto o ensino presencial quanto o EAD.

Quando comparamos esse cenário com a realidade do Instituto Federal do Ceará (IFCE), no campus Aracati, observa-se que a situação é igualmente crítica. De acordo com o site IFCE em Números (2025), entre 2013 e 2025, 2110 estudantes ingressaram nos cursos superiores da instituição. Desse total, 1152 alunos evadiram-se e apenas 304 concluíram a graduação, o que representa uma taxa de evasão de aproximadamente 54,6%. Esses dados mostram que, mesmo em instituições públicas com políticas de assistência estudantil, o problema persiste e demanda atenção imediata.

Diante desse cenário, é essencial compreender os fatores associados à evasão escolar e desenvolver estratégias eficazes de prevenção. Neste sentido, técnicas de mineração de dados educacionais vêm ganhando destaque, como demonstrado no trabalho de SANTOS (2021), que utilizou mineração de dados para identificar padrões de evasão e apoiar ações institucionais de intervenção.

Este trabalho tem como objetivo geral aplicar técnicas de *Machine Learning* (ML) para analisar e prever a evasão nos cursos superiores do IFCE – campus Aracati, que incluem: Bacharelado em Ciência da Computação, Bacharelado em Engenharia de Aquicultura, Licenciatura em Química e Tecnologia em Hotelaria. Os objetivos específicos são: (i) Avaliar a capacidade de diferentes modelos de *machine learning* em prever a evasão escolar; (ii) Identificar as características que mais afetam a evasão; (iii) Propor um método que possibilite identificar o mais cedo possível os alunos que correm risco de evasão; e (iv) Fornecer subsídios para a elaboração de estratégias institucionais de apoio.

2 REFERENCIAL TEÓRICO

Este capítulo aborda os conceitos de *Knowledge Discovery in Databases* (KDD), Mineração de Dados e ML, que são utilizados neste trabalho como ferramentas de investigação e apoio no combate à evasão escolar. Além disso, o capítulo explica com mais detalhes o que é evasão, apresentando, em seguida, um recorte específico do contexto do IFCE, campus Aracati.

2.1 Evasão Escolar

A evasão escolar é um problema complexo que vai além dos números. Conforme o Ministério da Educação (BRASIL, 1997), ela é definida como a saída definitiva do aluno do curso de origem sem a sua conclusão, medida pela diferença entre ingressantes e concluintes. No

entanto, como destacam Vieira, Gallindo e Cruz (2017), a evasão também envolve fatores sociais (como a pobreza), institucionais (como a falta de estrutura) e pessoais (como a dificuldade de adaptação). Essa complexidade exige uma análise contextualizada, especialmente em instituições públicas como os Institutos Federais, onde políticas de assistência e o perfil socioeconômico dos discentes influenciam diretamente a permanência estudantil (PEREIRA; LUZ; LIMA, 2020).

2.2 Ambientação da Evasão no IFCE - Campus Aracati

A evasão no IFCE campus Aracati revela cenários distintos de acordo com a área de formação. De acordo com os dados do site IFCE em Números (2025), o curso de Tecnologia em Hotelaria apresenta a maior taxa de evasão: dos 632 alunos ingressantes, 458 evadiram-se (72,5%), enquanto apenas 153 concluíram o curso (24,2%). Em contrapartida, o Bacharelado em Engenharia de Aquicultura, ofertado desde 2017, chama a atenção pela taxa de conclusão mais baixa: apenas 11 alunos (4,3%) finalizaram a graduação, e 145 evadiram-se (56,6%) entre os 256 ingressantes.

Por sua vez, o curso de Bacharelado em Ciência da Computação, com 817 ingressantes, registrou 421 evadidos (51,5%) e 103 concluintes (12,6%). Apesar de uma evasão moderada, a baixa taxa de conclusão sugere que muitos alunos levam mais tempo para concluir os projetos pedagógicos do curso do que o planejado, o que pode aumentar o risco de abandono do curso após anos de dedicação, possivelmente devido à complexidade das disciplinas e às contínuas reprovações. Já a Licenciatura em Química, com 368 ingressantes, teve 204 evadidos (55,4%) e apenas 37 concluintes (10%). A Tabela 1 apresenta um resumo comparativo dessas taxas entre os cursos citados.

Esses dados reforçam os desafios enfrentados pela instituição para garantir a permanência dos estudantes. Apesar de iniciativas como políticas de assistência estudantil e programas institucionais de apoio, o abandono dos cursos ainda ocorre com frequência.

Tabela 1 – Taxas de Evasão e Conclusão por Curso

Cursos	**Evasão**	**Conclusão**
Hotelaria	72,5%	24,2%
Eng. de Aquicultura	56,6%	4,3%
Ciência da Computação	51,5%	12,6%
Química	55,4%	10%

Fonte: Elaborado pelo autor.

2.3 Fatores Determinantes da Evasão

O Quadro 1 resume os principais fatores associados à evasão, categorizados em internos (relacionados à instituição) e externos (relacionados ao contexto do aluno), conforme proposto por Filho, Siqueira e Leal (2020):

Quadro 1 – Fatores relacionados à evasão no contexto educacional

Fatores internos	**Fatores externos**
Falta de recursos e de segurança nas escolas	Falta de transporte local/municipal ao estudante
O excesso de alunos nas salas de aula	Vulnerabilidade socioeconômica do estudante
Falta de qualificação dos professores	Ausência de ambiente/condições de estudo em casa
Matrizes curriculares e projetos de curso desatualizados e desalinhados com as necessidades atuais do mercado	Falta de oportunidades de trabalho na área do curso do discente

Fonte: FILHO; SIQUEIRA; LEAL (2020).

