



**INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO CEARÁ**  
**IFCE *CAMPUS* FORTALEZA**  
**PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO**  
**MESTRADO ACADÊMICO EM CIÊNCIA DA COMPUTAÇÃO**

**NEILA TEMÓTEO DE MATOS BASTOS**

**MINERAÇÃO DE DADOS EDUCACIONAIS PARA IDENTIFICAÇÃO DE PADRÕES  
DE EVASÃO NO PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA  
COMPUTAÇÃO (PPGCC) DO INSTITUTO FEDERAL DO CEARÁ (IFCE)**

**FORTALEZA**

**2026**

NEILA TEMÓTEO DE MATOS BASTOS

MINERAÇÃO DE DADOS EDUCACIONAIS PARA IDENTIFICAÇÃO DE PADRÕES DE  
EVASÃO NO PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO  
(PPGCC) DO INSTITUTO FEDERAL DO CEARÁ (IFCE)

Dissertação apresentada ao curso de Mestrado Acadêmico em Ciência da Computação do Programa de Pós-Graduação em Ciência da Computação do *Campus* Fortaleza do Instituto Federal de Educação, Ciência e Tecnologia do Ceará, como requisito parcial à obtenção do título de mestre em Ciência da Computação. Área de Concentração: Computação Aplicada.

Orientadora: Profa. Dra. Carina Teixeira de Oliveira.

Coorientadora: Profa. Dra. Francisca Raquel de Vasconcelos Silveira

FORTALEZA

2026

Dados Internacionais de Catalogação na Publicação  
Instituto Federal do Ceará - IFCE  
Sistema de Bibliotecas - SIBI

Ficha catalográfica elaborada pelo SIBI/IFCE, com os dados fornecidos pelo(a) autor(a)

---

B327m Bastos, Neila Temóteo de Matos.

Mineração de dados educacionais para identificação de padrões de evasão no programa de pós-graduação em ciência da computação (ppgcc) do instituto fed / Neila Temóteo de Matos Bastos. - 2026.  
114 f. : il. color.

Dissertação (Mestrado) – Instituto Federal do Ceará, Mestrado em Ciência da Computação, Campus Fortaleza, 2026.  
Orientação: Profa. Dra. Carina Teixeira de Oliveira.  
Coorientação: Profa. Dra. Francisca Raquel de Vasconcelos Silveira.

1. Evasão. 2. Predição. 3. Pós-graduação. 4. Computação. 5. Mineração de dados educacionais. I. Título.

CDD 004

---

**NEILA TEMÓTEO DE MATOS BASTOS**

**MINERAÇÃO DE DADOS EDUCACIONAIS PARA IDENTIFICAÇÃO DE PADRÕES  
DE EVASÃO NO PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA  
COMPUTAÇÃO (PPGCC) DO INSTITUTO FEDERAL DO CEARÁ (IFCE)**

**Dissertação apresentada ao curso de Mestrado Acadêmico em Ciência da Computação do Programa de Pós-Graduação em Ciência da Computação do *Campus* Fortaleza do Instituto Federal de Educação, Ciência e Tecnologia do Ceará, como requisito parcial à obtenção do título de mestre em Ciência da Computação. Área de Concentração: Computação Aplicada.**

**Aprovado em:** \_\_\_\_ / \_\_\_\_ / \_\_\_\_.

**BANCA EXAMINADORA**

---

Profa. Dra. Carina Teixeira de Oliveira (Orientadora)  
Instituto Federal de Educação, Ciência e Tecnologia do Ceará

---

Prof. Dr. Flávio Rubens de Carvalho Sousa  
Universidade Federal do Ceará

---

Prof. Dr. Anderson Passos de Aragão  
Instituto Federal de Educação, Ciência e Tecnologia do Ceará

---

Prof. Dr. Roger Moura Sarmento  
Instituto Federal de Educação, Ciência e Tecnologia do Ceará

## AGRADECIMENTOS

A Deus e à Nossa Senhora Aparecida, agradeço pela força, proteção e perseverança concedidas ao longo de toda esta caminhada, especialmente nos momentos em que desistir parecia a opção mais fácil.

Aos meus avós Temóteo e Zimar (*in memoriam*) e à minha tia Laíla, que não mediram esforços para investir na minha educação, oferecendo apoio, incentivo e exemplos que contribuíram de forma decisiva para a minha formação pessoal e acadêmica, deixo minha eterna gratidão.

À minha família e aos amigos, agradeço pelo carinho, compreensão e constante incentivo, fundamentais para que eu pudesse seguir em frente e concluir esta importante etapa da minha vida acadêmica.

Aos colegas de pesquisa e de curso, agradeço pela troca de conhecimentos, pelas discussões enriquecedoras e pelo apoio mútuo ao longo dessa trajetória, que tornaram o processo mais colaborativo e significativo.

Às minhas orientadoras, Carina Oliveira e Raquel Silveira, expresso minha sincera gratidão pela orientação, dedicação, paciência e pelas valiosas contribuições ao longo do desenvolvimento deste trabalho, que foram essenciais para o meu crescimento acadêmico e para a qualidade desta pesquisa.

À banca examinadora, agradeço pela disponibilidade, pela leitura atenta e pelas valiosas contribuições apresentadas durante a avaliação deste trabalho. As considerações, críticas e sugestões foram fundamentais para o aprimoramento da pesquisa e para o meu crescimento acadêmico, enriquecendo significativamente este estudo.

Ao Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE), agradeço pela oportunidade de formação acadêmica, bem como pelo suporte institucional e estrutural necessário à realização deste estudo.

## RESUMO

A evasão na pós-graduação *stricto sensu* constitui um fenômeno complexo, com impactos acadêmicos e institucionais relevantes, especialmente em instituições públicas. No Programa de Pós-Graduação em Ciência da Computação (PPGCC) do Instituto Federal do Ceará (IFCE), compreender os fatores associados à permanência e à evasão é fundamental para subsidiar políticas de acompanhamento discente. Nesse contexto, esta pesquisa aplica técnicas de Mineração de Dados Educacionais e Aprendizado de Máquina para identificar padrões associados à evasão no programa. A metodologia está estruturada em 7 etapas: (i) entendimento do domínio de estudo, (ii) coleta dos dados, (iii) análise visual dos dados, (iv) pré-processamento dos dados, (v) treinamento dos modelos, (vi) avaliação dos modelos e (vii) aplicação prática do melhor modelo. A base analisada contempla 255 matrículas registradas entre 2014/2 e 2025/1, contendo inicialmente 144 atributos acadêmicos, demográficos e socioeconômicos. A análise visual da base forneceu uma primeira visão sistemática sobre o conjunto de dados, permitindo a identificação de limitações intrínsecas (ex: presença de valores ausentes e inconsistências) e a caracterização dos atributos disponíveis para o estudo. Após o pré-processamento, foram treinados e avaliados três modelos supervisionados de classificação (*Decision Tree*, *Random Forest* e *XGBoost*), que foram combinados com diferentes técnicas de seleção de atributos, imputação de dados e balanceamento de dados. Um total de 3.600 experimentos foi realizado. O melhor desempenho foi obtido com o modelo *XGBoost* combinado com *Principal Component Analysis* (PCA), alcançando Acurácia global de 0,8810 e para os “Egressos sem Êxito” Precisão de 0,84, *Recall* de 0,9545 e *F1-score* de 0,8936. Os resultados indicam que a evasão está associada a fatores interdependentes, sendo melhor capturada por representações que exploram combinações latentes dos atributos. Embora o uso de PCA limite a interpretação direta, a análise de importância dos atributos, realizada com um modelo interpretável (sem a utilização do PCA) de desempenho competitivo, revelou três dimensões principais: (i) acadêmica (nível de instrução, matriz curricular e anos após a graduação); (ii) institucional (instituição de origem); e (iii) socioeconômica (tipo de cota, profissão e contexto geográfico, incluindo naturalidade). A idade também se destaca como uma variável relevante para caracterizar o perfil e o êxito dos estudantes. Como contribuição prática, foi desenvolvida uma aplicação que integra o modelo treinado, permitindo estimar a probabilidade de evasão e apoiar decisões institucionais baseadas em dados.

**Palavras-chave:** evasão; predição; pós-graduação; computação; mineração de dados educacionais; aprendizado de máquina.

## ABSTRACT

Dropout in stricto sensu graduate programs is a complex phenomenon with significant academic and institutional impacts, especially in public institutions. In the Graduate Program in Computer Science (PPGCC) at the Federal Institute of Ceará (IFCE), understanding the factors associated with student retention and dropout is essential to support student monitoring policies. In this context, this research applies Educational Data Mining and Machine Learning techniques to identify patterns associated with dropout in the program. The methodology is structured into seven stages: (i) understanding the study domain, (ii) data collection, (iii) data visualization, (iv) data preprocessing, (v) model training, (vi) model evaluation, and (vii) practical application of the best model. The analyzed dataset comprises 255 enrollments recorded between 2014/2 and 2025/1, initially containing 144 academic, demographic, and socioeconomic attributes. The exploratory data analysis provided an initial systematic overview of the dataset, enabling the identification of intrinsic limitations (e.g., missing values and inconsistencies) and the characterization of available attributes. After preprocessing, three supervised classification models (Decision Tree, Random Forest, and XGBoost) were trained and evaluated, combined with different techniques for feature selection, data imputation, and data balancing. A total of 3,600 experiments were conducted. The best performance was achieved with the XGBoost model combined with Principal Component Analysis (PCA), reaching an overall accuracy of 0.8810. For the “Unsuccessful Graduates” class, the model achieved a precision of 0.84, recall of 0.9545, and F1-score of 0.8936. The results indicate that dropout is associated with interdependent factors, being better captured through representations that explore latent combinations of attributes. Although the use of PCA limits direct interpretability, feature importance analysis performed with a competitive and interpretable model (without PCA) revealed three main dimensions: (i) academic (education level, curriculum structure, and years since undergraduate graduation); (ii) institutional (institution of origin); and (iii) socioeconomic (quota type, profession, and geographic context, including place of birth). Age also stands out as a relevant variable for characterizing student profiles and success. As a practical contribution, an application was developed integrating the trained model, allowing the estimation of dropout probability and supporting data-driven institutional decision-making.

**Keywords:** dropout; graduate education; educational data mining; machine learning; prediction.

## LISTA DE FIGURAS

|                                                                                                                      |    |
|----------------------------------------------------------------------------------------------------------------------|----|
| Figura 1 – Ferramentas e bibliotecas utilizadas pelos trabalhos. . . . .                                             | 41 |
| Figura 2 – Algoritmos utilizados pelos trabalhos. . . . .                                                            | 42 |
| Figura 3 – Melhores algoritmos apontados pelos trabalhos. . . . .                                                    | 43 |
| Figura 4 – Tamanho da base. . . . .                                                                                  | 44 |
| Figura 5 – Atributos. . . . .                                                                                        | 45 |
| Figura 6 – Tipos de dados. . . . .                                                                                   | 46 |
| Figura 7 – Fluxo metodológico do trabalho. . . . .                                                                   | 49 |
| Figura 8 – Matrículas por ano letivo. . . . .                                                                        | 52 |
| Figura 9 – Matrículas por sexo. . . . .                                                                              | 53 |
| Figura 10 – Matrículas por cor/raça. . . . .                                                                         | 54 |
| Figura 11 – Matrículas por renda. . . . .                                                                            | 54 |
| Figura 12 – Situação de matrícula. . . . .                                                                           | 55 |
| Figura 13 – Situações de matrícula de estudantes egressos sem êxito e com êxito. . . . .                             | 58 |
| Figura 14 – Combinação dos cenários experimentais. . . . .                                                           | 68 |
| Figura 15 – Matriz de confusão do melhor modelo. . . . .                                                             | 82 |
| Figura 16 – Curva ROC para melhor modelo . . . . .                                                                   | 83 |
| Figura 17 – Importância dos atributos para o segundo melhor. . . . .                                                 | 85 |
| Figura 18 – Fluxo Operacional da Aplicação. . . . .                                                                  | 86 |
| Figura 19 – Interface inicial da aplicação de predição de evasão no PPGCC. . . . .                                   | 87 |
| Figura 20 – Interface para entrada de dados de estudantes matriculados. . . . .                                      | 88 |
| Figura 21 – Interface de saída da aplicação com sintetizações geradas pelo modelo. . . . .                           | 89 |
| Figura 22 – Interface de saída da aplicação com distribuição dos estudantes por nível de<br>risco de evasão. . . . . | 91 |
| Figura 23 – Interface de saída da aplicação com lista de estudantes com maior risco de<br>evasão. . . . .            | 91 |
| Figura 24 – Interface de saída da aplicação com a previsão de evasão para todos os<br>estudantes. . . . .            | 92 |

## LISTA DE TABELAS

|                                                                                                                                                     |     |
|-----------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| Tabela 1 – Comparação entre mapeamentos sistemáticos sobre evasão escolar e Mineração de Dados Educacionais . . . . .                               | 35  |
| Tabela 2 – Trabalhos nacionais selecionados. . . . .                                                                                                | 37  |
| Tabela 3 – Situações de matrícula e suas definições. . . . .                                                                                        | 55  |
| Tabela 4 – Melhores resultados do modelo <i>Decision Tree</i> , sem seleção de atributos. . .                                                       | 71  |
| Tabela 5 – Melhores resultados do modelo <i>Random Forest</i> , sem seleção de atributos. .                                                         | 72  |
| Tabela 6 – Melhores resultados do modelo <i>XGBoost</i> , sem seleção de atributos. . . . .                                                         | 73  |
| Tabela 7 – Melhores resultados para cada modelo sem seleção de atributos. . . . .                                                                   | 75  |
| Tabela 8 – Melhores resultados do modelo <i>Decision Tree</i> , com seleção de atributos. . .                                                       | 77  |
| Tabela 9 – Melhores resultados do modelo <i>Random Forest</i> , com seleção de atributos. .                                                         | 78  |
| Tabela 10 – Melhores resultados do modelo <i>XGBoost</i> , com seleção de atributos. . . . .                                                        | 79  |
| Tabela 11 – Melhores resultados por modelo com seleção de atributos. . . . .                                                                        | 80  |
| Tabela 12 – Atributos derivados. . . . .                                                                                                            | 106 |
| Tabela 13 – Atributos eliminados da fonte de dados do PPGCC por conterem todas as linhas com valores integralmente ausentes. . . . .                | 107 |
| Tabela 14 – Atributos eliminados da fonte de dados do PPGCC por conterem o mesmo valor em todas as linhas. . . . .                                  | 108 |
| Tabela 15 – Atributos removidos da fonte de dados do PPGCC por baixo percentual de preenchimento. . . . .                                           | 109 |
| Tabela 16 – Grupos de atributos com indícios de duplicidade na base de dados do PPGCC.                                                              | 111 |
| Tabela 17 – Atributos eliminados da fonte de dados do PPGCC por serem diretamente relacionadas ao alvo. . . . .                                     | 112 |
| Tabela 18 – Atributos eliminados da fonte de dados do PPGCC por serem referentes a eventos ocorridos após o ingresso do estudante no curso. . . . . | 113 |
| Tabela 19 – Atributos e completude dos registros. . . . .                                                                                           | 114 |

## LISTA DE SIGLAS

|          |                                                              |
|----------|--------------------------------------------------------------|
| AdaBoost | <i>Adaptive Boosting</i>                                     |
| CAPES    | Coordenação de Aperfeiçoamento de Pessoal de Nível Superior  |
| CBIE     | Congresso Brasileiro de Informática na Educação              |
| CSV      | <i>Comma-Separated Values</i>                                |
| DM       | <i>Data Mining</i>                                           |
| DT       | <i>Decision Tree</i>                                         |
| EDM      | <i>Educational Data Mining</i>                               |
| ETC      | <i>Extra Trees Classifier</i>                                |
| GB       | <i>Gradient Boosting</i>                                     |
| GLM      | <i>Generalized Linear Model</i>                              |
| IFCE     | Instituto Federal de Educação, Ciência e Tecnologia do Ceará |
| KNN      | <i>K-Nearest Neighbors</i>                                   |
| LGPD     | Lei Geral de Proteção de Dados                               |
| MEC      | Ministério da Educação                                       |
| ML       | <i>Machine Learning</i>                                      |
| MLP      | <i>MultiLayer Perceptron</i>                                 |
| NB       | <i>Naive Bayes</i>                                           |
| PCA      | <i>Principal Component Analysis</i>                          |
| PNPG     | Plano Nacional de Pós-Graduação                              |
| PNS      | Plano de Normalização Setorial                               |
| PPGCC    | Programa de Pós-Graduação em Ciência da Computação           |
| RBIE     | Revista Brasileira de Informática na Educação                |
| RF       | <i>Random Forest</i>                                         |
| RL       | Regressão Logística                                          |
| SBC      | Sociedade Brasileira de Computação                           |
| SOL      | <i>SBC OpenLib</i>                                           |
| SLM      | <i>Systematic Literature Mapping</i>                         |
| SMOTE    | <i>Synthetic Minority Over-sampling Technique</i>            |
| SNPG     | Sistema Nacional de Pós-Graduação                            |
| SVM      | <i>Support Vector Machine</i>                                |
| XGBoost  | <i>eXtreme Gradient Boosting</i>                             |

## SUMÁRIO

|                |                                                                            |           |
|----------------|----------------------------------------------------------------------------|-----------|
| <b>1</b>       | <b>INTRODUÇÃO . . . . .</b>                                                | <b>13</b> |
| <b>1.1</b>     | <b>Contextualização . . . . .</b>                                          | <b>13</b> |
| <b>1.2</b>     | <b>Motivação . . . . .</b>                                                 | <b>15</b> |
| <b>1.3</b>     | <b>Justificativa . . . . .</b>                                             | <b>17</b> |
| <b>1.4</b>     | <b>Objetivos . . . . .</b>                                                 | <b>18</b> |
| <b>1.4.1</b>   | <b><i>Objetivo Geral . . . . .</i></b>                                     | <b>18</b> |
| <b>1.4.2</b>   | <b><i>Objetivos Específicos . . . . .</i></b>                              | <b>18</b> |
| <b>1.5</b>     | <b>Produção Científica . . . . .</b>                                       | <b>19</b> |
| <b>1.6</b>     | <b>Estrutura do documento . . . . .</b>                                    | <b>19</b> |
| <b>1.7</b>     | <b>Declaração de uso de Inteligência Artificial . . . . .</b>              | <b>20</b> |
| <b>2</b>       | <b>FUNDAMENTAÇÃO TEÓRICA . . . . .</b>                                     | <b>21</b> |
| <b>2.1</b>     | <b>Aprendizado de Máquina . . . . .</b>                                    | <b>21</b> |
| <b>2.1.1</b>   | <b><i>Aprendizado supervisionado . . . . .</i></b>                         | <b>22</b> |
| <b>2.1.1.1</b> | <b><i>Decision Tree . . . . .</i></b>                                      | <b>23</b> |
| <b>2.1.1.2</b> | <b><i>Random Forest . . . . .</i></b>                                      | <b>24</b> |
| <b>2.1.1.3</b> | <b><i>Extreme Gradient Boosting (XGBoost) . . . . .</i></b>                | <b>24</b> |
| <b>2.1.2</b>   | <b><i>Aprendizado não supervisionado . . . . .</i></b>                     | <b>25</b> |
| <b>2.1.3</b>   | <b><i>Aprendizado por reforço . . . . .</i></b>                            | <b>26</b> |
| <b>2.2</b>     | <b>Seleção de atributos . . . . .</b>                                      | <b>26</b> |
| <b>2.2.1</b>   | <b><i>SelectKBest com ANOVA F . . . . .</i></b>                            | <b>27</b> |
| <b>2.2.2</b>   | <b><i>Seleção baseada na correlação com a variável alvo . . . . .</i></b>  | <b>27</b> |
| <b>2.2.3</b>   | <b><i>Principal Component Analysis (PCA) . . . . .</i></b>                 | <b>28</b> |
| <b>2.3</b>     | <b>Balanceamento de Dados . . . . .</b>                                    | <b>28</b> |
| <b>2.3.1</b>   | <b><i>Random OverSampler . . . . .</i></b>                                 | <b>30</b> |
| <b>2.3.2</b>   | <b><i>Synthetic Minority Over-sampling Technique (SMOTE) . . . . .</i></b> | <b>30</b> |
| <b>2.3.3</b>   | <b><i>BorderlineSMOTE . . . . .</i></b>                                    | <b>30</b> |
| <b>2.4</b>     | <b>Métricas de avaliação de desempenho . . . . .</b>                       | <b>31</b> |
| <b>2.4.1</b>   | <b><i>Acurácia . . . . .</i></b>                                           | <b>31</b> |
| <b>2.4.2</b>   | <b><i>Precisão . . . . .</i></b>                                           | <b>31</b> |
| <b>2.4.3</b>   | <b><i>Recall . . . . .</i></b>                                             | <b>32</b> |
| <b>2.4.4</b>   | <b><i>F1-score . . . . .</i></b>                                           | <b>32</b> |

|              |                                                                                                |           |
|--------------|------------------------------------------------------------------------------------------------|-----------|
| <b>3</b>     | <b>MAPEAMENTO SISTEMÁTICO DA LITERATURA . . . . .</b>                                          | <b>33</b> |
| <b>3.1</b>   | <b>Trabalhos Relacionados . . . . .</b>                                                        | <b>33</b> |
| <b>3.2</b>   | <b>Metodologia . . . . .</b>                                                                   | <b>36</b> |
| <b>3.2.1</b> | <i>Etapa 1: Definição das Questões de Pesquisa (QP) . . . . .</i>                              | <i>36</i> |
| <b>3.2.2</b> | <i>Etapa 2: Identificação dos Estudos . . . . .</i>                                            | <i>36</i> |
| <b>3.2.3</b> | <i>Etapa 3: Seleção e Avaliação dos Estudos . . . . .</i>                                      | <i>37</i> |
| <b>3.2.4</b> | <i>Etapa 4: Síntese dos Dados e Apresentação dos Resultados . . . . .</i>                      | <i>37</i> |
| <b>3.3</b>   | <b>Resultados e Discussões . . . . .</b>                                                       | <b>37</b> |
| <b>3.3.1</b> | <i>QP1: Quais trabalhos científicos utilizam EDM no contexto da evasão escolar? . . . . .</i>  | <i>37</i> |
| <b>3.3.2</b> | <i>QP2: Quais as ferramentas/bibliotecas utilizadas pelos trabalhos? . . . . .</i>             | <i>41</i> |
| <b>3.3.3</b> | <i>QP3: Quais os algoritmos de aprendizagem de máquina utilizados nos trabalhos? . . . . .</i> | <i>42</i> |
| <b>3.3.4</b> | <i>QP4: Quais os algoritmos com melhor desempenho? . . . . .</i>                               | <i>43</i> |
| <b>3.3.5</b> | <i>QP5: Quantos registros existem nas bases de dados analisadas? . . . . .</i>                 | <i>44</i> |
| <b>3.3.6</b> | <i>QP6: Quantos atributos estão presentes nas bases de dados analisadas? . . . . .</i>         | <i>45</i> |
| <b>3.3.7</b> | <i>QP7: Quais os tipos de dados analisados nos trabalhos? . . . . .</i>                        | <i>46</i> |
| <b>3.4</b>   | <b>Conclusões . . . . .</b>                                                                    | <b>46</b> |
| <b>4</b>     | <b>PROPOSTA . . . . .</b>                                                                      | <b>49</b> |
| <b>4.1</b>   | <b>Etapa 1: Entendimento do domínio de estudo . . . . .</b>                                    | <b>49</b> |
| <b>4.2</b>   | <b>Etapa 2: Coleta dos dados . . . . .</b>                                                     | <b>51</b> |
| <b>4.3</b>   | <b>Etapa 3: Análise visual dos dados . . . . .</b>                                             | <b>51</b> |
| <b>4.4</b>   | <b>Etapa 4: Pré-processamento dos dados . . . . .</b>                                          | <b>56</b> |
| <b>4.4.1</b> | <i>Construção da variável alvo . . . . .</i>                                                   | <i>56</i> |
| <b>4.4.2</b> | <i>Criação de atributos . . . . .</i>                                                          | <i>58</i> |
| <b>4.4.3</b> | <i>Limpeza de dados . . . . .</i>                                                              | <i>60</i> |
| <b>4.4.4</b> | <i>Preenchimento de Valores Ausentes . . . . .</i>                                             | <i>61</i> |
| <b>4.4.5</b> | <i>Codificação de Variáveis Categóricas . . . . .</i>                                          | <i>63</i> |
| <b>4.5</b>   | <b>Etapa 5: Treinamento dos Modelos . . . . .</b>                                              | <b>64</b> |
| <b>4.5.1</b> | <i>Divisão em Treino e Testes . . . . .</i>                                                    | <i>65</i> |
| <b>4.5.2</b> | <i>Balanceamento das Classes . . . . .</i>                                                     | <i>65</i> |
| <b>4.5.3</b> | <i>Seleção de Atributos . . . . .</i>                                                          | <i>66</i> |
| <b>4.6</b>   | <b>Etapa 6: Avaliação dos Modelos . . . . .</b>                                                | <b>67</b> |

|                |                                                                                                                                             |            |
|----------------|---------------------------------------------------------------------------------------------------------------------------------------------|------------|
| <b>4.7</b>     | <b>Etapa 7: Aplicação Prática do Modelo</b>                                                                                                 | <b>69</b>  |
| <b>5</b>       | <b>RESULTADOS DA AVALIAÇÃO DOS MODELOS</b>                                                                                                  | <b>70</b>  |
| <b>5.1</b>     | <b>Avaliação dos modelos preditivos</b>                                                                                                     | <b>70</b>  |
| <b>5.1.1</b>   | <b>Modelagem sem seleção de atributos</b>                                                                                                   | <b>70</b>  |
| <b>5.1.1.1</b> | <i>Decision Tree</i>                                                                                                                        | <b>71</b>  |
| <b>5.1.1.2</b> | <i>Random Forest</i>                                                                                                                        | <b>72</b>  |
| <b>5.1.1.3</b> | <i>XGBoost</i>                                                                                                                              | <b>73</b>  |
| <b>5.1.1.4</b> | <i>Comparação entre os modelos</i>                                                                                                          | <b>75</b>  |
| <b>5.1.2</b>   | <b>Modelagem com seleção de atributos</b>                                                                                                   | <b>75</b>  |
| <b>5.1.2.1</b> | <i>Decision Tree</i>                                                                                                                        | <b>76</b>  |
| <b>5.1.2.2</b> | <i>Random Forest</i>                                                                                                                        | <b>77</b>  |
| <b>5.1.2.3</b> | <i>XGBoost</i>                                                                                                                              | <b>78</b>  |
| <b>5.1.2.4</b> | <i>Comparação entre modelos</i>                                                                                                             | <b>80</b>  |
| <b>5.2</b>     | <b>Discussões sobre Melhor Modelo</b>                                                                                                       | <b>81</b>  |
| <b>5.2.1</b>   | <i>Matriz de Confusão</i>                                                                                                                   | <b>81</b>  |
| <b>5.2.2</b>   | <i>Curva ROC</i>                                                                                                                            | <b>82</b>  |
| <b>5.2.3</b>   | <i>Análise da importância dos atributos</i>                                                                                                 | <b>83</b>  |
| <b>6</b>       | <b>APLICAÇÃO PRÁTICA DO MELHOR MODELO</b>                                                                                                   | <b>86</b>  |
| <b>6.1</b>     | <b>Fluxo Operacional da Aplicação</b>                                                                                                       | <b>86</b>  |
| <b>6.1.1</b>   | <i>Etapa 1: Entrada</i>                                                                                                                     | <b>88</b>  |
| <b>6.1.2</b>   | <i>Etapa 2: Engine</i>                                                                                                                      | <b>88</b>  |
| <b>6.1.3</b>   | <i>Etapa 3: Saída</i>                                                                                                                       | <b>89</b>  |
| <b>6.2</b>     | <b>Limitações da Solução Proposta</b>                                                                                                       | <b>91</b>  |
| <b>7</b>       | <b>CONCLUSÕES</b>                                                                                                                           | <b>94</b>  |
| <b>7.1</b>     | <b>Trabalhos futuros</b>                                                                                                                    | <b>95</b>  |
|                | <b>REFERÊNCIAS</b>                                                                                                                          | <b>97</b>  |
|                | <b>APÊNDICE A – ATRIBUTOS DERIVADOS</b>                                                                                                     | <b>106</b> |
|                | <b>APÊNDICE B – ATRIBUTOS ELIMINADOS DA FONTE DE DADOS<br/>DO PPGCC POR CONTEREM TODAS AS LINHAS<br/>COM VALORES INTEGRALMENTE AUSENTES</b> | <b>107</b> |
|                | <b>APÊNDICE C – ATRIBUTOS ELIMINADOS DA FONTE DE DADOS<br/>DO PPGCC POR CONTEREM O MESMO VALOR<br/>EM TODAS AS LINHAS</b>                   | <b>108</b> |

|                                                                                                                                                                              |            |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|
| <b>APÊNDICE D – ATRIBUTOS ELIMINADOS DA FONTE DE DADOS<br/>DO PPGCC POR BAIXO PERCENTUAL DE PRE-<br/>ENCHIMENTO . . . . .</b>                                                | <b>109</b> |
| <b>APÊNDICE E – ATRIBUTOS ELIMINADOS DA FONTE DE DADOS<br/>DO PPGCC POR SEREM DUPLICADOS . . . . .</b>                                                                       | <b>111</b> |
| <b>APÊNDICE F – ATRIBUTOS ELIMINADOS DA FONTE DE DADOS<br/>DO PPGCC POR SEREM DIRETAMENTE RELA-<br/>CIONADOS AO ALVO . . . . .</b>                                           | <b>112</b> |
| <b>APÊNDICE G – ATRIBUTOS ELIMINADOS DA FONTE DE DADOS<br/>DO PPGCC POR SEREM REFERENTES A EVEN-<br/>TOS OCORRIDOS APÓS O INGRESSO DO ESTU-<br/>DANTE NO CURSO . . . . .</b> | <b>113</b> |
| <b>APÊNDICE H – ATRIBUTOS DA FONTE DE DADOS DO PPGCC<br/>APÓS LIMPEZA . . . . .</b>                                                                                          | <b>114</b> |

## 1 INTRODUÇÃO

Neste primeiro capítulo, são apresentados os elementos centrais que contextualizam, motivam e justificam o desenvolvimento deste trabalho. Além disso, são delineados o objetivo geral e os objetivos específicos a serem alcançados, assim como é estruturada a dissertação.

### 1.1 Contextualização

A educação é um dos pilares fundamentais para o desenvolvimento humano, social e econômico em escala global. Ela desempenha um papel central na formação de indivíduos capacitados, na promoção da inovação e no fortalecimento da competitividade das nações em um mundo cada vez mais interconectado e tecnológico.

A educação superior, em particular, é essencial para a geração de conhecimento avançado, o desenvolvimento de habilidades especializadas e a consolidação de uma base científica que impulse o progresso em diversas áreas do saber.

A pós-graduação *stricto sensu* no Brasil teve seu marco regulatório estabelecido pelo Parecer n.º 977, de 3 de dezembro de 1965, do então Conselho Federal de Educação (CAPES, 1965). A partir desse momento, consolidou-se o Sistema Nacional de Pós-Graduação (SNPG), que tem por objetivo oferecer alta qualificação científica e tecnológica aos cidadãos brasileiros (CAPES, 2025).

Os Planos Nacionais de Pós-Graduação (PNPGs) reconhecem a evasão escolar e o baixo desempenho acadêmico como problemas estruturais do sistema educacional brasileiro, manifestando-se tanto no âmbito interno dos cursos de mestrado e doutorado quanto nas etapas anteriores da formação educacional (CAPES, 2022). Os primeiros planos, em especial o PNPG I (1975–1979) (CAPES, 1975) e o PNPG III (1986–1989) (CAPES, 1986), já diagnosticavam a evasão na pós-graduação como um ponto de estrangulamento do sistema, associada às demandas do mercado de trabalho não atendidas adequadamente pelos cursos de especialização, à baixa proporção de estudantes bolsistas em regime de dedicação integral e à consequente ineficiência dos programas, os quais eram considerados pouco eficazes em seu conjunto, uma vez que a maioria dos candidatos admitidos não concluía os créditos e, no caso do mestrado, apenas cerca de 15% atingiam a titulação.

