

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Sávio Dias Araújo

**Mineração de dados educacionais para
identificação do risco de evasão: uma análise
comparativa de algoritmos de classificação no
curso de Sistemas de Informação da UFU**

Uberlândia, Brasil

2026

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Sávio Dias Araújo

**Mineração de dados educacionais para identificação do
risco de evasão: uma análise comparativa de algoritmos
de classificação no curso de Sistemas de Informação da
UFU**

Trabalho de conclusão de curso apresentado
à Faculdade de Computação da Universidade
Federal de Uberlândia, como parte dos requi-
sitos exigidos para a obtenção título de Ba-
charel em Sistemas de Informação.

Orientador: Paulo Henrique Ribeiro Gabriel

Universidade Federal de Uberlândia – UFU

Faculdade de Computação

Bacharelado em Sistemas de Informação

Uberlândia, Brasil

2026

Resumo

A evasão acadêmica constitui um desafio relevante para as instituições de ensino superior, especialmente nos cursos da área de Computação. Este trabalho tem como objetivo avaliar comparativamente o desempenho de algoritmos de classificação supervisionada na identificação de um indicador operacional de risco de evasão entre estudantes do curso de Bacharelado em Sistemas de Informação da Universidade Federal de Uberlândia. Foi utilizada uma base de dados de discentes da Faculdade de Computação, originalmente organizada por disciplina cursada. Após as etapas de limpeza, transformação e agregação, os registros passaram a representar o desempenho de cada estudante por ano e período letivo, considerando a quantidade de disciplinas cursadas, aprovações, reprovações e o Coeficiente de Rendimento Acadêmico. A variável de classe *evadiu* recebeu o valor 1 quando o número de reprovações foi superior ao de aprovações e o valor 0 nos demais casos, devendo ser interpretada como indicador de risco, e não como confirmação institucional de evasão. Foram avaliados os algoritmos Naive Bayes, J48, Random Forest, IBk, Logistic e JRip, cada um com duas parametrizações, por meio de validação cruzada estratificada com 10 partições. O Logistic obteve 100% de acurácia nas duas configurações; Random Forest, JRip e J48 superaram 99,9%; o IBk apresentou desempenho próximo de 100%; e o Naive Bayes obteve acurácia entre 95,7% e 97,0%. Entretanto, como a classe foi construída a partir de atributos também utilizados como preditores, os resultados elevados refletem a forte relação entre essa regra e a representação dos dados. Assim, a comparação identifica o Logistic como o classificador de melhor desempenho na configuração estudada, mas não comprova sua capacidade de prever a evasão institucional definitiva.

Palavras-chave: Mineração de Dados Educacionais. Evasão acadêmica. Aprendizado de máquina. Classificação supervisionada. Indicador de risco.

Abstract

Academic dropout is a relevant challenge for higher education institutions, especially in Computer Science programs. This study aims to comparatively evaluate the performance of supervised classification algorithms in identifying an operational indicator of dropout risk among students enrolled in the Bachelor's Degree in Information Systems at the Federal University of Uberlândia. A dataset comprising student records from the School of Computer Science was used; these records were originally organized by individual courses taken. After cleaning, transformation, and aggregation, each record represented a student's performance by academic year and semester, considering the number of courses taken, courses passed, courses failed, and the academic performance coefficient. The class variable *evadiu* was assigned the value 1 when the number of failed courses exceeded the number of passed courses and 0 otherwise; therefore, it represents a risk indicator rather than institutionally confirmed dropout. The Naive Bayes, J48, Random Forest, IBk, Logistic, and JRip algorithms were evaluated with two parameter settings each using stratified 10-fold cross-validation. Logistic achieved 100% accuracy in both settings; Random Forest, JRip, and J48 exceeded 99.9%; IBk achieved performance close to 100%; and Naive Bayes obtained accuracy between 95.7% and 97.0%. However, because the class was constructed from attributes that were also used as predictors, the high scores reflect the strong relationship between this rule and the data representation. Thus, the comparison identifies Logistic as the best-performing classifier under the investigated setup, but it does not demonstrate an ability to predict definitive institutional dropout.

Keywords: Educational Data Mining. Academic dropout. Machine learning. Supervised classification. Risk indicator.

Lista de tabelas

Tabela 1	– Grupos de atributos da base original DadosFACOM.	18
Tabela 2	– Atributos do conjunto de dados processado.	20
Tabela 3	– Resumo do conjunto de dados final utilizado nos experimentos.	20
Tabela 4	– Parametrizações testadas nos algoritmos de classificação	28
Tabela 5	– Resultados agregados da validação cruzada com 10 partições.	29
Tabela 6	– Resumo das matrizes de confusão considerando a classe 1 como positiva	31

Lista de abreviaturas e siglas

ABNT	Associação Brasileira de Normas Técnicas
CRA	Coeficiente de Rendimento Acadêmico
CSV	<i>Comma-Separated Values</i>
EDM	<i>Educational Data Mining</i>
UFU	Universidade Federal de Uberlândia
FACOM	Faculdade de Computação
FN	Falso Negativo
FP	Falso Positivo
IBk	<i>Instance-Based Learner</i>
IFES	Instituições Federais de Ensino Superior
J48	Implementação do algoritmo C4.5 na ferramenta Weka
JRip	Implementação do algoritmo RIPPER na ferramenta Weka
KNN	<i>K-Nearest Neighbors</i>
MD	Mineração de Dados
RF	<i>Random Forest</i>
VP	Verdadeiro Positivo
VN	Verdadeiro Negativo
Weka	<i>Waikato Environment for Knowledge Analysis</i>
CTIC	Centro de Tecnologia da Informação e Comunicação

Sumário

1	INTRODUÇÃO	8
2	REVISÃO BIBLIOGRÁFICA	10
2.1	Evasão e Retenção Acadêmica	10
2.2	Mineração de Dados Educacionais	11
2.3	Algoritmos de Classificação em Dados Educacionais	13
2.4	Trabalhos Relacionados	14
3	DESENVOLVIMENTO	17
3.1	Método de pesquisa	17
3.2	Pré-processamento de Dados	17
3.2.1	Origem, abrangência e estrutura da base	17
3.2.2	Seleção dos atributos e mudança de granularidade	18
3.2.3	Construção da classe e conjunto final	19
3.3	Algoritmos de Classificação	21
3.3.1	Introdução aos Algoritmos	21
3.3.2	Naive Bayes	21
3.3.3	J48	22
3.3.4	Random Forest	22
3.3.5	IBk	23
3.3.6	Logistic	24
3.3.7	JRip	24
3.4	Configuração dos Experimentos	25
4	RESULTADOS E DISCUSSÃO	29
4.1	Desempenho Geral	29
4.2	Efeito das Parametrizações	30
4.3	Análise das Matrizes de Confusão	30
4.4	Discussão dos Resultados	31
4.4.1	O que os resultados efetivamente medem	31
4.4.2	Construção de uma classe institucional	32
4.4.3	Limitações da comparação	33
5	CONCLUSÃO	34
	REFERÊNCIAS	36

APÊNDICES	39
APÊNDICE A – DECLARAÇÃO DE USO DE IA GENERATIVA . .	40

1 Introdução

A evasão acadêmica constitui um problema relevante para as instituições de ensino superior, pois interrompe trajetórias formativas e produz consequências acadêmicas, sociais e institucionais. Trata-se de um fenômeno multifatorial, relacionado tanto às características anteriores ao ingresso quanto às condições vivenciadas durante o curso. Evidências recentes indicam que a evasão ocorre com maior intensidade nos períodos iniciais e que fatores como desempenho acadêmico, apoio social e identificação com o curso estão associados à permanência do estudante (NIEROTKA; SALATA; MARTINS, 2023). Em cursos nos quais há elevada retenção e reprovação, o acompanhamento da trajetória acadêmica torna-se especialmente importante para a identificação precoce de situações de dificuldade.

No curso noturno de Sistemas de Informação da Universidade Federal de Uberlândia, Dutra (2019) identificou turmas com taxas de evasão superiores a 70% e destacou a relação entre reprovações e dificuldades de progressão acadêmica. Estudos brasileiros mais recentes também mostram que informações de desempenho dos primeiros períodos podem contribuir para identificar estudantes em maior risco, embora a evasão não deva ser explicada exclusivamente por notas ou reprovações (CHAGAS; GOMIDE; SANTANA, 2025). Esses resultados reforçam a importância de investigar o que pode ser extraído dos registros acadêmicos, sem desconsiderar as dimensões sociais, econômicas e institucionais do fenômeno.

Nesse contexto, a Mineração de Dados Educacionais permite organizar e analisar registros históricos para comparar métodos de classificação e investigar indicadores associados à trajetória discente. Revisões sistemáticas recentes mostram o uso recorrente de algoritmos de aprendizado de máquina em estudos de retenção e evasão, incluindo árvores de decisão, métodos probabilísticos, regressão logística, métodos baseados em vizinhança e métodos de conjunto (SHAFIQ et al., 2022; ANDRADE-GIRÓN et al., 2023). Esses estudos também evidenciam que o desempenho obtido depende da definição da variável de classe, dos atributos disponíveis e da estratégia de validação, de modo que resultados elevados precisam ser interpretados à luz da configuração experimental adotada.

Diante disso, este trabalho tem como objetivo avaliar comparativamente o desempenho dos algoritmos Naive Bayes, J48, Random Forest, IBk, Logistic e JRip na classificação de um indicador operacional de risco de evasão entre estudantes do curso de Bacharelado em Sistemas de Informação da Universidade Federal de Uberlândia. A análise utiliza a base institucional DadosFACOM, fornecida diretamente pelo Centro de Tecnologia da Informação e Comunicação da UFU já anonimizada. A base não foi obtida

de um repositório público e contém registros acadêmicos originalmente organizados por disciplina cursada. Após o pré-processamento, os dados foram agregados por aluno, ano e período letivo, utilizando como atributos preditores o ano, o período, a quantidade de disciplinas cursadas, as aprovações, as reprovações e o Coeficiente de Rendimento Acadêmico.

A variável de classe *evadiu* não corresponde à confirmação de que o estudante abandonou o curso. Ela foi construída como um indicador binário, recebendo o valor 1 quando a quantidade de reprovações no período foi superior à quantidade de aprovações e o valor 0 nos demais casos. Como aprovações e reprovações também foram utilizadas como atributos preditores, a tarefa experimental mede, em grande parte, a capacidade dos classificadores de reproduzir essa regra operacional a partir dos dados de entrada. Portanto, neste trabalho, classificar corretamente a variável *evadiu* significa identificar a categoria definida por esse critério acadêmico, e não prever diretamente a evasão institucional futura.