Esses fatores interagem de forma dinâmica. Por exemplo, a falta de transporte (externo) pode agravar a exclusão digital (interna), especialmente em regiões rurais, como o entorno do Campus Aracati. Estudos como o de Alba (2018) reforçam a importância de programas institucionais que ofereçam apoio acadêmico, assistência estudantil e ações voltadas à permanência dos alunos. Uma estratégia que combina diferentes soluções, como esta, é fundamental para reduzir o distanciamento entre estudantes e instituições.

Diante da complexidade dos fatores que influenciam a evasão, torna-se necessário o uso de abordagens capazes de lidar com variáveis diversas. Nesse contexto, técnicas de ML têm sido cada vez mais aplicadas no meio educacional, oferecendo caminhos para identificar padrões de evasão e apoiar decisões institucionais baseadas em dados.

2.4 Machine Learning (ML)

Machine Learning consolida-se como uma estratégia consistente para a análise e previsão da evasão em institutos federais, permitindo identificar padrões e variáveis relevantes para o problema, facilitando a implementação de intervenções direcionadas (BARBOSA et al., 2023).

Neste contexto, a ferramenta *scikit-learn*, uma biblioteca *Python* para classificação e regressão, destaca-se pela capacidade de treinar modelos preditivos. Sua versatilidade tem sido apresentada em diferentes estudos, como o de Rajamani e Iyer (2023), que a utilizou para permitir que usuários desenvolvam e executem modelos de ML diretamente em seus celulares Android.

A aplicação dos algoritmos requer, no entanto, critérios claros para a avaliação de desempenho, a fim de garantir a confiabilidade dos resultados preditivos. Por isso, é necessário compreender as principais métricas envolvidas.

2.5 Métricas de Avaliação de *machine learning*

Em problemas de classificação binária, como prever se um aluno evadirá (classe positiva) ou não (classe negativa), a análise parte da matriz de confusão, que contabiliza os verdadeiros positivos (VP), que são os alunos corretamente identificados como evadidos; os falsos positivos (FP), que são os alunos erroneamente classificados como evadidos; os verdadeiros negativos

(VN), que são os alunos corretamente identificados como não evadidos; e os falsos negativos (FN), que são os alunos evadidos não detectados pelo modelo. Com esses valores, calculam-se as métricas: acurácia, que mede a proporção de acertos do modelo (Equação 1); precisão, que avalia a proporção de evadidos corretos entre os classificados como tais (Equação 2); *recall*, que indica a proporção de evadidos corretamente identificados (Equação 3); e F1-score, que mede a média harmônica entre a precisão e o *recall* (Equação 4).

$$Acurácia = \frac{VP+VN}{VP+FP+VN+FN} \quad (1) \qquad Precisão = \frac{VP}{VP+FP} \quad (2)$$

$$Recall = \frac{VP}{VP+FN} \quad (3) \qquad F1\text{-}Score = 2 \cdot \frac{\text{Precisão} \cdot Recall}{\text{Precisão} + Recall} \quad (4)$$

No contexto da evasão, priorizar um *recall* possivelmente alto é estratégico, pois é preferível identificar um aluno não evadido como risco (falso positivo) do que ignorar um aluno que realmente evadirá (falso negativo). Essa abordagem permite intervenções preventivas, mesmo com margem de erro, alinhando-se a estudos como o de Adnan et al. (2021), que defendem a utilização de um modelo preditivo para identificar precocemente um aluno em risco, permitindo que os instrutores atuem de forma proativa antes que a evasão ocorra.

Para que os modelos preditivos sejam eficazes, é fundamental garantir a qualidade dos dados utilizados no processo. Nesse sentido, o processo de Knowledge Discovery in Databases (KDD) fornece uma estrutura sistemática para transformar dados brutos em conhecimento útil.

2.6 KDD

Para Fayyad, Piatetsky-Shapiro e Smyth (1996), o processo de KDD é compreendido como etapas estruturadas para transformar grandes volumes de dados em conhecimento útil. O processo de KDD, conforme os autores, compreende as seguintes etapas: a) Seleção dos dados; b) Pré-processamento (limpeza e normalização); c) Transformação (adequação para análise); d) Mineração de dados (aplicação de algoritmos); e) Interpretação dos resultados.

Em projetos como o presente estudo, a adoção de práticas de KDD é essencial para garantir a qualidade dos dados que alimentarão os modelos de *machine learning*. Com elas, conseguimos extrair informações relevantes que, muitas vezes, estão ocultas nos conjuntos de dados (BARBOSA et al., 2023). Como demonstrado por Machado et al. (2024) e SANTOS (2021), a preparação adequada dos dados é decisiva para o bom desempenho dos algoritmos preditivos, uma vez que dados brutos frequentemente contêm ruídos, valores ausentes ou inconsistências que podem enviesar os modelos ou reduzir sua acurácia.

3 TRABALHOS RELACIONADOS

Machado et al. (2024) realizaram um estudo com o objetivo de identificar o algoritmo mais eficiente para prever a evasão escolar no curso de Tecnologia em Análise e Desenvolvimento de Sistemas (TADS) do IFPA Campus Altamira. Utilizando dados extraídos da plataforma SIGAA referentes ao período de 2018 a 2021, os autores conduziram um processo de pré-processamento que envolveu a conversão dos dados para o formato CSV, ajustes em caracteres incompatíveis com editores de planilhas, padronização do separador decimal e número de casas decimais.