Em uma perspectiva sistêmica, os PNPGs posteriores ampliaram essa análise ao evidenciar que a evasão e o baixo desempenho nas etapas da educação básica e do ensino médio comprometem diretamente a formação de recursos humanos qualificados para a pós-graduação,

visto que o processo de perda e fracasso escolar tem início precoce e se acumula ao longo da trajetória educacional. Documentos como os PNPGs 2005–2010 (CAPES, 2005), 2011–2020 (CAPES, 2010) e 2025-2029 (CAPES, 2025) ressaltam que as elevadas taxas de evasão no ensino médio, associadas à baixa qualidade do ensino, às desigualdades sociais e a fatores macroestruturais, limitam a expansão e a consolidação do SNPG.

Nesse contexto, os PNPGs atribuem à pós-graduação um papel estratégico no enfrentamento indireto da evasão escolar, defendendo a articulação entre o SNPG, coordenado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), e a educação básica, por meio da formação qualificada de professores, do fortalecimento de programas *stricto sensu* profissionais e do desenvolvimento de pesquisas que subsidiem políticas públicas voltadas à permanência, ao êxito escolar e à redução das desigualdades educacionais.

Destaca-se como relevante a constatação das dificuldades encontradas para reduzir a evasão de pós-graduandos nas áreas tecnológicas, a falta de mobilidade e as dificuldades de inserção em alguns setores do mercado (CAPES, 2025). De acordo com os documentos, essa carência de mobilidade refere-se ao reduzido fluxo de pós-graduandos, pesquisadores e docentes entre instituições de diferentes regiões do país (mobilidade intranacional) e entre o Brasil e o exterior (mobilidade internacional).

No cenário nacional, a mobilidade é considerada pouco significativa em função da dimensão continental do país e dos elevados custos de deslocamento e manutenção, fatores que dificultam a troca de conhecimentos e impedem que o SNPG atue de forma a mitigar as assimetrias regionais. Como consequência, o sistema tende à formação acadêmica, restringindo a diversidade na pesquisa e a colaboração interinstitucional. Para reverter esse quadro, o PNPG 2025-2029 (CAPES, 2025) propõe o incentivo à “intranacionalização” e à solidariedade acadêmica, expandindo as estratégias de mobilidade também para os orientadores, visando fortalecer redes de cooperação e aumentar a visibilidade global da ciência brasileira, que também foi afetada pela queda no contingente de doutorados plenos realizados no exterior.

Considerando que a pesquisa acadêmico-científica demanda recursos escassos e de alto custo (CUNHA, 2004), a evasão compromete o retorno esperado pela sociedade, já que o investimento realizado em estudantes que não concluem seus cursos deixa de se converter em novos saberes e contribuições científicas (VELHO, 2005).

No contexto brasileiro, os programas de pós-graduação *stricto sensu*, especialmente em áreas estratégicas como a Ciência da Computação, têm um papel crucial na formação de profissionais qualificados e na produção de pesquisas que respondam aos desafios contemporâneos.

Em 2024, o gasto corrente por matrícula equivalente do Instituto Federal do Ceará (IFCE) foi estimado em R\$ 17.030,41, valor que representa o investimento médio anual por estudante, incluindo despesas com pessoal e custeio. Essa informação foi disponibilizada pela Plataforma Nilo Peçanha (PNP), sistema oficial do Ministério da Educação (MEC) destinado ao monitoramento e à avaliação da Rede Federal de Educação Profissional, Científica e Tecnológica (MEC, 2026). Esse valor evidencia o significativo investimento público realizado na formação acadêmica dos estudantes, o que torna a evasão escolar um fenômeno de elevado impacto não apenas social e educacional, mas também econômico. A interrupção do percurso formativo antes da conclusão dos cursos implica a não consolidação dos resultados esperados desse investimento, reforçando a necessidade de estudos que identifiquem fatores associados à evasão e subsidiem a formulação de políticas institucionais voltadas à permanência e ao êxito discente, especialmente em instituições públicas de ensino como o IFCE.

## 1.2 Motivação

Nos últimos anos, a revolução tecnológica tem transformado de forma significativa o cenário educacional, ampliando as possibilidades de análise e gestão de dados educacionais. A crescente digitalização dos processos acadêmicos e administrativos tem resultado na geração de grandes volumes de dados, oriundos de ambientes virtuais de aprendizagem, sistemas acadêmicos institucionais e plataformas educacionais digitais. Quando adequadamente tratados e analisados, esses dados podem fornecer informações relevantes para a melhoria da qualidade do ensino e da aprendizagem, além de subsidiar o enfrentamento de desafios persistentes, como a evasão e a retenção escolar (HAN; KAMBER; PEI, 2012).

Para processar esse volume de informações, a Mineração de Dados (*Data Mining* – DM) apresenta-se como um conjunto fundamental de técnicas e métodos computacionais destinados à descoberta de padrões, relações e estruturas relevantes em grandes volumes de dados. O objetivo central da DM é transformar dados brutos em conhecimento útil, capaz de apoiar processos de análise e tomada de decisão utilizando abordagens descritivas, que exploram a estrutura dos dados para identificar tendências, descritivas ou preditivas, que visam estimar comportamentos futuros com base em dados históricos. Assim, a DM constitui uma etapa fundamental do processo de descoberta de conhecimento em bases de dados, possibilitando a identificação automatizada de padrões que dificilmente seriam percebidos por meio de análises convencionais (BRAGA, 2005).

No domínio específico da educação, essa aplicação deu origem ao campo da Mineração de Dados Educacionais (*Educational Data Mining* – EDM). A EDM constitui uma área de pesquisa interdisciplinar que integra métodos provenientes da ciência de dados, estatística e Inteligência Artificial para analisar grandes volumes de dados educacionais, com o objetivo de compreender e melhorar processos educacionais (ROMERO; VENTURA, 2010).

Diferentemente da DM aplicada a outros domínios, a EDM considera as particularidades do contexto educacional, incluindo fatores pedagógicos, institucionais e socioeconômicos que influenciam a trajetória acadêmica dos estudantes. A análise desses dados possibilita investigar fenômenos como desempenho acadêmico, engajamento discente, permanência estudantil e evasão escolar, contribuindo para o desenvolvimento de estratégias institucionais voltadas à melhoria da qualidade da educação e ao apoio à tomada de decisão baseada em evidências.

No contexto da EDM, a evasão escolar tem sido amplamente investigada por meio de abordagens analíticas que combinam análise descritiva e modelagem preditiva. A abordagem descritiva permite identificar padrões coletivos e perfis recorrentes de estudantes, contribuindo para uma compreensão mais abrangente dos fatores associados à evasão. Por sua vez, a abordagem preditiva utiliza modelos de aprendizado supervisionado para estimar, por exemplo, em nível individual, a probabilidade de um estudante evadir ou concluir o curso. A integração dessas duas perspectivas possibilita não apenas compreender o fenômeno da evasão, mas também desenvolver mecanismos de monitoramento e intervenção precoce voltados à promoção da permanência estudantil (ROMERO; VENTURA, 2010; BAKER; INVENTADO, 2014).

Desta forma, o Aprendizado de Máquina (*Machine Learning* – ML), aplicado na EDM, oferece um conjunto de técnicas capazes de aprender padrões a partir dos dados, realizar previsões e construir modelos analíticos que se adaptam automaticamente a novas informações (MITCHELL, 1997). Essas abordagens possibilitam a análise de grandes volumes de dados educacionais, incluindo registros acadêmicos — como notas, frequência, reprovações e desempenho em avaliações —, dados demográficos — como idade, gênero, cor/raça, etnia e estado civil — e informações socioeconômicas — como renda familiar, nível de escolaridade dos responsáveis e condições de moradia. A integração dessas múltiplas dimensões de dados permite a identificação de padrões complexos de comportamento discente, bem como a antecipação de situações de risco acadêmico, favorecendo intervenções pedagógicas mais precisas e personalizadas (BAKER; INVENTADO, 2014).

No âmbito da educação superior, a aplicação de técnicas de EDM e ML tem sido amplamente explorada para fins como a predição de evasão, a análise do desempenho acadêmico,

o apoio à gestão institucional e a avaliação de políticas educacionais. Estudos indicam que tais técnicas contribuem para a otimização da gestão acadêmica e para o aprimoramento dos processos decisórios, ao fornecer evidências baseadas em dados que orientam ações institucionais mais eficazes (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; WITTEN; FRANK; HALL, 2016).

Dessa forma, o uso de técnicas avançadas de análise de dados no contexto educacional representa uma oportunidade estratégica para enfrentar desafios estruturais do sistema educacional, promovendo a melhoria da qualidade do ensino, o sucesso acadêmico dos estudantes e o alinhamento das práticas educacionais às demandas de uma sociedade e de um mercado de trabalho em constante transformação.

### **1.3 Justificativa**

A evasão escolar na pós-graduação *stricto sensu* é um fenômeno complexo e multifatorial, com impactos negativos significativos para discentes, instituições e a própria sociedade como um todo. Para os discentes, a evasão do curso pode resultar em frustração pessoal, prejuízos profissionais e perda de oportunidades de qualificação. No âmbito institucional, a desistência de estudantes implica desperdício de recursos financeiros e humanos, além de comprometer indicadores de desempenho e reputação acadêmica. No âmbito social, a evasão reduz a formação de profissionais altamente qualificados, limitando a capacidade de inovação e o desenvolvimento econômico e tecnológico do país.

A ausência de indicadores objetivos para identificar precocemente situações de risco dificulta a implementação de estratégias eficazes de acompanhamento e prevenção. Muitas iniciativas institucionais ainda são reativas, baseadas em percepções subjetivas ou intervenções tardias, o que limita sua eficácia.

A EDM e o ML destacam-se como ferramentas essenciais nesse processo, permitindo a identificação de padrões complexos, a previsão de comportamentos e a proposição de soluções personalizadas para o ambiente de aprendizagem (CECHINEL; CAMARGO, 2020). Por meio dessas técnicas, é possível analisar dados acadêmicos, demográficos e socioeconômicos para prever quais estudantes estão em risco de evasão, fornecendo subsídios para a formulação de políticas e práticas de gestão acadêmica mais eficazes, capazes de promover a permanência e o sucesso discente. A capacidade de processar grandes bases de dados e extrair informações relevantes torna essas abordagens indispensáveis para otimizar a gestão educacional e melhorar

os resultados de aprendizagem.

Neste contexto, a justificativa para esta pesquisa está ancorada na relevância do tema da evasão na pós-graduação, especialmente em áreas como a Ciência da Computação. Este estudo busca aplicar técnicas de mineração de dados aos registros acadêmicos do Programa de Pós-Graduação em Ciência da Computação (PPGCC) do Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE), com o objetivo de identificar indicadores de atenção associados à evasão, delinear perfis de sucesso e não sucesso, e propor ações preventivas que favoreçam a permanência e o êxito discente. Assim, pretende-se contribuir tanto para a literatura científica quanto para a prática institucional, oferecendo subsídios para a melhoria da qualidade e eficiência do programa.

## **1.4 Objetivos**

### ***1.4.1 Objetivo Geral***

Este trabalho tem como objetivo geral aplicar técnicas de mineração de dados educacionais para identificar padrões de evasão no Programa de Pós-Graduação em Ciência da Computação (PPGCC) do Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE).

### ***1.4.2 Objetivos Específicos***

Os objetivos específicos desta dissertação são:

- Coletar dados acadêmicos dos estudantes do PPGCC, garantindo sua integridade e consistência para análise.
- Aplicar técnicas de pré-processamento de dados, como tratamento de valores ausentes, normalização e seleção de atributos, visando preparar o conjunto de dados para mineração.
- Utilizar algoritmos de aprendizagem de máquina supervisionada para prever o êxito ou a evasão estudantil.
- Avaliar os modelos de predição quanto ao seu desempenho na identificação do êxito e da evasão estudantil, por meio de métricas adequadas.
- Implementar uma aplicação computacional para integrar e disponibilizar o modelo de aprendizado de máquina desenvolvido, viabilizando sua aplicação prática na predição do

status acadêmico de novos estudantes.

## 1.5 Produção Científica

Durante o mestrado, os seguintes artigos foram publicados:

- **BASTOS, N.; GOMES, L. B. S., LINHARES, J. F., SILVEIRA, R.; OLIVEIRA, C. T. Mapeamento Sistemático da Mineração de Dados Educacionais no Combate à Evasão Escolar no Brasil.** In: Congresso sobre Tecnologias na Educação (CTRL+E), 2025, Fortaleza/CE. Sociedade Brasileira de Computação (SBC).
- GOMES, L. B. S., LINHARES, J. F., **BASTOS, N.**; SILVEIRA, R.; OLIVEIRA, C. T. **Perfil das Bases de Dados Nacionais na Área de Mineração de Dados Educacionais.** In: XXVI Simpósio Brasileiro de Informática na Educação (SBIE), 2025, Curitiba/PR. Sociedade Brasileira de Computação (SBC).

## 1.6 Estrutura do documento

O restante do trabalho está organizado da seguinte maneira:

O Capítulo 2 apresenta a fundamentação teórica que sustenta a pesquisa, abordando fundamentos do Aprendizado de Máquina, técnicas de seleção de atributos, estratégias de balanceamento de classes e métricas de avaliação adotadas para análise do desempenho dos modelos.

O Capítulo 3 descreve o mapeamento sistemático da literatura realizado sobre evasão escolar no contexto da Mineração de Dados Educacionais no Brasil. São apresentados o protocolo metodológico, os critérios de inclusão e exclusão dos estudos e a síntese dos resultados, permitindo identificar tendências, lacunas, algoritmos mais utilizados e boas práticas na área, situando esta pesquisa no cenário científico nacional.

O Capítulo 4 detalha a metodologia da proposta, desenvolvida em 7 etapas principais: (i) entendimento do domínio de estudo, (ii) coleta dos dados, (iii) análise visual dos dados, (iv) pré-processamento dos dados, (v) treinamento dos modelos, (vi) avaliação dos modelos e (vii) aplicação prática do melhor modelo.

O Capítulo 5 apresenta e discute os resultados obtidos a partir dos experimentos realizados. São analisadas as métricas de avaliação e discutido o impacto das decisões de pré-processamento

e seleção de atributos na capacidade preditiva, incluindo a interpretação dos fatores associados ao risco de evasão.

O Capítulo 6 descreve a aplicação prática do modelo selecionado como cenário final, detalhando sua implementação computacional e fluxo de funcionamento. O capítulo evidencia a viabilidade de utilização da solução em contexto institucional real, demonstrando seu potencial como ferramenta de apoio à identificação precoce de estudantes em risco.

Por fim, o Capítulo 7 apresenta as conclusões da pesquisa, destacando as principais contribuições, limitações identificadas e perspectivas para trabalhos futuros.

## **1.7 Declaração de uso de Inteligência Artificial**

Durante a elaboração desta dissertação, foram utilizadas ferramentas baseadas em Inteligência Artificial, especificamente o ChatGPT, desenvolvido pela OpenAI, e o Gemini, desenvolvido pelo Google, exclusivamente como apoio técnico e instrumental.

O uso dessas ferramentas limitou-se às seguintes atividades:

- Revisão gramatical e aprimoramento da clareza textual (ChatGPT e Gemini);
- Tradução de trechos entre português e inglês (ChatGPT);
- Sugestões de organização e estruturação textual (ChatGPT);
- Suporte ao desenvolvimento e à implementação da aplicação construída com o *framework Streamlit* (Gemini);
- Suporte na elaboração e ajuste de comandos em linguagem Python durante o desenvolvimento experimental, incluindo o uso do ambiente Google Colab (Gemini).

Assim, não houve utilização de IA na definição do problema de pesquisa, na construção da metodologia, na realização dos experimentos, na análise dos resultados ou na elaboração das conclusões. Todas as decisões científicas e técnicas foram conduzidas integralmente pela autora.

Todo o conteúdo gerado com apoio da ferramenta foi criteriosamente revisado, validado e, quando necessário, adaptado, sendo a autora integralmente responsável pela originalidade, veracidade e integridade deste trabalho.

## 2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta a fundamentação teórica que sustenta a abordagem metodológica adotada neste trabalho. Inicialmente, são apresentados os fundamentos do Aprendizado de Máquina, com ênfase nos métodos supervisionados utilizados nesse estudo. Na sequência, abordam-se as técnicas de seleção de atributos, destacando sua importância na redução de dimensionalidade, na mitigação de redundâncias e no aprimoramento do desempenho dos modelos. Posteriormente, são discutidas as estratégias de balanceamento de classes, fundamentais em cenários caracterizados por desbalanceamento entre categorias, como ocorre em estudos sobre evasão. Por fim, são apresentadas as métricas de avaliação adotadas neste trabalho, as quais permitem analisar de forma consistente e comparável o desempenho dos modelos avaliados. O capítulo fornece o embasamento teórico necessário para a compreensão das escolhas metodológicas e dos resultados obtidos ao longo da pesquisa.

### 2.1 Aprendizado de Máquina

O Aprendizado de Máquina (*Machine Learning* – ML) é um subcampo da Ciência da Computação que se dedica ao desenvolvimento de métodos e algoritmos capazes de aprender padrões a partir de dados, sem a necessidade de programação explícita para cada situação específica. De forma geral, um sistema de ML é capaz de melhorar seu desempenho em determinada tarefa à medida que é exposto a um volume maior de dados ou experiências anteriores (MITCHELL, 1997).

De maneira intuitiva, o ML pode ser comparado ao processo de aprendizagem humana: assim como uma pessoa aprende a reconhecer padrões após observar diversos exemplos, os algoritmos aprendem relações entre variáveis analisando dados históricos. Em vez de seguir regras fixas definidas manualmente, o modelo identifica regularidades nos dados e as utiliza para realizar previsões ou classificações sobre novas observações.

Segundo Witten, Frank e Hall (2016), o ML tem sido amplamente aplicado em áreas como reconhecimento de padrões, diagnóstico médico, recomendação de conteúdos, análise financeira e, mais recentemente, na educação, especialmente em estudos de predição de desempenho e evasão escolar.

Os métodos de ML podem ser classificados, de forma geral, em três grandes categorias, de acordo com a forma como os dados são apresentados ao algoritmo e com o tipo de tarefa a ser realizada: aprendizado supervisionado, aprendizado não supervisionado e aprendizado por

reforço. Como o aprendizado supervisionado será utilizado neste trabalho, ele é apresentado, na sequência deste capítulo, com maior detalhamento em relação aos demais.

### **2.1.1 *Aprendizado supervisionado***

No aprendizado supervisionado, o algoritmo é treinado a partir de um conjunto de dados rotulados, no qual cada instância possui um conjunto de atributos de entrada e uma variável alvo previamente conhecida (rótulo). O objetivo do modelo é aprender uma função que estabeleça a relação entre os atributos de entrada e o rótulo correspondente, permitindo que o sistema generalize esse conhecimento para realizar previsões em novos dados não observados anteriormente (MITCHELL, 1997). Nesse contexto, diferentes modelos e algoritmos podem ser empregados para realizar essa tarefa de aprendizado, incluindo métodos estatísticos e computacionais capazes de capturar padrões complexos nos dados (HASTIE; FRIEDMAN; TIBSHIRANI, 2006).

Esse tipo de aprendizado é amplamente utilizado em tarefas de classificação e regressão, sendo particularmente adequado quando o fenômeno de interesse já ocorreu no passado e seus resultados são conhecidos, como no caso da evasão escolar (HAN; KAMBER; PEI, 2012).

O funcionamento do ML supervisionado pode ser compreendido como um fluxo que inicia nos dados rotulados e vai até a geração de previsões para novos registros. Mais especificamente, o processo parte de um conjunto de dados de treinamento, no qual cada instância é descrita por um conjunto de atributos e associada a uma variável alvo previamente conhecida, como estudantes concluintes e não concluintes. Esses dados são utilizados na etapa de treinamento de um modelo supervisionado, no qual algoritmos de ML ajustam seus parâmetros internos a fim de aprender uma função capaz de mapear corretamente os atributos de entrada para as classes correspondentes. Durante esse processo, podem ser realizadas etapas de ajuste de hiperparâmetros e avaliação do desempenho do modelo. Após o treinamento, o modelo é aplicado a novos dados, produzindo previsões ou estimativas probabilísticas como, por exemplo, sobre o desfecho acadêmico. Esse fluxo caracteriza a essência do aprendizado supervisionado, no qual o conhecimento extraído a partir de exemplos rotulados é utilizado para generalizar decisões em dados não previamente observados.

Existem diversos algoritmos de ML supervisionado. Para este trabalho, conforme justificado no Capítulo 3, foram selecionados e avaliados três algoritmos consolidados na literatura por sua robustez, desempenho em dados tabulares e adequação a problemas com relações não lineares entre atributos: Árvore de Decisão (*Decision Tree*), Floresta Aleatória (*Random Forest*)

e *Extreme Gradient Boosting (XGBoost)*. Esses algoritmos são apresentados a seguir.

#### 2.1.1.1 *Decision Tree*

*Decision Tree* (DT) é um modelo preditivo que organiza o processo de decisão em uma estrutura hierárquica composta por nós internos (testes sobre atributos), ramos (resultados desses testes) e folhas (classes previstas). O algoritmo busca particionar iterativamente o conjunto de dados de modo a maximizar a pureza das classes em cada subdivisão (QUINLAN, 1993).

Do ponto de vista intuitivo, uma árvore de decisão pode ser comparada a um fluxograma de perguntas sucessivas, no qual cada resposta conduz a um novo critério até que se chegue a uma decisão final. Essa característica confere ao modelo elevada interpretabilidade, aspecto particularmente relevante em contextos educacionais, nos quais a explicação dos fatores associados à evasão é tão importante quanto a previsão em si (ROMERO; VENTURA, 2010).

Entretanto, árvores muito profundas tendem a se ajustar excessivamente aos dados de treinamento, fenômeno conhecido como sobreajuste (*overfitting*). Por esse motivo, o controle de sua complexidade por meio de hiperparâmetros é essencial (QUINLAN, 1993; HASTIE; FRIEDMAN; TIBSHIRANI, 2006; KUHN; JOHNSON, 2013).

O hiperparâmetro *max\_depth* define a profundidade máxima permitida para a construção das árvores de decisão que compõem os modelos avaliados. Esse parâmetro exerce papel fundamental no controle da complexidade do modelo, uma vez que limita o número de divisões sucessivas realizadas na estrutura da árvore. Dessa forma, sua configuração influencia diretamente o equilíbrio entre a capacidade do modelo de capturar padrões presentes nos dados e sua capacidade de generalização para novos conjuntos de dados, contribuindo para a mitigação do *overfitting* (KUHN; JOHNSON, 2013; HASTIE; FRIEDMAN; TIBSHIRANI, 2006).

De modo geral, valores mais baixos de *max\_depth* tendem a produzir modelos mais simples, reduzindo o risco de *overfitting*, porém podem levar ao subajuste (*underfitting*), quando o modelo se torna incapaz de capturar padrões relevantes presentes nos dados. Por outro lado, valores mais elevados aumentam a flexibilidade do modelo e permitem representar relações mais complexas entre os atributos, embora também elevem a probabilidade de o modelo se ajustar excessivamente aos dados de treinamento (KUHN; JOHNSON, 2013).

### 2.1.1.2 *Random Forest*

*Random Forest* (RF) é um método de *ensemble* que combina diversas árvores de decisão treinadas de forma independente, a partir de subconjuntos aleatórios de dados e atributos. A previsão final do modelo é obtida por meio de um processo de votação entre as árvores individuais (BREIMAN, 2001).

A principal vantagem dessa abordagem reside na redução da variância do modelo, tornando-o mais robusto a ruídos e menos suscetível ao sobreajuste em comparação a uma única árvore. Em problemas educacionais complexos, nos quais múltiplos fatores interagem de forma não linear, o *Random Forest* tem se mostrado particularmente eficaz (HAN; KAMBER; PEI, 2012).

Apesar de apresentar menor interpretabilidade global em relação a uma árvore isolada, o *Random Forest* mantém boa capacidade explicativa por meio da análise de importância dos atributos.

### 2.1.1.3 *Extreme Gradient Boosting (XGBoost)*

Do ponto de vista conceitual, o *XGBoost* pode ser compreendido como um processo de aprendizado iterativo baseado em *gradient boosting*, no qual modelos são construídos sequencialmente de modo que cada nova árvore de decisão seja ajustada para reduzir os erros cometidos pelos modelos anteriores. Nesse processo, o algoritmo aprende progressivamente a partir dos resíduos gerados nas etapas anteriores, refinando o modelo de forma incremental. Essa abordagem destaca-se por sua elevada eficiência computacional, capacidade de lidar com grandes volumes de dados e excelente desempenho preditivo em diferentes domínios de aplicação (CHEN; GUESTRIN, 2016; FRIEDMAN, 2001).

No *XGBoost*, árvores de decisão do tipo *Classification and Regression Trees* (CART) são utilizadas como modelos base e combinadas por meio do processo de *boosting*, resultando em um modelo *ensemble* mais robusto e com elevada capacidade preditiva. Além disso, o algoritmo incorpora técnicas de regularização e otimização computacional que contribuem para melhorar a generalização do modelo e reduzir o risco de *overfitting*, o que tem tornado o *XGBoost* amplamente utilizado em tarefas de classificação e regressão nos mais diversos domínios de aplicação (CHEN; GUESTRIN, 2016).

Entretanto, apesar de sua elevada capacidade preditiva, o *XGBoost* tende a apresentar menor interpretabilidade quando comparado a modelos mais simples, como árvores de deci-

são individuais, exigindo técnicas adicionais de interpretação para análise mais detalhada dos resultados (MOLNAR, 2020).

### **2.1.2 *Aprendizado não supervisionado***

No aprendizado não supervisionado, os dados não possuem rótulos previamente definidos. Nesse tipo de abordagem, o objetivo do algoritmo é identificar estruturas, padrões ou agrupamentos naturais presentes nos dados, sem conhecimento prévio das classes ou categorias existentes. Dessa forma, os métodos buscam revelar relações e regularidades ocultas a partir da organização intrínseca dos dados (HAN; KAMBER; PEI, 2012; HASTIE; FRIEDMAN; TIBSHIRANI, 2006).

O funcionamento do ML não supervisionado pode ser evidenciado pelo processo de extração de padrões a partir de dados não rotulados. Diferentemente do aprendizado supervisionado, nesse paradigma os dados de entrada não possuem uma variável alvo previamente definida, sendo compostos apenas por atributos descritivos. Nessa abordagem, os algoritmos buscam identificar estruturas inerentes ao conjunto de dados por meio de técnicas como agrupamento (*clustering*) e redução de dimensionalidade. Como resultado, os dados podem ser organizados em grupos de instâncias com características semelhantes ou projetados em espaços de menor dimensão, facilitando a análise e interpretação dos padrões identificados. No contexto educacional, tais técnicas possibilitam a identificação de perfis de estudantes e padrões de comportamento acadêmico, contribuindo para análises exploratórias que podem anteceder ou complementar modelos preditivos (HAN; KAMBER; PEI, 2012; WITTEN; FRANK; HALL, 2016; HASTIE; FRIEDMAN; TIBSHIRANI, 2006).

De modo geral, a principal diferença entre o aprendizado supervisionado e o aprendizado não supervisionado está na presença de rótulos nos dados utilizados durante o treinamento. Enquanto no aprendizado supervisionado o modelo aprende a partir de exemplos previamente rotulados, buscando prever a variável alvo para novos dados, no aprendizado não supervisionado não há rótulos disponíveis, e os algoritmos têm como objetivo identificar estruturas, padrões ou agrupamentos inerentes ao conjunto de dados (MITCHELL, 1997; HAN; KAMBER; PEI, 2012; HASTIE; FRIEDMAN; TIBSHIRANI, 2006).

### 2.1.3 Aprendizado por reforço

O aprendizado por reforço baseia-se na interação contínua de um agente com um ambiente, no qual ações são realizadas e recompensas ou penalidades são recebidas em função dessas ações (SUTTON; BARTO, 2018). Mais especificamente, no ML por reforço o agente observa o estado atual do ambiente e, com base em uma política de decisão, seleciona uma ação a ser executada. Como resposta, o ambiente retorna um novo estado e um sinal de recompensa ou punição, que indica a qualidade da ação realizada. O ML por reforço evidencia-se também pelo equilíbrio entre exploração, quando o agente testa novas ações, e exploração, quando utiliza estratégias já aprendidas, elemento central para a convergência do processo de aprendizagem.

Esse tipo de aprendizado é amplamente utilizado em problemas de controle, robótica e jogos (SUTTON; BARTO, 2018). No contexto educacional, embora menos explorado, o aprendizado por reforço pode ser aplicado em sistemas adaptativos e recomendações personalizadas, nos quais as decisões são ajustadas dinamicamente a partir do *feedback* obtido.

## 2.2 Seleção de atributos

Em problemas de ML aplicados a dados reais, é comum que os conjuntos de dados apresentem um grande número de atributos, muitos dos quais podem ser redundantes, irrelevantes ou ruidosos. A utilização indiscriminada de todos os atributos disponíveis pode comprometer o desempenho dos modelos preditivos, aumentar o custo computacional e dificultar a interpretação dos resultados (HAN; KAMBER; PEI, 2012).

Essa característica é particularmente frequente em bases de dados educacionais institucionais, que costumam integrar informações provenientes de diferentes sistemas administrativos e acadêmicos, originalmente coletadas para fins operacionais e não necessariamente para análise preditiva ou exploratória (HAN; KAMBER; PEI, 2012; KUHN; JOHNSON, 2013).

Nesse contexto, a seleção de atributos (*feature selection*) surge como uma etapa fundamental do pré-processamento, cujo objetivo é identificar um subconjunto de variáveis que represente adequadamente o fenômeno estudado, preservando ou melhorando o desempenho do modelo e reduzindo a complexidade do espaço de atributos (KOHAVI; JOHN, 1997).

Na sequência desta seção, são apresentadas as três técnicas de seleção de atributos aplicadas nos experimentos desta dissertação: *SelectKBest* com ANOVA F, seleção baseada na correlação com a variável alvo e, por fim, Análise de Componentes Principais (*Principal Component Analysis* – PCA) (técnica de redução de dimensionalidade utilizada nesta dissertação

na etapa de seleção de atributos, conforme descrito a seguir).

### 2.2.1 *SelectKBest com ANOVA F*

O *SelectKBest* é uma técnica de seleção de atributos do tipo filtro supervisionado, que avalia individualmente cada atributo em relação à variável alvo, atribuindo uma pontuação que representa seu grau de relevância para a tarefa de classificação.

Na presente dissertação, foi utilizado o teste estatístico ANOVA F como função de avaliação empregada pelo *SelectKBest* para calcular essa pontuação. Esse teste compara a variância entre grupos com a variância dentro dos grupos, permitindo identificar atributos que apresentam maior capacidade de discriminar entre as classes analisadas. Dessa forma, os atributos que obtêm as maiores pontuações são selecionados como os mais relevantes para a construção do modelo (WITTEN; FRANK; HALL, 2016; KUHN; JOHNSON, 2013).

A estatística F assume valores elevados quando as médias do atributo diferem significativamente entre as classes, indicando maior poder discriminativo. Essa técnica permite selecionar diretamente os  $k$  atributos mais relevantes, mantendo a interpretabilidade das variáveis originais (WITTEN; FRANK; HALL, 2016).

O *SelectKBest* com ANOVA F é amplamente utilizado em problemas de classificação supervisionada por sua simplicidade, eficiência computacional e fundamentação estatística sólida (KUHN; JOHNSON, 2013). No contexto da seleção de atributos (*features*), o procedimento do teste F da ANOVA avalia a significância estatística de cada atributo, utilizando o *F-score*, para reduzir efetivamente a dimensionalidade dos dados.

Segundo Raufi e Longo (2024), o método *SelectKBest* permite selecionar as  $k$  melhores características com base em um ranqueamento. Quando combinado com o teste F da ANOVA, o *SelectKBest* avalia todos os atributos e seleciona aqueles que apresentam os maiores escores, ou seja, as características consideradas mais relevantes para a tarefa de classificação ou análise dos dados.

### 2.2.2 *Seleção baseada na correlação com a variável alvo*

A seleção baseada na correlação consiste na identificação dos atributos que apresentam maior associação estatística com a variável alvo, considerando-se o valor absoluto do coeficiente de correlação. Essa abordagem busca selecionar variáveis que variam de forma consistente com o desfecho analisado.

Embora simples, esse método é eficaz como técnica de filtragem inicial, permitindo identificar rapidamente atributos potencialmente relevantes e eliminar variáveis pouco informativas. Segundo (HAN; KAMBER; PEI, 2012), técnicas baseadas em correlação são particularmente úteis em etapas exploratórias e em cenários nos quais a interpretabilidade dos resultados é prioritária.