Os resultados refletiram essa relação direta entre os atributos e a classe. O Logistic alcançou 100% de acurácia nas duas parametrizações, enquanto Random Forest, JRip e J48 superaram 99,9%; o IBk apresentou desempenho próximo de 100%; e o Naive Bayes obteve acurácia entre 95,7% e 97,0%. Assim, os valores elevados permitem comparar o comportamento dos algoritmos dentro da configuração proposta, mas não demonstram que os modelos estejam prontos para prever a evasão real. Uma aplicação institucional exigiria uma classe oficial de evasão, definida independentemente dos atributos preditores, além de uma estratégia de validação que separasse os registros por estudante e por período temporal.

Este trabalho está organizado da seguinte forma: o Capítulo 2 apresenta a revisão bibliográfica, abordando conceitos relacionados à evasão acadêmica, Mineração de Dados Educacionais e trabalhos relacionados. O Capítulo 3 descreve o desenvolvimento da pesquisa, incluindo o método adotado, o pré-processamento dos dados, os algoritmos utilizados e a configuração dos experimentos. O Capítulo 4 apresenta os resultados e sua discussão. Por fim, o Capítulo 5 apresenta as conclusões do trabalho e possibilidades de continuidade.

2 Revisão Bibliográfica

2.1 Evasão e Retenção Acadêmica

A evasão acadêmica é um fenômeno relacionado à interrupção da trajetória formativa do estudante antes da conclusão do curso. No ensino superior, esse problema assume relevância por envolver impactos institucionais, sociais e econômicos, uma vez que a saída do aluno representa perda de investimento público ou privado, redução da eficiência acadêmica e possíveis prejuízos à formação profissional. Em cursos de graduação, a evasão pode estar associada a fatores diversos, como dificuldades de adaptação, desempenho acadêmico insuficiente, reprovações, condições socioeconômicas, distância da instituição, perfil do estudante e aspectos institucionais (DUTRA, 2019; MÁXIMO et al., 2024).

A retenção acadêmica, por sua vez, refere-se à permanência prolongada do estudante no curso, geralmente associada ao atraso na integralização da matriz curricular. Embora retenção e evasão não representem o mesmo fenômeno, elas estão relacionadas, pois sucessivas reprovações e dificuldades de progressão podem aumentar o risco de abandono. Já a permanência está associada à continuidade do estudante no curso e à possibilidade de concluir sua formação, envolvendo condições materiais, pedagógicas e institucionais que favoreçam sua trajetória acadêmica (ARAÚJO, 2013; ARAÚJO; MARIANO; OLIVEIRA, 2021; MÁXIMO et al., 2024).

A literatura brasileira destaca que a evasão no ensino superior é um fenômeno complexo e multifatorial. Baggi e Lopes (2011) ressaltam que a evasão não deve ser analisada de forma isolada ou reduzida apenas à falta de recursos financeiros, pois envolve fatores sociais, acadêmicos, econômicos e institucionais. De forma complementar, Maciel, Júnior e Lima (2019) observam que as discussões sobre evasão e permanência ganharam relevância no Brasil especialmente no contexto de expansão do ensino superior, uma vez que essas temáticas auxiliam na formulação de políticas de acompanhamento, permanência e redução do abandono acadêmico.

No Brasil, os estudos sobre evasão no ensino superior passaram a receber maior atenção a partir da década de 1990, especialmente com a criação da Comissão Especial de Estudos sobre Evasão, vinculada à Secretaria de Educação Superior do Ministério da Educação. A literatura também destaca modelos teóricos anteriores, como o modelo de Tinto, que discute a influência de fatores individuais, sociais e institucionais na permanência ou abandono dos estudantes (MEIRELES, 2019). Esse contexto evidencia que a evasão deve ser analisada como resultado de múltiplas dimensões, e não como consequência de uma única causa.

Nas últimas décadas, o ensino superior brasileiro passou por transformações relevantes, incluindo ampliação do acesso, expansão de vagas e mudanças no perfil socioeconômico dos estudantes. Políticas como o Reuni, o Enem, o SiSU, o Pnaes e a Lei de Cotas contribuíram para democratizar o ingresso no ensino superior, mas também reforçaram a necessidade de políticas institucionais voltadas à permanência e ao acompanhamento acadêmico (NIEROTKA; SALATA; MARTINS, 2023). Nesse cenário, métodos de análise de dados podem apoiar a identificação de padrões associados à retenção e ao risco de evasão.

No curso noturno de Sistemas de Informação da Universidade Federal de Uberlândia, a evasão já foi identificada como um problema expressivo, com turmas que apresentaram taxas superiores a 70% (DUTRA, 2019). Embora fatores sociais, econômicos e institucionais também participem desse fenômeno, há evidências de que informações acadêmicas possuem relevância para sua análise. Em estudo longitudinal, Nierotka, Salata e Martins (2023) observaram que estudantes com melhor desempenho antes e durante o curso apresentaram menor risco de evasão. De modo semelhante, Chagas, Gomide e Santana (2025) verificaram que atributos do primeiro período, especialmente os relacionados ao rendimento em disciplinas, permitiram identificar alunos evadidos com AUC próxima de 0,80. No contexto internacional, Palacios et al. (2021) também obtiveram resultados superiores a 80% em diversos cenários que combinaram informações de ingresso, características socioeconômicas e desempenho universitário.

Essas evidências não indicam que atributos acadêmicos sejam suficientes para explicar toda a evasão, mas mostram que eles podem fornecer sinais relevantes e mensuráveis da trajetória discente. Por essa razão, o presente estudo delimita sua análise aos atributos disponíveis na base DadosFACOM, como ano, período letivo, quantidade de disciplinas, aprovações, reprovações e Coeficiente de Rendimento Acadêmico. Essa delimitação permite avaliar a informação contida nos registros acadêmicos, mas impede que os resultados sejam generalizados como uma explicação completa ou causal da evasão institucional.

2.2 Mineração de Dados Educacionais

A Mineração de Dados Educacionais, ou *Educational Data Mining* (EDM), é uma área de pesquisa voltada à aplicação de técnicas de mineração de dados em bases educacionais, com o objetivo de identificar padrões, relações e informações úteis para a compreensão de processos de ensino e aprendizagem. No contexto do ensino superior, essas técnicas podem ser utilizadas para analisar registros acadêmicos, históricos de desempenho, disciplinas cursadas, aprovações, reprovações, retenção e evasão. Dessa forma, a EDM permite transformar dados acadêmicos em conhecimento que pode apoiar tanto a tomada de decisão institucional quanto a criação de estratégias de acompanhamento

estudantil (BAKER; YACEF, 2009; BAKER; ISOTANI; CARVALHO, 2011; ROMERO; VENTURA, 2013).

Entre as principais aplicações da EDM estão a previsão de desempenho acadêmico, a identificação de estudantes em situação de risco e a predição de evasão. Agrusti, Bonavolontà e Mezzini (2019) destacam que a aplicação de técnicas de mineração de dados à evasão universitária busca extrair informações úteis de grandes conjuntos de dados educacionais, permitindo reconhecer padrões associados ao abandono e subsidiar ações de acompanhamento acadêmico. Nesse sentido, a evasão acadêmica pode ser tratada como um problema de classificação supervisionada, no qual modelos computacionais são treinados a partir de dados históricos para distinguir registros associados a maior ou menor risco de evasão.

A literatura recente mostra que técnicas supervisionadas de aprendizado de máquina são frequentemente utilizadas em problemas de retenção e evasão no ensino superior. Revisões sistemáticas apontam o uso recorrente de algoritmos como árvores de decisão, classificadores bayesianos, regressão logística, métodos baseados em vizinhança e métodos de conjunto, como Random Forest (SHAFIQ et al., 2022; ANDRADE-GIRÓN et al., 2023). Esses estudos também indicam que métricas derivadas da matriz de confusão, como acurácia, precisão, revocação e *F-Measure*, são amplamente utilizadas para avaliar o desempenho dos modelos preditivos.

Neste trabalho, a Mineração de Dados Educacionais é utilizada para analisar registros acadêmicos da base de dados de discentes da FACOM e comparar o desempenho de algoritmos de classificação supervisionada. A proposta não busca explicar todos os fatores sociais, econômicos ou institucionais envolvidos na evasão, mas avaliar como diferentes classificadores reproduzem o indicador operacional definido a partir do desempenho acadêmico. Assim, a EDM é empregada como abordagem metodológica para apoiar a análise experimental e a comparação entre classificadores.

O desempenho global dos modelos, contudo, não informa isoladamente quanto cada atributo contribuiu para as decisões de classificação. Uma análise complementar poderia empregar medidas de importância, seleção de atributos ou experimentos de ablação, nos quais os modelos seriam executados após a retirada de grupos de variáveis. Esse procedimento permitiria examinar o papel de atributos como ano, período, aprovações, reprovações e CRA e verificar em que medida os resultados dependem de cada um deles. Como essa etapa não foi realizada no presente estudo, ela é indicada como possibilidade de continuidade da pesquisa.

2.3 Algoritmos de Classificação em Dados Educacionais

Algoritmos de classificação são métodos de aprendizado supervisionado utilizados para atribuir registros a classes previamente definidas. Em problemas educacionais, esses algoritmos podem ser empregados para prever categorias como aprovação, reprovação, desempenho satisfatório, baixo desempenho, retenção ou risco de evasão. Para isso, os modelos são treinados a partir de exemplos históricos, nos quais os atributos de entrada representam características dos estudantes ou de sua trajetória acadêmica, e a classe representa o resultado que se deseja prever (FILHO; VINUTO; LEAL, 2020; LUDERMIR, 2021).

No contexto da Mineração de Dados Educacionais, a comparação entre classificadores é uma estratégia recorrente, pois o desempenho dos modelos pode variar conforme a estrutura da base de dados, os atributos disponíveis, a distribuição das classes e a estratégia de avaliação utilizada. Estudos da área indicam o uso frequente de diferentes famílias de algoritmos, como métodos probabilísticos, árvores de decisão, métodos baseados em vizinhança, modelos estatísticos, classificadores baseados em regras e métodos de conjunto (DUTRA, 2019; ANDRADE-GIRÓN et al., 2023).

Neste trabalho, foram selecionados seis algoritmos principais: Naive Bayes, J48, Random Forest, IBk, Logistic e JRip. Essa seleção permite comparar abordagens com diferentes formas de aprendizado e interpretação. O Naive Bayes representa os métodos probabilísticos; o J48 representa os algoritmos baseados em árvores de decisão; o Random Forest representa os métodos de conjunto; o IBk representa os métodos baseados em instâncias e vizinhança; o Logistic representa os modelos estatísticos aplicados à classificação; e o JRip representa os classificadores baseados em regras (BERRAR, 2019; QUINLAN, 1993; LIAW; WIENER, 2002; CUNNINGHAM; DELANY, 2007; BEWICK; CHEEK; BALL, 2005; COHEN, 1995).