A partir disso, aplicaram todos os algoritmos disponíveis na ferramenta WEKA, selecionando os cinco melhores com base na acurácia. Em seguida, métricas adicionais foram utilizadas para comparar os modelos, sendo o algoritmo J48 o que apresentou os melhores resultados, com uma acurácia de 97,76% e F1-score de 94,2%. Os autores destacam como limitação o desbalanceamento das classes no conjunto de dados, o que pode impactar a performance dos modelos.

Teodoro e Kappel (2020) propõem a aplicação de técnicas de *machine learning* para prever o risco de evasão em instituições públicas de ensino superior no Brasil. Utilizando dados do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) e implementações realizadas com a biblioteca *Scikit-learn*, os autores testaram diferentes algoritmos, como *Naive Bayes*, *K-Nearest Neighbors*, Árvores de Decisão, *Random Forest* e Redes Neurais. Após um pré-processamento criterioso, que incluiu a remoção de dados faltantes, a desconsideração de atributos irrelevantes e o balanceamento das classes, os dados foram utilizados na fase de treinamento.

O Random Forest destacou-se como o modelo mais eficaz, alcançando uma taxa de acerto de 80% nas previsões e apresentando uma performance estável e confiável em métricas como a curva ROC. Entre os atributos mais relevantes para a previsão da evasão, destacam-se a idade do aluno, a participação em atividades extracurriculares e a carga horária total do curso. Além disso, o estudo realizou uma análise aprofundada das trinta variáveis mais determinantes, promovendo um diálogo com outras pesquisas sobre a evasão no ensino superior brasileiro.

SANTOS (2021) desenvolveu um estudo com foco na aplicação de técnicas de mineração de dados para a predição da evasão escolar em cursos de graduação do IFSC – Campus Caçador. O autor utilizou as técnicas de Redes Neurais Artificiais e Árvore de Decisão, implementadas com a linguagem *Python* e as bibliotecas *Scikit-Learn* e *Pandas*. O processo de obtenção dos dados envolveu uma série de etapas administrativas e técnicas, incluindo a consulta a diagramas do banco de dados institucional e o desenvolvimento de scripts SQL para a extração dos dados relevantes.

Após o tratamento e a limpeza dos dados, foram selecionadas variáveis como idade, faltas no último semestre e status acadêmico, as quais se mostraram altamente influentes nas previsões. A validação cruzada foi utilizada devido ao pequeno volume de dados no *dataset* (380 registros). Embora o modelo de Rede Neural tenha obtido maior *recall* (82%), a Árvore de Decisão apresentou melhor desempenho geral, com acurácia de 84% e precisão de 87%, sendo

escolhida para compor o protótipo final de predição da evasão. A ferramenta resultante oferece uma análise explicativa dos fatores mais impactantes por meio da biblioteca SHAP.

Os estudos relacionados analisados neste trabalho empregam diferentes estratégias metodológicas para a predição de evasão, com recorrente destaque para modelos baseados em árvores, como *Decision Tree* e *Random Forest*, que têm demonstrado bom desempenho em bases institucionais. Parte da literatura também utiliza técnicas de balanceamento de classes, como SMOTE ou *Random Oversampling*, devido ao fato de que a evasão normalmente ocorre em menor proporção em diversas instituições. No entanto, essa característica não se aplica ao contexto do IFCE – Campus Aracati: na base analisada, as classes apresentam distribuição relativamente equilibrada, com 1.152 estudantes evadidos e 958 concluintes ou ainda matriculados. Assim, optou-se por não realizar qualquer procedimento de balanceamento, uma vez que a proporção natural das classes já é adequada para o treinamento dos modelos. Essa decisão diferencia este estudo de grande parte dos trabalhos anteriores e permite avaliar o desempenho dos algoritmos em um cenário institucional menos afetado por desbalanceamento, oferecendo uma análise mais fiel ao comportamento real dos dados da instituição. Deste modo, a Tabela 2 demonstra como esta pesquisa almeja unir as diferentes abordagens dos autores relacionados.

Tabela 2 – Comparação de Critérios com Trabalhos Relacionados

Critérios	**Teodoro (2020)**	**Santos (2021)**	**Machado (2024)**	**Este Trabalho**
Focado em uma instituição		X	X	X
Análise de múltiplos cursos		X		X
Scikit-Learn	X	X		X
Comparou vários modelos	X		X	X

Fonte: Elaborado pelo autor.

4 METODOLOGIA

Este trabalho adotou uma metodologia baseada no processo de KDD (*Knowledge Discovery in Databases*), conforme descrito por Fayyad, Piatetsky-Shapiro e Smyth (1996). Este processo é composto por etapas sequenciais e interativas que visam extrair conhecimento útil a partir de grandes volumes de dados. As etapas aplicadas neste estudo foram: Seleção, Pré-processamento, Transformação, Mineração e Interpretação dos dados.

4.1 Seleção dos dados

A etapa inicial consistiu na definição e coleta do *dataset* a ser utilizado. A fonte utilizada foi a plataforma institucional de dados abertos "IFCE em Números". Após uma análise dos dados completos do site, foram selecionadas 27 colunas consideradas potencialmente relevantes para o fenômeno da evasão, abrangendo dados socioeconômicos, acadêmicos e de ingresso. O conjunto inicial totalizou 13.225 registros.