### **2.2.3 *Principal Component Analysis (PCA)***

A *Principal Component Analysis* (PCA) é uma técnica estatística de redução de dimensionalidade, cujo objetivo é transformar o conjunto original de atributos em um novo conjunto de variáveis ortogonais, denominadas componentes principais. Cada componente é uma combinação linear dos atributos originais e é ordenado de acordo com a quantidade de variância dos dados que explica (JOLLIFFE, 1986).

Do ponto de vista intuitivo, a PCA busca concentrar a maior parte da informação presente nos atributos de entrada em um número reduzido de componentes principais. Esses componentes correspondem a combinações lineares das variáveis originais e são construídos de forma a preservar a maior variância possível dos dados, reduzindo assim a dimensionalidade do conjunto de atributos. Embora não seja uma técnica de seleção direta de atributos, a PCA é amplamente utilizada para avaliar o impacto da redução de dimensionalidade no desempenho de modelos de ML (HAN; KAMBER; PEI, 2012; JOLLIFFE, 2002).

Entretanto, uma limitação importante da PCA é a perda de interpretabilidade, uma vez que os componentes resultantes não correspondem diretamente a atributos observáveis, aspecto relevante em estudos educacionais.

## **2.3 Balanceamento de Dados**

Em problemas de classificação supervisionada, é comum que as classes não estejam distribuídas de forma equilibrada, fenômeno conhecido como desbalanceamento de classes. Esse cenário ocorre quando uma das classes possui número significativamente maior de observações em relação à outra.

O desbalanceamento de classes representa um desafio relevante em tarefas de classificação, pois muitos algoritmos de ML são projetados para maximizar a acurácia global e, conseqüentemente, tendem a favorecer a classe majoritária. Como resultado, modelos aparentemente precisos podem apresentar baixa capacidade de identificar corretamente instâncias da

classe minoritária (HE; GARCIA, 2009).

Uma característica frequente em bases educacionais é o desbalanceamento entre as classes, no qual a quantidade de registros de uma classe é significativamente maior do que da outra. Em estudos sobre evasão em cursos de tecnologia como Computação, por exemplo, pode haver predominância de estudantes não concluintes, enquanto os concluintes constituem a classe minoritária (SARAIVA *et al.*, 2021a; LINHARES *et al.*, 2025). Esse é o caso da base utilizada nos experimentos do presente trabalho.

Diante desse desafio, diversas técnicas de balanceamento de classes têm sido propostas na literatura, podendo ser classificadas, de forma geral, em três categorias principais:

- Técnicas de subamostragem (*undersampling*): As técnicas de subamostragem consistem na remoção de instâncias da classe majoritária, com o objetivo de equilibrar a proporção entre as classes. Embora essa abordagem possa reduzir o viés em favor da classe majoritária, ela implica a perda de informações potencialmente relevantes, especialmente em bases de dados com volume limitado. Em contextos educacionais, essa perda pode ser crítica, pois cada registro representa uma trajetória acadêmica única (HAN; KAMBER; PEI, 2012).
- Técnicas de sobreamostragem (*oversampling*): As técnicas de sobreamostragem buscam equilibrar as classes por meio do aumento do número de instâncias da classe minoritária, sem remover exemplos da classe majoritária. Esse aumento pode ocorrer por meio da replicação de instâncias existentes (sobreamostragem aleatória) ou da geração de novos exemplos sintéticos (sobreamostragem sintética). De forma intuitiva, a sobreamostragem pode ser comparada ao reforço do estudo de conteúdos mais difíceis: ao dedicar mais tempo e exemplos a esses conteúdos, busca-se equilibrar o aprendizado, sem descartar conhecimentos já adquiridos. Entre as abordagens mais conhecidas estão os métodos baseados na geração de exemplos sintéticos, como o *Synthetic Minority Over-sampling Technique* (SMOTE), amplamente utilizado para lidar com conjuntos de dados desbalanceados (CHAWLA *et al.*, 2002; HE; GARCIA, 2009).
- Técnicas híbridas e baseadas em custo: Além das abordagens anteriores, existem técnicas híbridas que combinam estratégias de subamostragem e sobreamostragem, buscando aproveitar as vantagens de ambas para melhorar o equilíbrio entre as classes. Também existem métodos baseados em custo (*cost-sensitive learning*), nos quais erros cometidos na classe minoritária recebem penalidades maiores durante o treinamento, incentivando o modelo a prestar maior atenção a essas instâncias (HE; GARCIA, 2009; FERNÁNDEZ *et*

*al.*, 2018). Embora eficazes em determinados cenários, essas técnicas geralmente exigem ajustes adicionais de parâmetros e maior complexidade na modelagem, razão pela qual não foram adotadas neste estudo.

Neste estudo, optou-se pela utilização as seguintes técnicas de sobreamostragem, aplicadas exclusivamente ao conjunto de treino, por razões metodológicas e práticas justificadas no Capítulo 4.

### **2.3.1 *Random OverSampler***

Consiste na replicação aleatória de exemplos da classe minoritária até que se atinja um equilíbrio entre as classes. Trata-se de uma abordagem simples, pois não altera a distribuição original dos atributos, apenas aumenta a frequência dos exemplos minoritários. No entanto, conforme apontado por He e Garcia (2009), essa técnica pode aumentar o risco de sobreajuste, uma vez que os mesmos exemplos são repetidamente apresentados ao modelo durante o treinamento.

### **2.3.2 *Synthetic Minority Over-sampling Technique (SMOTE)***

Propõe a geração de novos exemplos sintéticos da classe minoritária a partir da interpolação entre observações vizinhas no espaço de atributos. Diferentemente da simples replicação de registros, essa técnica busca ampliar a região ocupada pela classe minoritária, promovendo uma fronteira de decisão mais generalizável. De acordo com Chawla *et al.* (2002), o SMOTE é eficaz para reduzir o viés em favor da classe majoritária e melhorar a capacidade dos modelos de identificar corretamente instâncias da classe minoritária.

### **2.3.3 *BorderlineSMOTE***

Uma variação do SMOTE que concentra a geração de exemplos sintéticos nas regiões próximas à fronteira de decisão entre as classes. Essa abordagem parte do pressuposto de que instâncias minoritárias localizadas próximas à classe majoritária são mais difíceis de classificar e, portanto, mais informativas para o processo de aprendizagem. Segundo Han, Wang e Mao (2005), o foco nessas regiões críticas contribui para o refinamento da fronteira de decisão e pode resultar em ganhos de desempenho, especialmente em termos de revocação da classe minoritária.

## 2.4 Métricas de avaliação de desempenho

A avaliação do desempenho dos modelos de classificação é uma etapa fundamental para determinar sua utilidade prática. Neste estudo, foram utilizadas quatro métricas amplamente reconhecidas na literatura na avaliação dos modelos de classificação: Acurácia (*Accuracy*), Precisão (*Precision*), Revocação ou Cobertura (*Recall*) e *F1-score*. Essas métricas são derivadas da Matriz de Confusão, que apresenta a distribuição das predições do modelo em termos de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos.

### 2.4.1 Acurácia

A acurácia mede a proporção total de previsões corretas em relação ao número total de observações. Embora seja uma métrica intuitiva, sua interpretação pode ser enganosa em cenários de desbalanceamento de classes, nos quais um modelo pode obter alta acurácia simplesmente ao prever sempre a classe majoritária (HE; GARCIA, 2009).

Essa medida é calculada segundo a Equação 2.1. Em resumo, a acurácia irá indicar qual a proporção de egressos com êxito e egressos sem êxito que foram corretamente classificados.

$$Acuracia = \frac{VP + VN}{VP + VN + FP + FN}, \quad (2.1)$$

onde, considerando que a classe positiva é a egresso sem êxito, *VP* representa os verdadeiro positivos, ou seja, os egressos sem êxito classificados corretamente; *VN* os verdadeiro negativos, ou seja, os egressos com êxito classificados corretamente; *FP* os falso positivos, ou seja, os egressos com êxito classificados incorretamente como egressos sem êxito; enquanto que *FN* representa os falso negativos, ou seja, os egressos sem êxito classificados incorretamente como egressos como êxito.

### 2.4.2 Precisão

A precisão representa a proporção de instâncias corretamente classificadas em uma determinada classe de interesse em relação ao total de instâncias que o modelo atribuiu a essa mesma classe. Essa métrica é relevante quando o custo de falsos positivos é elevado, ou seja, quando é importante evitar classificar incorretamente instâncias como pertencentes a essa classe. Por exemplo, ao identificar estudantes em risco de evasão, a precisão indica quantos dos

alunos classificados como em risco de evasão realmente apresentam esse perfil (SOKOLOVA; LAPALME, 2009).

Para o problema em questão, a precisão irá indicar qual a proporção dos discentes classificados como egressos sem êxito era realmente de egressos sem êxito. Essa medida é calculada conforme a Equação 2.2.

$$Precision = \frac{VP}{VP + FP} \quad (2.2)$$

### 2.4.3 Recall

*Recall* mede a capacidade do modelo de identificar corretamente todas as instâncias pertencentes à classe de interesse. Em estudos sobre evasão escolar, essa métrica é particularmente importante, pois reflete a capacidade do modelo de não deixar de identificar estudantes em risco de evasão, mesmo que isso implique algum aumento de falsos positivos (TINTO, 2012).

No caso em estudo, o *Recall* é a fração de verdadeiros positivos, ou seja, ele irá determinar entre todas as amostras que realmente eram de egressos sem êxito, qual a proporção classificada como egressos sem êxito. Essa métrica é calculada conforme a Equação 2.3.

$$Recall = \frac{VP}{VP + FN} \quad (2.3)$$

### 2.4.4 F1-score

O *F1-score* corresponde à média harmônica entre Precisão e *Recall*, oferecendo uma medida equilibrada do desempenho do modelo. Essa métrica é especialmente indicada em problemas com classes desbalanceadas, pois penaliza tanto falsos positivos quanto falsos negativos (SOKOLOVA; LAPALME, 2009).

O *F1-score* combina, em uma única métrica, os valores de precisão e *recall*. Essa métrica é calculada de acordo com a Equação 2.4.

$$F1Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (2.4)$$

Diante do contexto desta pesquisa, no qual a identificação adequada de estudantes não concluintes é fundamental para subsidiar ações preventivas, o *F1-score* foi priorizado como critério principal para a seleção do melhor cenário experimental.

### 3 MAPEAMENTO SISTEMÁTICO DA LITERATURA

Este capítulo foi aprovado como artigo completo no IX Congresso sobre Tecnologias na Educação (Ctrl+E), apoiado pela Sociedade Brasileira de Computação (SBC) (BASTOS *et al.*, 2025). No contexto desta dissertação, o texto foi enriquecido com conteúdos adicionais, de modo a aprofundar conceitos, explicitar escolhas metodológicas e ampliar a discussão teórica apresentada.

O capítulo apresenta um Mapeamento Sistemático da Literatura (*Systematic Literature Mapping* - SLM) do uso de Mineração de Dados Educacionais (*Educational Data Mining* - EDM) e algoritmos de Aprendizado de Máquina (*Machine Learning* - ML) na identificação e prevenção da evasão escolar no Brasil. A pesquisa busca caracterizar o estado da arte da área, mapear tendências, destacar boas práticas e apontar lacunas que possam orientar futuras investigações, contribuindo para o avanço da aplicação de EDM e ML no enfrentamento da evasão escolar.

Além disso, os resultados deste SLM serviram de referência para a proposta desta dissertação. Os resultados do mapeamento subsidiaram diretamente a definição da abordagem metodológica desta dissertação, orientando a escolha do problema como uma tarefa de classificação supervisionada, a seleção dos algoritmos avaliados, o uso de métricas adequadas para cenários desbalanceados e a adoção de técnicas de pré-processamento recorrentes nos estudos analisados. Dessa forma, o SLM não apenas contextualizou a pesquisa, mas serviu como base empírica para a construção e fundamentação da proposta desenvolvida neste trabalho.

O SLM foi estruturado com base em diretrizes consolidadas para revisões sistemáticas, adaptadas às especificidades da área de tecnologia educacional. Essa abordagem garante rigor metodológico, transparência e reprodutibilidade na análise.

#### 3.1 Trabalhos Relacionados

Nesta seção, foram buscados por trabalhos que realizam mapeamento sistemático com foco em mineração de dados educacionais no contexto da evasão escolar.

O artigo de Colpo *et al.* (2020) realiza um mapeamento a partir das publicações do Congresso Brasileiro de Informática na Educação (CBIE), buscando responder a questões sobre a natureza da evasão (se em nível de curso, disciplina, instituição ou nível de ensino), os dados utilizados (tipos e volume de dados) e as técnicas aplicadas (algoritmos e ferramentas). Os autores destacam o uso expressivo de tarefas de classificação, com predominância de algoritmos

baseados em árvores de decisão. O estudo revela que a maior parte das investigações foca no ensino superior presencial, utilizando principalmente dados acadêmicos, sociais e econômicos. No entanto, o texto aponta lacunas críticas, como a baixa exploração do ensino básico e a falta de sistemas que auxiliem gestores na tomada de decisão em tempo real. Por fim, observa-se um crescimento recente no interesse pelo tema, reforçando a importância de transformar dados em estratégias preventivas eficazes nas instituições brasileiras. Apesar da relevância e da proximidade temática com o presente estudo, a análise se limita às publicações do CBIE e a um período anterior (2006–2019), distinto do foco desta pesquisa (2020–2024).

No mesmo contexto, o trabalho de Morais *et al.* (2021) conduz uma revisão sistemática focada na aplicação de Aprendizagem de Máquina para combater a evasão escolar no ensino básico. Os autores investigam como diferentes modelos computacionais, especialmente os de regressão, ajudam a prever a desistência dos alunos e a identificar padrões socioeconômicos críticos. A análise revela que, embora técnicas como a árvore de decisão sejam populares, o uso específico de modelos de regressão ainda é limitado na área educacional. O estudo destaca que a baixa renda familiar é o principal fator que impulsiona o abandono escolar, seguido por questões de idade e estrutura institucional. Por fim, o texto sugere a implementação de sistemas de detecção precoce como uma ferramenta essencial para apoiar a criação de políticas públicas eficazes.

O trabalho de Santos, Saraiva e Oliveira (2021) consiste em uma revisão sistemática da literatura focada na predição de evasão no ensino superior, utilizando as bases *Scopus* e *IEEE Xplore* para selecionar 21 artigos relevantes publicados a partir de 2015. A pesquisa identificou que os atributos mais utilizados para treinar modelos preditivos são de natureza pessoal/social (90,4%) e acadêmica (76,1%), como notas e registros escolares. Entre as técnicas de aprendizagem de máquina, as Árvores de Decisão foram as mais populares, presentes em 66,6% dos estudos, seguidas por métodos Bayesianos e Regressão Logística. Em termos de algoritmos específicos, o *Naive Bayes* (47,6%) foi o mais implementado, e a acurácia consolidou-se como a métrica de avaliação preferida pelos pesquisadores (42,8%) .

O estudo de Silva e Roman (2021) destaca-se por realizar uma análise comparativa entre trabalhos nacionais e internacionais de Mineração de Dados Educacionais (EDM), sendo o único a incluir bases brasileiras como a SBC OpenLib (SOL) e a RBIE. Analisando um total de 50 artigos, os autores observaram que a ferramenta Weka domina historicamente o cenário, embora a biblioteca Scikit-learn tenha crescido em uso desde 2019. No cenário nacional, os algoritmos com melhor desempenho foram o SVM e a Árvore de Decisão, enquanto no internacional

destacaram-se o ADTree e o *Random Forest*. Além disso, a pesquisa revelou que a maioria das bases de dados analisadas foca em dados acadêmicos (43 trabalhos) e possui entre 501 e 5.000 registros.

O mapeamento sistemático de Jesus e Gusmão (2024) investigou o estado da arte da aplicação de Mineração de Dados e Aprendizagem de Máquina na evasão escolar, analisando 71 estudos relevantes de cinco bases bibliográficas internacionais. Os resultados indicam que o interesse pela área é crescente, com o ano de 2020 apresentando o maior volume de publicações e a China liderando o ranking por países. O estudo confirmou a predominância de pesquisas voltadas ao ensino superior (78%) e à modalidade presencial (70%), reforçando o uso majoritário de algoritmos baseados em árvores de decisão (mencionados em 35 estudos) e regressão logística. As variáveis de maior destaque nos modelos foram o desempenho acadêmico ao longo do tempo, disciplinas específicas e o tipo de curso.

Tabela 1 – Comparação entre mapeamentos sistemáticos sobre evasão escolar e Mineração de Dados Educacionais

| Trabalho                          | Tipo de estudo                             | Período analisado | Bases de dados                         | Nível de ensino | Algoritmos predominantes               | Diferencial / Limitações                                |
|-----------------------------------|--------------------------------------------|-------------------|----------------------------------------|-----------------|----------------------------------------|---------------------------------------------------------|
| Colpo <i>et al.</i> (2020)        | Mapeamento sistemático                     | 2006–2019         | CBIE                                   | Ensino superior | Árvore de Decisão                      | Restrito ao CBIE; ausência de sistemas práticos         |
| Morais <i>et al.</i> (2021)       | Revisão sistemática                        | Até 2020          | Bases internacionais                   | Ensino básico   | Regressão, Árvore de Decisão           | Foco restrito ao ensino básico                          |
| Santos, Saraiva e Oliveira (2021) | Revisão sistemática                        | 2015–2020         | Scopus, IEEE Xplore                    | Ensino superior | Árvore de Decisão, Naive Bayes         | Pouca ênfase em aplicações práticas                     |
| Silva e Roman (2021)              | Mapeamento sistemático                     | Até 2021          | Internacionais + nacionais (SOL, RBIE) | Ensino superior | SVM, Árvore de Decisão, Random Forest  | Bases nacionais ainda pouco exploradas                  |
| Jesus e Gusmão (2024)             | Mapeamento sistemático                     | Até 2023          | Bases internacionais                   | Ensino superior | Árvore de Decisão, Regressão Logística | Ênfase internacional; cenário nacional secundário       |
| <b>Presente estudo</b>            | Mapeamento sistemático + aplicação prática | 2020–2024         | Bases nacionais e internacionais       | Pós-graduação   | Decision Tree, Random Forest, XGBoost  | Abordagem atual, nacional e com aplicação computacional |

Fonte: Elaborada pela autora (2026).

A Tabela 1 sintetiza as principais características dos mapeamentos sistemáticos e revisões da literatura apresentados, permitindo uma comparação direta entre os estudos analisados e o presente trabalho. Observa-se que, embora os estudos anteriores tenham contribuído significativamente para a consolidação da área, grande parte deles apresenta limitações quanto ao período analisado, ao escopo geográfico ou à ausência de aplicações práticas. Em especial, nota-se a predominância de investigações focadas no ensino superior, com ênfase em algoritmos baseados em árvores de decisão, mas com menor exploração de abordagens mais recentes e de cenários nacionais atualizados. Nesse contexto, o presente estudo se diferencia por contemplar

um período mais recente (2020–2024), incorporar bases nacionais e internacionais e avançar além do mapeamento teórico, propondo e avaliando modelos preditivos aplicados ao contexto da pós-graduação, culminando no desenvolvimento de uma aplicação computacional para apoio à tomada de decisão institucional.

## 3.2 Metodologia

Essa seção detalha as etapas da metodologia da SLM.

### 3.2.1 Etapa 1: Definição das Questões de Pesquisa (QP)

Nesta etapa, as QP foram formuladas para responder aos objetivos do mapeamento, abrangendo aspectos centrais da aplicação de EDM e ML no combate à evasão escolar. Um processo iterativo de revisão da literatura e análise de demandas práticas foi realizado para garantir a relevância e a abrangência das questões. Assim, as QP elaboradas foram:

- **QP1:** Quais trabalhos científicos utilizam EDM no contexto da evasão escolar?
- **QP2:** Quais as ferramentas/bibliotecas utilizadas pelos trabalhos?
- **QP3:** Quais os algoritmos de aprendizagem de máquina utilizados nos trabalhos?
- **QP4:** Quais os algoritmos com melhor desempenho?
- **QP5:** Quantos registros existem nas bases de dados analisadas?
- **QP6:** Quantos atributos estão presentes nas bases de dados analisadas?
- **QP7:** Quais os tipos de dados analisados nos trabalhos?

### 3.2.2 Etapa 2: Identificação dos Estudos

Na segunda etapa, foi estruturada a *string* de busca com termos-chave das áreas de tecnologia da informação e educação: ((*Prediction* OR *Classification*) AND “*Student Dropout*” AND (“*Machine Learning*” OR “*Data Mining*” OR “*Data Science*”)).

As buscas, realizadas em dezembro de 2024, abrangeram seis repositórios científicos de relevância nacional e internacional: *ACM Digital Library*, *IEEE Xplore*, *ScienceDirect*, *SpringerLink*, *SBC OpenLib* (SOL) e *Revista Brasileira de Informática na Educação* (RBIE). Esses repositórios foram selecionados por sua representatividade em publicações de tecnologia

educacional e informática, com ênfase em fontes que abrigam estudos brasileiros, como SOL e RBIE. Ao final dessa fase, foram identificados 67 artigos.

3.2.3 *Etapa 3: Seleção e Avaliação dos Estudos*

Com o objetivo de selecionar os artigos mais adequados para responder às QP, foram definidos critérios de inclusão e exclusão.

Os critérios de inclusão exigiam que os trabalhos fossem nacionais, publicados entre 2020 e 2024, utilizassem EDM e abordassem a evasão escolar em qualquer nível de ensino.

Foram excluídos estudos não escritos em português ou inglês, trabalhos em andamento, duplicados ou que estivessem fora do escopo da pesquisa.

Após a aplicação desses critérios, 46 artigos foram incluídos na análise.

3.2.4 *Etapa 4: Síntese dos Dados e Apresentação dos Resultados*

Nesta etapa, foram desenvolvidas visualizações para responder a cada QP da Seção 3.2.1, a fim de analisar os atributos de forma mais específica e direcionada. Essas visualizações foram construídas utilizando a ferramenta *Business Intelligence Power BI*, versão Desktop de maio de 2025. Os resultados são apresentados e discutidos na próxima seção.

3.3 Resultados e Discussões

3.3.1 *QP1: Quais trabalhos científicos utilizam EDM no contexto da evasão escolar?*

A Tabela 2 apresenta os 46 trabalhos selecionados. O ano com mais trabalhos foi 2021 (12), e com menos trabalhos foi o de 2020 (7).

Tabela 2 – Trabalhos nacionais selecionados.

| ID | Referências                                                                                                                                                                                                                                      | ID  | Referências                                                                                                                                                                                                                                                                                                              |
|----|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| E1 | Sonnenstrahl, T. S., Bernardi, G., and Pertile, S. (2021). <b>Análise de interações do ambiente virtual de aprendizagem para predição de evasão em cursos no ensino a distância.</b> EaD em Foco, 11(1). (SONNENSTRAHL; BERNARDI; PERTILE, 2021) | E24 | Silva, D., Tamayo, S., Pessoa, M., Pires, F., Oliveira, D., Oliveira, E., and Carvalho, L. (SONNENSTRAHL; BERNARDI; PERTILE, 2021) (2020). <b>Minerando dados de um juiz on-line para prever a evasão de estudantes em disciplinas introdutórias de programação.</b> XXXI Simpósio Brasileiro de Informática na Educação |

Tabela 2 – Trabalhos nacionais selecionados (continuação)

| ID | Referências                                                                                                                                                                                                                                                                                                                  | ID  | Referências                                                                                                                                                                                                                                                                                                                          |
|----|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| E2 | Noetzold, E. and de L. Pertile, S. (2021). <b>Análise e predição de evasão dos alunos de um curso de graduação em sistemas de Informação por meio da mineração de dados educacionais</b> . RENOTE. (NOETZOLD; PERTILE, 2021)                                                                                                 | E25 | Kantorski, G., Martins, R., Balejo, A., and Frick, M. (2023). <b>Mineração de Dados Educacionais para Predição da Evasão em Cursos de Graduação Presenciais no Ensino Superior</b> . XXXIV Simpósio Brasileiro de Informática na Educação (KANTORSKI <i>et al.</i> , 2023)                                                           |
| E3 | Fonseca Silveira, R., Holanda, M., Ramos, G. N., Victorino, M., and Da Silva, D. (2022). <b>Analysis of Student Performance and Social-economic Data in Introductory Computer Science Courses at the University of Brasília</b> 2022 IEEE Frontiers in Education Conference (FIE) (SILVEIRA <i>et al.</i> , 2022)            | E26 | Carvalho, C., Mattos, J., and Aguiar, M. (2023). <b>Avaliação da interpretabilidade de modelos por meio da clusterização de explicações no contexto da predição de evasão no ensino superior</b> . XXXIV Simpósio Brasileiro de Informática na Educação.(CARVALHO; MATTOS; AGUIAR, 2023)                                             |
| E4 | Silva, J. C. L. (2023). <b>Definição de modelos de aprendizado de máquina para predição de evasão de alunos do curso técnico</b> Refas - Revista Fatec Zona Sul. (SILVA, 2023)                                                                                                                                               | E27 | Falcão, A., Villwock, R., and Miloca, S. (2023). <b>Análise de dados pré-universidade para predizer a evasão de alunos ingressantes em uma instituição de ensino superior</b> . XXXIV Simpósio Brasileiro de Informática na Educação (FALCÃO; VILLWOCK; MILOCA, 2023)                                                                |
| E5 | Bitencourt, W. A., Silva, D. M., and Xavier, G. d. C. (2022). <b>Pode a inteligência artificial apoiar ações contra evasão escolar universitária?</b> . Ensaio: Avaliação e Políticas Públicas em Educação (BITENCOURT; SILVA; XAVIER, 2022)                                                                                 | E28 | Erica Carmo, Gasparini, I., and Oliveira, E. (2022). <b>Identificação de Trajetórias de Aprendizagem em um Curso de Graduação e sua relação com a Evasão Escolar</b> . XXXIII Simpósio Brasileiro de Informática na Educação (CARMO; GASPARINI; OLIVEIRA, 2022)                                                                      |
| E6 | Nóbrega, B. S. d., Maia, J. d. S., Filho, M. A. S., Júnior, E. E. A., and Alves, M. B. (2022). <b>Sistema para identificação e monitoramento de estudantes em risco de evasão: System for identifying and monitoring students at risk of circumventin</b> . Brazilian Journal of Development. (NÓBREGA <i>et al.</i> , 2022) | E29 | Viana, F., Santana, A., and Rabêlo, R. (2022). <b>Avaliação de Classificadores para Predição de Evasão no Ensino Superior Utilizando Janela Semestral</b> . XX-XIII Simpósio Brasileiro de Informática na Educação (VIANA; SANTANA; RABÊLO, 2022)                                                                                    |
| E7 | Mathews de, N. S. L., Fachini Gomes, J. B., Holanda, M., Koike, C. C., and Leao Costa, M. T. (2023). <b>Study on Computer Science Undergraduate Students Dropout at the University of Brasilia</b> 2023 IEEE Frontiers in Education Conference (FIE) (DE <i>et al.</i> , 2023)                                               | E30 | da Silva Garcia, L. L., Lara, D., Gomes, R., and Cazella, S. (2022). <b>Mineração de Dados Educacionais na Predição do Desempenho Acadêmico: um prognóstico a partir do percurso curricular realizado</b> . XXXIII Simpósio Brasileiro de Informática na Educação (GARCIA <i>et al.</i> , 2022)                                      |
| E8 | Oliveira, J. L., Paula Ambrósio, A., Silva, U., Brancher, J., and Franco, J. J. (2020). <b>Undergraduate Students' Effectiveness in an Institution With High Dropout Index</b> . 2020 IEEE Frontiers in Education Conference (FIE) (OLIVEIRA <i>et al.</i> , 2020)                                                           | E31 | Nascimento, P., Junior, A. S., Schulz, C., Santos, M., Maciel, A., Rodrigues, R., Nascimento, R., and Alencar, F. (2021). <b>Análise dos Impactos da Gestão do Tempo no Desempenho Acadêmico Através da Mineração de Dados Educacionais</b> . XXXII Simpósio Brasileiro de Informática na Educação (NASCIMENTO <i>et al.</i> , 2021) |

Tabela 2 – Trabalhos nacionais selecionados (continuação)

| ID  | Referências                                                                                                                                                                                                                                                                            | ID  | Referências                                                                                                                                                                                                                                                                                           |
|-----|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| E9  | Santos, G., Souza, A., Mantovani, R., Cruz, R., Cordeiro, T., and Souza, F. (2024). <b>An Exploratory Analysis on Gender-Related Dropout Students in Distance Learning Higher Education using Machine Learning</b> . Association for Computing Machinery (SANTOS <i>et al.</i> , 2024) | E32 | Santos, C. H., Martins, S., and Plastino, A. (2021a). <b>É Possível Prever Evasão com Base Apenas no Desempenho Acadêmico?</b> . XXXII Simpósio Brasileiro de Informática na Educação (SANTOS; MARTINS; PLASTINO, 2021)                                                                               |
| E10 | Villar, A. and de Andrade, C. R. V. (2024). <b>Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study</b> . Discover Artificial Intelligence (VILLAR; ANDRADE, 2024)                                                          | E33 | Santos, J., Sousa, J. D., Mello, R., Cristino, C., and Alves, G. (2021a). <b>Um Modelo para Análise do Impacto da Retenção e Evasão no Ensino Superior Utilizando Cadeias de Markov Absorventes</b> . XXXII Simpósio Brasileiro de Informática na Educação (SANTOS <i>et al.</i> , 2021)              |
| E11 | Rabelo, A. and Zárate, L. (2024). <b>A Model for Predicting Dropout of Higher Education Students</b> . Data Science and Management (RABELO; ZÁRATE, 2024)                                                                                                                              | E34 | Colpo, M., Primo, T., and Aguiar, M. (2021). <b>Predição da evasão estudantil: uma análise comparativa de diferentes representações de treino na aprendizagem de modelos genéricos</b> . SBC (COLPO; PRIMO; AGUIAR, 2021)                                                                             |
| E12 | Krüger, J. G. C., de Souza Britto, A., and Barddal, J. P. (2023). <b>An explainable machine learning approach for student dropout prediction</b> . Expert Systems with Applications (KRÜGER; BRITTO; BARDDAL, 2023)                                                                    | E35 | Souza, A. L. and Braga, A. (2021). <b>Uma análise dos algoritmos de classificação com base na evasão dos estudantes dos cursos técnicos integrados ao Ensino Médio do Campus Ceres do IF Goiano</b> . XXXII Simpósio Brasileiro de Informática na Educação (SOUZA; BRAGA, 2021)                       |
| E13 | de Jesus, H. O., Rodriguez, L. C., and Costa Junior, A. d. O. (2021). <b>Predição de Evasão Escolar na Licenciatura em Computação</b> . Revista Brasileira de Informática na Educação (JESUS; RODRIGUEZ; JUNIOR, 2021)                                                                 | E36 | Brito, B., Mello, R., and Alves, G. (2020). <b>Identificação de Atributos Relevantes na Evasão no Ensino Superior Público Brasileiro</b> . XXXI Simpósio Brasileiro de Informática na Educação (BRITO; MELLO; ALVES, 2020)                                                                            |
| E14 | Oliveira, R. d. S. and Medeiros, F. P. A. d. (2024). <b>Modelo de Predição de Evasão Escolar com Base em Dados de Autoavaliação de Cursos de Graduação</b> . Revista Brasileira de Informática na Educação (OLIVEIRA; MEDEIROS, 2024)                                                  | E37 | Marques, L., Marques, B., Rocha, R., e Silva, L., de Castro, A., and Queiroz, P. G. (2020). <b>Evasão Acadêmica e suas Causas em Cursos de Bacharelado em Ciência da Computação: Um Estudo de Caso na UFERSA</b> . XXXI Simpósio Brasileiro de Informática na Educação (MARQUES <i>et al.</i> , 2020) |
| E15 | Teodoro, L. d. A. and Kappel, M. A. A. (2020). <b>Aplicação de Técnicas de Aprendizado de Máquina Para Predição de Risco de Evasão Escolar em Instituições Públicas de Ensino Superior no Brasil</b> . Revista Brasileira de Informática na Educação (TEODORO; KAPPEL, 2020)           | E38 | Filho, F., Vinuto, T., and Leal, B. (2020). <b>Análise de Classificadores para Predição de Evasão dos Campi de uma Instituição de Ensino Federal</b> . XXXI Simpósio Brasileiro de Informática na Educação (FILHO; VINUTO; LEAL, 2020)                                                                |