A escolha desses algoritmos também encontra apoio em estudos aplicados à previsão de desempenho acadêmico, retenção e evasão. Hussain et al. (2018) utilizaram o Weka para analisar o desempenho acadêmico de estudantes com dados acadêmicos, socioeconômicos e demográficos, comparando classificadores como J48, PART, Random Forest e BayesNet, com destaque para o desempenho do Random Forest. Jedidi et al. (2024) compararam regressão logística, árvore de decisão, Random Forest e Naive Bayes na previsão de desempenho acadêmico, utilizando divisão percentual e validação cruzada em 10 partes, mostrando que os resultados podem variar conforme os atributos utilizados e a estratégia de validação. Além disso, revisões recentes sobre predição de evasão indicam a presença recorrente de Random Forest, árvores de decisão, regressão logística, Naive Bayes, KNN e classificadores baseados em regras em estudos de classificação educacional (ANDRADE-GIRÓN et al., 2023).

Dessa forma, a seleção dos algoritmos adotados busca equilibrar diversidade metodológica e recorrência na literatura de Mineração de Dados Educacionais. Essa variedade permite avaliar diferentes formas de representação do conhecimento e comparar o desempenho dos classificadores na identificação de padrões associados ao risco de evasão acadêmica.

2.4 Trabalhos Relacionados

[Manhães e Cruz \(2019\)](#) propuseram um estudo voltado ao apoio de gestores acadêmicos na previsão do desempenho de estudantes e na identificação de risco de evasão por meio da arquitetura WAVE, baseada em Mineração de Dados Educacionais. A pesquisa utilizou dados acadêmicos de cursos da Universidade Federal do Rio de Janeiro ao longo de 16 anos e aplicou algoritmos classificadores para identificar aqueles com maior precisão na previsão do abandono acadêmico. Os resultados indicaram a viabilidade da abordagem proposta, com destaque para o algoritmo Naive Bayes. Esse estudo apresenta relação com a presente pesquisa por utilizar dados acadêmicos e algoritmos classificadores para apoiar a identificação de estudantes em risco.

[Dutra \(2019\)](#) investigou fatores associados à reprovação e à evasão no curso de Bacharelado em Sistemas de Informação da Universidade Federal de Uberlândia. O trabalho aplicou técnicas de mineração de dados com o auxílio do Weka, utilizando atributos sociais, econômicos e acadêmicos dos estudantes. Entre os algoritmos avaliados, destacam-se KNN, JRIP e Naive Bayes, sendo o JRIP apontado como um dos modelos com melhor desempenho na identificação de alunos em risco. A relação com esta pesquisa é direta, pois ambos os trabalhos estão inseridos no contexto do curso de Sistemas de Informação da UFU e utilizam técnicas de classificação para analisar padrões relacionados ao desempenho acadêmico e ao risco de evasão. A principal diferença é que o presente estudo concentra-se na base de dados de discentes da FACOM e em atributos acadêmicos agregados por aluno, ano e período letivo.

[Silva \(2023\)](#) analisou a previsão da evasão acadêmica em cursos de graduação presencial da Universidade de Brasília. A metodologia incluiu uma revisão sistemática da literatura para identificar fatores associados à evasão e a aplicação de algoritmos de aprendizado de máquina com base em dados acadêmicos disponíveis nos sistemas institucionais. Os resultados apontaram a relevância de indicadores como pontuação média, gênero e notas, além de comparar diferentes modelos preditivos. Esse estudo relaciona-se com a presente pesquisa por tratar da previsão de evasão no ensino superior e por utilizar algoritmos de aprendizado de máquina para identificar estudantes com maior risco de abandono.

Entre os trabalhos recentes relacionados à previsão de evasão no ensino superior,

Chagas, Gomide e Santana (2025) investigaram o risco de evasão universitária nos cursos de Engenharia da UFRJ-Macaé, no Brasil. O estudo utilizou técnicas de ciência de dados e aprendizado de máquina com base na metodologia CRISP-DM, treinando modelos de Regressão Logística, Árvore de Decisão e XGBoost. Os autores utilizaram atributos relacionados ao desempenho acadêmico dos estudantes, especialmente informações do primeiro período, e observaram que esses dados foram suficientes para identificar alunos evadidos com AUC próxima de 0,80. Além disso, a análise exploratória revelou que 75% das evasões ocorreram até o terceiro semestre e que houve forte relação entre baixo desempenho acadêmico inicial e risco de evasão, com destaque para disciplinas como Cálculo e Física. Esse trabalho se aproxima da presente pesquisa por utilizar dados acadêmicos institucionais e por tratar a evasão como um problema de classificação, embora empregue algoritmos diferentes e concentre a análise em cursos de Engenharia.

No contexto internacional, Palacios et al. (2021) desenvolveram um estudo de caso no Chile voltado à retenção de estudantes no ensino superior a partir de técnicas de mineração de dados e aprendizado de máquina. O trabalho utilizou a metodologia KDD e avaliou algoritmos como árvores de decisão, KNN, regressão logística, Naive Bayes, Random Forest e máquinas de vetores de suporte, implementados no software Weka. Os autores propuseram modelos para prever a retenção ou evasão em diferentes momentos da trajetória acadêmica, considerando variáveis socioeconômicas, desempenho escolar anterior, dados de ingresso e desempenho universitário. Os resultados indicaram acurácia superior a 80% na maioria dos cenários avaliados, com melhor desempenho geral do Random Forest. O estudo também destacou a importância de variáveis como desempenho na educação secundária e índice de pobreza da comunidade de origem do estudante.

Tedeschi (2025) investigou a previsão de evasão estudantil no curso de Bacharelado em Ciência da Computação da Faculdade de Computação da Universidade Federal de Uberlândia. O estudo utilizou uma base institucional com registros acadêmicos de aproximadamente 1.500 estudantes, abrangendo um período de 27 anos, e concentrou a análise no desempenho dos discentes durante os três primeiros semestres do curso. Após as etapas de limpeza, transformação e agregação dos dados, foram utilizados como atributos preditivos os valores da Média Geral Acadêmica obtidos em cada um desses semestres, enquanto a variável de classe representava a situação final do estudante como evadido ou formado. A autora comparou os algoritmos Floresta Aleatória, Árvore de Decisão, Máquina de Vetores de Suporte, Regressão Logística, Regressão Linear, Naive Bayes e Perceptron Multicamadas, considerando também diferentes estratégias de balanceamento das classes, como sobreamostragem aleatória, SMOTE, ponderação de classes e manutenção da distribuição original. Os resultados indicaram bom desempenho geral dos classificadores, com destaque para o Perceptron Multicamadas, o Naive Bayes e a Floresta Aleatória, embora as diferenças entre os métodos e as técnicas de balanceamento tenham sido pouco expressivas.

A relação com o presente estudo está no uso de dados acadêmicos da FACOM e na comparação de classificadores no contexto da evasão. As diferenças metodológicas mais relevantes, entretanto, estão na unidade de análise, nos atributos e na definição da classe. [Tedeschi \(2025\)](#) analisou estudantes de Ciência da Computação a partir do desempenho nos três primeiros semestres e utilizou como classe a situação institucional final de evadido ou formado, além de avaliar estratégias de balanceamento. O presente trabalho analisa registros de Sistemas de Informação agregados por aluno, ano e período, utiliza contagens de aprovações e reprovações e adota uma classe operacional derivada desses mesmos atributos. Consequentemente, os resultados dos dois trabalhos respondem a perguntas distintas e não devem ser comparados apenas pelos valores das métricas obtidas.

Esses trabalhos reforçam a relevância da Mineração de Dados Educacionais na análise da permanência estudantil, mas também mostram que a comparação entre pesquisas deve considerar mais do que o algoritmo ou a ferramenta utilizada. As principais diferenças estão na população analisada, na unidade de representação dos registros, nos atributos disponíveis, na definição da classe, nas estratégias de balanceamento e no procedimento de validação. Enquanto parte dos estudos relacionados utiliza a situação institucional final de evasão e incorpora variáveis socioeconômicas, de ingresso ou da trajetória completa, esta pesquisa trabalha com registros por aluno, ano e período e compara classificadores em relação a um indicador operacional construído a partir de aprovações e reprovações. Dessa forma, sua contribuição está na comparação experimental dos algoritmos dentro dessa representação específica, e não na apresentação de um modelo já validado para prever a evasão institucional.

3 Desenvolvimento

3.1 Método de pesquisa

Esta pesquisa caracteriza-se como um estudo aplicado, quantitativo e experimental. Seu objetivo é comparar o desempenho de seis algoritmos de classificação supervisionada sobre uma mesma representação de dados acadêmicos e sob uma configuração comum de avaliação. Portanto, a unidade principal de comparação é o algoritmo, e não a estimativa da influência individual de cada atributo nem a explicação causal da evasão.

O procedimento foi organizado em quatro etapas: (i) inspeção da origem, da estrutura e da qualidade dos dados; (ii) seleção manual dos campos necessários e transformação da granularidade de atividade curricular para aluno-ano-período; (iii) construção da variável de classe e preparação do conjunto experimental; e (iv) execução dos classificadores no Weka e comparação por métricas de classificação. Os detalhes da origem dos dados e das transformações são apresentados na Seção 3.2; os algoritmos e a configuração experimental são descritos nas seções seguintes.

É importante delimitar o que o experimento mede. A variável de classe foi definida por uma regra conhecida que compara reprovações e aprovações do próprio período, e esses dois valores também permanecem entre os preditores. Assim, a tarefa avalia a capacidade dos algoritmos de reproduzir esse indicador operacional. Ela não constitui uma predição temporal de evasão institucional, não confirma que um estudante abandonou o curso e não inclui análise de importância dos atributos.

3.2 Pré-processamento de Dados

3.2.1 Origem, abrangência e estrutura da base

A base DadosFACOM foi fornecida pelo Centro de Tecnologia da Informação e Comunicação (CTIC) da UFU já anonimizada. O arquivo principal recebido, *DadosFACOM.xlsx*, contém uma planilha denominada *Dados*, com 174.667 registros acadêmicos e 33 atributos. Cada linha representa uma atividade curricular vinculada a um identificador de estudante em determinado ano e período letivo; portanto, o arquivo não chegou agregado por aluno ou em uma tabela pronta para classificação.