Embora a base contenha variáveis relevantes para a análise, como idade, sexo, etnia, renda familiar, forma de ingresso, tipo de cota, curso, turno e coeficiente de rendimento, deve-se considerar que se trata de uma base administrativa, apresentando limitações quanto à profundidade dos aspectos determinantes para a evasão. Informações socioeconômicas detalhadas, como renda exata, situação laboral, tempo de deslocamento e participação em programas de assistência estudantil, não estão disponíveis, apesar de serem amplamente reconhecidas na literatura como fatores decisivos. Além disso, dados institucionais importantes, como histórico de reprovações, fluxo curricular e indicadores de engajamento, também estão ausentes. A falta dessas variáveis pode introduzir vieses e limitar o poder explicativo dos modelos, visto que os algoritmos aprendem apenas os padrões presentes nos registros disponíveis. Adicionalmente, a possibilidade de inconsistências no preenchimento ou atrasos na atualização dos dados reforça a necessidade de interpretar os resultados com a devida cautela.

Nesta fase, a análise exploratória revelou também desafios estruturais significativos nos dados brutos. O principal deles foi a granularidade temporal: os dados estavam organizados por matrícula semestral, gerando múltiplos registros para um mesmo aluno (uma linha para cada semestre cursado). Identificou-se também a ausência de informações críticas (valores nulos) em atributos como o turno.

4.2 Pré-processamento dos dados

O pré-processamento teve como objetivo transformar a base semestral em um conjunto consolidado em que cada estudante fosse representado por uma única linha. As colunas numéricas, originalmente registradas com vírgula como separador decimal, foram convertidas para o formato numérico padrão. As datas que indicavam o início de cada período letivo foram padronizadas e utilizadas para ordenar cronologicamente os registros de cada matrícula, o que permitiu identificar a evolução de cada aluno ao longo dos períodos.

Com as linhas devidamente ordenadas, foi realizada a agregação dos dados por matrícula. Cada variável foi processada conforme sua natureza: informações estáveis, como forma de ingresso ou turno, foram extraídas da primeira ocorrência, já que não sofrem alteração após o ingresso; indicadores de evolução acadêmica, como rendimento geral e trancamentos, foram obtidos do registro mais recente, representando o estágio final da trajetória do discente; e medidas de desempenho por período, como coeficientes semestrais, foram sintetizadas através de média, mínimo e máximo, de modo a preservar variações importantes ao longo do tempo. Embora esses critérios garantam a consistência na representação final dos estudantes, é importante ressaltar que a própria consolidação envolve a perda de nuances temporais, o que pode limitar a capacidade dos modelos de identificar padrões sequenciais que emergem ao longo dos semestres.

Após a agregação, o tratamento de valores ausentes adotou estratégias estatísticas distintas para preservar a integridade dos dados. Para as variáveis categóricas, utilizou-se a imputação pela moda (valor mais frequente), visando manter a característica mais representativa do perfil do estudante. Já para as variáveis numéricas, optou-se pela substituição pela mediana, uma medida

de tendência central mais robusta a *outliers* do que a média, evitando que valores extremos distorçam a representação do desempenho acadêmico.

4.3 Transformação dos dados

Na etapa de transformação, os dados foram preparados para a fase de treinamento dos algoritmos. As variáveis categóricas passaram pelo processo de codificação utilizando a técnica de *One-Hot Encoding*, que converteu cada variável em novas colunas binárias correspondentes a cada valor único existente. Para as variáveis numéricas, aplicaram-se técnicas de normalização para colocar todos os atributos na mesma escala, evitando que variáveis com grandes magnitudes dominassem o cálculo dos modelos.

Adicionalmente, buscou-se refinar a base de dados através da detecção e remoção de *outliers*. O critério adotado baseou-se no método do intervalo interquartil (IQR), identificando como atípicos os registros cuja amplitude ultrapassasse 1,5 vezes o intervalo entre o primeiro e o terceiro quartis. Esse procedimento ocorreu principalmente sobre atributos numéricos relacionados ao rendimento acadêmico. Embora a exclusão desses valores contribua para a estabilidade dos modelos, reconhece-se que tal ação é uma decisão metodológica que busca equilibrar a robustez estatística com a preservação da variabilidade original dos dados.

Por fim, a definição da variável alvo seguiu a categorização institucional presente na coluna "Situação de Matrícula (grupo)". Adotou-se uma abordagem binária na qual apenas os registros marcados estritamente como "Evadida" foram atribuídos à classe positiva (classe 1). Todos os demais status, "Concluída", "Em curso" e "Em fase de conclusão", foram agrupados na classe negativa (classe 0), visto que representam situações de vínculo mantido ou êxito acadêmico. Essa classificação alinha-se à definição operacional de evasão do MEC, considerando apenas a saída definitiva do estudante da trajetória formativa, resultando em um *dataset* final com 37 colunas e 2.110 registros.