Tabela 2 – Trabalhos nacionais selecionados (continuação)

| ID  | Referências                                                                                                                                                                                                                                                                                                                                    | ID  | Referências                                                                                                                                                                                                                                                                                                                               |
|-----|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| E16 | Souza, V. F. d. and Santos, T. C. B. d. (2021). <b>Processo de Mineração de Dados Educacionais aplicado na Previsão do Desempenho de Alunos: Uma comparação entre as Técnicas de Aprendizagem de Máquina e Aprendizagem Profunda</b> . Revista Brasileira de Informática na Educação (SOUZA; SANTOS, 2021)                                     | E39 | Barbosa, D., Cabral, L., Dwan, F., Feitas, E., and Mello, R. (2023). <b>Previsão da Evasão Escolar através da Análise de Dados e Aprendizagem de Máquina: Um estudo de caso</b> . II Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil (BARBOSA <i>et al.</i> , 2023)                             |
| E17 | Carneiro, M. G., Dutra, B. L., Paiva, J. G. S., Gabriel, P. H. R., and Araújo, R. D. (2022). <b>Educational data mining to support identification and prevention of academic retention and dropout: a case study in introductory programming</b> . Revista Brasileira de Informática na Educação (CARNEIRO <i>et al.</i> , 2022)               | E40 | Oliveira, R., Medeiros, F., and Alves, K. (2023). <b>Predição de Evasão por meio de um Instrumento Sistemático de Avaliação Institucional</b> . II Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil (OLIVEIRA; MEDEIROS; ALVES, 2023)                                                            |
| E18 | Rodrigues, H., Moraes, L., Santiago, E., Campos, J., Júnior, E. G., Wanderley, G., Garcia, A., Mello, C., Alvares, R., and Santos, R. (2024). <b>Predicting Student Dropout on the Information Systems Undergraduate Program of UNIRIO Using Decision Trees</b> . XXXII Workshop sobre Educação em Computação (RODRIGUES <i>et al.</i> , 2024) | E41 | Teodoro, L., Ferreira, A., and Kappel, M. (2023). <b>GraduAI – Sistema com Aprendizagem de Máquina para Avaliação de Risco de Evasão</b> . Anais Estendidos do XII Congresso Brasileiro de Informática na Educação (TEODORO; FERREIRA; KAPPEL, 2023)                                                                                      |
| E19 | Freire, J., Landim, F., Moraes, L., Delgado, C., and Pedreira, C. (2024). <b>Modelo para previsão precoce de abandono de uma disciplina de introdução à programação</b> . XXXII Workshop sobre Educação em Computação (FREIRE <i>et al.</i> , 2024)                                                                                            | E42 | Schoeffel, P., Ramos, V., and Wazlawick, R. (2020). <b>A Method to Predict At-risk Students in Introductory Computing Courses Based on Motivation</b> . Anais Estendidos do IX Congresso Brasileiro de Informática na Educação (SCHOEFFEL; RAMOS; WAZLAWICK, 2020)                                                                        |
| E20 | Sousa, R., Fachini-Gomes, J., Holanda, M., and Leão, M. (2024). <b>Um Estudo da Evasão no Curso de Licenciatura em Computação da Universidade de Brasília</b> . XXXII Workshop sobre Educação em Computação (SOUSA <i>et al.</i> , 2024)                                                                                                       | E43 | Magalhães dos Santos, J. K., da Rocha, H. O., Rodrigues Okamura, E. M., Araujo Dias, V. A., Pessoa de Melo, L. H., Oliveira Viana, G. B., Rodrigues, V. C., and da Silva, D. A. (2023). <b>A Review of IA Use in Education Analysis</b> . 2023 Workshop on Communication Networks and Power Systems (WCNPS) (SANTOS <i>et al.</i> , 2023) |
| E21 | Correia, R., Mendonça, H., Silva, C., and Toledo, D. (2024). <b>Análise dos principais fatores que influenciam a evasão no ensino superior utilizando técnicas de mineração de dados educacionais</b> . XXXII Workshop sobre Educação em Computação (CORREIA <i>et al.</i> , 2024)                                                             | E44 | De Lima, L. M. and Krohling, R. A. (2021). <b>Discovering an Aid Policy to Minimize Student Evasion Using Offline Reinforcement Learning</b> . 2021 International Joint Conference on Neural Networks (IJCNN) (LIMA; KROHLING, 2021)                                                                                                      |
| E22 | Souza, J., Komati, K., and Andrade, J. (2022). <b>Análise de Sobrevivência: um estudo de caso em um Curso de Sistemas de Informação</b> . XXX Workshop sobre Educação em Computação (SOUZA; KOMATI; ANDRADE, 2022)                                                                                                                             | E45 | Leite, D., Filho, E., de Oliveira, J. F. L., Carneiro, R. E., and Maciel, A. (2021). <b>Early detection of students at risk of failure from a small dataset</b> . 2021 International Conference on Advanced Learning Technologies (ICALT) (LEITE <i>et al.</i> , 2021)                                                                    |

Tabela 2 – Trabalhos nacionais selecionados (continuação)

| ID  | Referências                                                                                                                                                                                                            | ID  | Referências                                                                                                                                                                                                                                                                                                                       |
|-----|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| E23 | Saraiva, D., Pereira, S., Braga, R., and Oliveira, C. (2021). <b>Análise de Agrupamentos para Caracterização de Indicadores de Evasão</b> . XXIX Workshop sobre Educação em Computação (SARAIVA <i>et al.</i> , 2021b) | E46 | Dias, J. C., Da Silva, T. L., Juliatto, M. A., Da Paix~ao, A. N., and Prata, D. N. (2023). <b>School dropout in the Federal Network Education of Brazil: is it an inherent individual attribute or it lies on setting conditions?</b> . 2023 International Symposium on Computers in Education (SIIE) (DIAS <i>et al.</i> , 2023) |

Fonte: Elaborada pela autora (2025).

### 3.3.2 QP2: Quais as ferramentas/bibliotecas utilizadas pelos trabalhos?

A Figura 1 apresenta 55 registros de uso de ferramentas e bibliotecas de Mineração de Dados ao longo dos anos, com destaque para 2021, que teve o maior número de ocorrências (20). Os anos de 2020 e 2024 registraram menor uso (8 e 6, respectivamente).

Figura 1 – Ferramentas e bibliotecas utilizadas pelos trabalhos.

| Ano          | CRISP-DM | Excel    | Matplotlib | Não Informado | Numpy    | Outros    | Pandas   | Scikit-learn | Scipy    | Tableau  | Weka     | Total     |
|--------------|----------|----------|------------|---------------|----------|-----------|----------|--------------|----------|----------|----------|-----------|
| 2020         |          |          | 1          | 2             |          | 1         | 2        | 1            | 1        |          |          | 8         |
| 2021         | 2        | 1        |            | 3             | 1        | 5         | 2        | 2            |          | 1        | 3        | 20        |
| 2022         |          |          |            | 3             |          | 3         |          | 2            |          |          | 1        | 9         |
| 2023         | 3        |          |            | 3             | 1        | 1         | 1        | 2            |          |          | 1        | 12        |
| 2024         | 1        |          |            |               |          | 3         |          | 2            |          |          |          | 6         |
| <b>Total</b> | <b>6</b> | <b>1</b> | <b>1</b>   | <b>11</b>     | <b>2</b> | <b>13</b> | <b>5</b> | <b>9</b>     | <b>1</b> | <b>1</b> | <b>5</b> | <b>55</b> |

Fonte: Elaborada pela autora (2025).

A biblioteca *Scikit-learn*, amplamente utilizada em ML com Python, apareceu em 9 estudos, mantendo presença constante entre 2021 e 2024. A metodologia CRISP-DM foi citada em 6 trabalhos, com destaque para 2021 e o biênio 2023-2024, mostrando sua relevância como estrutura de projeto. O *Pandas* (5 usos) teve maior destaque em 2020 e 2021, enquanto *Weka* também foi utilizada 5 vezes, principalmente em 2021 e 2024.

Bibliotecas auxiliares como *NumPy* (2 usos), *Excel*, *Matplotlib*, *Scipy* e *Tableau* (1 uso cada) foram pouco frequentes. A categoria “Outros”, que inclui diversas ferramentas não especificadas (ex.: KDD, Shap), apareceu em 13 trabalhos, especialmente em 2021. Além disso, 11 estudos não especificaram as ferramentas usadas, o que compromete a reprodutibilidade dos resultados.

A análise dos resultados da Q2 evidenciou a predominância de ferramentas consolidadas

no ecossistema Python, bem como a recorrência de metodologias estruturadas para condução de projetos de Mineração de Dados. Esses achados influenciaram diretamente as escolhas metodológicas adotadas nesta dissertação, que utiliza bibliotecas amplamente difundidas e validadas na literatura, como *Scikit-learn*, *Pandas* e *NumPy*, assegurando robustez, transparência e reprodutibilidade ao processo experimental. Da mesma forma, a recorrência da metodologia CRISP-DM nos estudos analisados reforçou sua adoção como referência conceitual para a organização das etapas do trabalho. Assim, as evidências levantadas na Q2 serviram como subsídio empírico para a definição das ferramentas, bibliotecas e estrutura metodológica empregadas na sequência da pesquisa.

### 3.3.3 QP3: Quais os algoritmos de aprendizagem de máquina utilizados nos trabalhos?

A Figura 2 evidencia que o algoritmo mais utilizado foi a Árvore de Decisão (*Decision Tree* - DT), com 27 usos, destacando-se por sua simplicidade e boa performance, especialmente em 2021 e 2024 (7 usos cada). A seguir, a *Random Forest* (RF) apareceu 23 vezes, com maior incidência em 2022 (6) e 2023 (5). O *Naive Bayes* (NB), eficiente para dados categóricos e textos, teve 14 usos, principalmente entre 2020 e 2022. O *Support Vector Machine* (SVM) foi moderadamente usado (12 vezes), possivelmente limitado pelo alto custo computacional. Técnicas de *Boosting* cresceram, somando 12 usos, com destaque em 2024 (4 usos), indicando maior adoção de métodos *ensemble*. Redes Neurais foram menos frequentes (7 usos), talvez por sua complexidade. Algoritmos tradicionais como Regressão Logística (LR), *K-Nearest Neighbors* (KNN) e *MultiLayer Perceptron* (MLP) tiveram uso moderado (7 a 10 vezes). A categoria “Outros” agrupou 14 algoritmos não especificados.

Figura 2 – Algoritmos utilizados pelos trabalhos.

| Ano          | Boost     | DT        | GB       | KNN      | LR       | MLP       | Não Informado | NB        | Outros    | Redes Neurais | RF        | SVM       | Total      |
|--------------|-----------|-----------|----------|----------|----------|-----------|---------------|-----------|-----------|---------------|-----------|-----------|------------|
| 2020         | 2         | 2         | 1        | 2        |          |           | 1             | 3         | 2         | 1             | 3         | 2         | 19         |
| 2021         | 2         | 7         | 1        |          | 2        | 5         | 1             | 5         | 6         | 2             | 5         | 4         | 40         |
| 2022         | 2         | 5         | 2        | 3        | 1        | 4         | 2             | 4         | 4         |               | 6         | 2         | 35         |
| 2023         | 2         | 6         |          | 2        | 2        | 1         |               | 2         | 1         | 3             | 4         | 1         | 24         |
| 2024         | 4         | 7         | 1        |          | 4        |           |               |           | 1         | 1             | 5         | 3         | 26         |
| <b>Total</b> | <b>12</b> | <b>27</b> | <b>5</b> | <b>7</b> | <b>9</b> | <b>10</b> | <b>4</b>      | <b>14</b> | <b>14</b> | <b>7</b>      | <b>23</b> | <b>12</b> | <b>144</b> |

Fonte: Elaborada pela autora (2025).

Por ano, observa-se um aumento do uso de ML a partir de 2021, com destaque para DT e

RF. Em 2022, o uso foi mais equilibrado, e em 2023-2024 houve diversificação, com maior foco em *Boosting* e DT.

Os resultados da QP3 evidenciaram a predominância de algoritmos baseados em árvores de decisão nos estudos analisados, em especial a *Decision Tree* e a *Random Forest*, destacando-se pela recorrência, simplicidade de implementação, boa performance e interpretabilidade em contextos educacionais. Essas evidências subsidiaram diretamente a escolha dos algoritmos avaliados na sequência desta dissertação, priorizando métodos amplamente consolidados na literatura nacional sobre evasão escolar. A seleção da *Decision Tree* e da *Random Forest* permitiu alinhar a proposta às práticas mais recorrentes identificadas no mapeamento, enquanto a inclusão do XGBoost — representante das técnicas de *Boosting*, cuja adoção mostrou crescimento nos estudos mais recentes — possibilitou avaliar um modelo de maior complexidade e capacidade preditiva. Dessa forma, a escolha dos algoritmos buscou equilibrar interpretabilidade, desempenho e alinhamento com as tendências identificadas no mapeamento sistemático.

3.3.4 QP4: Quais os algoritmos com melhor desempenho?

A Figura 3 mostra os algoritmos de ML com melhor desempenho nos estudos.

Figura 3 – Melhores algoritmos apontados pelos trabalhos.

| Ano   | Boost | DT | ETC | GLM | KNN | Não informado | RF | SVM | Total |
|-------|-------|----|-----|-----|-----|---------------|----|-----|-------|
| 2020  |       |    | 1   |     |     |               |    |     | 1     |
| 2021  |       | 2  |     |     |     | 1             | 2  | 1   | 6     |
| 2022  |       | 1  |     | 1   | 1   |               | 2  | 1   | 6     |
| 2023  | 1     | 1  |     |     |     | 1             | 4  |     | 7     |
| 2024  | 3     |    |     |     |     |               | 1  |     | 4     |
| Total | 4     | 4  | 1   | 1   | 1   | 2             | 9  | 2   | 24    |

Fonte: Elaborada pela autora (2025).

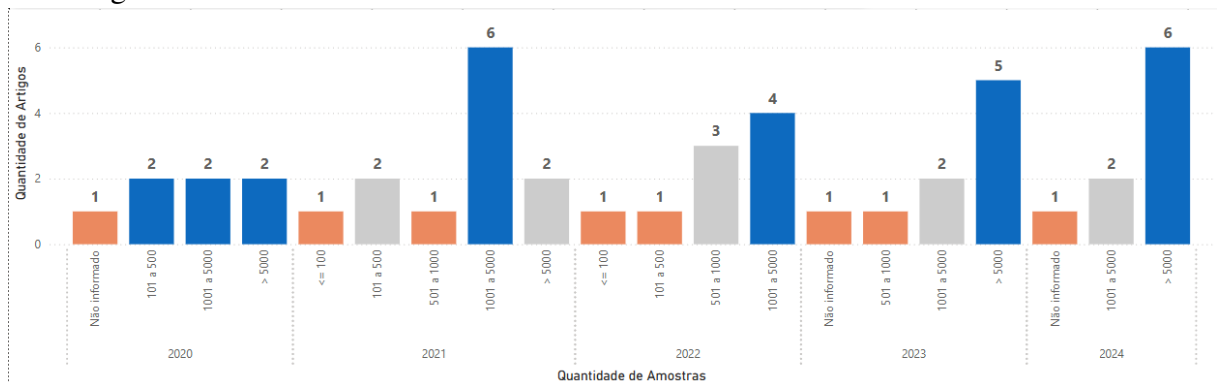
O algoritmo RF lidera, indicado por 9 artigos entre 2021 e 2024, refletindo sua eficácia e popularidade. Algoritmos baseados em árvores (DT, RF, *Boosting*) predominam, somando 17 usos, evidenciando preferência por métodos interpretáveis. *Boosting* ganha destaque em 2024 (3 artigos), sugerindo tendência de crescimento. SVM tem baixa frequência (2 artigos), possivelmente pela complexidade e necessidade de ajuste. Os algoritmos *Extra Trees Classifier* (ETC), *Generalized Linear Model* (GLM) e KNN aparecem pontualmente, com pouca representatividade.

Os resultados apresentados na QP4 reforçam a eficácia dos algoritmos baseados em árvores de decisão, em especial a Random Forest, que se destacou como o modelo de melhor desempenho na maioria dos estudos analisados. Essas evidências tiveram impacto direto na definição dos algoritmos avaliados nesta dissertação, justificando a escolha da *Random Forest* como um dos modelos centrais do experimento. A predominância de métodos interpretáveis também reforçou a adoção da Árvore de Decisão, permitindo análises explicativas sobre os fatores associados à evasão discente. Além disso, o crescimento observado no uso de técnicas de Boosting, sobretudo nos estudos mais recentes, motivou a inclusão do XGBoost como representante dessa abordagem, possibilitando a avaliação de modelos de maior capacidade preditiva. Dessa forma, as escolhas metodológicas realizadas na sequência do trabalho refletem não apenas tendências de uso, mas também evidências empíricas de desempenho identificadas no mapeamento sistemático.

### 3.3.5 QP5: Quantos registros existem nas bases de dados analisadas?

A Figura 4 apresenta a quantidade de artigos por faixa de tamanho da amostra, distribuídos por ano (de 2020 a 2024).

Figura 4 – Tamanho da base.



Fonte: Elaborada pela autora (2025).

A faixa mais recorrente ao longo dos anos foi a de 1001 a 5000 amostras, com destaque em 2021 (6 artigos) e 2022 (4 artigos), isso sugere uma preferência por bases de dados de porte médio a grande, equilibrando disponibilidade de dados e viabilidade de análise. Em alta nos anos mais recentes, a faixa acima de 5000 amostras, utilizada em 5 artigos em 2023 e 6 artigos em 2024, indica uma tendência crescente de uso de grandes bases de dados, possivelmente relacionadas a sistemas educacionais completos ou bancos de dados integrados. Amostras

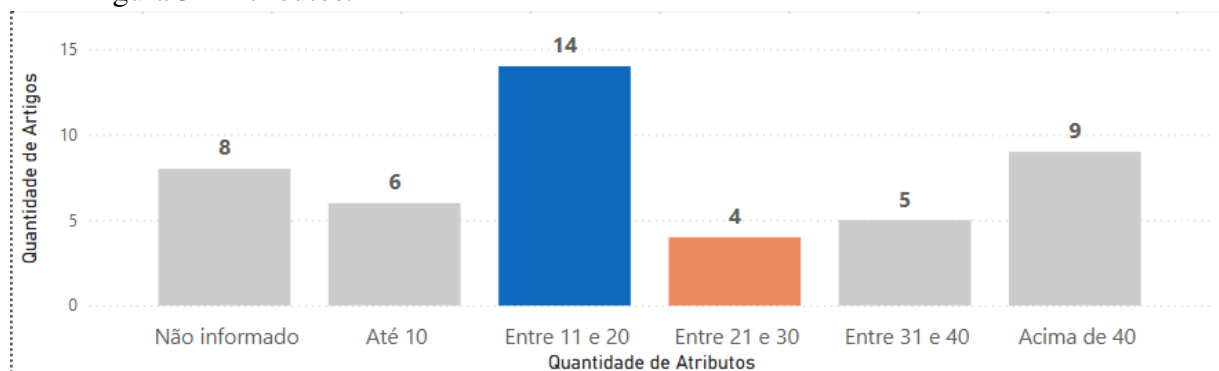
menores ou iguais a 100 ou até 500 amostras, foram pouco frequentes e concentradas entre 2020 e 2022, o que pode refletir estudos de caso, pilotos ou pesquisas com menor acesso a dados. Aparecem em todos os anos, com pelo menos 1 artigo por ano omitindo o tamanho da amostra, o que pode comprometer a avaliação da robustez metodológica da parte dos estudos.

Destaca-se, portanto, que há uma tendência clara de aumento no uso de amostras maiores ao longo do tempo, especialmente a partir de 2021, o que pode indicar uma melhora no acesso a dados educacionais, como também uma maior preocupação com validade estatística dos resultados.

### 3.3.6 QP6: *Quantos atributos estão presentes nas bases de dados analisadas?*

A Figura 5 responde a QP6 mostrando informações sobre a quantidade de atributos presentes nas bases de dados utilizadas nos trabalhos selecionados.

Figura 5 – Atributos.



Fonte: Elaborada pela autora (2025).

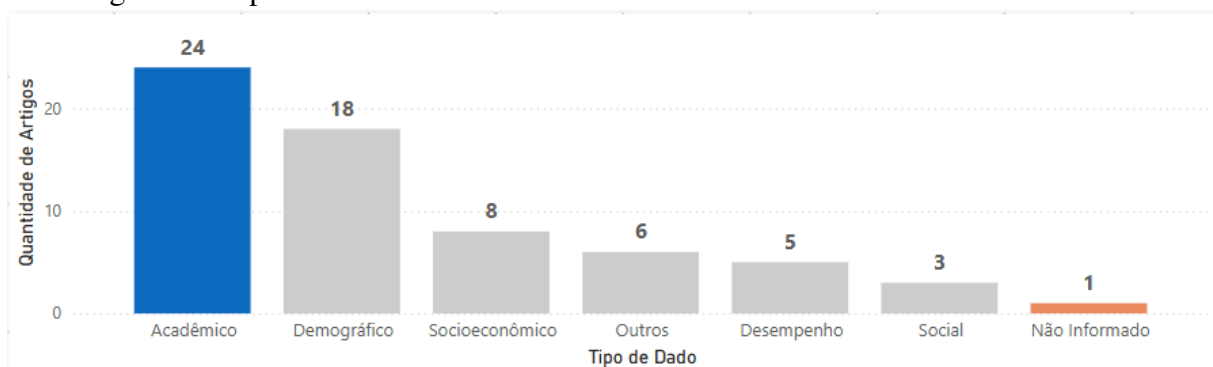
A faixa de quantidade de atributos mais frequente foi a entre 11 e 20 atributos (14 artigos), o que indica uma preferência dos pesquisados por um número moderado de variáveis, possivelmente equilibrando complexidade e interpretabilidade. A faixa acima de 40 atributos (9 artigos) também representa uma parcela significativa, mostra que alguns estudos optam por uma abordagem mais robusta, com grande volume de dados. Estudos mais simples ou com foco restritos limitaram-se a utilizar até 10 atributos (6 artigos). Entre 31 e 40 atributos (5 artigos), foi pouco frequente, mas ainda representativo para estudos mais detalhados. A faixa menos utilizada foi a que utilizou entre 21 e 30 atributos (4 artigos), possivelmente evitada por ser uma zona intermediária que nem garante simplicidade nem profundidade analítica. Tivemos um número considerável de artigos (8) que não informaram o tamanho de suas amostras, o que pode indicar uma falta de detalhamento metodológico ou limitações na apresentação dos dados. A presença

de artigos sem essa informação pode dificultar a avaliação da robustez de certos estudos. A maioria dos artigos utiliza entre 11 e 20 atributos, sugerindo uma preferência por modelos de análise com um número controlado de variáveis. Contudo, há também uma boa quantidade de estudos que trabalham com um número elevado de atributos (mais de 40), o que pode indicar o uso de técnicas mais avançadas, como algoritmos de aprendizado de máquina.

### 3.3.7 QP7: Quais os tipos de dados analisados nos trabalhos?

A Figura 6 mostra a quantidade de artigos segundo os tipos de dados utilizados, totalizando mais que 46, pois alguns trabalhos analisaram múltiplos tipos. A maior parte (24 artigos) utiliza dados acadêmicos, como notas, frequência e histórico escolar, refletindo o foco principal das pesquisas. Dados demográficos, como idade, sexo, etnia e local de residência, aparecem em 18 estudos, indicando interesse no perfil dos alunos. Dados socioeconômicos, envolvendo renda familiar e ocupação dos pais, são considerados em menor escala (8 artigos). Métricas de desempenho externo, como avaliações padronizadas, foram usadas em 5 artigos, e dados sociais, como redes de relacionamento e convivência escolar, aparecem em 3. Um artigo não especificou os dados usados, e 6 se enquadram numa categoria genérica, que pode incluir variáveis institucionais ou ambientais.

Figura 6 – Tipos de dados.



Fonte: Elaborada pela autora (2025).

## 3.4 Conclusões

O estudo analisou 46 trabalhos nacionais sobre Mineração de Dados Educacionais focados na evasão escolar, com visualização de resultados no Power BI. Observou-se diversidade de ferramentas, com destaque para bibliotecas Python (*Scikit-learn*, *Pandas*, *Numpy*) e uso frequente

da metodologia CRISP-DM. Muitos estudos não informam detalhadamente as ferramentas utilizadas.

A maioria dos artigos utiliza entre 11 e 20 atributos, sugerindo uma preferência por modelos de análise com um número controlado de variáveis. Contudo, há também uma boa quantidade de estudos que trabalham com um número elevado de atributos (mais de 40), o que pode indicar o uso de técnicas mais avançadas, como algoritmos de ML. A presença de artigos sem essa informação pode dificultar a avaliação da robustez de certos estudos.

A diversidade de algoritmos observada ao longo dos anos indica um amadurecimento técnico nas pesquisas analisadas, especialmente a partir de 2021, com o aumento do uso de técnicas de ML. O *Random Forest* se destaca como o algoritmo mais utilizado, o que se justifica por sua eficácia em tarefas de classificação e sua interpretabilidade, especialmente em contextos educacionais. Algoritmos baseados em árvores (como Árvores de Decisão, *Random Forest* e *Boosting*) são predominantes, refletindo uma preferência por modelos compreensíveis. O crescimento do uso de técnicas ensemble nos últimos anos também reforça essa evolução metodológica. Quanto aos dados utilizados, há predominância de informações acadêmicas e demográficas, revelando um foco na análise do histórico escolar e perfil do estudante, enquanto dados sociais e de desempenho aparecem menos, indicando uma possível lacuna para investigações futuras.

Os estudos estão evoluindo em quantidade e qualidade, com uso crescente de métodos mais sofisticados, bases maiores e ferramentas mais modernas. Porém, ainda há lacunas importantes na descrição de metodologias, como linguagem usada e ferramentas, o que compromete a reprodutibilidade.

Os achados deste SLM não apenas caracterizam o estado da arte nacional sobre Mineração de Dados Educacionais aplicada à evasão escolar, mas também fundamentam diretamente as escolhas metodológicas adotadas nesta dissertação. A predominância de algoritmos baseados em árvores, como *Random Forest* e técnicas de *Boosting*, bem como o uso recorrente de bibliotecas Python e da metodologia CRISP-DM, orientaram a definição das abordagens empregadas no desenvolvimento do modelo preditivo deste trabalho.

Além disso, a identificação de lacunas, como a insuficiente descrição de ferramentas e procedimentos metodológicos em parte dos estudos analisados, reforçou a preocupação desta pesquisa com a transparência, a reprodutibilidade e o detalhamento das etapas experimentais. Da mesma forma, a constatação de que a maioria dos estudos concentra-se em dados acadêmicos e demográficos motivou a ampliação da análise para variáveis derivadas por meio de engenharia de atributos, buscando enriquecer a modelagem. Assim, a SLM não se limita a um levantamento

bibliográfico, mas constitui base estruturante para a construção, avaliação e aplicação prática dos modelos desenvolvidos nas etapas subsequentes desta dissertação.

## 4 PROPOSTA

Conforme discutido ao longo dos capítulos anteriores, este trabalho tem como objetivo geral aplicar a Mineração de Dados Educacionais (MDE) para identificar padrões de evasão no Programa de Pós-Graduação em Ciência da Computação (PPGCC) do Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE).

Este capítulo apresenta a metodologia adotada para o desenvolvimento desta pesquisa, estruturada de modo a viabilizar o alcance desse objetivo geral, assim como dos objetivos específicos definidos na Seção 1.4.2.

A metodologia foi estruturada em sete etapas interdependentes, que abrangem: 1) entendimento do domínio de estudo; 2) coleta de dados; 3) análise visual dos dados; 4) pré-processamento dos dados; 5) treinamento dos modelos; 6) avaliação dos modelos; e 7) aplicação prática do modelo. A Figura 7 ilustra essa metodologia.

Figura 7 – Fluxo metodológico do trabalho.



Fonte: Elaborada pela autora (2026).

Cada etapa desempenha um papel específico na construção do processo analítico. Embora apresentadas de forma sequencial para fins de organização e clareza, tais etapas possuem caráter iterativo e dinâmico, permitindo revisões e refinamentos ao longo do desenvolvimento da pesquisa. Essa natureza cíclica é coerente com abordagens metodológicas amplamente utilizadas em projetos de DM, nas quais a retroalimentação entre as fases contribui para o refinamento contínuo das decisões analíticas e para o alinhamento permanente com os objetivos do estudo.

### 4.1 Etapa 1: Entendimento do domínio de estudo

A metodologia tem início com a etapa de Entendimento do domínio de estudo, na qual se busca elaborar a motivação do trabalho, compreender a problemática em profundidade e estabelecer os objetivos a serem alcançados com a dissertação.

Nesse sentido, esta primeira etapa compreendeu, inicialmente às pesquisas realizadas nos documentos institucionais de referência, como o Plano Estratégico para Permanência e Êxito dos Estudantes do IFCE 2017-2024, publicado pela Pró-Reitoria de Ensino do IFCE (IFCE, 2017).

Além desses documentos institucionais, destacam-se as pesquisas e leituras realizadas para a construção da fundamentação teórica e para a revisão de trabalhos relacionados à área de MDE. Dessa forma, destaca-se aqui o Mapeamento Sistemático da Literatura do uso de MDE no contexto da evasão escolar apresentado e discutido no Capítulo 3.

Nesta etapa, definiu-se também o Programa de Pós-Graduação em Ciência da Computação (PPGCC) do IFCE Campus Fortaleza como estudo de caso desta pesquisa. Conforme descrito em seu regimento (PPGCC, 2017), o PPGCC realiza curso de pós-graduação *stricto sensu* em nível de mestrado acadêmico. O programa tem por finalidade a formação de pessoal qualificado para o exercício de atividades de ensino, pesquisa e desenvolvimento e para a produção de conhecimento científico e tecnológico, na área de concentração de Sistemas de Computação, sendo destinado a candidatos que tenham concluído curso de graduação em Engenharias, Computação e Ciências Exatas.

A escolha do PPGCC/IFCE como estudo de caso neste trabalho fundamenta-se em múltiplos aspectos que reforçam a relevância e a viabilidade da pesquisa. Primeiramente, por ser o programa ao qual esta dissertação está vinculada, há uma compreensão aprofundada do contexto institucional e pedagógico, o que permite uma análise mais precisa e contextualizada. Em segundo lugar, destaca-se a possibilidade de acesso direto e eticamente autorizado aos dados acadêmicos do PPGCC, condição essencial para o desenvolvimento desta pesquisa. Adicionalmente, os resultados deste estudo têm o potencial de subsidiar políticas institucionais no IFCE, contribuindo para a melhoria dos indicadores do programa, o fortalecimento da formação de pesquisadores e a otimização de recursos, evitando o desperdício de talentos. Outro fator relevante refere-se às características dos cursos da área de Ciência da Computação, que historicamente apresenta disparidades significativas na proporção de matrículas entre homens e mulheres, configurando um contexto oportuno para a investigação de questões relacionadas a gênero, diversidade e inclusão em ambientes acadêmicos de tecnologia. Assim, a escolha do PPGCC alinha-se aos objetivos da pesquisa, promovendo contribuições relevantes tanto para a instituição quanto para a literatura científica na área.

## 4.2 Etapa 2: Coleta dos dados

A segunda etapa da metodologia consistiu na obtenção e preparação dos dados relevantes ao domínio de estudo. Para o desenvolvimento desta dissertação, os dados foram extraídos do sistema Q-Acadêmico, plataforma oficial de gestão acadêmica do IFCE Campus Fortaleza. Os dados foram disponibilizados em formato *Comma Separated Values* (CSV) no dia 11 de agosto de 2025 e fornecidos de forma anonimizada, não contendo identificadores pessoais, como nome, CPF ou número de matrícula, em conformidade com a Lei Geral de Proteção de Dados (LGPD) e com as diretrizes éticas institucionais.

A leitura do conjunto de dados foi realizada a partir do arquivo *dados\_originais.csv*, utilizando a biblioteca *Pandas* da linguagem *Python*. Nessa etapa inicial, procedeu-se à verificação da quantidade de registros (linhas) e do número total de atributos (colunas), com o objetivo de compreender a dimensionalidade da base e subsidiar as decisões subsequentes relacionadas ao pré-processamento, à seleção de atributos e à modelagem preditiva.

Conforme detalhado nos resultados da análise visual na próxima seção, a base de dados utilizada no estudo é composta por 255 matrículas de ingressantes, distribuídas entre os semestres 2014.2 (primeira turma) e 2025.1 (última turma), abrangendo um total de 144 atributos<sup>1</sup>.