Os registros do arquivo bruto abrangem três ofertas de graduação: Ciência da Computação, bacharelado integral; Sistemas de Informação, bacharelado noturno; e Sistemas de Informação, bacharelado integral no campus Monte Carmelo. A cobertura acadêmica

observada concentra-se entre 1995 e 2023. O arquivo também contém dois códigos de ano anômalos, 200 e 212, com um registro cada. A data exata em que o CTIC realizou a extração não consta no material recebido; por isso, o intervalo informado neste trabalho se refere aos anos letivos registrados, e não à data de coleta.

Foram encontrados 1.035 códigos e 734 nomes distintos de atividades curriculares. Esse conjunto também inclui ocorrências administrativas, como trancamento total e vínculo. Como não foi fornecido um catálogo externo das matrizes curriculares vigentes em cada ano, não é possível afirmar que todas as disciplinas de todas as versões curriculares estejam presentes. No pré-processamento, foram consideradas todas as ocorrências avaliadas disponíveis no extrato; não foi realizada uma verificação de completude contra os currículos oficiais.

A Tabela 1 resume os grupos de informações presentes nas 33 colunas originais. A base possui uma estrutura tabular única, mas reúne, na mesma linha, identificadores técnicos, dados temporais, informações curriculares, desempenho acadêmico e situação final do vínculo.

Tabela 1 – Grupos de atributos da base original DadosFACOM.

Grupo	Atributos e finalidade no arquivo original
Identificação	Identificador, ID_CURRIC_ALUNO, ID_ATIV_CURRIC, ID_ESTRUTURA_CUR e outros códigos internos.
Temporal	ANO, PERIODO_ITEM, PERIODO, ANO_INGRESSO e PERIODO_INGRESSO.
Curricular	Códigos e nomes de atividade, curso e unidade, versão e descrição da estrutura, créditos e cargas horárias.
Desempenho	MEDIA_FINAL, SITUACAO_ITEM, SITUACAO, CONCEITO e ID_NOTA.
Vínculo e curso	FORMA_EVASAO, nome do curso de diploma, modalidade e metadados de reconhecimento.

3.2.2 Seleção dos atributos e mudança de granularidade

A seleção de atributos foi manual e orientada pela construção da unidade aluno–ano–período; não foi aplicado um método estatístico de seleção de características. Foram mantidos: (i) o identificador anonimizado, o ano e o período, necessários ao agrupamento; (ii) a situação da atividade, necessária para contar aprovações e reprovações; e (iii) a média final, usada no cálculo da média do período. Os demais campos não foram utilizados como preditores porque eram identificadores técnicos, descrições redundantes, variáveis muito específicas da disciplina ou metadados não contemplados pela representação agregada escolhida.

O atributo original FORMA_EVASAO também não foi empregado como preditor nem transformado na classe deste experimento. Ele descreve uma situação final do vínculo,

como formado, abandono ou desistência oficial, que pode ser conhecida depois dos períodos acadêmicos usados como entrada. Seu uso direto como preditor introduziria informação posterior ao momento de interesse. A construção de uma classe institucional independente a partir desse campo exigiria alinhar temporalmente cada trajetória e definir quais formas de desligamento representam evasão, procedimento que não foi realizado neste trabalho.

A base original era incompatível com a unidade de análise proposta porque um mesmo contexto semestral aparecia fragmentado em várias linhas, uma para cada atividade. Se os classificadores fossem aplicados diretamente nessa estrutura, classificariam atividades isoladas, e não o resumo do estudante no período. Para resolver essa incompatibilidade, os registros foram agrupados pela combinação *id*, *ano* e *periodo*.

As situações acadêmicas foram interpretadas a partir da correspondência entre `SITUACAO_ITEM` e `SITUACAO`. Aprovações, reprovações por nota e reprovações por frequência contribuíram para as respectivas contagens. Situações sem resultado acadêmico comparável, como matrícula em andamento, trancamento, dispensa, equivalência e vínculo, não foram contabilizadas como aprovação ou reprovação. Os valores numéricos foram preparados com ponto como separador decimal.

Para cada grupo aluno–ano–período, foram produzidas a quantidade total de registros de atividades (*disciplinas*), as contagens de atividades aprovadas (*aprovadas*) e reprovadas (*reprovadas*) e a média das notas finais válidas, denominada *cra* no arquivo processado. Foram removidos grupos com zero atividades ou sem média válida. Essa transformação reduz as 174.667 linhas originais para 30.549 instâncias agregadas.

A agregação evita que cada disciplina seja tratada como uma observação independente do estudante, mas elimina informação: as contagens não preservam quais disciplinas foram cursadas, suas cargas horárias, dificuldade, posição na matriz curricular, pré-requisitos ou a sequência completa da trajetória. Assim, estudantes com as mesmas contagens podem ter históricos qualitativamente diferentes.

3.2.3 Construção da classe e conjunto final

A variável *evadiu* foi construída depois da agregação, segundo a regra apresentada na Equação (3.1).

$$evadiu = \begin{cases} 1, & \text{se } reprovadas > aprovadas; \\ 0, & \text{caso contrário.} \end{cases} \quad (3.1)$$

Embora seja armazenada com os rótulos 0 e 1, a classe foi declarada como nominal e binária no Weka. O valor 1 indica que, naquele período, as reprovações superaram as aprovações; o valor 0 reúne todos os demais casos, inclusive empates. Esses valores não representam graus de risco e não correspondem a uma confirmação institucional de

abandono. A conferência do arquivo final mostrou que todas as 30.549 instâncias obedecem à regra da Equação (3.1).

Como *aprovadas* e *reprovadas* permanecem entre os preditores, a classe é deterministicamente derivada de duas entradas do modelo. Essa dependência explica por que a tarefa pode produzir resultados próximos de 100% e restringe a conclusão à reprodução do indicador. Para uma predição futura de evasão, seria necessário utilizar uma classe institucional definida independentemente dos preditores e respeitar a ordem temporal dos dados.

A Tabela 2 detalha o tratamento dos atributos do conjunto final. Todos os preditores foram importados como numéricos. Em particular, *periodo* foi mantido com os códigos administrativos 100, 201, 202, 601 e 602; portanto, o experimento atribui a esses códigos uma representação numérica, embora a distância entre eles não corresponda necessariamente a intervalos acadêmicos equivalentes.

Tabela 2 – Atributos do conjunto de dados processado.

Atributo	Tipo	Construção e uso
<i>id</i>	Numérico	Identificador anonimizado usado no agrupamento e na conferência; excluído dos classificadores.
<i>ano</i>	Numérico	Ano letivo; valores de 1995 a 2023, além de uma instância anômala com valor 212.
<i>periodo</i>	Numérico	Código do período: 100, 201, 202, 601 ou 602.
<i>disciplinas</i>	Numérico	Número de registros de atividades no grupo; variação de 1 a 62.
<i>aprovadas</i>	Numérico	Contagem de aprovações; variação de 0 a 19.
<i>reprovadas</i>	Numérico	Contagem de reprovações; variação de 0 a 12.
<i>cra</i>	Numérico	Média das notas finais válidas no grupo; variação de 0 a 100.
<i>evadiu</i>	Nominal	Classe $\{0,1\}$, definida pela Equação (3.1).

A Tabela 3 apresenta um resumo reproduzível da base utilizada nos experimentos. Dos 4.245 estudantes, 3.769 possuem mais de uma linha, pois cada período letivo gera uma instância própria. O arquivo final possui, em média, 7,20 registros por estudante e máximo de 31.

Tabela 3 – Resumo do conjunto de dados final utilizado nos experimentos.

Característica	Valor
Unidade de análise	aluno-ano-período
Número de instâncias	30.549
Número de estudantes distintos	4.245
Estudantes com múltiplas instâncias	3.769
Preditores utilizados	6
Classe 0	21.983 (71,96%)
Classe 1	8.566 (28,04%)
Cobertura acadêmica predominante	1995–2023

Por fim, o conjunto foi salvo em CSV, com nomes simplificados e separador por vírgula. O atributo *id* foi preservado somente para rastreabilidade do pré-processamento e removido antes do treinamento. A instância anômala com ano 212 permaneceu no arquivo experimental; sua presença é registrada aqui como limitação de qualidade dos dados, sem atribuir a ela significado temporal.

3.3 Algoritmos de Classificação

3.3.1 Introdução aos Algoritmos

Foram selecionados os algoritmos Naive Bayes, J48, Random Forest, IBk, Logistic e JRip. O conjunto abrange abordagens probabilística, baseada em árvore, baseada em instâncias, estatística, de conjunto e baseada em regras. A diversidade foi usada para comparar as métricas de classificação obtidas sob a mesma base e o mesmo protocolo experimental.

O experimento não mediu tempo de execução, custo computacional, interpretabilidade dos modelos, importância ou sensibilidade dos atributos e tampouco realizou testes específicos de generalização externa. Assim, a seleção não sustenta comparações sobre esses aspectos; as conclusões entre algoritmos se restringem às métricas e às parametrizações efetivamente reportadas.

3.3.2 Naive Bayes

O algoritmo Naive Bayes é um classificador probabilístico baseado no teorema de Bayes, utilizado para estimar a probabilidade posterior de uma classe a partir dos atributos de entrada. De forma geral, o classificador estima a probabilidade de uma classe C_k dado um conjunto de atributos $x = (x_1, x_2, \dots, x_n)$, conforme apresentado na Equação (3.2) (BERRAR, 2019).

$$P(C_k | x) = \frac{P(C_k) \prod_{i=1}^n P(x_i | C_k)}{P(x)} \quad (3.2)$$

Nessa formulação, $P(C_k | x)$ representa a probabilidade posterior da classe C_k , $P(C_k)$ representa a probabilidade prévia da classe, $P(x_i | C_k)$ representa a probabilidade condicional de cada atributo em relação à classe e $P(x)$ corresponde à probabilidade dos atributos observados.

Sua principal característica é a suposição de independência condicional entre os atributos, o que simplifica os cálculos e torna o algoritmo computacionalmente eficiente. Esse método é amplamente utilizado em problemas de classificação, especialmente em cenários com dados de alta dimensionalidade. Entre suas principais vantagens destacam-se

a simplicidade, a rapidez de execução e a boa capacidade de generalização, mesmo na presença de atributos irrelevantes. Por outro lado, sua principal limitação está na suposição de independência entre os atributos, que pode não ser válida em muitos problemas reais, impactando negativamente o desempenho do modelo (BERRAR, 2019).