4.4 Mineração de dados

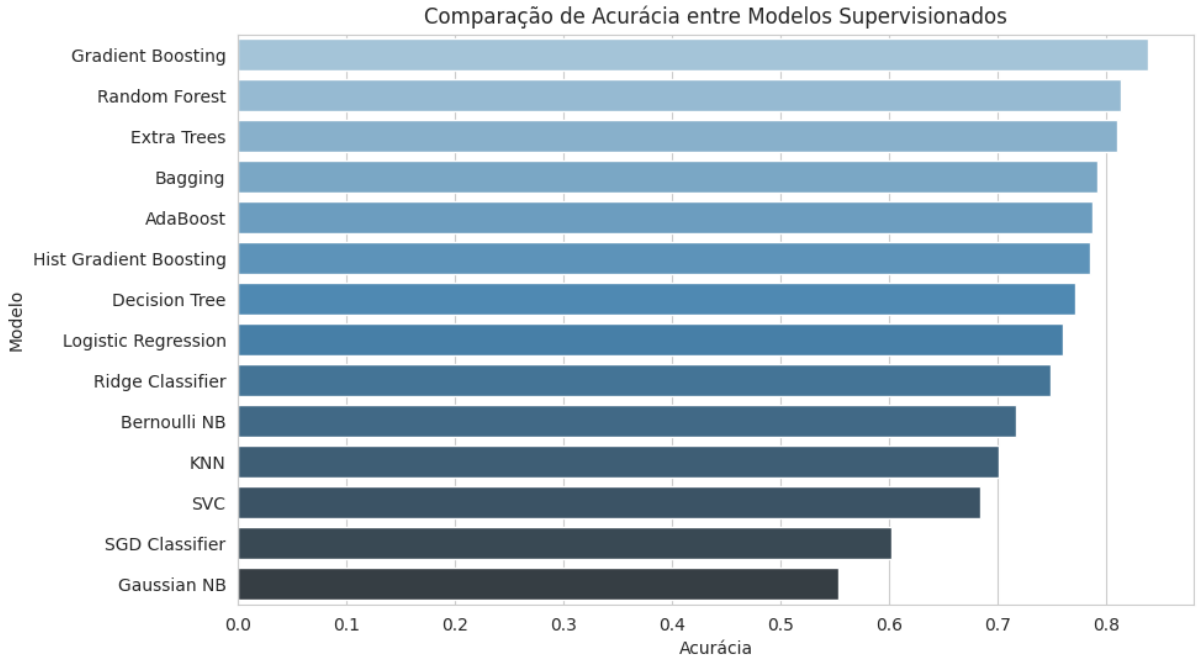
A seleção dos algoritmos fundamentou-se tanto na revisão da literatura quanto nas características estruturais da base de dados processada. Estudos correlatos indicam que modelos baseados em árvores, como *Decision Tree*, *Random Forest* e *Gradient Boosting*, apresentam desempenho superior em dados educacionais por capturarem eficientemente interações complexas e não lineares entre os atributos.

Entretanto, para evitar a dependência exclusiva dessa família de modelos, adotou-se uma estratégia exploratória abrangente. Optou-se por especificar um conjunto de 14 algoritmos distintos disponibilizados pelo *Scikit-Learn*, em detrimento de uma abordagem genérica de testar todos os classificadores disponíveis. Essa delimitação metodológica visa garantir a transparência e a reprodutibilidade do estudo, priorizando modelos com documentação consolidada e aplicabilidade reconhecida na literatura. Assim, o conjunto selecionado abrange uma diversidade

de técnicas, incluindo métodos lineares, probabilísticos, baseados em distância e *ensembles*, permitindo identificar, de forma robusta, qual abordagem melhor se ajusta aos dados do IFCE.

Dessa forma, a etapa de mineração consistiu na construção e avaliação prática desses modelos preditivos. Adotou-se uma abordagem exploratória inicial, na qual os 14 algoritmos foram treinados e testados, tendo como objetivo selecionar os dois melhores modelos de acordo com a acurácia apresentada. A Figura 1 ilustra o *ranking* de desempenho obtido nesta etapa preliminar.

Figura 1 – Comparação de Acurácia entre Modelos Supervisionados



Fonte: Elaborado pelo autor.

A análise dos resultados preliminares revela uma clara distinção no desempenho dos classificadores. Observa-se que um grupo de algoritmos conseguiu superar a marca de 80% de acurácia, demonstrando boa aderência aos padrões da base de dados. Embora modelos como *Extra Trees* e *Bagging* tenham apresentado métricas competitivas e muito próximas ao topo, o *Gradient Boosting* e o *Random Forest* destacaram-se com as maiores pontuações absolutas. Com base nessa hierarquia, esses dois algoritmos líderes foram selecionados para a fase de refinamento.

Para extrair o máximo potencial dos finalistas, o ajuste fino dos hiperparâmetros foi realizado por meio do algoritmo *GridSearchCV* com validação cruzada interna. No caso do *Random Forest*, foram testadas configurações variando o número de árvores, profundidade máxima e critérios de divisão. Para o *Gradient Boosting*, a busca incluiu diferentes taxas de aprendizado, profundidades e proporções de subamostragem. A validação cruzada garante que a seleção não dependa de uma única divisão dos dados, reduzindo o risco de *overfitting*. As melhores combinações obtidas, com *n_estimators*=100 e *max_features*='sqrt' para o *Random*

Forest, e *n_estimators*=300, *learning_rate*=0.1 e *subsample*=0.8 para o *Gradient Boosting*, foram utilizadas nas avaliações finais, assegurando que cada modelo fosse analisado em sua forma otimizada.