A combinação entre a quantidade de registros e a diversidade de variáveis disponíveis confere robustez ao conjunto, permitindo não apenas a caracterização detalhada do perfil dos estudantes, mas também a exploração de diferentes dimensões associadas ao fenômeno da evasão. Essa variedade de atributos favorece a aplicação de técnicas de Mineração de Dados e Aprendizado de Máquina, ampliando a possibilidade de identificar padrões e relações relevantes para a análise.

## 4.3 Etapa 3: Análise visual dos dados

Inicialmente, foi conduzida uma análise descritiva e exploratória dos dados acadêmicos e socioeconômicos obtidos na coleta de dados, com o propósito de avaliar a estrutura, a completude e a qualidade da base. Essa fase é fundamental, pois fornece a primeira visão sistemática sobre o conjunto de dados, permitindo tanto a identificação de limitações intrínsecas (como a presença de valores ausentes e inconsistências), quanto a caracterização dos atributos disponíveis para o estudo da evasão escolar.

---

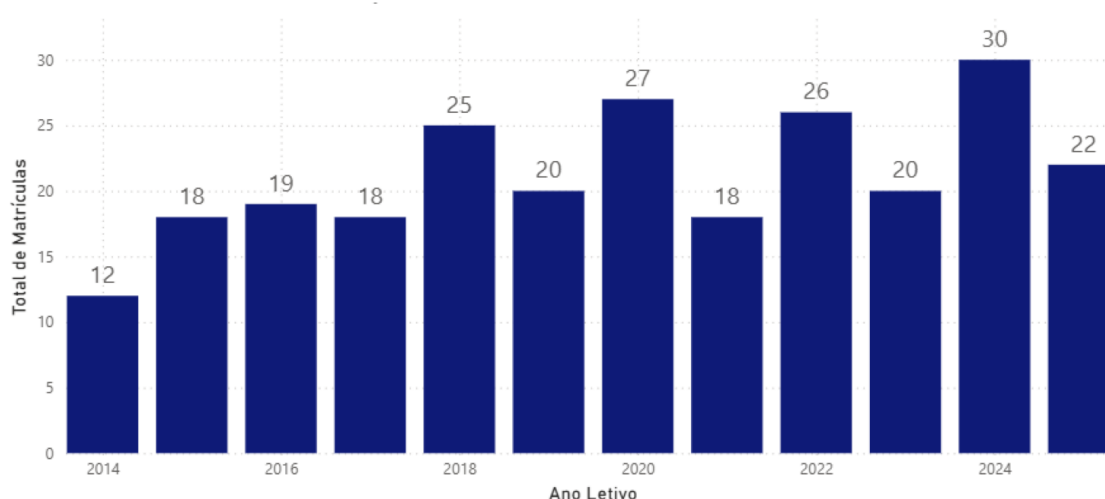
<sup>1</sup> O sistema Qacademico disponibiliza um total de 243 atributos, mas 99 atributos não foram disponibilizados para essa pesquisa em razão das restrições da LGPD.

Esta etapa foi realizada por meio da construção de visualizações gráficas, as quais possibilitam avaliar o grau de completude dos dados, a frequência das categorias presentes nas variáveis categóricas (como sexo, cor/raça, renda familiar, situação de matrícula) e a evolução das matrículas ao longo dos anos letivos.

As visualizações a seguir foram elaboradas com o auxílio da ferramenta de *Business Intelligence* Power BI<sup>2</sup>, proporcionando uma representação gráfica clara e objetiva das informações analisadas e permitindo identificar padrões, relações e tendências presentes nos conjuntos de dados, contribuindo para uma interpretação mais precisa e fundamentada dos resultados obtidos.

A Figura 8 apresenta a evolução do total de matrículas por ano letivo entre 2014 e 2025. Observa-se uma tendência geral de crescimento ao longo do período, ainda que com oscilações. O menor número de matrículas ocorreu em 2014, com 12 estudantes, enquanto o pico foi registrado em 2024, com 30 matrículas. Entre os anos, destacam-se aumentos expressivos, como em 2018 (25) e 2020 (27), seguidos por quedas, como em 2021 (18) e 2023 (20). Apesar dessas variações, nota-se que nos anos mais recentes os valores se mantêm em patamares mais altos do que no início da série, indicando uma ampliação gradativa da procura ao longo do tempo.

Figura 8 – Matrículas por ano letivo.



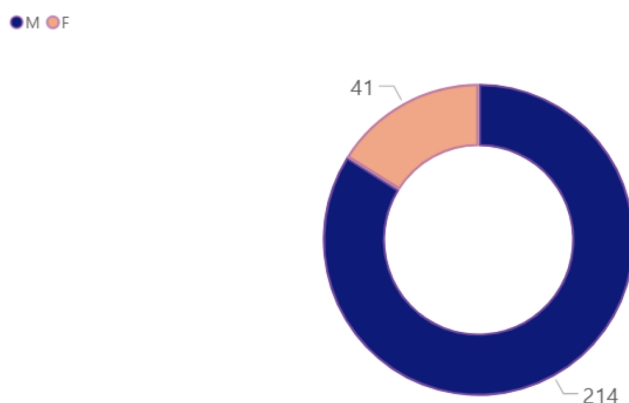
Fonte: Elaborada pela autora (2025).

A Figura 9 apresenta a distribuição do total de matrículas por sexo. Observa-se que a maior parte dos estudantes matriculados é do sexo masculino (M), representando 214 matrículas, enquanto o sexo feminino (F) corresponde a 41 matrículas. Essa diferença evidencia uma predominância masculina significativa no conjunto analisado, já que os homens representam

<sup>2</sup> <https://www.microsoft.com/pt-br/power-platform/products/power-bi>

83.92% do total de matrículas. Esse desequilíbrio pode refletir fatores sociais, culturais ou relacionados à área de estudo em questão, sendo um ponto relevante para discussões sobre inclusão e equilíbrio de gênero.

Figura 9 – Matrículas por sexo.



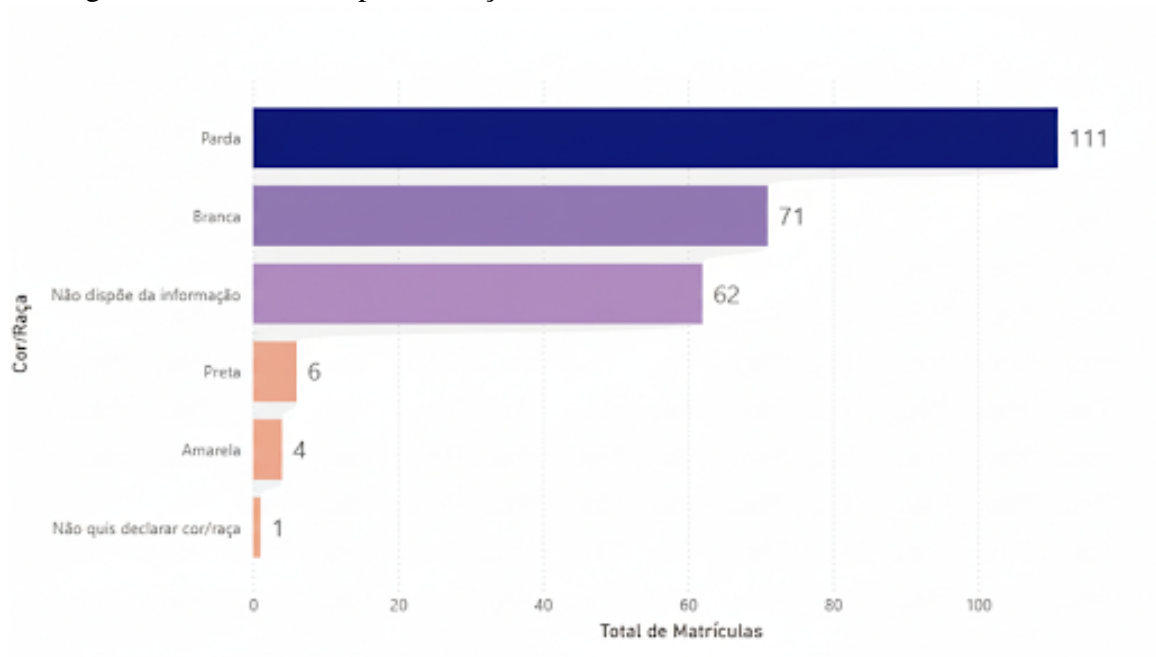
Fonte: Elaborada pela autora (2025).

A Figura 10 apresenta o total de matrículas por cor/raça e revela que a maioria dos estudantes se autodeclara parda, somando 111 matrículas. Em seguida, observa-se uma quantidade expressiva de estudantes que se identificam como brancos (71) e também um número considerável de registros sem informação declarada (62). Já os grupos que se autodeclararam pretos (6), amarelos (4) e aqueles que optaram por não declarar (1) representam uma parcela reduzida do total. Nenhum estudante se autodeclarou na cor/raça Indígena. Esses dados evidenciam a predominância de estudantes pardos no conjunto analisado, ao mesmo tempo em que mostram a relevância de informações faltantes, que poderiam impactar na compreensão mais detalhada da diversidade racial entre os matriculados.

Analisando a Figura 11, pode-se observar a distribuição das matrículas de acordo com a renda familiar dos estudantes. Nota-se que a maior parte não informou a renda, totalizando 135 registros, o que representa um dado importante a ser considerado na análise. Entre os que declararam, observa-se uma predominância de estudantes com renda de até 1 salário mínimo (34) e de 1 a 2 salários (30), evidenciando um perfil socioeconômico mais concentrado nas faixas de baixa renda. Já os estudantes com renda entre 2 e 5 salários somam 33 matrículas, enquanto rendas mais altas (de 5 a 20 salários) apresentam números reduzidos, e apenas um estudante declarou renda superior a 20 salários. Esses resultados reforçam a relevância de políticas de apoio à permanência estudantil voltadas principalmente a estudantes de baixa renda.

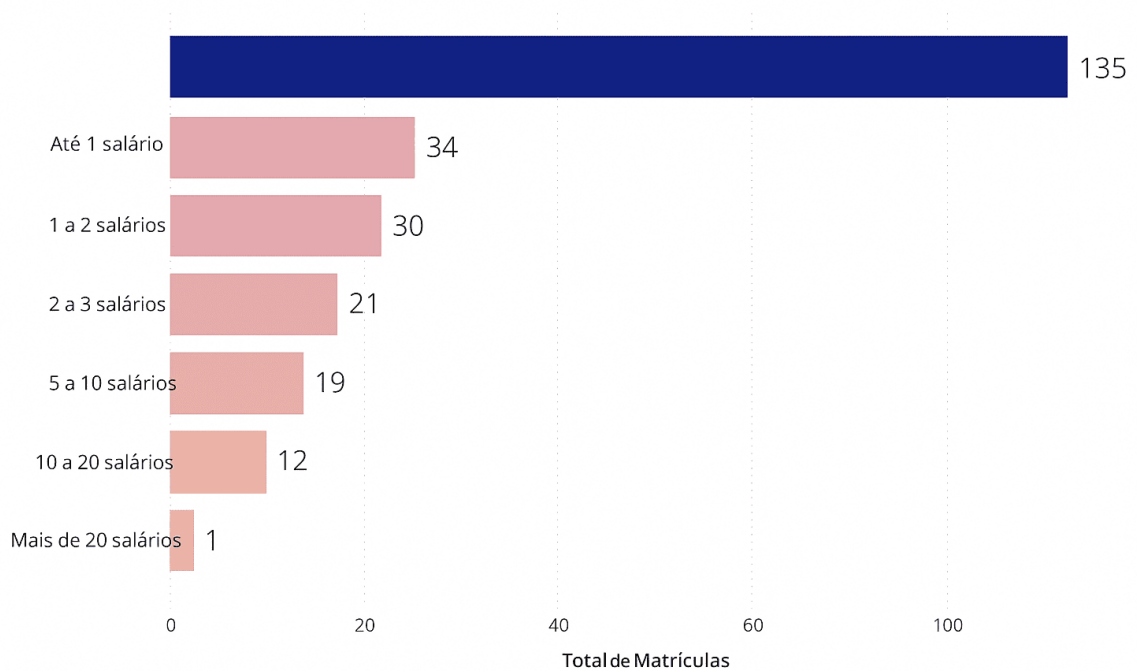
A visualização da Figura 12 mostra a situação de matrícula dos estudantes e evidencia que

Figura 10 – Matrículas por cor/raça.



Fonte: Elaborada pela autora (2025).

Figura 11 – Matrículas por renda.

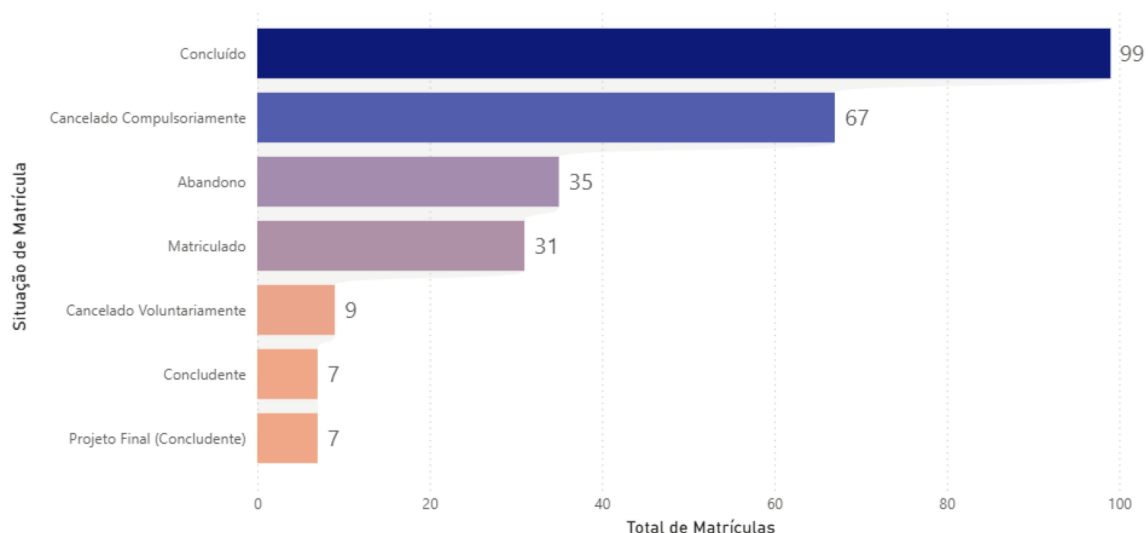


Fonte: Elaborada pela autora (2025).

a maior parte concluiu o curso, totalizando 99 alunos (38.82%). Em seguida, destaca-se o número de cancelamentos compulsórios (67), seguido pelos casos de abandono (35) e pelos estudantes ainda matriculados (31). Situações menos frequentes incluem o cancelamento voluntário (9)

e os estudantes que estão em fase final, seja como concluintes (7) ou em projeto final (7). Os dados revelam, portanto, um cenário em que há uma quantidade expressiva de desligamentos compulsórios e abandonos, indicando desafios na permanência estudantil ao longo do curso.

Figura 12 – Situação de matrícula.



Fonte: Elaborada pela autora (2025).

Com o objetivo de alinhar a análise às diretrizes institucionais, a Tabela 3 apresenta as situações de matrícula segundo o Plano de Dados Abertos (IFCE, 2024) e o Plano de Permanência e Êxito (IFCE, 2017).

Tabela 3 – Situações de matrícula e suas definições.

| Situação                    | Definição                                                                                                                          |
|-----------------------------|------------------------------------------------------------------------------------------------------------------------------------|
| Abandono                    | Representa que o estudante não renovou a matrícula sendo lançada a situação de abandono                                            |
| Cancelado Compulsoriamente  | Representa que o IFCE cancelou a matrícula do estudante naquele período letivo                                                     |
| Cancelado Voluntariamente   | Representa que o estudante solicitou o cancelamento da sua matrícula                                                               |
| Concludente                 | Representa que o estudante concluiu todas as disciplinas do curso, mas falta pagar estágio e/ou TCC e/ou atividades complementares |
| Concluído                   | Representa que o estudante finalizou todos os requisitos para conclusão do curso                                                   |
| Matriculado                 | Representa que o estudante está ativo no curso                                                                                     |
| Projeto Final (Concludente) | Representa que o estudante finalizou todas as disciplinas do curso e está cursando o projeto final obrigatório                     |

Fonte: Elaborada pela autora (2025).

#### 4.4 Etapa 4: Pré-processamento dos dados

Os conjuntos de dados podem apresentar características heterogêneas, variando em dimensionalidade e formato (FACELI *et al.*, 2021). Esses conjuntos podem conter anomalias, como valores incorretos, inconsistentes, duplicados ou ausentes, comprometendo a integridade dos dados. Os atributos podem apresentar-se de forma independente ou correlacionada, enquanto os registros podem variar desde poucos objetos com reduzido número de atributos até bases extensas com elevada dimensionalidade.

Diante desse cenário, torna-se necessária a aplicação de técnicas de pré-processamento, com o objetivo de melhorar a qualidade do conjunto e mitigar os problemas mencionados. Essa etapa é fundamental para que os algoritmos de mineração consigam gerar modelos mais representativos e próximos da distribuição real dos dados (FACELI *et al.*, 2021).

O pré-processamento dos dados foi conduzido de forma estruturada, seguindo as etapas descritas nas subseções subsequentes. A aplicação sistemática dessas etapas possibilitou a construção de uma base estruturada e apropriada para os experimentos conduzidos, contribuindo para a confiabilidade dos resultados obtidos.

##### 4.4.1 Construção da variável alvo

A variável alvo da pesquisa, denominada *Status*, foi construída a partir do atributo Desc\_Sit\_Matricula, responsável por registrar o desfecho acadêmico do vínculo do estudante ao curso. A definição da variável alvo constitui etapa central em estudos de predição de evasão, pois determina o tipo de problema a ser modelado e influencia diretamente a escolha dos algoritmos de ML e das métricas de avaliação adotadas.

Para fins de modelagem preditiva, optou-se por formular o problema como uma classificação binária, abordagem amplamente utilizada na literatura sobre evasão no ensino superior, por permitir distinguir, de forma objetiva, estudantes que concluíram o curso (Egressos com Êxito), daqueles estudantes que interromperam sua trajetória acadêmica antes da conclusão (Egressos sem Êxito) (ROMERO; VENTURA, 2010). Nesse contexto, a variável Status foi definida conforme os seguintes critérios:

- *Status* = 1 (Egressos com Êxito): quando Desc\_Sit\_Matricula = "Concluído";
- *Status* = 0 (Egressos sem Êxito): quando Desc\_Sit\_Matricula assume os valores: "Abandono", "Cancelado Compulsoriamente" ou "Cancelado Voluntariamente".

A agregação das diferentes situações de não conclusão (evasão) em uma única classe foi adotada com o objetivo de concentrar a análise no fenômeno da evasão em sentido amplo, estratégia comum em estudos educacionais que buscam identificar fatores associados à permanência ou à saída precoce do estudante do sistema de ensino (TAN; STEINBACH; KUMAR, 2018).

Destaca-se que a base original contempla sete situações distintas de matrícula, conforme apresentado na Figura 12 e detalhado na Tabela 3. Entretanto, para fins de modelagem supervisionada, foram mantidos apenas os registros com desfecho acadêmico consolidado, sendo excluídas as situações correspondentes a estudantes ainda em curso, como “Matriculado”, “Concludente” e “Projeto Final”. Assim, registros que não se enquadravam nas categorias de conclusão (Egressos com Êxito) ou evasão (Egressos sem Êxito) tiveram o valor da variável alvo *Status* definido como nulo e foram posteriormente removidos da base de dados. Essa decisão fundamenta-se na necessidade de trabalhar exclusivamente com desfechos acadêmicos finalizados e observáveis, uma vez que registros com matrícula ativa representam trajetórias em andamento, cujo resultado final — conclusão ou evasão — é desconhecido no momento da análise (TINTO, 2012). A inclusão desses casos poderia introduzir incerteza e viés no processo de aprendizagem do modelo, pois tais estudantes podem futuramente tanto concluir quanto abandonar o curso (BAKER; YACEF, 2009; WITTEN; FRANK; HALL, 2016).

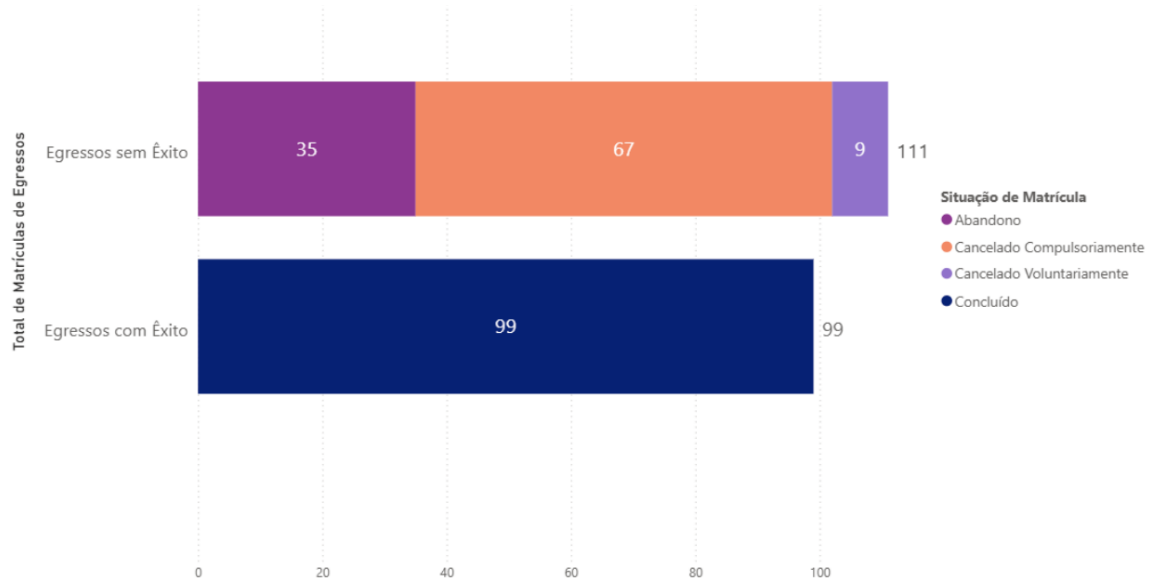
Importante destacar que, embora excluídos do processo de treinamento, os registros com situação ainda em andamento foram preservados para utilização posterior na fase de aplicação prática do modelo preditivo, permitindo estimar a probabilidade de evasão ou conclusão desses estudantes em um cenário real de apoio à tomada de decisão institucional.

Dessa forma, a definição da variável *Status* permitiu estruturar o problema de forma clara, metodologicamente consistente e alinhada ao objetivo da pesquisa, viabilizando tanto o treinamento de modelos supervisionados quanto sua aplicação prática na identificação de estudantes com maior risco de evasão.

A Figura 13 ilustra a distribuição do *Status* ao final dessa etapa. Dentre o total de 210 matrículas, 99 matrículas (47,1%) são de Egressos com Êxito e 111 matrículas são de Egressos sem Êxito (52,9%). As matrículas dos Egressos sem Êxito estão distribuídas entre Cancelamento Compulsório (67 casos, 31,9%), Abandono (35 casos, 16,7%) e Cancelamento Voluntário (9 casos, 4,3%). Nota-se, portanto, uma predominância do insucesso acadêmico, ainda que por pequena margem, indicando que mais da metade dos discentes não conseguiu concluir o curso. Esses resultados reforçam a necessidade de aprofundar a análise sobre os motivos do insucesso e de formular estratégias institucionais voltadas à redução da evasão e ao fortalecimento da

permanência discente.

Figura 13 – Situações de matrícula de estudantes egressos sem êxito e com êxito.



Fonte: Elaborada pela autora (2025).

#### 4.4.2 Criação de atributos

Essa etapa teve como objetivo enriquecer o conjunto de dados original, transformando variáveis brutas em representações mais informativas e alinhadas ao problema de predição da evasão discente. Em bases educacionais institucionais, atributos diretamente coletados nem sempre capturam de forma explícita aspectos relevantes do percurso acadêmico do estudante, sendo necessária a construção de novas variáveis que sintetizem informações temporais, acadêmicas e comportamentais. Segundo Kuhn e Johnson (2013), a engenharia de atributos é uma das etapas mais críticas do processo de modelagem preditiva, frequentemente exercendo maior impacto no desempenho dos modelos do que a própria escolha do algoritmo.

No presente estudo, foram construídos 6 atributos derivados (Apêndice A), com base em informações sociodemográficas e acadêmicas, conforme descrito a seguir.

- Idade do estudante ao ingressar no curso (Idade): obtida a partir da diferença entre o ano letivo inicial e o ano de nascimento do estudante. Essa variável busca capturar características relacionadas à fase de vida no momento do ingresso no curso. Estudos na área de MDE indicam que diferenças etárias podem estar associadas a distintos padrões de engajamento, disponibilidade de tempo, responsabilidades familiares e adaptação

ao ambiente acadêmico, fatores que influenciam diretamente a permanência ou evasão discente (ROMERO; VENTURA, 2010).

- Intervalo (em anos) entre a graduação e a pós-graduação (Anos\_Apos\_Graduacao): calculado a partir do intervalo entre o ano de conclusão da graduação e o ano letivo de ingresso no curso analisado. Esse atributo visa representar o distanciamento temporal entre a formação acadêmica anterior e o retorno aos estudos. A literatura aponta que períodos prolongados fora do ambiente acadêmico podem impactar o desempenho inicial, a adaptação às exigências do curso e a probabilidade de evasão, especialmente em cursos de pós-graduação (TINTO, 2012).
- Histórico acadêmico do estudante a partir do identificador da pessoa como base para agregações:
  - Número de matrículas associadas à mesma pessoa(n\_matriculas\_pessoa), representando a recorrência do estudante no sistema acadêmico;
  - Identificação de primeira matrícula(primeira\_matricula), indicando se o vínculo analisado corresponde ao primeiro ingresso do estudante no curso analisado;
  - Indicador de reingresso(reingresso), sinalizando trajetórias acadêmicas não lineares com múltiplos ciclos de ingresso.
  - Indicador de mais de uma matrícula(n\_matriculas\_maior\_que\_1), indicando se uma pessoa teve mais de uma matrícula ao longo do tempo.

Esses atributos permitem capturar padrões de persistência, retorno ou interrupção dos estudos, frequentemente associados a experiências acadêmicas prévias, dificuldades de adaptação ou mudanças de objetivos educacionais. Conforme destacado por Han, Kamber e Pei (2012), a utilização de atributos agregados é uma estratégia eficaz para representar comportamentos históricos em tarefas de classificação supervisionada.

As variáveis construídas foram incluídas no conjunto final por apresentarem potencial explicativo amplamente reconhecido na literatura sobre permanência e evasão no ensino superior, além de contribuírem para uma representação mais rica e contextualizada do perfil discente. Dessa forma, a engenharia de atributos possibilitou ampliar a capacidade dos modelos de identificar padrões relevantes associados à conclusão ou evasão, reforçando a robustez e a interpretabilidade dos resultados obtidos.

#### 4.4.3 Limpeza de dados

O processo de limpeza de dados constituiu uma etapa fundamental desta pesquisa, uma vez que bases de dados educacionais institucionais tendem a apresentar elevado número de variáveis redundantes, inconsistentes ou irrelevantes para fins preditivos.

A base do PPGCC do IFCE integra diversos atributos relacionados aos perfis socioeconômicos e ao desempenho acadêmico dos estudantes, mas apresentou lacunas significativas decorrentes de campos incompletos ou ausentes. Conforme destacam Han, Kamber e Pei (2012), a presença excessiva de atributos pouco informativos pode aumentar a complexidade do modelo, elevar o custo computacional e favorecer o sobreajuste (*overfitting*), comprometendo a capacidade de generalização.

Nesta etapa, foram adotados critérios baseados, principalmente, em análise exploratória, visando a construção de um conjunto de atributos consistente e informativo. Os critérios foram:

- Eliminação de 27 atributos com valores integralmente ausentes (*null*). O Apêndice B detalha os atributos eliminados, informação crucial para subsidiar ações internas do PPGCC voltadas ao incentivo do preenchimento desses dados. A omissão de variáveis relevantes, como indicadores de necessidades especiais (incluindo suas categorias específicas) e grupo étnico, podem comprometer a completude e a representatividade da base para análises estatísticas e preditivas.
- Eliminação de 20 atributos por apresentarem baixa variabilidade, ou seja, possuírem o mesmo valor em todos os registros. Como esses atributos não contribuem para a distinção entre os casos analisados, assim como também não agregam informações relevantes ao modelo, optou-se por suas exclusões. A lista completa desses atributos, acompanhada de seus respectivos valores únicos, encontra-se detalhada no Apêndice C. Conforme Witten, Frank e Hall (2016), atributos constantes não possuem poder discriminativo e devem ser descartados nas fases iniciais do pré-processamento.
- Através da análise do percentual de preenchimento de cada atributo, colunas com menos de 50% de valores preenchidos foram eliminadas, uma vez que exigiriam imputações excessivas, aumentando o risco de introdução de ruído e viés. Um total de 35 atributos foram removidos (Apêndice D). Conforme Little e Rubin (2019), altos percentuais de dados ausentes dificultam inferências confiáveis, especialmente quando o mecanismo de ausência não é completamente aleatório. Para atributos com pequena quantidade de valores

faltantes, optou-se por mantê-los na base, adotando posteriormente técnicas de imputação, como substituição por média, moda, zero ou -1, conforme os cenários experimentais avaliados.

- Identificação e eliminação de atributos duplicados, isto é, colunas que apresentavam valores similares linha a linha, ainda que com nomes distintos. O Apêndice E apresenta os 7 grupos de atributos com indícios de duplicidade, seja por diferença apenas na nomenclatura, padronização de maiúsculas/minúsculas ou representação alternativa da mesma informação. A remoção dessas redundâncias contribuiu para a redução da dimensionalidade e evitou possíveis distorções em algoritmos sensíveis à correlação entre variáveis (HAN; KAMBER; PEI, 2012).
- Eliminação de 10 atributos determinísticos em relação à variável alvo, ou seja, aqueles cujos valores estavam univocamente associados a uma única classe (Apêndice F). Esse tipo de vazamento determinístico inviabiliza a avaliação real da capacidade preditiva do modelo (KUHN; JOHNSON, 2013).
- Eliminação de atributos referentes a eventos ocorridos após o ingresso do estudante no curso, assim como códigos administrativos e identificadores institucionais (Apêndice G). Essa decisão assegura que os modelos sejam treinados apenas com informações disponíveis no momento inicial do curso, alinhando-se ao objetivo preditivo da pesquisa e evitando relações espúrias decorrentes de identificadores sem significado semântico direto (WITTEN; FRANK; HALL, 2016).

Após a aplicação de todas essas etapas, obteve-se um conjunto final composto por 32 atributos (Apêndice H). Esse conjunto consolidado representa uma base mais adequada para a modelagem preditiva, preservando apenas informações consideradas relevantes e confiáveis.

#### **4.4.4 Preenchimento de Valores Ausentes**

A presença de valores ausentes é uma característica recorrente em bases de dados educacionais, especialmente aquelas provenientes de sistemas institucionais integrados, nos quais informações são coletadas em diferentes momentos e com distintos níveis de obrigatoriedade. Segundo Little e Rubin (2019), dados ausentes podem ocorrer por falhas no processo de coleta, mudanças nos instrumentos de registro ou ausência de resposta por parte dos indivíduos, configurando um desafio relevante para análises estatísticas e aplicações de ML.

A aplicação direta de algoritmos de aprendizado supervisionado sobre dados contendo valores ausentes é, em geral, inviável, uma vez que a maioria dos modelos utilizados neste estudo — como *Decision Tree*, *Random Forest* e *XGBoost* — pressupõe entradas numéricas completas. Além disso, a simples exclusão de registros incompletos pode resultar em perda significativa de informação e introdução de viés, especialmente quando os valores ausentes não ocorrem de forma completamente aleatória (HAN; KAMBER; PEI, 2012).

Nesse contexto, apesar da eliminação dos atributos com menos de 50% dos valores não preenchidos realizados na etapa de limpeza (Seção 4.4.3), ainda restaram atributos com valores ausentes, conforme apresentado no Apêndice H. Por isso, foi necessária a adoção de técnicas de imputação de valores ausentes com o objetivo de preservar o volume de dados disponíveis, manter a representatividade do conjunto amostral e possibilitar a aplicação consistente dos modelos preditivos.