3.3.3 J48

O algoritmo J48 é uma implementação do algoritmo C4.5, proposto por Quinlan (1993). Trata-se de um método baseado em árvores de decisão, no qual o conjunto de dados é dividido recursivamente a partir dos atributos que melhor separam as classes. O modelo resultante possui estrutura hierárquica, formada por nós de decisão, ramos e folhas, o que facilita sua interpretação. Em árvores de decisão, a escolha dos atributos pode ser baseada em medidas de impureza ou ganho de informação. Uma medida comumente utilizada é a entropia, apresentada na Equação (3.3).

$$Entropia(S) = - \sum_{i=1}^c p_i \log_2(p_i) \quad (3.3)$$

Na Equação (3.3), S representa o conjunto de instâncias, c corresponde ao número de classes e p_i representa a proporção de instâncias pertencentes à classe i . Quanto menor a entropia, maior a homogeneidade do conjunto em relação às classes.

O ganho de informação de um atributo A pode ser representado pela Equação (3.4).

$$Ganho(S, A) = Entropia(S) - \sum_{v \in \mathcal{V}(A)} \frac{|S_v|}{|S|} Entropia(S_v) \quad (3.4)$$

Nesse caso, $\mathcal{V}(A)$ representa o conjunto de valores possíveis do atributo A , e S_v representa o subconjunto de instâncias de S para determinado valor v desse atributo. Atributos com maior ganho de informação tendem a ser mais úteis para separar as classes durante a construção da árvore. Além disso, o C4.5 permite trabalhar com atributos nominais e numéricos, bem como aplicar técnicas de poda para reduzir o risco de sobreajuste aos dados de treinamento (QUINLAN, 1993).

3.3.4 Random Forest

O Random Forest é um método de conjunto que combina diversas árvores de decisão para produzir uma classificação final. Cada árvore é construída a partir de amostras dos dados de treinamento, e em cada divisão apenas um subconjunto aleatório de atributos é considerado. A decisão final é obtida, em problemas de classificação, por votação entre as árvores (LIAW; WIENER, 2002).

A classificação final do Random Forest pode ser representada como uma votação majoritária entre as árvores do conjunto, conforme apresentado na Equação (3.5).

$$\hat{y} = \text{moda}\{h_1(x), h_2(x), \dots, h_T(x)\} \quad (3.5)$$

Na Equação (3.5), $\hat{y}$ representa a classe prevista, $h_t(x)$ representa a predição da árvore t para a instância x e T corresponde ao número total de árvores do conjunto. Dessa forma, a classe final é definida pela classe mais frequentemente prevista pelas árvores individuais.

Essa estratégia tende a reduzir a variância do modelo e melhorar sua capacidade de generalização em comparação com uma única árvore de decisão. Além disso, o método pode fornecer medidas de importância dos atributos, embora seja menos interpretável do que modelos individuais, como árvores simples (LIAW; WIENER, 2002).

3.3.5 IBk

O algoritmo IBk é uma implementação do método *k-Nearest Neighbors*, também conhecido como k-NN (CUNNINGHAM; DELANY, 2007). Esse algoritmo é classificado como um método baseado em instâncias, pois armazena os exemplos de treinamento e classifica novas instâncias a partir da proximidade com os exemplos já conhecidos. A classe atribuída a uma nova instância é definida com base nos seus vizinhos mais próximos no espaço de atributos.

Como o IBk é baseado na comparação entre instâncias, uma das medidas de distância mais utilizadas é a distância euclidiana, apresentada na Equação (3.6).

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (3.6)$$

Na Equação (3.6), x e y representam duas instâncias da base de dados e n corresponde ao número de atributos considerados. Quanto menor o valor de $d(x, y)$, maior a similaridade entre as instâncias.

A classe atribuída a uma nova instância pode ser definida pela classe predominante entre os K vizinhos mais próximos, conforme indicado na Equação (3.7).

$$\hat{y} = \text{moda}\{y_1, y_2, \dots, y_K\} \quad (3.7)$$

Nesse caso, $\hat{y}$ representa a classe prevista e $y_1, y_2, \dots, y_K$ representam as classes dos K vizinhos mais próximos.

Sua principal vantagem é a simplicidade conceitual; entretanto, seu desempenho pode ser afetado pela escolha do valor de k , pela métrica de distância utilizada, pela escala dos atributos e pelo custo computacional durante a etapa de classificação (CUNNINGHAM; DELANY, 2007).

3.3.6 Logistic

A regressão logística é um método estatístico utilizado para problemas de classificação, especialmente quando a variável de saída é binária (BEWICK; CHEEK; BALL, 2005). Apesar do nome, seu objetivo não é prever um valor contínuo, mas estimar a probabilidade de uma instância pertencer a determinada classe. Para isso, o modelo combina os atributos de entrada por meio de uma função linear e aplica uma função logística, produzindo uma probabilidade entre 0 e 1.

No caso da classificação binária, a regressão logística estima a probabilidade de uma instância pertencer à classe positiva por meio da função sigmoide, apresentada na Equação (3.8).

$$P(y = 1 | x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}} \quad (3.8)$$

Na Equação (3.8), $P(y = 1 | x)$ representa a probabilidade de a instância pertencer à classe positiva, β_0 é o intercepto, β_i representa o peso associado ao atributo x_i e n corresponde ao número de atributos utilizados no modelo.

Entre suas vantagens estão a simplicidade, a eficiência computacional e a interpretabilidade dos coeficientes. No entanto, o método pode apresentar limitações quando há relações não lineares complexas entre os atributos e a classe (BEWICK; CHEEK; BALL, 2005).

3.3.7 JRip

O JRip é um classificador baseado em regras, correspondente a uma implementação do algoritmo RIPPER (COHEN, 1995). Esse método aprende um conjunto ordenado de regras do tipo “se-então”, que são aplicadas sequencialmente para classificar novas instâncias. Quando nenhuma regra é satisfeita, uma classe padrão é atribuída. A principal vantagem desse tipo de algoritmo está na interpretabilidade, pois as regras geradas podem ser analisadas diretamente pelo pesquisador. Por outro lado, em problemas com relações complexas entre atributos, conjuntos simples de regras podem não capturar todos os padrões presentes nos dados (COHEN, 1995).

3.4 Configuração dos Experimentos

Após o pré-processamento, a base resultante foi utilizada para comparar a capacidade dos algoritmos de classificar a variável nominal definida pela Equação (3.1). Para isso, foram especificados a ferramenta, o tratamento dos atributos, a estratégia de avaliação, as métricas e as parametrizações. Os resultados medem a reprodução desse indicador operacional na base DadosFACOM processada, e não a predição de uma situação institucional futura.

Os experimentos foram realizados na ferramenta Weka, versão 3.8.6. O Weka é um ambiente de mineração de dados que reúne algoritmos de aprendizado de máquina, recursos de pré-processamento, seleção de atributos, visualização e avaliação de modelos. Além disso, a ferramenta permite a comparação de diferentes métodos de classificação em um mesmo ambiente experimental, facilitando a execução de testes com diferentes configurações de parâmetros (FRANK et al., 2005). Dessa forma, sua utilização mostrou-se adequada ao objetivo deste trabalho, que consiste em comparar classificadores supervisionados aplicados a dados acadêmicos.

Antes da execução, o atributo *id* foi removido da entrada dos classificadores. Os preditores foram *ano*, *periodo*, *disciplinas*, *aprovadas*, *reprovadas* e *cra*, todos tratados como numéricos, e *evadiu* foi definida como classe nominal. Essa codificação reproduz a configuração dos experimentos, mas deve ser interpretada com cautela no caso de *periodo*, cujos valores são códigos administrativos, não uma escala contínua.

Os atributos possuem ordens de grandeza diferentes: *ano* está próximo de 2.000, *periodo* varia de 100 a 602, as contagens têm valores bem menores e *cra* varia de 0 a 100. Para Naive Bayes, J48, Random Forest, Logistic e JRip, foram usados os valores originais. Nos experimentos com IBk, os atributos numéricos foram padronizados, pois a distância euclidiana é diretamente afetada pela escala e permitiria que atributos com maior amplitude dominassem a vizinhança.

A padronização transforma cada valor com base na média e no desvio padrão do atributo, conforme a Equação (3.9).

$$z = \frac{x - \mu}{\sigma} \quad (3.9)$$

Na Equação (3.9), x é o valor original, μ é a média, σ é o desvio padrão e z é o valor padronizado. A padronização foi escolhida em vez da reescala mínimo-máximo porque os atributos não possuem limites naturais comuns e há valores extremos nas contagens. O procedimento centraliza os atributos e os expressa em unidades de desvio padrão sem fixá-los aos extremos observados na amostra. Essa escolha corrige diferenças de escala para o IBk, mas não altera o significado dos códigos de período nem resolve a dependência entre

a classe e as contagens usadas para construí-la.

A estratégia de avaliação adotada foi a validação cruzada estratificada com 10 partições, também conhecida como *10-fold cross-validation*. Nesse procedimento, a base de dados é dividida em dez subconjuntos, sendo que, a cada iteração, nove subconjuntos são utilizados para treinamento e um subconjunto é utilizado para teste. Esse processo é repetido até que todos os subconjuntos tenham sido utilizados uma vez como teste. A validação cruzada é uma técnica amplamente utilizada para estimar o desempenho e a capacidade de generalização de modelos de aprendizado de máquina, reduzindo a dependência de uma única divisão entre treino e teste (ALLGAIER; PRYSS, 2024). A estratificação, por sua vez, contribui para preservar a distribuição das classes nos diferentes subconjuntos avaliados.

Na validação cruzada com k partições, o desempenho final do modelo pode ser obtido pela média dos resultados alcançados em cada partição, conforme apresentado na Equação (3.10).

$$M = \frac{1}{k} \sum_{i=1}^k M_i \quad (3.10)$$

Na Equação (3.10), M representa o desempenho médio obtido, k corresponde ao número de partições utilizadas e M_i representa o valor da métrica avaliada na partição i . Neste trabalho, foi adotado $k = 10$, correspondente à validação cruzada com 10 partições.

Uma limitação dessa validação decorre do fato de 3.769 estudantes possuírem múltiplas instâncias. Embora o identificador tenha sido removido, as linhas de um mesmo estudante não são observações completamente independentes: diferentes períodos compartilham características de uma única trajetória acadêmica. Como a divisão foi feita por linha, registros do mesmo estudante podem aparecer simultaneamente no treinamento e no teste. Nesse caso, a avaliação se aproxima da classificação de um novo período de um estudante já representado no treinamento, e não da classificação de um estudante inteiramente não observado; correlações repetidas da trajetória podem, portanto, tornar a estimativa otimista. A comparação padronizada entre os algoritmos foi mantida, mas uma validação voltada a estudantes novos deveria agrupar as partições por *id*, mantendo todas as linhas de cada estudante em apenas um dos lados da avaliação. Uma divisão temporal também seria necessária para avaliar previsão futura.