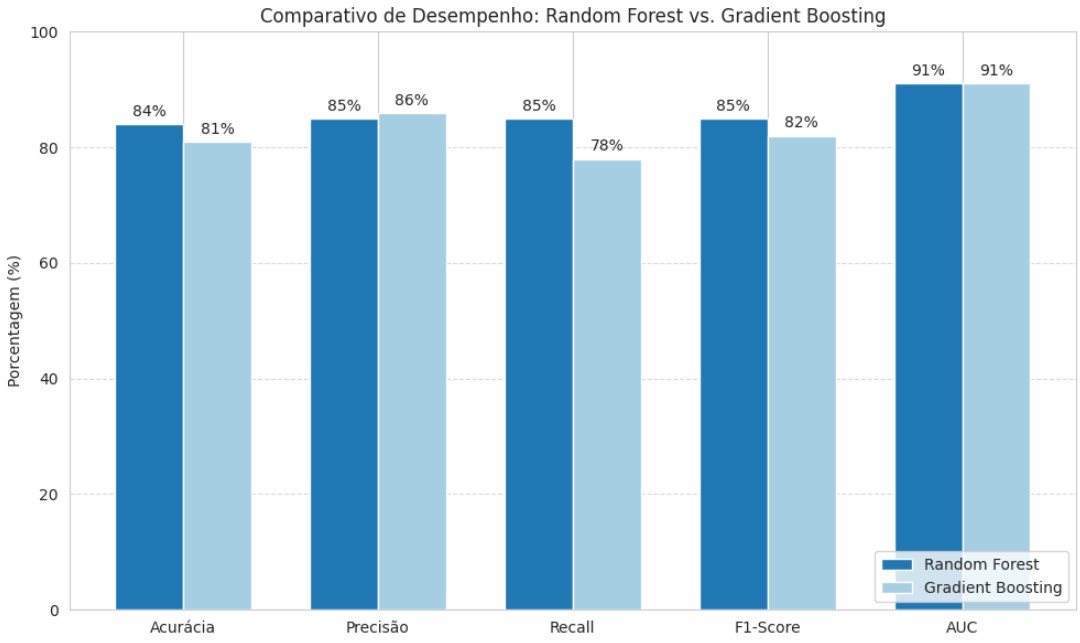
5 RESULTADOS

Nesta seção, são apresentados e discutidos os resultados obtidos pelos modelos selecionados após a otimização de hiperparâmetros. Para a validação definitiva, o conjunto de dados foi dividido na proporção de 70% para treinamento (1.477 registros) e 30% reservados exclusivamente para testes (633 registros), garantindo a avaliação da capacidade de generalização dos modelos.

A análise comparativa entre os modelos finalistas, ilustrada na Figura 2, revela nuances importantes. O modelo *Random Forest* consolidou-se como a opção mais robusta, superando o *Gradient Boosting* em quatro das cinco métricas avaliadas. Embora o *Gradient Boosting* tenha apresentado uma leve vantagem na precisão (86% contra 85%), o *Random Forest* demonstrou maior consistência geral, com um *F1-Score* superior (85% contra 82%) e uma Acurácia de 84% (frente aos 81% do concorrente). Ambos alcançaram um AUC idêntico de 91%, sugerindo capacidade similar de discriminação entre classes.

A decisão de selecionar o *Random Forest* como modelo final não se baseou apenas na acurácia, mas na robustez do conjunto de métricas. Como o objetivo central do trabalho é apoiar ações institucionais, a prioridade recai sobre métricas sensíveis à classe positiva, garantindo que alunos em risco não sejam ignorados. Nesse sentido, a superioridade do *Random Forest* no *Recall* (85% contra 78%) e seu equilíbrio entre os indicadores sugerem menor variância e maior estabilidade nas predições. Portanto, este modelo oferece a solução mais confiável para o problema da evasão no contexto analisado.

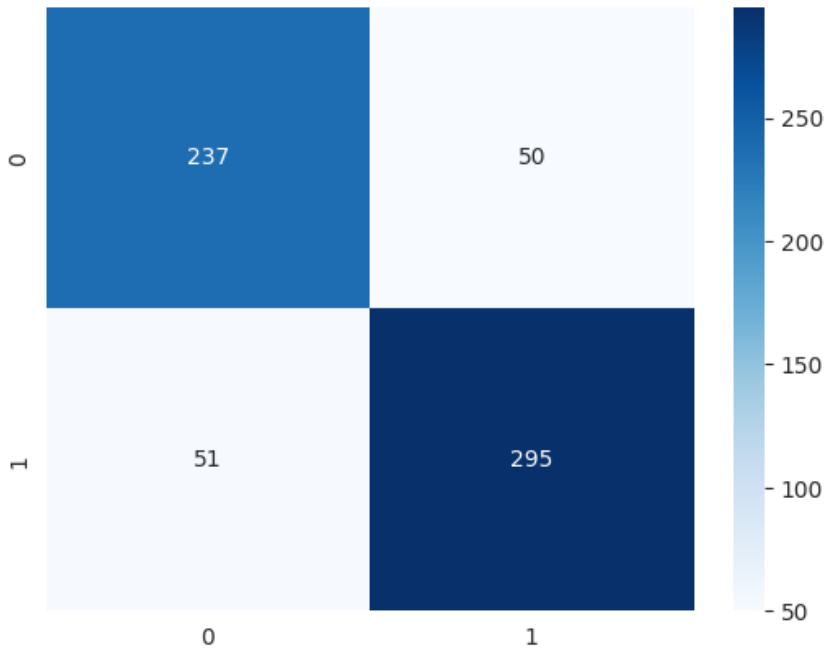
Figura 2 – Comparativo de Desempenho: Random Forest vs. Gradient Boosting



Fonte: Elaborado pelo autor.