Neste trabalho, foram avaliadas quatro técnicas de imputação, todas baseadas em estratégias determinísticas simples de substituição dos valores ausentes:

- Zero: estratégia simples de preenchimento dos valores ausentes pelo valor zero, permitindo avaliar o impacto mínimo da imputação sobre os dados. Embora possa introduzir distorções em determinadas variáveis, sua inclusão possibilita a comparação com métodos mais sofisticados, sendo adotada como abordagem inicial em estudos de DM e ML, conforme discutido em trabalhos clássicos da área (HASTIE; FRIEDMAN; TIBSHIRANI, 2006; WITTEN; FRANK; HALL, 2016).
- Menos um: utilização de um valor sentinela, no caso -1, fora do intervalo típico das variáveis, com o intuito de sinalizar explicitamente a ausência de informação. Essa abordagem é particularmente relevante em algoritmos baseados em árvores, que conseguem criar regras específicas para identificar e tratar esses valores durante o processo de decisão (QUINLAN, 1993).
- Média: aplicada exclusivamente a atributos numéricos, essa técnica busca preservar a tendência central dos dados, reduzindo alterações na distribuição global das variáveis. De acordo com Little e Rubin (2019), trata-se de uma abordagem amplamente utilizada quando se assume que os valores ausentes ocorrem de forma aproximadamente aleatória.
- Moda: aplicada tanto a atributos categóricos quanto a atributos numéricos discretos, essa estratégia consiste na substituição dos valores ausentes pelo valor mais frequente observado

na variável, preservando a estrutura predominante dos dados e evitando a introdução de valores artificiais, conforme recomendado em Han, Kamber e Pei (2012).

A imputação utilizando os valores fixos zero e -1 foi adotada como forma de testar o impacto de valores fixos no processo de aprendizagem, sendo que o valor -1 atua como marcador explícito de ausência quando fora do domínio natural da variável. Já as imputações pela Média (para variáveis numéricas) e pela Moda (para variáveis categóricas) representam estratégias estatísticas clássicas, amplamente utilizadas na literatura, por preservarem parcialmente a distribuição original dos dados.

A avaliação das quatro técnicas de imputação permitiu investigar, de forma sistemática, o impacto do tratamento de dados ausentes no desempenho dos modelos de classificação, assim como identificar as configurações mais robustas quando combinadas às técnicas de balanceamento de classes e seleção de atributos.

#### **4.4.5 Codificação de Variáveis Categóricas**

Segundo Han, Kamber e Pei (2012), a transformação de variáveis categóricas em representações numéricas constitui uma etapa essencial do pré-processamento, pois permite que informações qualitativas sejam incorporadas ao processo de aprendizagem dos modelos.

As variáveis categóricas presentes na base de dados foram convertidas para formato numérico por meio de mapeamento ordinal, por meio da criação de dicionários de correspondência entre cada categoria observada e um valor inteiro distinto. Essa etapa é necessária porque a maioria dos algoritmos de aprendizado de máquina utilizada neste estudo — como *Decision Tree*, *Random Forest* e *XGBoost* — operam exclusivamente sobre dados numéricos, não sendo capaz de processar diretamente atributos textuais ou simbólicos.

A codificação foi realizada antes da divisão do conjunto de dados em treino e teste, garantindo que todas as categorias observadas fossem mapeadas de forma consistente em todo o conjunto. Essa decisão é fundamental para evitar discrepâncias na representação dos dados, como a atribuição de códigos diferentes para a mesma categoria em subconjuntos distintos, o que poderia comprometer a validade dos experimentos e introduzir viés nos resultados (WITTEN; FRANK; HALL, 2016).

Dessa forma, a etapa de codificação das variáveis categóricas contribuiu para a padronização dos dados, viabilizou a aplicação dos algoritmos de ML e assegurou a reprodutibilidade e consistência dos resultados obtidos.

## 4.5 Etapa 5: Treinamento dos Modelos

Após o pré-processamento dos dados, iniciou-se a etapa de treinamento dos modelos preditivos no contexto de aprendizado supervisionado. O objetivo desta etapa consiste em ajustar os modelos de forma que sejam capazes de aprender padrões subjacentes aos dados históricos e, a partir deles, realizar previsões sobre novos registros não observados.

A seleção dos algoritmos fundamentou-se nas evidências identificadas no mapeamento sistemático da literatura apresentado no Capítulo 3. Mais especificamente, na análise visual da Seção 3.3.4. Assim, foram utilizados os algoritmos *Decision Tree*, *Random Forest* e *XGBoost*, apresentados na Seção 2.1.1.

Esses algoritmos foram treinados a partir de um conjunto de exemplos rotulados, no qual cada instância apresenta um conjunto de variáveis explicativas (atributos independentes) e uma variável alvo previamente definida. Conforme apresentado na Seção 4.4.1, neste estudo a variável alvo corresponde ao *Status* acadêmico do estudante, sendo utilizada como referência para o processo de classificação.

Para cada modelo, foram avaliadas diferentes configurações de hiperparâmetros, com o objetivo de analisar o impacto da complexidade no desempenho preditivo:

- *Decision Tree*: foi variado o parâmetro *max\_depth*, responsável por limitar a profundidade da árvore, assumindo os valores *None*, 2, 3, 4, 5, 6, 7, 10, 15, 20. A inclusão do valor *None* permite o crescimento irrestrito da árvore, enquanto os demais valores controlam progressivamente a complexidade do modelo.
- *Random Forest*: além da variação do *max\_depth* nos mesmos valores acima, foi fixado o número de árvores em *n\_estimators* = 300, visando tornar o modelo mais robusto, pois as previsões passam a ser baseadas na combinação de várias árvores, reduzindo a influência de decisões individuais e tornando os resultados mais consistentes.
- *XGBoost*: foram testados os mesmos valores de *max\_depth*, mantendo-se *n\_estimators* = 300, o que possibilita avaliar a influência da profundidade das árvores no contexto do *boosting*, considerando ainda seus mecanismos de regularização. Essa estratégia experimental permite investigar diferentes níveis de complexidade dos modelos, favorecendo a identificação de configurações que proporcionem melhor equilíbrio entre viés e variância.

#### 4.5.1 *Divisão em Treino e Testes*

O conjunto de dados foi particionado em dois subconjuntos distintos: 80% dos registros foram destinados ao treinamento dos modelos e 20% reservados para a etapa de teste. Essa proporção é amplamente adotada na literatura de ML por proporcionar um equilíbrio adequado entre a quantidade de dados disponíveis para o aprendizado dos modelos e a necessidade de uma avaliação independente de seu desempenho (KUHN; JOHNSON, 2013).

A divisão foi realizada por meio de amostragem estratificada com base na variável alvo *Status*, garantindo que a proporção entre Egressos com Êxito e Egressos sem Êxito fosse preservada em ambos os subconjuntos. Conforme discutido anteriormente, essa estratégia é particularmente relevante em problemas de predição de evasão escolar, nos quais o conjunto de dados costuma apresentar desbalanceamento entre as classes.

A preservação das proporções originais das classes no conjunto de teste é fundamental para garantir que as métricas de desempenho reflitam de forma realista a capacidade dos modelos em generalizar para dados não vistos. Métricas como *recall* e *F1-score*, especialmente relevantes em contextos de evasão escolar, são fortemente influenciadas pela distribuição das classes, tornando a estratificação uma etapa essencial para a confiabilidade dos resultados (HAN; KAMBER; PEI, 2012).

#### 4.5.2 *Balanceamento das Classes*

Segundo He e Garcia (2009), o desbalanceamento de classes é um dos principais desafios em tarefas de classificação supervisionada. Nesses casos, torna-se necessário adotar estratégias específicas para mitigar esse problema e permitir que os modelos aprendam padrões representativos de ambas as classes.

Em problemas de predição da evasão escolar, é recorrente a presença de desbalanceamento entre as classes. No conjunto de dados analisado (Figura 13), verificou-se uma leve predominância de Egressos sem Êxito (111) em relação aos Egressos com Êxito (99), caracterizando estes últimos como a classe minoritária. Tal configuração pode impactar negativamente o desempenho de modelos de ML, uma vez que diversos algoritmos tendem a favorecer a classe majoritária durante o processo de treinamento. Como consequência, obtêm-se modelos com elevada acurácia global, porém com desempenho limitado na identificação da classe minoritária.

Diante desse contexto, foram aplicadas técnicas de sobreamostragem exclusivamente no conjunto de treino, preservando-se o conjunto de teste com sua distribuição original de classes.

Essa decisão metodológica é fundamental para evitar viés na avaliação dos modelos, garantindo que as métricas de desempenho reflitam a capacidade real de generalização para dados não vistos (KUHN; JOHNSON, 2013).

Nesta etapa, foram avaliadas as técnicas de sobreamostragem apresentadas na Seção 2.3: *Random OverSampler*, SMOTE e BorderlineSMOTE. Essa avaliação constituiu uma etapa essencial para analisar os efeitos do desbalanceamento de classes, melhorar a sensibilidade dos modelos à classe minoritária e reforçar a validade dos resultados obtidos.

#### 4.5.3 Seleção de Atributos

Além do treinamento dos modelos utilizando o conjunto completo de atributos após o pré-processamento, foram avaliadas técnicas de seleção de atributos com o objetivo de reduzir a dimensionalidade dos dados, eliminar variáveis redundantes ou pouco informativas e analisar o impacto dessa redução no desempenho preditivo dos modelos. Em problemas de MDE, conjuntos de dados com grande número de atributos podem aumentar a complexidade do modelo, elevar o custo computacional e favorecer o sobreajuste, sem necessariamente resultar em ganhos de desempenho (HAN; KAMBER; PEI, 2012).

Neste estudo, foram avaliados os três métodos distintos de seleção de atributos apresentados na Seção 2.2:

- *SelectKBest* com teste ANOVA F: essa abordagem possibilita a seleção direta dos  $k$  atributos mais relevantes para a classificação, mantendo a interpretabilidade das variáveis e facilitando a análise dos fatores associados à evasão discente. Segundo (KUHN; JOHNSON, 2013), métodos baseados em testes estatísticos são particularmente úteis quando se deseja equilibrar desempenho preditivo e explicabilidade do modelo.
- Seleção baseada na correlação com a variável alvo: neste estudo, essa abordagem foi empregada com o objetivo de selecionar os  $k$  atributos diretamente relacionados ao desfecho acadêmico, mantendo a interpretabilidade dos resultados e possibilitando análises qualitativas sobre os fatores associados à conclusão ou evasão.
- PCA: no contexto desta pesquisa, o PCA foi utilizado para avaliar se a compactação da informação em  $k$  componentes latentes seria capaz de manter o desempenho preditivo dos modelos, mesmo com a perda de interpretabilidade direta dos atributos originais. Essa

técnica é amplamente empregada em estudos exploratórios e comparativos, especialmente quando o foco está no desempenho preditivo (HAN; KAMBER; PEI, 2012).

Com o intuito de analisar o impacto do tamanho do subconjunto de atributos no desempenho dos métodos de seleção de atributos, foram considerados diferentes valores para o parâmetro  $k$  — 10, 15 e 20 atributos ou componentes. A avaliação de múltiplos valores de  $k$  permite investigar o equilíbrio entre a redução da dimensionalidade e a preservação das informações relevantes presentes nos dados. A escolha de um subconjunto muito pequeno pode resultar na perda de atributos informativos, enquanto a manutenção de um número elevado de variáveis pode introduzir redundância e ruído, afetando negativamente o desempenho dos algoritmos de aprendizado. Dessa forma, a análise comparativa entre diferentes valores de  $k$  contribui para identificar a configuração que proporciona melhor capacidade preditiva para os modelos avaliados (KOHAVI; JOHN, 1997).

A combinação dessas técnicas de seleção de atributos com diferentes cenários de imputação, balanceamento de classes e algoritmos de classificação possibilitou uma análise abrangente do efeito da redução do espaço de atributos sobre o desempenho preditivo.

#### 4.6 Etapa 6: Avaliação dos Modelos

Essa etapa da metodologia consiste na avaliação dos modelos de MDE. Essa etapa é fundamental para conferir sentido prático às descobertas, possibilitando sua utilização em diferentes contextos de análise.

Uma das principais abordagens empregadas para favorecer a compreensão e a interpretação dos resultados é a utilização de métricas para avaliação dos modelos preditivos desenvolvidos. Neste sentido, para avaliação do desempenho dos modelos deste trabalho, foram adotadas as quatro métricas apresentadas na Seção 2.4: Acurácia, Precisão, *Recall* e *F1-score*.

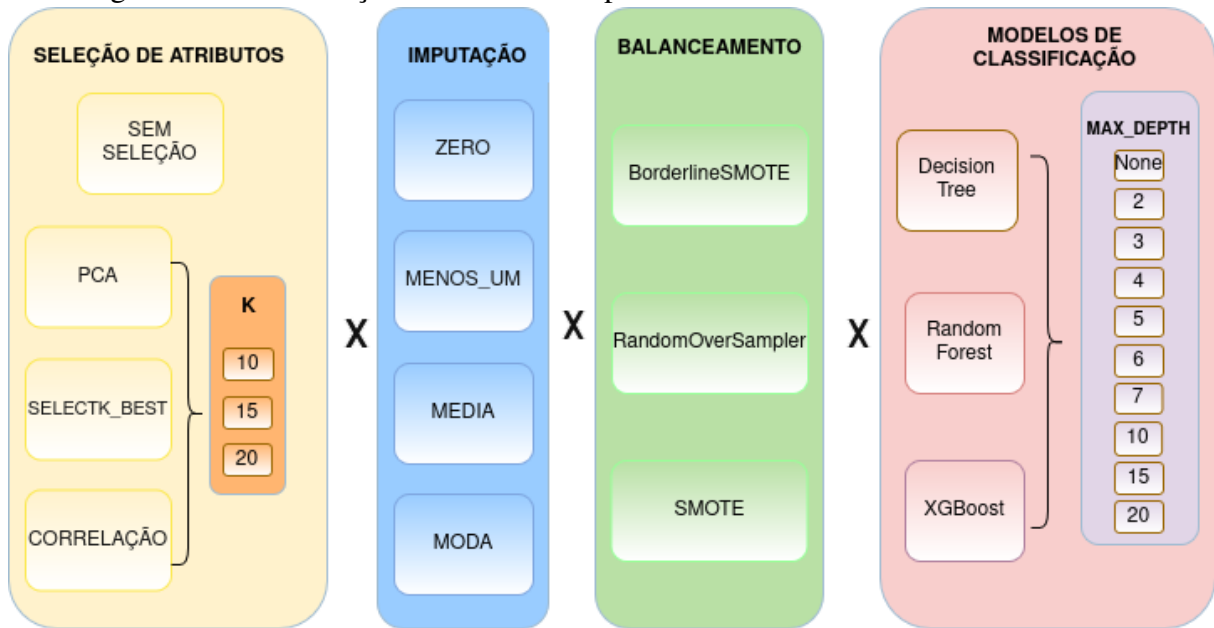
Destaca-se que, para Precisão, *Recall* e *F1-score*, foi utilizada a configuração padrão da biblioteca *Scikit-learn* (*average = binary* e *pos\_label = 0*), de modo que essas métricas foram calculadas considerando exclusivamente a classe de interesse, correspondente aos casos de evasão acadêmica (Egressos sem Êxito), enquanto a acurácia representa o desempenho global do modelo considerando ambas as classes. Essa escolha metodológica permite enfatizar a capacidade dos modelos em identificar corretamente os estudantes em risco de evasão, aspecto central para aplicações práticas no contexto educacional.

Assim, a partir de todas as decisões de pesquisa detalhadas até o momento neste capítulo,

obteve-se um total de **3600 cenários de experimentação**.

Para tornar essa combinação mais evidente, a Figura 14 ilustra a combinação sistemática de diversas estratégias de pré-processamento e treinamento dos modelos. Essa avaliação densa permite identificar não apenas a eficácia global dos algoritmos, mas também a sensibilidade de cada modelo na identificação correta da classe de evasão (Egressos sem Êxito).

Figura 14 – Combinação dos cenários experimentais.



Fonte: Elaborado pela autora.

O detalhamento estatístico desses resultados, a comparação entre os modelos *Decision Tree*, *Random Forest* e *XGBoost*, e a definição do cenário de melhor desempenho global são apresentados e discutidos no Capítulo 5. Naquele capítulo, os resultados são organizados para evidenciar o impacto das decisões metodológicas aqui descritas sobre a capacidade preditiva final do sistema.

A partir da análise comparativa das métricas, identificou-se o modelo com melhor desempenho preditivo, considerando não apenas a acurácia global, mas principalmente o equilíbrio entre Precisão e *Recall* e *F1-score* correspondente à classe Egressos sem Êxito. Essa avaliação possibilitou selecionar o cenário mais adequado para a predição do *Status* acadêmico dos estudantes, assegurando maior capacidade de generalização e robustez do modelo escolhido na predição do risco de evasão. O modelo com melhor desempenho foi aplicado na etapa seguinte.

#### 4.7 Etapa 7: Aplicação Prática do Modelo

Um diferencial central desta proposta, em contraste com a maioria dos estudos mapeados que se restringem ao campo teórico ou experimental, é a transposição do modelo preditivo para uma solução tecnológica aplicada ao contexto institucional.

Para viabilizar essa entrega, o modelo com melhor desempenho apontado na Etapa 6 (XGBoost combinado com PCA(k=20), imputação Zero, balanceador SMOTE,  $max\_depth = 6$  e  $n\_estimators = 300$ ) foi integrado a uma *Application Programming Interface* (API). A solução permite que a identificação do risco de evasão dos estudantes atualmente matriculados no PPGCC ocorra de forma automatizada e escalável, superando a análise estática e oferecendo uma ferramenta dinâmica de apoio à tomada de decisão institucional baseada em dados.

A API, desenvolvida com a biblioteca *Streamlit*, foi estruturada para receber como entrada os atributos acadêmicos e sociodemográficos dos estudantes e retornar como saída a predição do *Status* acadêmico, com o nível de risco (baixo, médio ou alto) e a probabilidade estimada de evasão. Os detalhes técnicos dessa aplicação prática do melhor modelo são apresentados e discutidos no Capítulo 6.

## 5 RESULTADOS DA AVALIAÇÃO DOS MODELOS

Este capítulo apresenta os resultados obtidos a partir da avaliação dos modelos de predição da evasão escolar estruturados neste estudo sob diferentes cenários experimentais, conforme detalhado na Etapa 6 da proposta (Seção 4.6).

Conforme citado anteriormente, a Figura 14 apresenta a estrutura dos cenários experimentais considerados neste estudo, evidenciando a combinação sistemática de diferentes etapas do processo de modelagem. Inicialmente, são aplicadas diferentes estratégias de seleção de atributos, seguidas por métodos de imputação para tratamento de dados ausentes e técnicas de balanceamento de classes, visando mitigar o desbalanceamento no conjunto de dados. Em seguida, diferentes algoritmos de classificação, com variadas configurações de hiperparâmetros, são empregados para a construção dos modelos preditivos. Por fim, o desempenho dos modelos é avaliado por meio de métricas consolidadas na literatura, como Acurácia, Precisão, *Recall* e *F1-score*. Essa abordagem permite explorar um amplo espaço de configurações, possibilitando a identificação das combinações mais adequadas para o problema investigado.

O restante deste capítulo está estruturado em duas seções principais. Inicialmente, na Seção 5.1, é apresentada a avaliação dos modelos preditivos. A Seção 5.2 é dedicada para discussões sobre o melhor modelo.

### 5.1 Avaliação dos modelos preditivos

Nesta seção, são apresentados e analisados os resultados da avaliação dos modelos preditivos. A avaliação dos experimentos está organizada em dois cenários principais: modelagem *sem* seleção de atributos (Seção 5.1.1) e modelagem *com* seleção de atributos (Seção 5.1.2).

#### 5.1.1 Modelagem sem seleção de atributos

Nesta seção, são apresentados os resultados dos experimentos a partir das seguintes configurações ilustradas na Figura 14: sem seleção de atributos, as quatro técnicas de imputação de dados, as três técnicas de balanceamento e os três modelos. Além disso, para cada modelo foram avaliadas diferentes configurações do hiperparâmetro *max\_depth*, considerando valores variados (2, 3, 4, 5, 6, 7, 10, 15, 20) e também sem limitação de profundidade (*None*), com o objetivo de analisar o impacto da complexidade dos modelos no desempenho. Dessa forma, considerando o elevado número de experimentos gerados a partir dessas combinações (total

de 360 experimentos), optou-se por apresentar apenas os melhores resultados de cada modelo quanto à variação dos hiperparâmetros, selecionados com base no *F1-score* da classe “Egresso sem Êxito”, por representar o equilíbrio entre Precisão e *Recall*.

Inicialmente, são discutidos os resultados individuais de cada modelo, permitindo uma compreensão de seu comportamento em diferentes cenários, para, posteriormente, realizar uma análise comparativa entre as abordagens.

#### 5.1.1.1 Decision Tree

A Tabela 4 apresenta os resultados para o modelo *Decision Tree*, sem seleção de atributos. Na tabela é dado destaque para o melhor cenário, no qual a imputação dos valores é realizada pela técnica *menos\_um*, o balanceamento pelo *RandomOverSampler* e com profundidade máxima da árvore de 6, alcançando Acurácia de 0,761905 e *F1-score* da classe “Egresso sem Êxito” de 0,782609, indicando um bom equilíbrio entre Precisão (0,75) e *Recall* (0,818182).

Tabela 4 – Melhores resultados do modelo *Decision Tree*, sem seleção de atributos.

| Imputação | Balanceamento     | Param.         | Acurácia | Precisão | Recall   | F1-score |
|-----------|-------------------|----------------|----------|----------|----------|----------|
| zero      | BorderlineSMOTE   | max_depth=2    | 0.571429 | 0.550000 | 1.000000 | 0.709677 |
| zero      | RandomOverSampler | max_depth=None | 0.619048 | 0.588235 | 0.909091 | 0.714286 |
| zero      | SMOTE             | max_depth=None | 0.690476 | 0.666667 | 0.818182 | 0.734694 |
| menos_um  | BorderlineSMOTE   | max_depth=5    | 0.690476 | 0.655172 | 0.863636 | 0.745098 |
| menos_um  | RandomOverSampler | max_depth=6    | 0.761905 | 0.750000 | 0.818182 | 0.782609 |
| menos_um  | SMOTE             | max_depth=3    | 0.642857 | 0.600000 | 0.954545 | 0.736842 |
| média     | BorderlineSMOTE   | max_depth=10   | 0.619048 | 0.600000 | 0.818182 | 0.692308 |
| média     | RandomOverSampler | max_depth=2    | 0.666667 | 0.681818 | 0.681818 | 0.681818 |
| média     | SMOTE             | max_depth=4    | 0.642857 | 0.606061 | 0.909091 | 0.727273 |
| moda      | BorderlineSMOTE   | max_depth=2    | 0.571429 | 0.550000 | 1.000000 | 0.709677 |
| moda      | RandomOverSampler | max_depth=6    | 0.595238 | 0.600000 | 0.681818 | 0.638298 |
| moda      | SMOTE             | max_depth=3    | 0.642857 | 0.600000 | 0.954545 | 0.736842 |

Fonte: Elaborada pela autora (2026).

A análise dos resultados da Tabela 4 evidencia que valores intermediários de profundidade (como *max\_depth* entre 3 e 6) apresentaram um melhor equilíbrio entre as métricas, com destaque para combinações como imputação *menos\_um* com *RandomOverSampler* (*max\_depth* = 6), que, conforme dito anteriormente, alcançou os melhores resultados gerais. Esses resultados sugerem que árvores com profundidade moderada conseguem capturar padrões relevantes sem incorrer significativamente em sobreajuste. De forma geral, observa-se que valores muito baixos de profundidade (como *max\_depth* = 2) tendem a produzir modelos mais simples, com menor capacidade de capturar padrões complexos, o que se reflete em desempenhos mais modestos em termos de Acurácia e *F1-score*. Embora, em alguns casos, apresentem *Recall* elevado (inclusive

igual a 1.0), indicando maior sensibilidade à classe “Egresso sem Êxito”. Já configurações com maior profundidade ou sem limite definido (*max\_depth = None*) não apresentaram ganhos expressivos de desempenho, o que pode indicar tendência ao sobreajuste, especialmente considerando o tamanho e características do conjunto de dados.

Observa-se que o *RandomOverSampler* é a técnica de balanceamento presente no cenário que alcançou a maior acurácia e o melhor *F1-score*. O SMOTE também apresentou desempenho competitivo, especialmente por alcançar altos valores de Acurácia e *F1-score* em diversos cenários, como em *zero* e *moda*. Por outro lado, o *BorderlineSMOTE* apresentou resultados mais instáveis, com destaque para cenários como *moda* e *zero*, nos quais atingiu *Recall* máximo (1,0), mas com menor precisão, o que impacta o equilíbrio geral do modelo.

Em relação às estratégias de imputação, o cenário com imputação *menos\_um* se destaca como o mais consistente, apresentando melhores métricas em comparação aos demais. Assim, os resultados indicam que, mesmo sem seleção de atributos, a combinação entre imputação, técnica de balanceamento e configuração do hiperparâmetro, influencia o desempenho do modelo *Decision Tree*.

#### 5.1.1.2 Random Forest

Os resultados apresentados na Tabela 5 evidenciam o desempenho do modelo *Random Forest*. Destaca-se o cenário com imputação pela *moda*, uso do *RandomOverSampler* e com os parâmetros de 300 estimadores e profundidade máxima da árvore de 6, obtendo a maior Acurácia (0,761905) e o melhor *F1-score* da classe “Egresso sem Êxito” (0,772727), além de apresentar bom equilíbrio entre Precisão (0,75) e *Recall* (0,75).

Tabela 5 – Melhores resultados do modelo *Random Forest*, sem seleção de atributos.

| Imputação | Balanceamento     | Param.         | Acurácia | Precisão | Recall   | F1-score |
|-----------|-------------------|----------------|----------|----------|----------|----------|
| zero      | BorderlineSMOTE   | max_depth=None | 0.738095 | 0.764706 | 0.650000 | 0.765957 |
| zero      | RandomOverSampler | max_depth=None | 0.714286 | 0.700000 | 0.700000 | 0.727273 |
| zero      | SMOTE             | max_depth=None | 0.714286 | 0.722222 | 0.650000 | 0.739130 |
| menos_um  | BorderlineSMOTE   | max_depth=7    | 0.738095 | 0.764706 | 0.650000 | 0.765957 |
| menos_um  | RandomOverSampler | max_depth=None | 0.738095 | 0.714286 | 0.750000 | 0.744186 |
| menos_um  | SMOTE             | max_depth=None | 0.738095 | 0.714286 | 0.750000 | 0.744186 |
| media     | BorderlineSMOTE   | max_depth=3    | 0.690476 | 0.652174 | 0.750000 | 0.682927 |
| media     | RandomOverSampler | max_depth=4    | 0.738095 | 0.695652 | 0.800000 | 0.731707 |
| media     | SMOTE             | max_depth=10   | 0.690476 | 0.666667 | 0.700000 | 0.697674 |
| moda      | BorderlineSMOTE   | max_depth=4    | 0.738095 | 0.736842 | 0.700000 | 0.755556 |
| moda      | RandomOverSampler | max_depth=6    | 0.761905 | 0.750000 | 0.750000 | 0.772727 |
| moda      | SMOTE             | max_depth=4    | 0.738095 | 0.714286 | 0.750000 | 0.744186 |

Fonte: Elaborada pela autora (2026).

A análise dos resultados da Tabela 5 indica que o parâmetro *max\_depth* do modelo *Random Forest* exerceu influência no desempenho. De modo geral, valores intermediários de profundidade (*max\_depth* entre 4 e 7) apresentaram os melhores resultados, com destaque para o melhor cenário citado anteriormente. Esses resultados sugerem que árvores com profundidade moderada conseguem capturar padrões relevantes sem comprometer a capacidade de generalização do conjunto.

De modo geral, observa-se que o uso da técnica de balanceamento pelo *RandomOverSampler* está presente no cenário que proporcionou os melhores resultados, conforme citado anteriormente. Ainda em relação ao melhor cenário, os valores de Precisão (0,75) e *Recall* (0,75) reforçam a estabilidade do modelo. Por outro lado, o uso do *BorderlineSMOTE* apresentou desempenho competitivo em alguns casos, especialmente nos cenários *menos\_um* e *zero*, embora com maior variação entre Precisão e *Recall*. Já o SMOTE apresentou resultados intermediários, sem se destacar significativamente em relação às demais técnicas.

Complementarmente, nota-se que os diferentes métodos de imputação impactam o desempenho do modelo, sendo a imputação pela *moda* a mais eficaz nesse contexto, possivelmente por preservar melhor a distribuição dos dados.

### 5.1.1.3 XGBoost

A Tabela 6 evidencia o desempenho do modelo *XGBoost*, sem seleção de atributos. Destaca-se o cenário com imputação pela *moda*, uso do *BorderlineSMOTE* e com os parâmetros de 300 estimadores e profundidade máxima da árvore de 15, que obteve a maior Acurácia (0,833333) e o melhor *F1-score* da classe “Egresso sem Êxito” (0,844444), além de apresentar bom equilíbrio entre Precisão (0,842105) e *Recall* (0,80).

Tabela 6 – Melhores resultados do modelo *XGBoost*, sem seleção de atributos.

| Imputação | Balanceamento     | Param.         | Acurácia | Precisão | Recall   | F1-score |
|-----------|-------------------|----------------|----------|----------|----------|----------|
| zero      | BorderlineSMOTE   | max_depth=4    | 0.809524 | 0.833333 | 0.750000 | 0.826087 |
| zero      | RandomOverSampler | max_depth=4    | 0.809524 | 0.800000 | 0.800000 | 0.818182 |
| zero      | SMOTE             | max_depth=5    | 0.809524 | 0.833333 | 0.750000 | 0.826087 |
| menos_um  | BorderlineSMOTE   | max_depth=3    | 0.809524 | 0.833333 | 0.750000 | 0.826087 |
| menos_um  | RandomOverSampler | max_depth=5    | 0.809524 | 0.800000 | 0.800000 | 0.818182 |
| menos_um  | SMOTE             | max_depth=4    | 0.809524 | 0.800000 | 0.800000 | 0.818182 |
| media     | BorderlineSMOTE   | max_depth=5    | 0.738095 | 0.736842 | 0.700000 | 0.755556 |
| media     | RandomOverSampler | max_depth=15   | 0.738095 | 0.714286 | 0.750000 | 0.744186 |
| media     | SMOTE             | max_depth=3    | 0.738095 | 0.736842 | 0.700000 | 0.755556 |
| moda      | BorderlineSMOTE   | max_depth=15   | 0.833333 | 0.842105 | 0.800000 | 0.844444 |
| moda      | RandomOverSampler | max_depth=2    | 0.761905 | 0.727273 | 0.800000 | 0.761905 |
| moda      | SMOTE             | max_depth=None | 0.785714 | 0.823529 | 0.700000 | 0.808511 |

Fonte: Elaborada pela autora (2026).

A análise dos resultados da Tabela 6 demonstra que o parâmetro *max\_depth* também desempenha papel importante no desempenho do modelo *XGBoost*, porém com um comportamento distinto em relação aos modelos anteriores. Como o *XGBoost* é um método de *boosting*, que constrói árvores sequencialmente corrigindo os erros das anteriores, ele tende a ser menos suscetível ao sobreajuste causado por árvores mais profundas.

Observa-se que o número de estimadores foi mantido constante ( $n_{estimators} = 300$ ), permitindo isolar o efeito da profundidade das árvores. De modo geral, tanto valores baixos quanto intermediários de *max\_depth* (entre 3 e 5) apresentaram bom desempenho e resultados consistentes, especialmente nas combinações com imputação *menos\_um* e *zero*, nas quais foram obtidas Acurácias de até 0.809524 e *F1-scores* de até 0.826087. Esses resultados indicam que árvores relativamente rasas já são suficientes para capturar padrões relevantes quando combinadas no processo iterativo do *boosting*.

Por outro lado, diferentemente do observado nos demais modelos, valores mais elevados de profundidade (como *max\_depth* = 15) também apresentaram desempenho competitivo, com destaque para o melhor cenário citado anteriormente e em destaque na Tabela 6. Isso sugere que o *XGBoost* consegue explorar maior complexidade sem comprometer significativamente a generalização, possivelmente devido aos mecanismos de regularização inerentes ao algoritmo.

Além disso, configurações sem limite explícito de profundidade (*max\_depth* = *None*) não se destacaram em relação às demais, indicando que há um ponto ótimo de complexidade além do qual não há ganhos expressivos. Assim, os resultados evidenciam que o *XGBoost* apresenta maior robustez em relação à variação de *max\_depth*, sendo capaz de manter bom desempenho em diferentes níveis de profundidade.