O desempenho dos classificadores foi analisado por meio das métricas disponibilizadas pelo Weka, incluindo acurácia, precisão, *recall*, *F-Measure* e matriz de confusão. A acurácia indica a proporção total de instâncias corretamente classificadas, enquanto a precisão representa a proporção de predições positivas que foram corretamente classificadas. O *recall*, também chamado de sensibilidade, mede a capacidade do modelo em identificar corretamente as instâncias pertencentes à classe positiva. A *F-Measure* corresponde

à média harmônica entre precisão e *recall*, sendo útil para avaliar o equilíbrio entre essas duas medidas. A matriz de confusão, por sua vez, permite observar os acertos e erros de classificação em cada classe, oferecendo uma análise mais detalhada do comportamento dos modelos. A utilização conjunta dessas métricas é importante porque a acurácia, isoladamente, pode não representar adequadamente o desempenho de um classificador em todas as classes (OBI, 2023).

As métricas de avaliação utilizadas podem ser expressas a partir dos valores da matriz de confusão: verdadeiros positivos (VP), verdadeiros negativos (VN), falsos positivos (FP) e falsos negativos (FN). A acurácia, apresentada na Equação (3.11), representa a proporção total de classificações corretas.

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \quad (3.11)$$

A precisão indica a proporção de instâncias classificadas como positivas que realmente pertencem à classe positiva, conforme a Equação (3.12).

$$\text{Precisão} = \frac{VP}{VP + FP} \quad (3.12)$$

O *recall*, também chamado de sensibilidade, representa a proporção de instâncias positivas corretamente identificadas pelo modelo, conforme a Equação (3.13).

$$\text{Recall} = \frac{VP}{VP + FN} \quad (3.13)$$

Por fim, a *F-Measure* corresponde à média harmônica entre precisão e *recall*, como apresentado na Equação (3.14).

$$F\text{-Measure} = 2 \cdot \frac{\text{Precisão} \cdot \text{Recall}}{\text{Precisão} + \text{Recall}} \quad (3.14)$$

A definição das parametrizações teve caráter exploratório: para cada algoritmo, foram comparadas duas configurações previamente especificadas. Essa comparação permite verificar se as métricas mudam entre os valores testados, mas não constitui busca exaustiva de hiperparâmetros nem avaliação isolada de sobreajuste, interpretabilidade, custo ou generalização externa. Estudos em Mineração de Dados Educacionais frequentemente comparam algoritmos e configurações em tarefas de desempenho acadêmico ou risco de evasão (PRADEEP; THOMAS, 2015).

A Tabela 4 apresenta as parametrizações utilizadas nos experimentos.

Tabela 4 – Parametrizações testadas nos algoritmos de classificação

Algoritmo	Parametrização 1	Parametrização 2
J48	confidenceFactor = 0.25 minNumObj = 2	confidenceFactor = 0.10 minNumObj = 10
Random Forest	numTrees = 100 maxDepth = 0	numTrees = 200 maxDepth = 10
IBk	K = 3 DistanceWeighting = None	K = 5 DistanceWeighting = Inverse
Logistic	ridge = 1.0E-8	ridge = 1.0E-4
Naive Bayes	useKernelEstimator = false	useKernelEstimator = true
JRip	minNo = 2 numFolds = 3	minNo = 5 numFolds = 5

No caso do algoritmo J48, foram avaliadas duas configurações relacionadas à poda da árvore e ao número mínimo de instâncias por folha. A primeira parametrização tende a permitir a geração de uma árvore mais detalhada, enquanto a segunda busca produzir uma árvore mais conservadora, reduzindo sua complexidade. Para o Random Forest, foram avaliadas configurações com diferentes quantidades de árvores e controle de profundidade, de modo a observar o impacto do tamanho do conjunto de árvores e da limitação estrutural dos modelos. No caso do algoritmo *IBk*, foram avaliadas configurações com diferentes valores de K e diferentes formas de ponderação por distância. No algoritmo Logistic, foram avaliados dois valores do parâmetro *ridge*, relacionado à regularização do modelo. No Naive Bayes, a comparação foi realizada com e sem o uso do estimador *kernel*, permitindo observar o impacto dessa opção em atributos numéricos. Por fim, no JRip, foram avaliadas configurações com diferentes valores para o número mínimo de instâncias por regra e para o número de partições utilizadas no processo de poda.

Com essa configuração experimental, cada algoritmo foi executado duas vezes, totalizando doze experimentos. Os resultados foram comparados pelas métricas definidas, permitindo identificar quais algoritmos e parametrizações reproduziram com maior acerto o indicador operacional adotado.

4 Resultados e Discussão

Este capítulo apresenta diretamente os resultados dos doze experimentos descritos no Capítulo 3 e discute o que eles permitem concluir. A interpretação mantém a distinção entre a classe nominal *evadiu*, definida por uma regra acadêmica, e uma situação institucional de evasão observada posteriormente.

4.1 Desempenho Geral

A Tabela 5 reúne os resultados das duas parametrizações de cada classificador. Na validação cruzada com 10 partições, cada uma das 30.549 instâncias é classificada uma vez enquanto pertence à partição de teste. Assim, a coluna *Erros* corresponde ao número total de predições incorretas acumuladas ao final das dez partições, e não à média de erros por partição. A acurácia também é calculada sobre o total de predições, isto é, $(30549 - \text{Erros})/30549$. Precisão, *recall* e *F-Measure* são as médias ponderadas exibidas pelo Weka na linha *Weighted Avg.*.

Tabela 5 – Resultados agregados da validação cruzada com 10 partições.

Algoritmo	Param.	Acurácia	Prec.	Recall	F-Measure	Erros
Naive Bayes	P1	95,7216%	0,962	0,957	0,958	1.307
	P2	96,9950%	0,971	0,970	0,970	918
J48	P1	99,9804%	1,000	1,000	1,000	6
	P2	99,9771%	1,000	1,000	1,000	7
Random Forest	P1	99,9967%	1,000	1,000	1,000	1
	P2	99,9967%	1,000	1,000	1,000	1
IBk	P1	99,9476%	0,999	0,999	0,999	16
	P2	99,9542%	1,000	1,000	1,000	14
Logistic	P1	100,0000%	1,000	1,000	1,000	0
	P2	100,0000%	1,000	1,000	1,000	0
JRip	P1	99,9967%	1,000	1,000	1,000	1
	P2	99,9902%	1,000	1,000	1,000	3

Os valores em negrito identificam o melhor resultado global. Algumas métricas arredondadas a três casas aparecem como 1,000 em outros classificadores, mas somente o Logistic apresentou zero erro e acurácia exatamente igual a 100%.

O Logistic obteve o melhor resultado global nas duas configurações, classificando corretamente todas as instâncias. Random Forest e JRip também ficaram próximos desse valor: o primeiro registrou um erro em cada configuração, e o segundo registrou um e três erros. J48 apresentou seis e sete erros, enquanto IBk apresentou 16 e 14. Portanto, as diferenças absolutas entre esses cinco classificadores são pequenas diante das 30.549

instâncias e não devem ser interpretadas, isoladamente, como evidência de superioridade prática ou estatística.

O Naive Bayes apresentou o menor desempenho relativo. Sua primeira configuração produziu 1.307 erros e acurácia de 95,7216%; com o estimador *kernel*, os erros foram reduzidos para 918 e a acurácia aumentou para 96,9950%. Embora inferior aos demais métodos, esse resultado permanece elevado em termos absolutos e deve ser analisado em conjunto com os tipos de erro apresentados na matriz de confusão.

4.2 Efeito das Parametrizações

A alteração de parâmetros não afetou Logistic nem Random Forest nas métricas observadas: as duas configurações de cada algoritmo produziram resultados idênticos. No J48, P1 apresentou um erro a menos que P2; no JRip, P1 apresentou dois erros a menos. Essas diferenças são numericamente muito pequenas e não permitem afirmar que uma configuração seja mais robusta fora da amostra avaliada.

No IBk, P2 reduziu os erros de 16 para 14, uma diferença de 0,0066 ponto percentual de acurácia. Como P2 altera simultaneamente o número de vizinhos e a ponderação por distância, o experimento não permite atribuir a melhora exclusivamente a um desses fatores. No Naive Bayes ocorreu a maior variação: o estimador *kernel* elevou a acurácia em 1,2734 ponto percentual e reduziu o total de erros em 389. Esse resultado indica que P2 se ajustou melhor à distribuição dos atributos numéricos nesta base, sem demonstrar que seja a melhor configuração possível.

Foram testadas somente duas configurações por algoritmo. Não foi executada uma terceira parametrização nem uma busca sistemática de hiperparâmetros; por isso, esta seção constitui uma comparação local entre os valores da Tabela 4, e não uma otimização exaustiva. Experimentos adicionais, preferencialmente com validação interna para seleção dos parâmetros, dariam maior suporte às diferenças observadas.

4.3 Análise das Matrizes de Confusão

Além das métricas gerais, as matrizes de confusão permitem observar os tipos de erro cometidos pelos classificadores. Neste trabalho, a classe 0 representa registros não associados ao indicador de risco de evasão, enquanto a classe 1 representa registros associados ao indicador de risco. Assim, considerando a classe 1 como classe positiva, os falsos negativos correspondem aos casos de risco classificados incorretamente como não risco, enquanto os falsos positivos correspondem aos casos de não risco classificados incorretamente como risco.

A Tabela 6 apresenta um resumo das matrizes de confusão dos experimentos. Nessa tabela, VP representa verdadeiros positivos, VN representa verdadeiros negativos, FP representa falsos positivos e FN representa falsos negativos, considerando a classe 1 como positiva.

Tabela 6 – Resumo das matrizes de confusão considerando a classe 1 como positiva

Algoritmo	Param.	VN	FP	FN	VP
Naive Bayes	P1	20.776	1.207	100	8.466
	P2	21.221	762	156	8.410
J48	P1	21.983	0	6	8.560
	P2	21.983	0	7	8.559
Random Forest	P1	21.982	1	0	8.566
	P2	21.982	1	0	8.566
IBk	P1	21.971	12	4	8.562
	P2	21.973	10	4	8.562
Logistic	P1	21.983	0	0	8.566
	P2	21.983	0	0	8.566
JRip	P1	21.982	1	0	8.566
	P2	21.982	1	2	8.564

O Logistic não apresentou falsos positivos nem falsos negativos. Random Forest também identificou todas as 8.566 instâncias da classe 1 e cometeu somente um falso positivo em cada configuração. J48 concentrou seus poucos erros em falsos negativos, enquanto JRip P2 registrou um falso positivo e dois falsos negativos.