Com base nesses indicadores, a análise detalhada das predições do modelo final pode ser observada na Matriz de Confusão apresentada na Figura 3. O modelo demonstrou uma boa eficácia ao classificar corretamente 295 alunos que evadiram (Verdadeiros Positivos) e 237 que não evadiram (Verdadeiros Negativos). Os erros foram equilibrados e relativamente baixos, com apenas 51 falsos negativos e 50 falsos positivos, confirmando a confiabilidade do preditor.

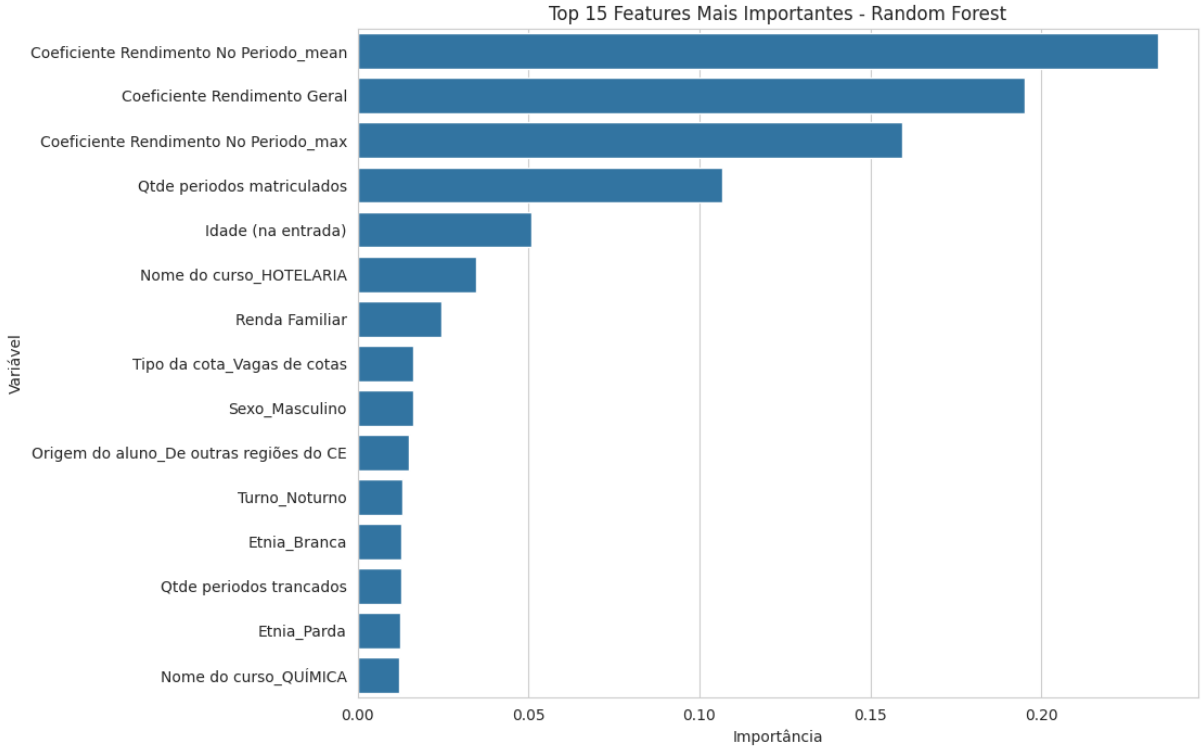
Figura 3 – Matriz de Confusão - Random Forest



Fonte: Elaborado pelo autor.

Para compreender os fatores que priorizaram a decisão do modelo final, realizou-se a análise de importância das variáveis utilizando exclusivamente o *Random Forest*. Os resultados, exibidos na Figura 4, evidenciam que atributos acadêmicos exercem papel central na predição, com destaque absoluto para o coeficiente de rendimento médio. Outras variáveis, como forma de ingresso, tipo de cota e origem do aluno, também exibiram relevância significativa, sugerindo a influência conjunta de fatores institucionais e socioeconômicos. É importante destacar, contudo, que a importância calculada reflete a redução de impureza nas divisões das árvores, não implicando causalidade direta. Além disso, técnicas baseadas em árvores podem favorecer atributos com maior número de divisões possíveis, especialmente após o processo de *One-Hot Encoding*, o que recomenda cautela na interpretação dos valores absolutos.

Figura 4 – Features mais importantes para o modelo



Fonte: Elaborado pelo autor.

6 CONCLUSÃO

O presente trabalho atingiu seu objetivo principal ao desenvolver um modelo preditivo capaz de identificar, com antecedência, estudantes em risco de evasão no IFCE - campus Aracati. Diante de um cenário em que a taxa histórica de evasão alcança aproximadamente 54,6%, a aplicação da metodologia KDD (*Knowledge Discovery in Databases*) provou-se eficaz para transformar dados administrativos brutos em conhecimento estratégico, permitindo uma análise profunda do perfil discente.

Na etapa de mineração de dados, a estratégia exploratória com 14 algoritmos distintos garantiu que a escolha do modelo final não fosse enviesada. Após a otimização de hiperparâme-

tros, o algoritmo Random Forest consolidou-se como a solução mais robusta para o problema. O modelo alcançou um equilíbrio superior entre as métricas, com destaque para a Acurácia de 84% e, principalmente, um Recall de 85%. Este último indicador é crítico para o contexto educacional, pois demonstra a alta sensibilidade do modelo em detectar corretamente a classe minoritária e mais importante: o aluno que irá evadir, minimizando a ocorrência de falsos negativos.