Observa-se que as três técnicas de balanceamento apresentaram desempenhos próximos em vários cenários, sugerindo que o *XGBoost* é menos sensível à escolha da técnica de balanceamento em comparação aos demais modelos.

Em relação às estratégias de imputação, a imputação pela *moda* novamente se destacou, proporcionando o melhor resultado geral. Assim, os achados indicam que, mesmo sem seleção de atributos, o *XGBoost* apresenta elevada capacidade preditiva, sendo robusto às variações de pré-processamento e se consolidando como o modelo mais eficaz entre os analisados.

#### 5.1.1.4 Comparação entre os modelos

A Tabela 7 apresenta, de forma consolidada, os melhores resultados obtidos por cada modelo de classificação sem a aplicação de técnicas de seleção de atributos. A organização das informações em uma única tabela tem como objetivo facilitar a visualização, a comparação e a compreensão do desempenho dos diferentes algoritmos avaliados, considerando as estratégias de imputação de dados, balanceamento de classes e parametrizações que proporcionaram os melhores resultados. Dessa forma, torna-se mais clara a análise das métricas de desempenho (Acurácia, Precisão, *Recall* e *F1-score*), permitindo identificar as configurações mais eficazes para cada modelo e compreender as diferenças de desempenho entre as abordagens investigadas.

Tabela 7 – Melhores resultados para cada modelo sem seleção de atributos.

| Modelo        | Imputação | Balanceamento     | Param.       | Acurácia | Precisão | Recall   | F1-score |
|---------------|-----------|-------------------|--------------|----------|----------|----------|----------|
| XGBoost       | moda      | BorderlineSMOTE   | max_depth=15 | 0.833333 | 0.842105 | 0.800000 | 0.844444 |
| RandomForest  | moda      | RandomOverSampler | max_depth=6  | 0.761905 | 0.750000 | 0.750000 | 0.772727 |
| Decision Tree | menos_um  | RandomOverSampler | max_depth=6  | 0.761905 | 0.750000 | 0.818182 | 0.782609 |

Fonte: Elaborada pela autora (2026).

De forma geral, a comparação entre os três modelos sem uso de métodos de seleção de atributos evidencia que o *XGBoost* apresentou o melhor desempenho global, com maiores valores de Acurácia e *F1-score* da classe “Egresso sem Êxito” e maior estabilidade entre os diferentes cenários. O *Random Forest* obteve resultados intermediários, destacando-se principalmente com o uso de *RandomOverSampler* e imputação pela *moda*. Já o *Decision Tree* apresentou o desempenho mais instável, embora também tenha alcançado bons resultados no cenário *menos\_um* com *RandomOverSampler*.

Em todos os modelos, observa-se que a escolha das técnicas de balanceamento e imputação influencia diretamente os resultados, com destaque recorrente para a imputação pela *moda* e o uso de *RandomOverSampler* ou *BorderlineSMOTE*.

Assim, conclui-se que, mesmo sem seleção de atributos, o *XGBoost* se sobressai como a abordagem mais robusta, seguido pelo *Random Forest*, enquanto o *Decision Tree* apresenta maior sensibilidade às variações dos cenários experimentais.

#### 5.1.2 Modelagem com seleção de atributos

Nesta seção, são apresentados os resultados dos experimentos a partir das seguintes configurações ilustradas na Figura 14: três métodos de seleção de atributos, quatro técnicas de imputação de dados, três técnicas de balanceamento e três modelos. É importante lembrar

que, conforme detalhado na Seção 4.5.3, para cada técnica de seleção de atributos, foi avaliada uma variação dos valores do parâmetro  $k$  que representa a quantidade de atributos selecionados (10, 15 e 20). Ademais, de acordo com o estabelecido na Seção 4.5, para cada modelo foram avaliadas diferentes configurações do hiperparâmetro *max\_depth*, considerando valores variados (2, 3, 4, 5, 6, 7, 10, 15, 20) e também sem limitação de profundidade (*None*).

Portanto, os experimentos conduzidos compreenderam um total de 3.240 cenários distintos. Em virtude do elevado volume de resultados gerados e das limitações de espaço inerentes a este trabalho, optou-se por apresentar os resultados em três tabelas distintas, uma para cada modelo avaliado (*Decision Tree*, *Random Forest* e *XGBoost*). Em cada tabela, são exibidos os melhores desempenhos para os parâmetros da técnica de seleção de atributo e dos modelos de classificação, obtidos para cada combinação de cenário de imputação, método de balanceamento e técnica de seleção de atributos, adotando-se o *F1-score* da classe “Egresso sem Êxito” como métrica principal de avaliação. Tal organização visa proporcionar uma análise mais clara, concisa e comparativa, evidenciando as configurações de maior desempenho no contexto investigado para cada abordagem de modelagem.

#### 5.1.2.1 *Decision Tree*

Os resultados apresentados na Tabela 8 evidenciam uma melhora significativa no desempenho do modelo *Decision Tree* com a aplicação de técnicas de seleção de atributos, quando comparados aos resultados sem essa etapa. Destaca-se o cenário com seleção de atributos por correlação, imputação *menos\_um* e uso da técnica de balanceamento SMOTE, que alcançou a maior acurácia (0,8571) e o melhor *F1-score* da classe “Egresso sem Êxito” (0,8750), além de um elevado *Recall* (0,9545), indicando excelente capacidade de identificar corretamente os casos de sem êxito.

De forma geral, observa-se que as combinações envolvendo SMOTE e *BorderlineSMOTE*, associadas principalmente à seleção por correlação e *SelectKBest*, proporcionaram os melhores resultados, sugerindo que a redução de atributos contribuiu para melhorar a capacidade de generalização do modelo. O cenário *menos\_um* novamente se destaca como o mais consistente, concentrando os melhores desempenhos, enquanto os cenários *media*, *moda* e *zero* apresentaram resultados mais modestos ou intermediários. A técnica PCA, por sua vez, mostrou desempenho inferior na maioria dos casos, possivelmente por reduzir a interpretabilidade e não preservar adequadamente as características mais relevantes para a classificação. Assim, os resultados

indicam que a combinação entre imputação adequada, balanceamento e seleção de atributos é fundamental para potencializar o desempenho do *Decision Tree*.

Tabela 8 – Melhores resultados do modelo *Decision Tree*, com seleção de atributos.

| Seleção de atributos | Imputação | Balanceamento     | Param.         | Acurácia | Precisão | Recall | F1-score |
|----------------------|-----------|-------------------|----------------|----------|----------|--------|----------|
| PCA (k=15)           | zero      | BorderlineSMOTE   | max_depth=2    | 0.6667   | 0.6818   | 0.6818 | 0.6818   |
| PCA (k=15)           | zero      | RandomOverSampler | max_depth=4    | 0.7143   | 0.7778   | 0.6364 | 0.7000   |
| PCA (k=20)           | zero      | SMOTE             | max_depth=3    | 0.6905   | 0.6667   | 0.8182 | 0.7347   |
| PCA (k=20)           | menos_um  | BorderlineSMOTE   | max_depth=15   | 0.6667   | 0.6429   | 0.8182 | 0.7200   |
| PCA (k=15)           | menos_um  | RandomOverSampler | max_depth=4    | 0.6905   | 0.7143   | 0.6818 | 0.6977   |
| PCA (k=20)           | menos_um  | SMOTE             | max_depth=4    | 0.7381   | 0.7200   | 0.8182 | 0.7660   |
| PCA (k=10)           | media     | BorderlineSMOTE   | max_depth=7    | 0.5714   | 0.5667   | 0.7727 | 0.6538   |
| PCA (k=10)           | media     | RandomOverSampler | max_depth=15   | 0.6190   | 0.6000   | 0.8182 | 0.6923   |
| PCA (k=15)           | media     | SMOTE             | max_depth=4    | 0.5714   | 0.5588   | 0.8636 | 0.6786   |
| PCA (k=10)           | moda      | BorderlineSMOTE   | max_depth=20   | 0.6667   | 0.6538   | 0.7727 | 0.7083   |
| PCA (k=20)           | moda      | RandomOverSampler | max_depth=5    | 0.7857   | 0.8095   | 0.7727 | 0.7907   |
| PCA (k=15)           | moda      | SMOTE             | max_depth=5    | 0.6905   | 0.6667   | 0.8182 | 0.7347   |
| SelectKBest (k=15)   | zero      | BorderlineSMOTE   | max_depth=4    | 0.8095   | 0.7692   | 0.9091 | 0.8333   |
| SelectKBest (k=20)   | zero      | RandomOverSampler | max_depth=None | 0.5952   | 0.5758   | 0.8636 | 0.6909   |
| SelectKBest (k=15)   | zero      | SMOTE             | max_depth=6    | 0.7857   | 0.8095   | 0.7727 | 0.7907   |
| SelectKBest (k=20)   | menos_um  | BorderlineSMOTE   | max_depth=4    | 0.7381   | 0.6897   | 0.9091 | 0.7843   |
| SelectKBest (k=20)   | menos_um  | RandomOverSampler | max_depth=2    | 0.6429   | 0.5946   | 1.0000 | 0.7458   |
| SelectKBest (k=15)   | menos_um  | SMOTE             | max_depth=7    | 0.8095   | 0.7917   | 0.8636 | 0.8261   |
| SelectKBest (k=15)   | media     | BorderlineSMOTE   | max_depth=3    | 0.7619   | 0.8000   | 0.7273 | 0.7619   |
| SelectKBest (k=15)   | media     | RandomOverSampler | max_depth=6    | 0.7857   | 0.8824   | 0.6818 | 0.7692   |
| SelectKBest (k=15)   | media     | SMOTE             | max_depth=4    | 0.7381   | 0.7037   | 0.8636 | 0.7755   |
| SelectKBest (k=15)   | moda      | BorderlineSMOTE   | max_depth=6    | 0.6667   | 0.6538   | 0.7727 | 0.7083   |
| SelectKBest (k=20)   | moda      | RandomOverSampler | max_depth=5    | 0.7143   | 0.7500   | 0.6818 | 0.7143   |
| SelectKBest (k=10)   | moda      | SMOTE             | max_depth=None | 0.7381   | 0.7037   | 0.8636 | 0.7755   |
| Correlação (k=10)    | zero      | BorderlineSMOTE   | max_depth=None | 0.6429   | 0.6296   | 0.7727 | 0.6939   |
| Correlação (k=20)    | zero      | RandomOverSampler | max_depth=10   | 0.6429   | 0.6129   | 0.8636 | 0.7170   |
| Correlação (k=20)    | zero      | SMOTE             | max_depth=20   | 0.6905   | 0.6667   | 0.8182 | 0.7347   |
| Correlação (k=20)    | menos_um  | BorderlineSMOTE   | max_depth=10   | 0.8333   | 0.8261   | 0.8636 | 0.8444   |
| Correlação (k=10)    | menos_um  | RandomOverSampler | max_depth=2    | 0.6429   | 0.5946   | 1.0000 | 0.7458   |
| Correlação (k=20)    | menos_um  | SMOTE             | max_depth=20   | 0.8571   | 0.8077   | 0.9545 | 0.8750   |
| Correlação (k=15)    | media     | BorderlineSMOTE   | max_depth=4    | 0.7381   | 0.7037   | 0.8636 | 0.7755   |
| Correlação (k=15)    | media     | RandomOverSampler | max_depth=4    | 0.7381   | 0.7619   | 0.7273 | 0.7442   |
| Correlação (k=15)    | media     | SMOTE             | max_depth=4    | 0.7381   | 0.7037   | 0.8636 | 0.7755   |
| Correlação (k=15)    | moda      | BorderlineSMOTE   | max_depth=7    | 0.6429   | 0.6207   | 0.8182 | 0.7059   |
| Correlação (k=20)    | moda      | RandomOverSampler | max_depth=5    | 0.7143   | 0.7500   | 0.6818 | 0.7143   |
| Correlação (k=10)    | moda      | SMOTE             | max_depth=15   | 0.7143   | 0.6923   | 0.8182 | 0.7500   |

Fonte: Elaborada pela autora (2026).

### 5.1.2.2 Random Forest

A Tabela 9 apresenta os melhores resultados obtidos pelo modelo *Random Forest* nos diferentes cenários experimentais avaliados.

Os resultados apresentados na Tabela 9 demonstram que o modelo *Random Forest* apresentou desempenho elevado e consistente com a aplicação de técnicas de seleção de atributos, superando seus resultados anteriores sem essa etapa. Destaca-se o cenário com seleção de atributos por correlação, imputação *menos\_um* e uso do SMOTE, que alcançou a maior acurácia (0,8571) e *F1-score* da classe “Egresso sem Êxito” (0,8636), com valores equilibrados de precisão

Tabela 9 – Melhores resultados do modelo *Random Forest*, com seleção de atributos.

| Seleção de atributos | Imputação | Balanceamento     | Param.                          | Acurácia | Precisão | Recall | F1-score |
|----------------------|-----------|-------------------|---------------------------------|----------|----------|--------|----------|
| PCA (k=20)           | zero      | BorderlineSMOTE   | n_estimators=300,max_depth=10   | 0.7857   | 0.7407   | 0.9091 | 0.8163   |
| PCA (k=20)           | zero      | RandomOverSampler | n_estimators=300,max_depth=6    | 0.8333   | 0.8571   | 0.8182 | 0.8372   |
| PCA (k=20)           | zero      | SMOTE             | n_estimators=300,max_depth=5    | 0.8333   | 0.8000   | 0.9091 | 0.8511   |
| PCA (k=20)           | menos_um  | BorderlineSMOTE   | n_estimators=300,max_depth=2    | 0.7619   | 0.7308   | 0.8636 | 0.7917   |
| PCA (k=20)           | menos_um  | RandomOverSampler | n_estimators=300,max_depth=2    | 0.8333   | 0.8000   | 0.9091 | 0.8511   |
| PCA (k=20)           | menos_um  | SMOTE             | n_estimators=300,max_depth=7    | 0.7857   | 0.7600   | 0.8636 | 0.8085   |
| PCA (k=15)           | media     | BorderlineSMOTE   | n_estimators=300,max_depth=7    | 0.6905   | 0.6667   | 0.8182 | 0.7347   |
| PCA (k=20)           | media     | RandomOverSampler | n_estimators=300,max_depth=None | 0.6667   | 0.6429   | 0.8182 | 0.7200   |
| PCA (k=20)           | media     | SMOTE             | n_estimators=300,max_depth=7    | 0.6667   | 0.6538   | 0.7727 | 0.7083   |
| PCA (k=20)           | moda      | BorderlineSMOTE   | n_estimators=300,max_depth=10   | 0.7857   | 0.7600   | 0.8636 | 0.8085   |
| PCA (k=20)           | moda      | RandomOverSampler | n_estimators=300,max_depth=15   | 0.7619   | 0.7308   | 0.8636 | 0.7917   |
| PCA (k=20)           | moda      | SMOTE             | n_estimators=300,max_depth=10   | 0.8333   | 0.8000   | 0.9091 | 0.8511   |
| SelectKBest (k=20)   | zero      | BorderlineSMOTE   | n_estimators=300,max_depth=10   | 0.8095   | 0.7917   | 0.8636 | 0.8261   |
| SelectKBest (k=15)   | zero      | RandomOverSampler | n_estimators=300,max_depth=10   | 0.7857   | 0.7407   | 0.9091 | 0.8163   |
| SelectKBest (k=20)   | zero      | SMOTE             | n_estimators=300,max_depth=7    | 0.8095   | 0.8500   | 0.7727 | 0.8095   |
| SelectKBest (k=15)   | menos_um  | BorderlineSMOTE   | n_estimators=300,max_depth=15   | 0.7857   | 0.7407   | 0.9091 | 0.8163   |
| SelectKBest (k=20)   | menos_um  | RandomOverSampler | n_estimators=300,max_depth=15   | 0.8333   | 0.8571   | 0.8182 | 0.8372   |
| SelectKBest (k=15)   | menos_um  | SMOTE             | n_estimators=300,max_depth=10   | 0.8095   | 0.7692   | 0.9091 | 0.8333   |
| SelectKBest (k=15)   | media     | BorderlineSMOTE   | n_estimators=300,max_depth=3    | 0.7619   | 0.7727   | 0.7727 | 0.7727   |
| SelectKBest (k=20)   | media     | RandomOverSampler | n_estimators=300,max_depth=None | 0.8095   | 0.7917   | 0.8636 | 0.8261   |
| SelectKBest (k=20)   | media     | SMOTE             | n_estimators=300,max_depth=5    | 0.8095   | 0.8500   | 0.7727 | 0.8095   |
| SelectKBest (k=20)   | moda      | BorderlineSMOTE   | n_estimators=300,max_depth=15   | 0.7857   | 0.7826   | 0.8182 | 0.8000   |
| SelectKBest (k=20)   | moda      | RandomOverSampler | n_estimators=300,max_depth=7    | 0.7857   | 0.8421   | 0.7273 | 0.7805   |
| SelectKBest (k=15)   | moda      | SMOTE             | n_estimators=300,max_depth=None | 0.8095   | 0.8182   | 0.8182 | 0.8182   |
| Correlação (k=20)    | zero      | BorderlineSMOTE   | n_estimators=300,max_depth=15   | 0.8095   | 0.8182   | 0.8182 | 0.8182   |
| Correlação (k=15)    | zero      | RandomOverSampler | n_estimators=300,max_depth=7    | 0.8333   | 0.8261   | 0.8636 | 0.8444   |
| Correlação (k=15)    | zero      | SMOTE             | n_estimators=300,max_depth=None | 0.7857   | 0.7826   | 0.8182 | 0.8000   |
| Correlação (k=15)    | menos_um  | BorderlineSMOTE   | n_estimators=300,max_depth=None | 0.7857   | 0.7600   | 0.8636 | 0.8085   |
| Correlação (k=20)    | menos_um  | RandomOverSampler | n_estimators=300,max_depth=15   | 0.8333   | 0.8261   | 0.8636 | 0.8444   |
| Correlação (k=20)    | menos_um  | SMOTE             | n_estimators=300,max_depth=5    | 0.8571   | 0.8636   | 0.8636 | 0.8636   |
| Correlação (k=20)    | media     | BorderlineSMOTE   | n_estimators=300,max_depth=7    | 0.7619   | 0.7500   | 0.8182 | 0.7826   |
| Correlação (k=20)    | media     | RandomOverSampler | n_estimators=300,max_depth=5    | 0.7619   | 0.8000   | 0.7273 | 0.7619   |
| Correlação (k=20)    | media     | SMOTE             | n_estimators=300,max_depth=None | 0.7857   | 0.7826   | 0.8182 | 0.8000   |
| Correlação (k=20)    | moda      | BorderlineSMOTE   | n_estimators=300,max_depth=5    | 0.7857   | 0.8421   | 0.7273 | 0.7805   |
| Correlação (k=15)    | moda      | RandomOverSampler | n_estimators=300,max_depth=10   | 0.7619   | 0.7500   | 0.8182 | 0.7826   |
| Correlação (k=15)    | moda      | SMOTE             | n_estimators=300,max_depth=15   | 0.7619   | 0.7727   | 0.7727 | 0.7727   |

Fonte: Elaborada pela autora (2026).

e *Recall*, indicando excelente capacidade de identificar corretamente os casos de sem êxito.

O cenário com imputação *menos\_um* novamente se destaca como o mais consistente, concentrando os melhores desempenhos. Além disso, diferentemente do observado no *Decision Tree*, a técnica PCA não comprometeu significativamente o desempenho do modelo em todos os cenários, mantendo métricas elevadas em diversas configurações. Assim, os resultados indicam que o *Random Forest*, aliado à seleção de atributos, apresenta alta capacidade preditiva e estabilidade.

### 5.1.2.3 XGBoost

A Tabela 10 apresenta os resultados, no qual o modelo *XGBoost* apresentou o melhor desempenho geral entre todos os modelos analisados, especialmente com a aplicação de técnicas de seleção de atributos. Destaca-se o cenário com redução por PCA, imputação *zero* e uso do SMOTE, que alcançou a maior acurácia (0,8810) e o melhor *F1-score* da classe “Egresso sem Êxito” (0,8936), além de um elevado *Recall* (0,9545), indicando excelente capacidade de

identificar corretamente os casos de sem êxito.

De forma geral, observa-se que o *XGBoost* manteve resultados elevados e consistentes em diversas combinações de técnicas de balanceamento e métodos de seleção, com destaque para o uso de PCA, que, diferentemente dos outros modelos, contribuiu para o aumento do desempenho. Os cenários com imputação com *zero*, *menos\_um* e *moda* apresentaram resultados bastante competitivos, enquanto o cenário *media* manteve desempenho inferior, porém ainda consistente. Além disso, nota-se que diferentes técnicas de balanceamento (*SMOTE*, *BorderlineSMOTE* e *RandomOverSampler*) produziram resultados próximos, reforçando a robustez do modelo. Assim, os resultados indicam que o *XGBoost*, aliado à seleção de atributos, apresenta elevada capacidade preditiva, estabilidade e maior eficácia na identificação dos casos de interesse, consolidando-se como o modelo mais promissor para o problema analisado.

Tabela 10 – Melhores resultados do modelo *XGBoost*, com seleção de atributos.

| Seleção de atributos | Imputação | Balanceamento     | Param.                          | Acurácia | Precisão | Recall | F1-score |
|----------------------|-----------|-------------------|---------------------------------|----------|----------|--------|----------|
| PCA (k=20)           | zero      | BorderlineSMOTE   | n_estimators=300,max_depth=4    | 0.8571   | 0.8333   | 0.9091 | 0.8696   |
| PCA (k=20)           | zero      | RandomOverSampler | n_estimators=300,max_depth=15   | 0.8571   | 0.8333   | 0.9091 | 0.8696   |
| PCA (k=20)           | zero      | SMOTE             | n_estimators=300,max_depth=6    | 0.8810   | 0.8400   | 0.9545 | 0.8936   |
| PCA (k=20)           | menos_um  | BorderlineSMOTE   | n_estimators=300,max_depth=None | 0.7857   | 0.7600   | 0.8636 | 0.8085   |
| PCA (k=20)           | menos_um  | RandomOverSampler | n_estimators=300,max_depth=4    | 0.8095   | 0.7917   | 0.8636 | 0.8261   |
| PCA (k=20)           | menos_um  | SMOTE             | n_estimators=300,max_depth=5    | 0.8333   | 0.8000   | 0.9091 | 0.8511   |
| PCA (k=15)           | media     | BorderlineSMOTE   | n_estimators=300,max_depth=3    | 0.7143   | 0.6786   | 0.8636 | 0.7600   |
| PCA (k=15)           | media     | RandomOverSampler | n_estimators=300,max_depth=2    | 0.7143   | 0.6786   | 0.8636 | 0.7600   |
| PCA (k=20)           | media     | SMOTE             | n_estimators=300,max_depth=2    | 0.6667   | 0.6667   | 0.7273 | 0.6957   |
| PCA (k=15)           | moda      | BorderlineSMOTE   | n_estimators=300,max_depth=4    | 0.7857   | 0.7600   | 0.8636 | 0.8085   |
| PCA (k=20)           | moda      | RandomOverSampler | n_estimators=300,max_depth=10   | 0.7857   | 0.7407   | 0.9091 | 0.8163   |
| PCA (k=20)           | moda      | SMOTE             | n_estimators=300,max_depth=20   | 0.8333   | 0.8261   | 0.8636 | 0.8444   |
| SelectKBest (k=15)   | zero      | BorderlineSMOTE   | n_estimators=300,max_depth=7    | 0.8095   | 0.7917   | 0.8636 | 0.8261   |
| SelectKBest (k=15)   | zero      | RandomOverSampler | n_estimators=300,max_depth=7    | 0.7857   | 0.7600   | 0.8636 | 0.8085   |
| SelectKBest (k=15)   | zero      | SMOTE             | n_estimators=300,max_depth=4    | 0.8095   | 0.7917   | 0.8636 | 0.8261   |
| SelectKBest (k=20)   | menos_um  | BorderlineSMOTE   | n_estimators=300,max_depth=5    | 0.8333   | 0.8571   | 0.8182 | 0.8372   |
| SelectKBest (k=20)   | menos_um  | RandomOverSampler | n_estimators=300,max_depth=3    | 0.8333   | 0.8571   | 0.8182 | 0.8372   |
| SelectKBest (k=20)   | menos_um  | SMOTE             | n_estimators=300,max_depth=10   | 0.8333   | 0.8261   | 0.8636 | 0.8444   |
| SelectKBest (k=15)   | media     | BorderlineSMOTE   | n_estimators=300,max_depth=3    | 0.7381   | 0.7200   | 0.8182 | 0.7660   |
| SelectKBest (k=15)   | media     | RandomOverSampler | n_estimators=300,max_depth=10   | 0.7619   | 0.7727   | 0.7727 | 0.7727   |
| SelectKBest (k=15)   | media     | SMOTE             | n_estimators=300,max_depth=15   | 0.7143   | 0.7083   | 0.7727 | 0.7391   |
| SelectKBest (k=15)   | moda      | BorderlineSMOTE   | n_estimators=300,max_depth=5    | 0.8333   | 0.8571   | 0.8182 | 0.8372   |
| SelectKBest (k=15)   | moda      | RandomOverSampler | n_estimators=300,max_depth=10   | 0.8333   | 0.8571   | 0.8182 | 0.8372   |
| SelectKBest (k=15)   | moda      | SMOTE             | n_estimators=300,max_depth=None | 0.8333   | 0.8571   | 0.8182 | 0.8372   |
| Correlação (k=15)    | zero      | BorderlineSMOTE   | n_estimators=300,max_depth=3    | 0.8095   | 0.7917   | 0.8636 | 0.8261   |
| Correlação (k=20)    | zero      | RandomOverSampler | n_estimators=300,max_depth=None | 0.8095   | 0.8889   | 0.7273 | 0.8000   |
| Correlação (k=20)    | zero      | SMOTE             | n_estimators=300,max_depth=20   | 0.8333   | 0.8947   | 0.7727 | 0.8293   |
| Correlação (k=20)    | menos_um  | BorderlineSMOTE   | n_estimators=300,max_depth=None | 0.8333   | 0.8571   | 0.8182 | 0.8372   |
| Correlação (k=20)    | menos_um  | RandomOverSampler | n_estimators=300,max_depth=3    | 0.8333   | 0.8571   | 0.8182 | 0.8372   |
| Correlação (k=20)    | menos_um  | SMOTE             | n_estimators=300,max_depth=2    | 0.8095   | 0.8500   | 0.7727 | 0.8095   |
| Correlação (k=15)    | media     | BorderlineSMOTE   | n_estimators=300,max_depth=3    | 0.7381   | 0.7391   | 0.7727 | 0.7556   |
| Correlação (k=15)    | media     | RandomOverSampler | n_estimators=300,max_depth=15   | 0.7619   | 0.8000   | 0.7273 | 0.7619   |
| Correlação (k=15)    | media     | SMOTE             | n_estimators=300,max_depth=3    | 0.7381   | 0.7391   | 0.7727 | 0.7556   |
| Correlação (k=15)    | moda      | BorderlineSMOTE   | n_estimators=300,max_depth=6    | 0.8571   | 0.8636   | 0.8636 | 0.8636   |
| Correlação (k=15)    | moda      | RandomOverSampler | n_estimators=300,max_depth=4    | 0.8095   | 0.8500   | 0.7727 | 0.8095   |
| Correlação (k=15)    | moda      | SMOTE             | n_estimators=300,max_depth=20   | 0.8333   | 0.8261   | 0.8636 | 0.8444   |

Fonte: Elaborada pela autora (2026).

#### 5.1.2.4 Comparação entre modelos

A Tabela 11 reúne os melhores resultados obtidos por cada modelo após a aplicação de técnicas de seleção de atributos, apresentando de forma sintetizada as configurações que alcançaram o melhor desempenho. A disposição dessas informações em uma única tabela visa proporcionar uma visualização mais clara e facilitar a compreensão dos impactos da seleção de atributos sobre os modelos avaliados. Além disso, a comparação conjunta das estratégias de seleção, imputação, balanceamento e parametrização permite identificar quais combinações contribuíram para o aprimoramento das métricas de desempenho. Observa-se que os modelos apresentaram resultados superiores aos obtidos sem seleção de atributos, destacando-se o *XGBoost* com seleção por PCA ( $k=20$ ), que alcançou os maiores valores de Acurácia, *Recall* e *F1-score*, evidenciando a relevância da etapa de seleção de atributos para a melhoria da capacidade preditiva dos classificadores.

Tabela 11 – Melhores resultados por modelo com seleção de atributos.

| Modelo        | Seleção de atributos  | Imputação | Balanceamento | Param.       | Acurácia | Precisão | Recall | F1-score |
|---------------|-----------------------|-----------|---------------|--------------|----------|----------|--------|----------|
| XGBoost       | PCA ( $k=20$ )        | zero      | SMOTE         | max_depth=6  | 0.8810   | 0.8400   | 0.9545 | 0.8936   |
| Random Forest | Correlação ( $k=20$ ) | menos_um  | SMOTE         | max_depth=5  | 0.8571   | 0.8636   | 0.8636 | 0.8636   |
| Decision Tree | Correlação ( $k=20$ ) | menos_um  | SMOTE         | max_depth=20 | 0.8571   | 0.8077   | 0.9545 | 0.8750   |

Fonte: Elaborada pela autora (2026).

De forma geral, a comparação entre os resultados dos três modelos *sem* e *com* a seleção de atributos evidencia que a aplicação da seleção de atributos apresenta melhoria de desempenho. Ademais, entre os modelos, o *XGBoost* se destacou como o mais eficaz, alcançando os maiores valores de acurácia e *F1-score* da classe “Egresso sem Êxito”, além de maior consistência entre diferentes configurações.

O *Random Forest* também apresentou resultados elevados e estáveis, com desempenho próximo ao *XGBoost* em diversos cenários, demonstrando robustez frente às variações de técnicas de seleção, imputação e balanceamento. Já o *Decision Tree*, embora tenha apresentado ganhos significativos em relação à ausência de seleção de atributos, ainda se mostrou inferior aos demais, com maior sensibilidade às escolhas de pré-processamento.

Em todos os modelos, observa-se que os cenários *menos\_um* e *zero* tendem a concentrar os melhores resultados, enquanto o cenário *media* apresenta desempenho mais limitado. Além disso, a utilização de técnicas de seleção de atributos, especialmente correlação e SelectKBest, contribuiu diretamente para a melhoria dos modelos com *Decision Tree* e *Random Forest*, enquanto o PCA apresentou impacto variável, sendo mais efetivo no *XGBoost*.

Assim, conclui-se que a combinação entre pré-processamento e modelos mais robustos, como o *XGBoost* e o *Random Forest*, é fundamental para alcançar melhores resultados na tarefa analisada.

## 5.2 Discussões sobre Melhor Modelo

Nesta seção, são analisados o desempenho preditivo e os fatores associados às decisões dos modelos. Inicialmente, apresenta-se a matriz de confusão do modelo de melhor desempenho, visando compreender os padrões de acerto e erro na identificação de estudantes em risco de evasão. Em seguida, realiza-se a análise da importância dos atributos, considerando o modelo de melhor desempenho que não utiliza PCA, devido à limitação dessa técnica quanto à interpretabilidade das variáveis originais.

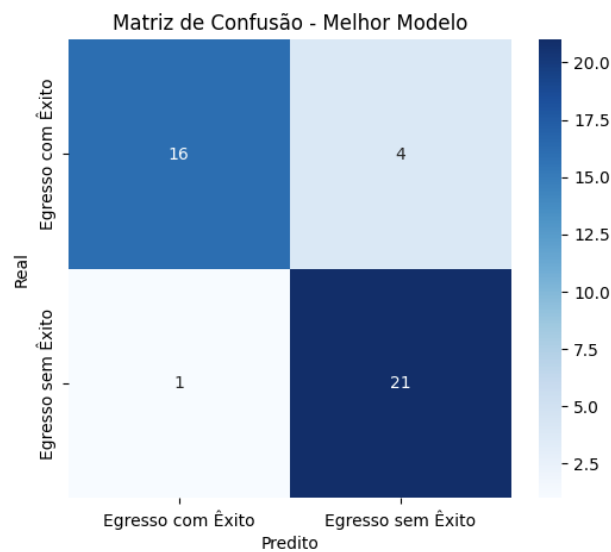
### 5.2.1 Matriz de Confusão

O melhor desempenho entre todas as variações experimentais foi obtido pelo modelo *XGBoost*, utilizando seleção de atributos com PCA ( $k = 20$ ), imputação por zero, balanceamento com SMOTE, 300 estimadores e profundidade máxima de árvore igual a 6, alcançando acurácia de 0,8810 e *F1-score* da classe “Egresso sem Êxito” de 0,8936.