No IBk, P2 reduziu os falsos positivos de 12 para 10 e manteve quatro falsos negativos. A mudança mais relevante ocorreu no Naive Bayes: P2 reduziu os falsos positivos em 445 casos, de 1.207 para 762, mas aumentou os falsos negativos em 56, de 100 para 156. A redução líquida explica a melhora da acurácia, mas revela uma troca entre os dois tipos de erro que não seria visível pela acurácia isoladamente.

Essas interpretações consideram a classe 1 como positiva apenas para avaliar o indicador operacional. Um falso negativo, nesse experimento, significa que a regra *reprovadas > aprovadas* não foi reproduzida pelo classificador; não significa que uma evasão institucional efetivamente ocorrida deixou de ser detectada.

4.4 Discussão dos Resultados

4.4.1 O que os resultados efetivamente medem

A classe *evadiu* é definida pela condição *reprovadas > aprovadas*, e os dois atributos que formam essa condição foram fornecidos aos classificadores. Consequentemente, existe uma relação determinística entre parte da entrada e a saída. Os valores elevados não de-

monstram a descoberta de fatores desconhecidos de evasão; demonstram, principalmente, que os algoritmos conseguiram reconstruir uma regra já presente nos dados.

Esse aspecto também ajuda a explicar o desempenho perfeito do Logistic. A fronteira entre as classes pode ser escrita como *reprovadas* – *aprovadas* > 0, isto é, uma separação linear entre duas variáveis. Um modelo logístico é capaz de representar diretamente uma fronteira dessa natureza. Árvores e regras também podem aproximar a comparação entre as contagens, o que é compatível com os poucos erros de J48, Random Forest e JRip. O desempenho inferior do Naive Bayes é coerente com sua suposição de independência condicional, pois *disciplinas*, *aprovadas*, *reprovadas* e *cra* possuem relações entre si e com a própria construção da classe.

Portanto, o Logistic foi o melhor classificador para reproduzir o indicador na configuração estudada, mas não há evidência de que seja o melhor modelo para prever evasão institucional. Como a regra pode ser calculada diretamente a partir das contagens, o classificador não acrescenta, nesta forma, informação operacional além da própria regra. A contribuição experimental está na comparação controlada dos métodos e na identificação das limitações dessa representação.

4.4.2 Construção de uma classe institucional

Uma aplicação real de apoio à decisão exigiria uma variável de saída independente dos preditores. A base original contém o atributo **FORMA_EVASAO**, com categorias como aluno com vínculo, formado, abandono, desistência oficial, jubilamento e transferência. Esse campo oferece um ponto de partida, mas não deve ser convertido automaticamente em uma classe binária: a instituição precisa definir quais situações contam como evasão, como tratar transferências, estudantes ainda ativos e trajetórias sem desfecho, e em que momento a classificação deve ser realizada.

Uma possível reformulação seria construir uma instância em um período de referência t usando apenas informações disponíveis até esse momento e definir como saída a ocorrência de evasão institucional em um horizonte posterior, por exemplo no período seguinte ou dentro de um ano. Estudantes formados poderiam compor a classe de não evasão, enquanto casos ativos sem tempo suficiente de acompanhamento precisariam ser tratados como observações sem desfecho, e não automaticamente como negativos. A data e a forma do desligamento deveriam ser validadas com o setor acadêmico responsável.

Nessa reformulação, aprovações, reprovações e CRA poderiam continuar como preditores porque a classe deixaria de ser calculada diretamente a partir deles. Entretanto, seria necessário impedir o uso de informações posteriores ao período de referência, retirar campos que revelem o desfecho e separar treinamento e teste por estudante e por tempo. A avaliação também deveria considerar o custo institucional dos erros: falsos negativos

deixam de sinalizar estudantes que poderiam necessitar de apoio, enquanto falsos positivos podem direcionar intervenções a estudantes que não abandonariam o curso.

4.4.3 Limitações da comparação

A validação cruzada foi estratificada por instância, não agrupada por estudante. Como 3.769 estudantes possuem múltiplos registros, períodos de uma mesma trajetória podem ter sido distribuídos entre treinamento e teste. Mesmo sem o identificador como preditor, essa dependência reduz a separação entre os conjuntos e pode produzir estimativas mais otimistas para estudantes já representados no treinamento.

Além disso, foi executada uma única validação cruzada com duas configurações por algoritmo, sem repetição com diferentes sementes, intervalos de confiança ou teste estatístico entre classificadores. Diferenças de um a poucos erros, como as observadas entre Random Forest, JRip e J48, não são suficientes para sustentar conclusões sobre superioridade estatística ou capacidade de generalização. A base também mistura registros de diferentes ofertas de graduação; uma validação mais rigorosa poderia reservar uma coorte, um período ou uma oferta inteira para teste externo.

Por fim, a distribuição das classes é de 71,96% para a classe 0 e 28,04% para a classe 1. O uso conjunto de acurácia, métricas ponderadas e matrizes de confusão reduz a dependência de uma única medida, mas não corrige o viés estrutural da classe. Trabalhos posteriores devem utilizar um desfecho institucional independente, validação agrupada por estudante, separação temporal e análise de atributos adicionais, como frequência, carga horária, trancamentos, tempo de permanência, forma de ingresso, assistência estudantil e informações socioeconômicas.

5 Conclusão

Este trabalho teve como objetivo comparar o desempenho de seis algoritmos de classificação supervisionada sobre a base DadosFACOM processada. Na configuração experimental adotada, o Logistic apresentou o melhor resultado, com 100% de acurácia nas duas parametrizações, enquanto Random Forest, JRip, J48 e IBk também superaram 99,9%. O Naive Bayes obteve o menor desempenho relativo, entre 95,7216% e 96,9950% de acurácia.

A conclusão central, entretanto, não é que esses modelos tenham previsto a evasão institucional. A classe *evadiu* foi definida pela condição *reprovadas > aprovadas*, e esses mesmos atributos foram fornecidos aos classificadores. Como essa condição estabelece uma fronteira linear conhecida, o resultado perfeito do Logistic e os valores próximos de 100% dos demais métodos refletem principalmente a capacidade de reproduzir a regra utilizada para construir a classe. Nessa representação, o classificador não acrescenta informação operacional além daquela que poderia ser obtida pela aplicação direta da própria regra.

O estudo atingiu o objetivo de realizar uma comparação controlada entre algoritmos e de documentar a transformação de registros por atividade curricular em uma base agregada por aluno, ano e período. Sua principal contribuição está em evidenciar como a construção da variável de saída e a representação dos dados influenciam a avaliação de classificadores. Esse resultado também demonstra a necessidade de distinguir um indicador acadêmico calculado de um desfecho institucional observado.

As principais limitações são a dependência determinística entre a classe e dois preditores, a possibilidade de períodos do mesmo estudante aparecerem no treinamento e no teste, a perda de detalhes causada pela agregação por contagens e a ausência de repetição da validação ou de testes estatísticos entre os algoritmos. A base também reúne diferentes ofertas de graduação e contém ocorrências anômalas de codificação temporal. Por essas razões, as diferenças de poucos erros entre os classificadores não comprovam superioridade fora da amostra, e os modelos atuais não devem ser utilizados para rotular estudantes ou orientar intervenções institucionais.

Como continuidade, recomenda-se construir uma classe independente a partir de situações institucionais de evasão validadas pela universidade, definir um horizonte temporal de previsão e utilizar somente informações disponíveis antes do desfecho. A avaliação deverá separar os dados por estudante e por tempo, reservar coortes ou ofertas para validação externa e incorporar atributos como frequência, carga horária, trancamentos, forma de ingresso, assistência estudantil e tempo de permanência. Também poderão ser realizados experimentos de importância e seleção de atributos, busca sistemática de hiperparâmetros

e comparação estatística entre modelos. Somente após essas etapas será possível avaliar se um classificador oferece informação adicional e confiável para apoiar o acompanhamento acadêmico.

Referências

- AGRUSTI, F.; BONAVALONTÀ, G.; MEZZINI, M. University dropout prediction through educational data mining techniques: A systematic review. **Journal of e-Learning and Knowledge Society**, v. 15, n. 3, p. 161–182, 2019. Citado na página 12.
- ALLGAIER, J.; PRYSS, R. Cross-validation visualized: A narrative guide to advanced methods. **Machine Learning and Knowledge Extraction**, 2024. Citado na página 26.
- ANDRADE-GIRÓN, D.; SANDIVAR-ROSAS, J.; MARÍN-RODRIGUEZ, W.; SUSANIBAR-RAMIREZ, E.; TORO-DEXTRE, E.; AUSEJO-SANCHEZ, J.; VILLARREAL-TORRES, H.; ANGELES-MORALES, J. Predicting student dropout based on machine learning and deep learning: A systematic review. **EAI Endorsed Transactions on Scalable Information Systems**, v. 10, n. 5, 2023. Citado 3 vezes nas páginas 8, 12 e 13.
- ARAÚJO, A. C. P. L. d.; MARIANO, F. Z.; OLIVEIRA, C. S. d. Determinantes acadêmicos da retenção no ensino superior. **Ensaio**, v. 29, n. 113, p. 1045–1066, 2021. Citado na página 10.
- ARAÚJO, C. B. Z. M. A permanência de estudantes nos cursos de graduação no brasil: uma categoria em construção. **Revista Temas em Educação**, v. 22, n. 2, p. 25–43, 2013. Citado na página 10.
- BAGGI, C. A. d. S.; LOPES, D. A. Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. **Avaliação: Revista da Avaliação da Educação Superior**, v. 16, n. 2, p. 355–374, 2011. Citado na página 10.
- BAKER, R.; ISOTANI, S.; CARVALHO, A. **Mineração de Dados Educacionais: Oportunidades para o Brasil**. 2011. <<http://dx.doi.org/10.5753/RBIE.2011.19.02.03>>. [Online; accessed 15-Setembro-2021]. Citado na página 12.
- BAKER, R.; YACEF, K. **The State of Educational Data Mining in 2009: A Review and Future Visions**. 2009. <<http://www.educationaldatamining.org/JEDM/index.php/JEDM/article/view/8>>. [Online; accessed 15-Setembro-2021]. Citado na página 12.
- BERRAR, D. Bayes' theorem and naive bayes classifier. In: **Encyclopedia of Bioinformatics and Computational Biology**. [S.l.]: Elsevier, 2019. p. 403–412. Citado 3 vezes nas páginas 13, 21 e 22.
- BEWICK, V.; CHEEK, L.; BALL, J. Statistics review 14: Logistic regression. **Critical Care**, v. 9, n. 1, p. 112–118, 2005. Citado 2 vezes nas páginas 13 e 24.
- CHAGAS, P. A.; GOMIDE, J. S.; SANTANA, L. E. A. d. S. Previsão do risco de evasão universitária utilizando aprendizado de máquina: Características e dinâmica da evasão nos cursos de engenharia da ufrj-macae. **Revista Brasileira de Informática na Educação**, v. 33, p. 1226–1247, 2025. Citado 3 vezes nas páginas 8, 11 e 15.