Além da capacidade preditiva, o estudo proporcionou interpretabilidade sobre os fatores associados ao abandono. A análise de importância das variáveis evidenciou que o desempenho acadêmico (representado pelos coeficientes de rendimento) é o indicador mais forte de permanência ou evasão. Paralelamente, fatores como o curso de origem, com destaque para a alta evasão em Hotelaria, e dados socioeconômicos também exerceram influência significativa na decisão do algoritmo. Esses achados corroboram a complexidade do fenômeno da evasão, que entrelaça desempenho pedagógico com questões estruturais e sociais.

Em resumo, a ferramenta desenvolvida oferece ao IFCE um mecanismo poderoso para transitar de uma gestão reativa para uma abordagem proativa. Ao identificar precocemente os discentes em situação de risco, a instituição pode direcionar intervenções pedagógicas e políticas de assistência estudantil de forma mais assertiva e personalizada. Portanto, a integração deste modelo aos processos de gestão acadêmica representa um passo tecnológico viável e necessário para o fortalecimento das políticas de permanência e êxito estudantil.

7 TRABALHOS FUTUROS

Como desdobramento desta pesquisa, sugere-se o enriquecimento da base de dados por meio da incorporação de novas variáveis, como notas detalhadas por disciplina, frequência escolar e dados socioeconômicos mais aprofundados, visando refinar a precisão do modelo. Além disso, propõe-se o desenvolvimento de um protótipo funcional, como um *dashboard* de gestão pedagógica, que integre o modelo treinado para gerar alertas em tempo real sobre a situação dos discentes.

REFERÊNCIAS

ADNAN, M. et al. Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models. **IEEE Access**, v. 9, p. 7519–7539, 2021. ISSN 2169-3536. Disponível em: <https://ieeexplore.ieee.org/abstract/document/9314000>.

ALBA, F. D. Evasão acadêmica em uma instituição de ensino superior privada na Região Sul do Brasil: do diagnóstico à proposição de um programa de permanência. jan. 2018. Disponível em: <http://repositorio.jesuita.org.br/handle/UNISINOS/6924>.

BARBOSA, D. et al. Previsão da Evasão Escolar através da Análise de Dados e Aprendizagem de Máquina: Um estudo de caso. In: **Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil (WAPLA)**. SBC, 2023. p. 42–50. ISSN: 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/wapla/article/view/26127>.

BRASIL. **Evasão no ensino superior: conceitos e metodologias**. Brasília, 1997. Disponível em: <http://portal.mec.gov.br/>.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 17, n. 3, p. 37–37, mar. 1996. ISSN 2371-9621. Number: 3. Disponível em: <https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1230>.

FILHO, F. H.; SIQUEIRA, D.; LEAL, B. Predição de Evasão Utilizando Técnicas de Classificação: Um Estudo de Caso do Instituto Federal do Ceará. In: **Escola Regional de Computação do Ceará, Maranhão e Piauí (ERCEMAPI)**. SBC, 2020. p. 141–148. ISSN: 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/ercemapi/article/view/11478>.

IFCE em Números. **Portal de estatísticas institucionais do Instituto Federal do Ceará**. 2025. Pró-Reitoria de Ensino do Instituto Federal do Ceará. Disponível em: <https://ifceemnumeros.ifce.edu.br/>.

Instituto Semesp. **Mapa do Ensino Superior no Brasil 2024**. 2024. <https://www.semesp.org.br>. Acesso em: 30 abr. 2025. Disponível em: <https://www.semesp.org.br/wp-content/uploads/2024/04/mapa-do-ensino-superior-no-brasil-2024.pdf>.

MACHADO, D. F. M. G. et al. Avaliação de algoritmos de aprendizado de máquina na previsão de evasão escolar: estudo de caso no IFPA campus Altamira. **Cuadernos de Educación y Desarrollo**, v. 16, n. 10, p. e5874–e5874, out. 2024. ISSN 1989-4155. Disponível em: <https://ojs.cuadernoseducacion.com/ojs/index.php/ced/article/view/5874>.

PEREIRA, L. M.; LUZ, A. B.; LIMA, C. D. Fatores de evasão em instituições públicas. **Educação e Sociedade**, 2020.

RAJAMANI, S. K.; IYER, R. S. Machine Learning-Based Mobile Applications Using Python and Scikit-Learn:. In: SAMANTA, D. (Ed.). **Advances in Wireless Technologies and Telecommunication**. IGI Global, 2023. p. 282–306. ISBN 978-1-6684-8582-8 978-1-6684-8584-2. Disponível em: <https://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-6684-8582-8.ch016>.

SANTOS, J. C. B. D. **Usando mineração de dados para predição da evasão escolar**. [S.l.], 2021. Trabalho acadêmico. Disponível em: <https://repositorio.ifsc.edu.br/handle/123456789/1818>.

TEODORO, L. d. A.; KAPPEL, M. A. A. Aplicação de Técnicas de Aprendizado de Máquina para Predição de Risco de Evasão Escolar em Instituições Públicas de Ensino Superior no Brasil. **Revista Brasileira de Informática na Educação**, v. 28, n. 0, p. 838–863, nov. 2020. ISSN 2317-6121. Number: 0. Disponível em: <http://milanesa.ime.usp.br/rbie/index.php/rbie/article/view/v28p838>.

VIEIRA, A. C. F.; GALLINDO, E. de L.; CRUZ, H. A. **Plano estratégico para permanência e êxitos dos estudantes do IFCE**. 2017. Disponível em: <https://ifce.edu.br/proen/plano-de-permanencia-e-exito>.