- COHEN, W. W. Fast effective rule induction. In: **Proceedings of the Twelfth International Conference on Machine Learning**. [S.l.]: Morgan Kaufmann, 1995. p. 115–123. Citado 2 vezes nas páginas 13 e 24.
- CUNNINGHAM, P.; DELANY, S. J. **k-Nearest Neighbour Classifiers**. [S.l.], 2007. Citado 3 vezes nas páginas 13, 23 e 24.
- DUTRA, B. L. **Mineração de dados para identificação de fatores de reprovação no ensino superior**. 2019. Trabalho de Conclusão de Curso (Graduação em Sistemas de Informação) — Universidade Federal de Uberlândia, Uberlândia. Citado 5 vezes nas páginas 8, 10, 11, 13 e 14.
- FILHO, F. W. B. H.; VINUTO, T. S.; LEAL, B. C. Análise de classificadores para predição de evasão dos campi de uma instituição de ensino federal. In: **Anais do Simpósio Brasileiro de Informática na Educação (SBIE)**, 31. [S.l.]: Sociedade Brasileira de Computação, 2020. p. 1132–1141. Citado na página 13.
- FRANK, E.; HALL, M.; HOLMES, G.; KIRKBY, R.; PFAHRINGER, B.; WITTEN, I. H.; TRIGG, L. E. Weka—a machine learning workbench for data mining. In: **Data Mining and Knowledge Discovery Handbook**. [S.l.]: Springer, 2005. Citado na página 25.
- HUSSAIN, S.; DAHAN, N. A.; BA-ALWI, F. M.; RIBATA, N. Educational data mining and analysis of students' academic performance using weka. **Indonesian Journal of Electrical Engineering and Computer Science**, v. 9, n. 2, p. 447–459, 2018. Citado na página 13.
- JEDIDI, Y.; IBRIZ, A.; BENSLIMANE, M.; HACHMOUD, A.; TMIMI, M.; HAJJIOUI, Y.; RAHHALI, M. Predicting students' academic performance and modeling using data mining techniques. **Mathematical Modeling and Computing**, v. 11, n. 3, p. 814–825, 2024. Citado na página 13.
- LIAW, A.; WIENER, M. Classification and regression by randomforest. **R News**, v. 2, n. 3, p. 18–22, 2002. ISSN 1609-3631. Disponível em: <<https://journal.r-project.org/articles/RN-2002-022/>>. Citado 3 vezes nas páginas 13, 22 e 23.
- LUDERMIR, T. B. Inteligência artificial e aprendizado de máquina: estado atual e tendências. **Estudos Avançados**, v. 35, n. 101, p. 85–94, 2021. Citado na página 13.
- MACIEL, C. E.; JÚNIOR, M. C.; LIMA, T. d. S. A produção científica sobre permanência e evasão na educação superior no brasil. **Educação e Pesquisa**, v. 45, p. 1–20, 2019. Citado na página 10.
- MANHÃES, L. M. B.; CRUZ, S. M. S. d. Predição do desempenho acadêmico de alunos da graduação utilizando mineração de dados. In: CENTRO DE ANÁLISES DE SISTEMAS NAVAIS. **Anais do XIX Simpósio de Pesquisa Operacional e Logística da Marinha**. Rio de Janeiro, RJ, 2019. Citado na página 14.
- MÁXIMO, R. d. O.; CARVALHO, L.; LOPES, J. E. F.; RIBEIRO, K. C. d. S. Evasão e retenção no ensino superior: fatores socioeconômicos regionais e gestão de custos como determinantes da trajetória acadêmica. **Revista de Gestão e Secretariado**, v. 15, n. 3, p. e3615, 2024. Citado na página 10.

- MEIRELES, R. S. **Fatores causais da evasão e retenção universitária: uma análise nos cursos de licenciatura da Universidade de Brasília no período de 2002 até 2018**. 53 p. Dissertação (Dissertação (Mestrado Profissional em Economia)) — Universidade de Brasília, Brasília, 2019. 53 f., il. Citado na página 10.
- NIEROTKA, R. L.; SALATA, A.; MARTINS, M. K. Fatores associados à evasão no ensino superior: um estudo longitudinal. **Cadernos de Pesquisas**, v. 53, p. e09961, 2023. Citado 2 vezes nas páginas 8 e 11.
- OBI, J. C. A comparative study of several classification metrics and their performances on data. **World Journal of Advanced Engineering Technology and Sciences**, 2023. Citado na página 27.
- PALACIOS, C. A.; REYES-SUAREZ, J. A.; BEARZOTTI, L. A.; LEIVA, V.; MARCHANT, C. Knowledge discovery for higher education student retention based on data mining: Machine learning algorithms and case study in chile. **Entropy**, v. 23, n. 4, p. 485, 2021. Citado 2 vezes nas páginas 11 e 15.
- PRADEEP, A.; THOMAS, J. M. Predicting college students dropout using edm techniques. **International Journal of Computer Applications**, 2015. Citado na página 27.
- QUINLAN, J. R. **C4.5: Programs for Machine Learning**. San Mateo, CA: Morgan Kaufmann Publishers, 1993. Citado 2 vezes nas páginas 13 e 22.
- ROMERO, C.; VENTURA, S. **Data Mining in Education**. 2013. <<https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1075>>. [Online; accessed 15-Setembro-2021]. Citado na página 12.
- SHAFIQ, D. A.; MARJANI, M.; HABEEB, R. A. A.; ASIRVATHAM, D. Student retention using educational data mining and predictive analytics: A systematic literature review. **IEEE Access**, v. 10, p. 72480–72503, 2022. Citado 2 vezes nas páginas 8 e 12.
- SILVA, E. Cosme da. **Previsão da evasão acadêmica no ensino superior: O caso dos cursos de graduação presencial da Universidade de Brasília**. 2023. Dissertação (Mestrado Profissional em Computação Aplicada) — Universidade de Brasília, Brasília. Citado na página 14.
- TEDESCHI, I. P. **Previsão de evasão estudantil na Faculdade de Computação utilizando mineração de dados**. 65 p. Monografia (Trabalho de Conclusão de Curso) — Universidade Federal de Uberlândia, Uberlândia, 2025. Bacharelado em Ciência da Computação. Citado 2 vezes nas páginas 15 e 16.

Apêndices

APÊNDICE A – Declaração de uso de inteligência artificial generativa

Para a produção deste trabalho, foi utilizado o *ChatGPT*, ferramenta de inteligência artificial generativa desenvolvida pela OpenAI, em diferentes versões disponibilizadas pela plataforma durante sessões de acesso realizadas, principalmente, entre 22 de abril e 16 de julho de 2026. Como a ferramenta pode ser atualizada automaticamente ao longo do tempo e foi utilizada em diversas sessões, não foi possível associar todo o processo de elaboração do trabalho a uma única versão específica do modelo.

O uso da ferramenta ocorreu como apoio à elaboração, revisão e organização do trabalho, abrangendo as seguintes atividades:

1. auxílio na delimitação e no refinamento do tema, do título, do objetivo geral e da organização dos capítulos e seções;
2. revisão gramatical, ortográfica e estilística de trechos do texto, com sugestões para melhorar clareza, coesão, coerência, precisão terminológica e adequação à linguagem acadêmica;
3. apoio na reescrita e estruturação de partes da introdução, revisão bibliográfica, desenvolvimento, resultados e discussão, conclusão, resumo, palavras-chave e *abstract* em língua inglesa;
4. auxílio na síntese e integração, ao texto do trabalho, de referências e trabalhos relacionados previamente selecionados pelo autor, incluindo sugestões de posicionamento de citações e formatação de referências em BibTeX;
5. apoio na descrição metodológica da base DadosFACOM e de suas etapas de pré-processamento, incluindo seleção de atributos, limpeza, transformação, agregação dos registros por estudante, ano e período letivo, criação dos atributos derivados e preparação do arquivo para utilização no Weka;
6. discussão e revisão da definição operacional da variável de classe *evadiu*, incluindo a necessidade de tratá-la como indicador de risco de evasão, e não como confirmação institucional definitiva de abandono do curso;
7. auxílio na seleção e justificativa dos algoritmos Naive Bayes, J48, Random Forest, IBk, Logistic e JRip, bem como na definição das duas parametrizações avaliadas para cada algoritmo;

8. orientação para a configuração e execução dos doze experimentos no Weka, incluindo o uso da validação cruzada estratificada com 10 partições, a configuração do filtro de padronização para o IBk e a identificação das métricas disponibilizadas pela ferramenta;
9. apoio na organização, conferência e interpretação dos resultados gerados pelo Weka, incluindo acurácia, precisão, *recall*, *F-Measure*, matrizes de confusão e comparação entre algoritmos e parametrizações;
10. auxílio na redação da discussão dos resultados e na identificação de limitações metodológicas, como a relação direta entre os atributos preditores e a regra de construção da classe, a possibilidade de registros de um mesmo estudante aparecerem em diferentes partições da validação cruzada e a ausência de uma variável institucional oficial de evasão;
11. apoio na elaboração de tabelas, fórmulas, listas de abreviaturas e siglas, trechos em LaTeX e ajustes de organização e formatação do documento;
12. revisão global da consistência entre título, objetivos, metodologia, resultados e conclusão, incluindo a remoção de referências a algoritmos que não foram utilizados nos experimentos finais.

A ferramenta não foi utilizada para produzir ou alterar os registros da base de dados original, executar autonomamente os experimentos, fabricar resultados ou substituir a análise acadêmica do autor. A preparação prática dos dados, a configuração final e a execução dos experimentos no Weka, assim como a conferência dos valores apresentados nas tabelas e matrizes de confusão, foram realizadas pelo autor. As respostas geradas pela inteligência artificial foram utilizadas como sugestões e material de apoio, sendo revisadas, adaptadas e validadas criticamente antes de sua incorporação ao texto.

O autor assume integral responsabilidade pelas escolhas metodológicas, pela interpretação dos resultados, pela verificação das referências utilizadas e pelo conteúdo final apresentado neste trabalho.