Com o objetivo de aprofundar a interpretação dos resultados da modelagem preditiva, apresenta-se a Matriz de Confusão (Figura 15) referente às predições no conjunto de teste, evidenciando o desempenho do modelo na classificação dos estudantes entre “Egresso com Êxito” e “Egresso sem Êxito”.

Observa-se que o modelo acertou 16 casos de estudantes com êxito e 21 casos de estudantes sem êxito, representando os acertos em cada classe. Por outro lado, ocorreram 4 erros em que estudantes com êxito foram classificados como sem êxito (falsos positivos) e 1 caso em que aluno sem êxito foi classificado como com êxito (falso negativo). Esse último é particularmente relevante no contexto da evasão, pois indica estudante em risco que não foi corretamente identificado pelo modelo. De forma geral, a matriz evidencia um desempenho equilibrado entre as classes, mas com presença de erros que podem impactar ações de intervenção educacional.

Figura 15 – Matriz de confusão do melhor modelo.



Fonte: Elaborado pela autora.

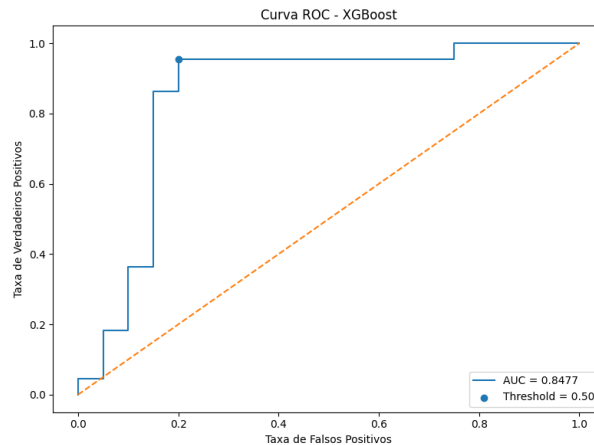
### 5.2.2 Curva ROC

A Curva *Receiver Operating Characteristic* (ROC) é uma das métricas mais utilizadas para avaliar o desempenho de modelos de classificação binária. Ela representa graficamente a relação entre a Taxa de Verdadeiros Positivos e a Taxa de Falsos Positivos para diferentes limiares de decisão (*thresholds*). Seu principal objetivo é demonstrar a capacidade do modelo de distinguir corretamente as diferentes classes, independentemente de um único valor de limiar adotado para a classificação.

A utilização da Curva ROC é particularmente importante em problemas de predição de evasão acadêmica, pois permite avaliar não apenas quantos estudantes em risco são corretamente identificados, mas também o custo associado aos falsos alarmes. Em outras palavras, a curva possibilita analisar o equilíbrio entre a capacidade de detectar alunos propensos à evasão e o risco de classificar incorretamente estudantes que não evadiriam. Dessa forma, a Curva ROC fornece uma visão mais abrangente do desempenho do modelo do que métricas isoladas, como acurácia.

No gráfico apresentado na Figura 16 para o modelo *XGBoost*, observa-se que a curva azul permanece significativamente acima da linha diagonal tracejada. Essa linha diagonal representa o desempenho de um classificador aleatório, cuja capacidade discriminatória é equivalente ao acaso. Quanto mais próxima a Curva ROC estiver do canto superior esquerdo do gráfico, melhor é o desempenho do modelo, pois isso indica altas taxas de verdadeiros positivos e baixas taxas de falsos positivos.

Figura 16 – Curva ROC para melhor modelo



Fonte: Elaborado pela autora.

O valor da área sob a Curva (*Area Under the Curve* - AUC) obtido foi de 0,8477, indicando uma excelente capacidade de discriminação entre estudantes que concluem o curso com êxito e aqueles que apresentam risco de evasão. Em termos práticos, esse resultado sugere que o modelo possui aproximadamente 84,77% de probabilidade de atribuir uma pontuação de risco maior a um estudante que evadirá do que a um estudante que não evadirá. Como valores de AUC acima de 0,80 são geralmente considerados muito bons na literatura, o resultado obtido reforça a eficácia do modelo proposto.

O ponto destacado na curva corresponde ao *threshold* de 0,50, utilizado como limiar para a classificação final. Nesse ponto, o modelo alcança uma Taxa de Verdadeiros Positivos próxima de 95%, indicando elevada capacidade de identificar corretamente os estudantes em situação de risco, ainda que com uma Taxa de Falsos Positivos em torno de 20%. Esse comportamento é desejável no contexto educacional, pois prioriza a identificação de possíveis casos de evasão, permitindo que a instituição adote ações preventivas e de acompanhamento antes que o desligamento acadêmico ocorra.

### 5.2.3 Análise da importância dos atributos

Complementarmente, a avaliação da importância dos atributos é uma prática recomendada em Mineração de Dados, pois permite compreender quais fatores mais contribuem para a capacidade discriminativa do modelo, além de favorecer a interpretabilidade dos resultados (HAN; KAMBER; PEI, 2012).

O modelo de melhor desempenho utiliza a técnica de seleção de atributos por PCA, que reduz a dimensionalidade ao projetar os dados em componentes não interpretáveis, o que dificulta

a análise direta da importância dos atributos, pois a importância refere-se aos componentes principais e não diretamente aos atributos originais.

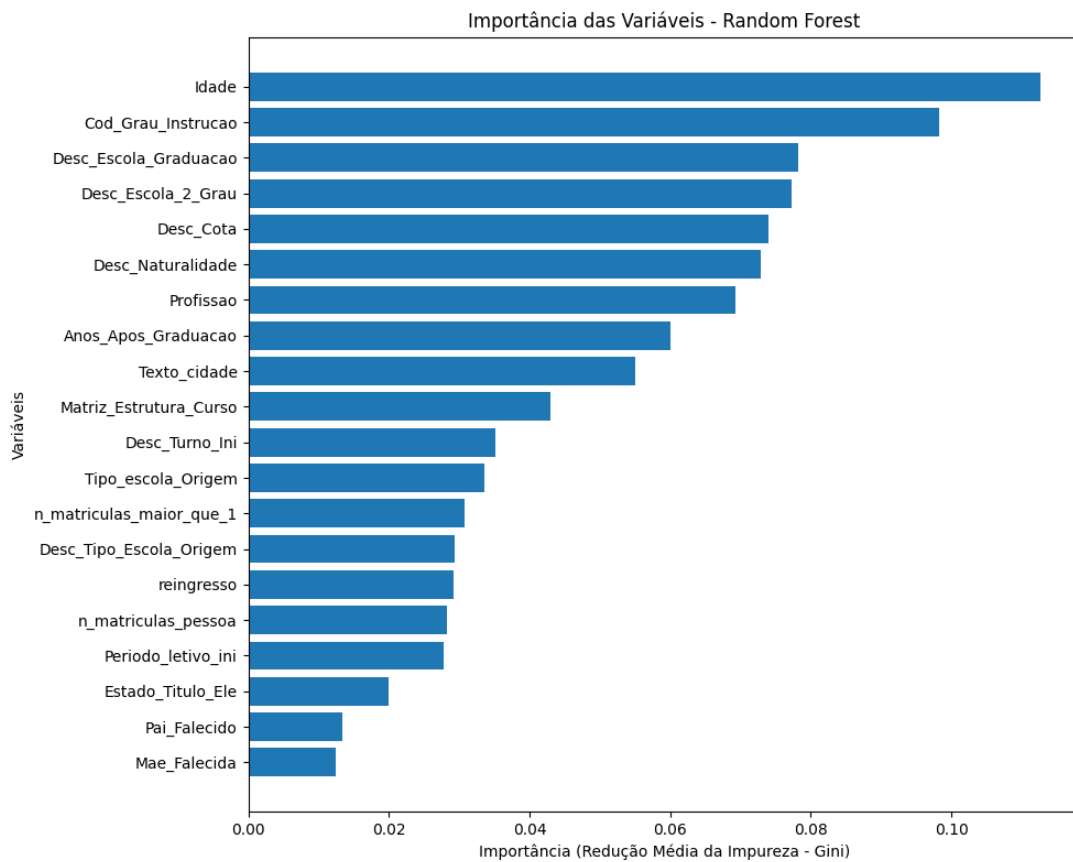
Diante dessa limitação, visando realizar uma análise interpretável, foi considerado o modelo de melhor desempenho que não utiliza PCA. A seleção foi realizada com base em um critério hierárquico, priorizando, nessa ordem, os maiores valores de Acurácia, Precisão, *Recall* e *F1-score* para a classe “Egresso sem Êxito”, seguido do valor de  $k$  utilizado no método de seleção de atributos. Portanto, nessa análise considera-se o cenário com o modelo *Random Forest*, seleção de atributos por correlação  $k = 20$ , imputação *menos\_um* e balanceamento via SMOTE, alcançando Acurácia de 0,8571, e para a classe “Egresso sem Êxito”, atingiu Precisão de 0,8636, *Recall* de 0,8636 e *F1-score* de 0,8636. Destaca-se, ainda, que cenário corresponde ao modelo com segundo melhor desempenho geral.

A importância dos atributos foi estimada por meio da redução média da impureza de Gini, métrica que quantifica a contribuição de cada atributo para a diminuição da heterogeneidade das classes ao longo das árvores do *ensemble* (HAN; KAMBER; PEI, 2012). Assim, quanto maior a redução acumulada da impureza proporcionada por uma variável, maior sua relevância para o modelo.

A Figura 17 apresenta o *ranking* das dez variáveis com maior contribuição para o desempenho preditivo do modelo no cenário adotado para a análise de importância dos atributos.

Observa-se que a variável *Idade* apresentou a maior contribuição individual para o modelo, indicando forte associação com a probabilidade de conclusão ou evasão. Esse resultado sugere que características relacionadas ao perfil etário dos estudantes podem influenciar sua permanência acadêmica, possivelmente em interação com fatores como trajetória acadêmica, tempo disponível para os estudos e demandas profissionais. Entretanto, essa relação não deve ser interpretada de forma causal, tampouco utilizada de maneira discriminatória em processos institucionais. Embora a idade tenha contribuído estatisticamente para a identificação de padrões associados à evasão escolar, o modelo desenvolvido neste trabalho não deve, em hipótese alguma, ser utilizado para fins de exclusão, restrição de acesso ou seleção de candidatos em processos acadêmicos. Seu propósito é exclusivamente preventivo e de apoio à permanência estudantil, buscando subsidiar ações de acompanhamento pedagógico, assistência estudantil e intervenção precoce para estudantes em situação de vulnerabilidade acadêmica. Dessa forma, a utilização do modelo deve observar princípios éticos, de transparência e não discriminação, garantindo que os resultados sejam empregados apenas para promoção do apoio educacional e redução da evasão escolar.

Figura 17 – Importância dos atributos para o segundo melhor.



Fonte: Elaborada pela autora (2026).

A elevada importância de *Cod\_Grau\_Instrucao* e das variáveis relacionadas à instituição de origem (*Desc\_Escola\_Graduacao* e *Desc\_Escola\_2\_Grau*) evidencia o papel da trajetória educacional prévia na adaptação ao curso. Esse achado dialoga com a literatura sobre permanência estudantil, que enfatiza a influência do capital acadêmico acumulado e da integração acadêmica no sucesso do estudante (TINTO, 2012).

Atributos de natureza socioeconômica, como *Desc\_Cota*, *Profissao*, *Texto\_cidade* e *Desc\_Naturalidade*, também figuram entre os mais relevantes, indicando que a evasão constitui fenômeno multifatorial, influenciado por condições estruturais e contextuais.

## 6 APLICAÇÃO PRÁTICA DO MELHOR MODELO

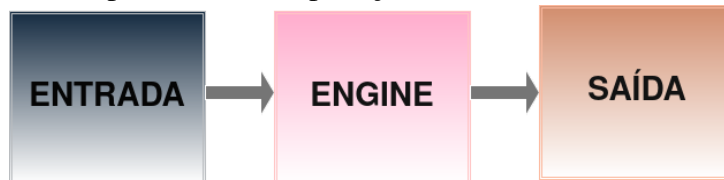
O objetivo principal deste capítulo é realizar a transposição do ambiente puramente experimental para uma solução tecnológica aplicada ao contexto institucional do PPGCC do IFCE. Assim, este capítulo detalha a ferramenta desenvolvida para dar sentido prático às descobertas. A aplicação visa oferecer um suporte dinâmico à coordenação e ao colegiado do programa, viabilizando o apoio à tomada de decisão baseada em evidências, permitindo intervenções pedagógicas precoces e informadas.

Ressalta-se que a solução possui caráter consultivo e de apoio, não substituindo o julgamento humano ou decisões acadêmicas da coordenação e colegiado do PPGCC. Assim, as predições devem ser interpretadas como subsídios adicionais para políticas de permanência e êxito, reforçando o compromisso ético de utilizar a IA para promover a equidade e o sucesso acadêmico no PPGCC.

### 6.1 Fluxo Operacional da Aplicação

A aplicação foi desenvolvida utilizando o framework *Streamlit*, um framework em Python especializado na criação de interfaces para ciência de dados, que permite a interação dinâmica com modelos complexos de forma simplificada e estruturada em três componentes principais: entrada de dados, mecanismo de processamento e apresentação dos resultados (Figura 18). O sistema recebe como entrada arquivos nos formatos CSV ou Excel contendo informações dos estudantes matriculados. Após o carregamento, os dados passam por uma etapa de validação e pré-processamento, na qual são removidos atributos não utilizados pelo modelo, criadas variáveis derivadas e aplicados os mesmos procedimentos de transformação empregados durante o treinamento do modelo preditivo.

Figura 18 – Fluxo Operacional da Aplicação.



Fonte: Elaborada pela autora (2026).

O mecanismo de processamento utiliza artefatos previamente treinados e serializados com a biblioteca *Joblib*, incluindo o *pipeline* de Aprendizado de Máquina, a lista de atributos esperados pelo modelo e os dicionários de codificação das variáveis categóricas. Durante a

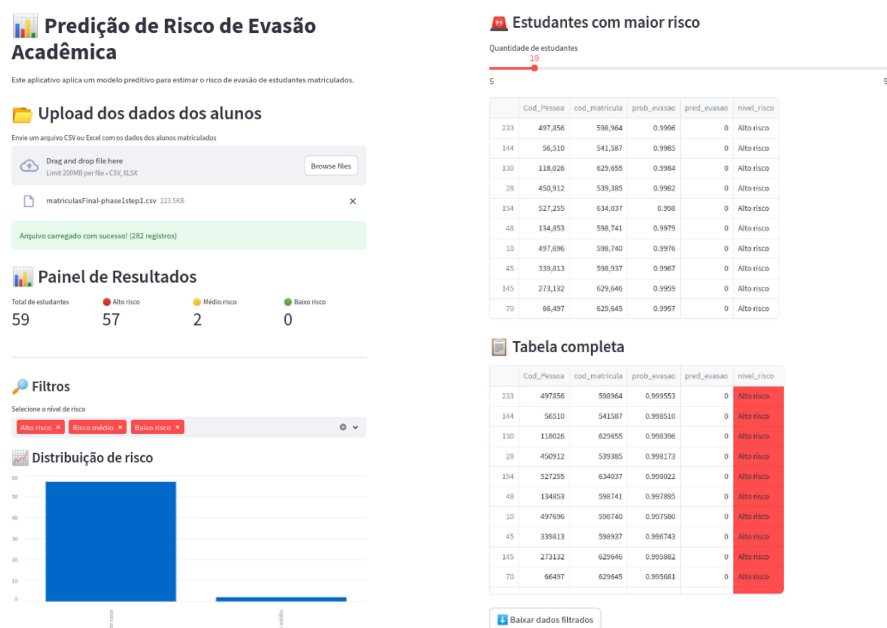
execução, as variáveis categóricas são convertidas para representações numéricas, os atributos são alinhados à estrutura utilizada no treinamento e os dados são submetidos ao modelo *XGBoost* responsável pela inferência.

Como saída, o sistema calcula a probabilidade de evasão para cada estudante e realiza sua classificação em níveis de risco (baixo, médio ou alto). Essa categorização é baseada em limiares previamente definidos e armazenados em arquivo de configuração, sendo o limiar de alto risco obtido por meio do Índice de Youden, que identifica o ponto de melhor equilíbrio entre sensibilidade e especificidade do modelo. Os resultados são apresentados por meio de indicadores estatísticos, gráficos de distribuição de risco, tabelas interativas e relatórios exportáveis em formato CSV, permitindo o acompanhamento individual e coletivo dos estudantes analisados.

A arquitetura adotada garante que todas as etapas de pré-processamento executadas durante o treinamento sejam reproduzidas automaticamente no momento da predição, assegurando consistência entre os dados utilizados para construção do modelo e aqueles empregados em sua aplicação prática.

A Figura 19 apresenta as interfaces geradas pela execução do fluxo. Essas interfaces são apresentadas em detalhes (com recortes) nas próximas seções.

Figura 19 – Interface inicial da aplicação de predição de evasão no PPGCC.



Fonte: Elaborada pela autora (2026).

### 6.1.1 Etapa 1: Entrada

O fluxo operacional da aplicação inicia-se com o carregamento de dados de estudantes (Figura 20). A aplicação permite o *upload* de arquivos nos formatos CSV ou Excel. Esse conjunto de dados corresponde ao subconjunto do *dataset* original composto por registros cuja variável alvo (*Status*) não pôde ser definida. Dessa forma, os registros representam estudantes cujo desfecho acadêmico ainda é desconhecido ou não se enquadra nas categorias previamente definidas.

O principal objetivo desse conjunto de dados é servir como base para aplicação do modelo preditivo treinado, permitindo estimar a probabilidade de evasão ou conclusão para esses estudantes. Assim, ele representa o conjunto de dados alvo para geração de previsões e apoio à tomada de decisão institucional.

Figura 20 – Interface para entrada de dados de estudantes matriculados.



Fonte: Elaborada pela autora (2026).

### 6.1.2 Etapa 2: Engine

Após o carregamento dos dados, inicia-se a etapa de *engine*, que foi projetada para operar exclusivamente no modo de inferência, utilizando o modelo treinado de melhor desempenho: *XGBoost*, com seleção de atributos com PCA ( $k = 20$ ), imputação por zero, balanceamento com SMOTE, 300 estimadores e profundidade máxima de árvore igual a 6.

Para garantir que o modelo opere em produção exatamente como no ambiente de treino, deve-se utilizar o *pipeline* completo de pré-processamento dos dados, a seleção dos atributos por meio do PCA e a classificação com o modelo treinado. Mais especificamente, os dados de entrada passam automaticamente pelo *pipeline* que inclui:

1. Criação de atributos: Criação de novos a partir dos dados existentes, conforme descrito na

#### Seção 4.4.2.

2. Limpeza de dados: Eliminação de atributos redundantes ou irrelevantes para a predição da evasão escolar, identificados anteriormente a partir da análise exploratória, conforme detalhado na Seção 4.4.3.
3. Preenchimento de valores ausentes: Preenchimento de possíveis valores ausentes com zero, conforme validado no melhor cenário e descrito na Seção 4.4.4.
4. Codificação de variáveis categóricas: Conversão de variáveis categóricas em numéricas, conforme descrito na Seção 4.4.5.
5. Projeção PCA: Redução da dimensionalidade para os  $k = 20$  componentes principais definidos na modelagem.
6. Classificação: Aplicação do modelo treinado para gerar a predição final.

Este processo garante que estudantes matriculados sejam avaliados sob o mesmo rigor metodológico aplicado à base histórica de matrículas.

#### 6.1.3 Etapa 3: Saída

A Figura 21 apresenta a parte do painel de resultados da aplicação na qual são sintetizadas as informações geradas pelo modelo preditivo. Essa interface fornece uma visão geral e imediata da distribuição dos estudantes em diferentes níveis de risco de evasão, facilitando a análise por parte da coordenação e colegiado do programa.

Figura 21 – Interface de saída da aplicação com sintetizações geradas pelo modelo.



Fonte: Elaborada pela autora (2026).

Os limiares utilizados para classificação dos níveis de risco foram definidos a partir das probabilidades previstas pelo modelo *XGBoost* para a classe correspondente à evasão escolar. Para determinação desses valores, foi utilizada a Curva ROC (Figura 16), considerando a evasão como classe de interesse. O modelo apresentou valor de AUC igual a 0,8477, indicando boa capacidade discriminativa na identificação de estudantes com maior propensão à evasão.

O limiar de alto risco foi obtido por meio do Índice de Youden, resultando no valor de 0,5025, correspondente ao ponto de melhor equilíbrio entre sensibilidade e especificidade do modelo. O Índice de Youden é uma medida amplamente utilizada para determinar o limiar de classificação mais adequado em problemas de diagnóstico e classificação binária. Ele é calculado pela expressão  $J = \text{Sensibilidade} + \text{Especificidade} - 1$ , assumindo valores entre 0 e 1, sendo que valores mais elevados indicam melhor capacidade discriminatória. O limiar que maximiza esse índice representa o ponto em que o modelo alcança simultaneamente alta capacidade de identificar corretamente os casos positivos e baixa taxa de classificações incorretas dos casos negativos. Dessa forma, a utilização do Índice de Youden permitiu selecionar um ponto de corte mais robusto e adequado para a identificação de estudantes com maior risco de evasão, favorecendo a adoção de ações preventivas pela instituição.

A partir desse valor, foi estabelecido um limiar intermediário de 0,3266 para caracterização do risco médio. Dessa forma, estudantes com probabilidades inferiores a 0,3266 foram classificados como baixo risco, aqueles com probabilidades entre 0,3266 e 0,5025 foram classificados como risco médio e, por fim, estudantes com probabilidades superiores a 0,5025 foram classificados como alto risco de evasão. Essa abordagem permitiu transformar as probabilidades previstas em categorias interpretáveis, facilitando a identificação de estudantes com maior vulnerabilidade à evasão escolar e apoiando ações institucionais de acompanhamento e permanência acadêmica.

A Figura 22 apresenta a interface de análise da distribuição dos estudantes por nível de risco de evasão, incorporando mecanismos de filtragem que permitem segmentar os resultados gerados pelo modelo preditivo. Essa funcionalidade amplia a capacidade de exploração dos dados, possibilitando análises mais específicas conforme o interesse do programa.

A Figura 23 apresenta a interface destinada à identificação dos estudantes com maior risco de evasão, permitindo a visualização ordenada dos casos mais críticos com base nas probabilidades estimadas pelo modelo. Essa funcionalidade foi projetada para apoiar diretamente a priorização de intervenções acadêmicas. A aplicação fornece ainda, para cada estudante, uma probabilidade estimada de evasão e o nível.

As variáveis identificadoras dos estudantes são preservadas exclusivamente para fins de rastreabilidade dos resultados, não sendo utilizadas como atributos de entrada do modelo.

A Figura 24 apresenta a visualização detalhada dos resultados gerados pelo modelo preditivo por meio de uma tabela completa contendo todos os estudantes analisados. Essa interface permite a inspeção individual dos registros, complementando as análises agregadas e

Figura 22 – Interface de saída da aplicação com distribuição dos estudantes por nível de risco de evasão.



Fonte: Elaborada pela autora (2026).

Figura 23 – Interface de saída da aplicação com lista de estudantes com maior risco de evasão.

**Estudantes com maior risco**

Quantidade de estudantes

10

5

|     | Cod_Pessoa | cod_matricula | prob_evasao | pred_evasao | nivel_risco |
|-----|------------|---------------|-------------|-------------|-------------|
| 233 | 497,856    | 598,964       | 0.9996      | 0           | Alto risco  |
| 144 | 56,510     | 541,587       | 0.9985      | 0           | Alto risco  |
| 130 | 118,026    | 629,655       | 0.9984      | 0           | Alto risco  |
| 28  | 450,912    | 539,385       | 0.9982      | 0           | Alto risco  |
| 154 | 527,255    | 634,037       | 0.998       | 0           | Alto risco  |
| 48  | 134,853    | 598,741       | 0.9979      | 0           | Alto risco  |
| 10  | 497,696    | 598,740       | 0.9976      | 0           | Alto risco  |
| 45  | 339,813    | 598,937       | 0.9967      | 0           | Alto risco  |
| 145 | 273,132    | 629,646       | 0.9959      | 0           | Alto risco  |
| 70  | 66,497     | 629,645       | 0.9957      | 0           | Alto risco  |

Fonte: Elaborada pela autora (2026).

visuais apresentadas anteriormente.

Além da visualização interativa dos resultados, a aplicação permite a exportação dos dados preditos em formato CSV, possibilitando sua integração com outros sistemas institucionais ou análises posteriores.

## 6.2 Limitações da Solução Proposta

Apesar da viabilidade de aplicação prática da solução desenvolvida, algumas limitações devem ser consideradas na interpretação e utilização dos resultados.

Figura 24 – Interface de saída da aplicação com a previsão de evasão para todos os estudantes.



### Tabela completa

|     | Cod_Pessoa | cod_matricula | prob_evasao | pred_evasao | nivel_risco |
|-----|------------|---------------|-------------|-------------|-------------|
| 233 | 497856     | 598964        | 0.999553    | 0           | Alto risco  |
| 144 | 56510      | 541587        | 0.998510    | 0           | Alto risco  |
| 130 | 118026     | 629655        | 0.998396    | 0           | Alto risco  |
| 28  | 450912     | 539385        | 0.998173    | 0           | Alto risco  |
| 154 | 527255     | 634037        | 0.998022    | 0           | Alto risco  |
| 48  | 134853     | 598741        | 0.997895    | 0           | Alto risco  |
| 10  | 497696     | 598740        | 0.997580    | 0           | Alto risco  |
| 45  | 339813     | 598937        | 0.996743    | 0           | Alto risco  |
| 145 | 273132     | 629646        | 0.995882    | 0           | Alto risco  |
| 70  | 66497      | 629645        | 0.995681    | 0           | Alto risco  |

↓ Baixar dados filtrados

Fonte: Elaborada pela autora (2026).

Em primeiro lugar, destaca-se que o modelo foi treinado com base em dados históricos do próprio PPGCC, o que implica que os padrões aprendidos refletem comportamentos passados. Dessa forma, mudanças institucionais, como alterações no currículo, políticas acadêmicas ou perfil dos ingressantes, podem impactar a capacidade de generalização do modelo ao longo do tempo. Esse fenômeno, conhecido como *data drift*, pode reduzir a acurácia das previsões caso não haja atualização periódica do modelo.

Outra limitação relevante refere-se ao conjunto de atributos utilizados. Embora tenham sido empregadas técnicas de engenharia de atributos e seleção de variáveis, o modelo não incorpora fatores externos que podem influenciar significativamente a evasão, como aspectos socioeconômicos, condições de trabalho, saúde mental ou apoio familiar. A ausência dessas variáveis pode limitar a capacidade explicativa do modelo, resultando em previsões que capturam apenas parcialmente a realidade dos estudantes.

Outra limitação importante está relacionada ao caráter probabilístico das previsões. O modelo fornece estimativas de risco, e não certezas sobre a evasão. Assim, existe a possibilidade de ocorrência de falsos positivos (estudantes classificados como risco elevado que não evadem) e

falsos negativos (estudantes não identificados como risco que eventualmente evadem). Essas situações podem impactar a priorização de intervenções e devem ser consideradas na utilização da ferramenta.

Do ponto de vista ético, é fundamental reconhecer o risco de uso inadequado das predições. A classificação de estudantes em níveis de risco pode, se mal interpretada, levar à estigmatização ou a decisões enviesadas. Por esse motivo, reforça-se que a solução deve ser utilizada exclusivamente como instrumento de apoio à decisão, sempre associada à análise qualitativa e ao acompanhamento individualizado dos estudantes.

Por fim, a aplicação foi desenvolvida com foco no contexto específico do PPGCC do IFCE, o que limita sua generalização direta para outros programas ou instituições sem adaptações. Diferenças estruturais, curriculares e institucionais podem exigir reconfiguração do modelo e reavaliação de desempenho.

Diante dessas limitações, recomenda-se a atualização periódica do modelo com novos dados, a ampliação do conjunto de atributos considerados e a utilização da ferramenta em conjunto com práticas institucionais já consolidadas de acompanhamento acadêmico.

## 7 CONCLUSÕES

Este trabalho investigou a aplicação de técnicas de Mineração de Dados Educacionais e Aprendizado de Máquina para a predição de evasão acadêmica no contexto do PPGCC/IFCE. A proposta desenvolvida foi estruturada em etapas, abrangendo desde a coleta e preparação dos dados acadêmicos até a construção e avaliação dos modelos preditivos. Inicialmente, foi realizada a integração e o pré-processamento dos dados, incluindo diferentes estratégias de tratamento de valores ausentes, balanceamento das classes e seleção de atributos. Em seguida, foram treinados e avaliados os algoritmos *Decision Tree*, *Random Forest* e *XGBoost* em diversos cenários experimentais, considerando diferentes combinações de técnicas e hiperparâmetros.

Através da análise de um conjunto de dados institucionais, foram avaliados de forma rigorosa 3.600 cenários experimentais, permitindo identificar as configurações metodológicas que melhor capturam os padrões de interrupção da trajetória discente. Por fim, os modelos foram comparados por meio de métricas de desempenho, possibilitando a identificação da configuração mais adequada para a predição da evasão acadêmica e para o apoio à tomada de decisão institucional voltada à permanência estudantil.

Os resultados demonstraram que o tratamento adequado dos dados é crucial. Embora modelos treinados sem seleção de atributos tenham apresentado desempenho satisfatório, a aplicação de técnicas de redução de dimensionalidade, especificamente com *Principal Component Analysis* (PCA), contribuiu para a melhoria da capacidade preditiva global. O cenário selecionado com melhor desempenho — *XGBoost* combinado com PCA ( $k = 20$ ), imputação por zero e balanceamento SMOTE — alcançou a melhor acurácia global, além de melhor *F1-score* para os “Egressos sem Êxito” e equilíbrio entre precisão e *Recall*, consolidando-se como o mais apto para a identificação de estudantes em risco de evasão.

O objetivo geral desta pesquisa foi plenamente alcançado, resultando não apenas em um modelo validado estatisticamente, mas na transposição desse conhecimento para uma aplicação funcional. Este diferencial prático permite que a predição deixe de ser um dado histórico e se torne um indicador ativo de permanência, oferecendo suporte dinâmico ao programa para intervenções pedagógicas precoces.

No que tange aos objetivos específicos, o Mapeamento Sistemático da Literatura fundamentou as escolhas metodológicas, enquanto a análise comparativa entre algoritmos confirmou a superioridade dos métodos *ensemble* (*boosting*), com o *XGBoost* superando a Árvore de Decisão e o *Random Forest* em termos de robustez e estabilidade.

Apesar do êxito, a pesquisa apresenta limitações intrínsecas. Destaca-se a dependência da qualidade e completude dos dados disponíveis, assim como a ausência de variáveis de natureza socioeconômica ou comportamental, que poderiam enriquecer a modelagem. Ademais, a utilização do PCA implica redução da interpretabilidade direta das variáveis originais.

## 7.1 Trabalhos futuros

A partir dos resultados obtidos e das limitações identificadas, diversas oportunidades de aprimoramento e expansão da solução proposta podem ser exploradas em trabalhos futuros.

Uma das principais possibilidades consiste na ampliação do conjunto de atributos utilizados pelo modelo, incorporando variáveis de natureza socioeconômica, comportamental e psicopedagógica. A inclusão desses fatores pode contribuir para uma compreensão mais abrangente do fenômeno da evasão, aumentando o poder explicativo e preditivo do modelo.

Outra direção promissora refere-se à atualização contínua do modelo por meio de estratégias de aprendizado incremental ou reentrenamento periódico. Essa abordagem permitiria mitigar efeitos de *data drift*, garantindo que o modelo permaneça aderente às mudanças no perfil dos estudantes e nas dinâmicas institucionais ao longo do tempo.

Adicionalmente, sugere-se a investigação de técnicas de explicabilidade mais avançadas, como métodos baseados em valores de Shapley (SHAP), que possibilitam interpretar a contribuição individual de cada variável nas predições. A adoção dessas técnicas pode aumentar a transparência do sistema e fortalecer a confiança dos usuários na ferramenta.

Outra possibilidade relevante é a integração da aplicação com sistemas institucionais já existentes, como ambientes acadêmicos e plataformas de gestão educacional. Em especial, destaca-se a possibilidade de integração futura com o Sistema Unificado de Administração Pública (SUAP), sistema amplamente utilizado pelos Institutos Federais para o gerenciamento de informações acadêmicas e administrativas. Essa integração permitiria automatizar o fluxo de dados, reduzir a necessidade de intervenções manuais na atualização das informações e viabilizar a identificação contínua de estudantes com risco de evasão. Além disso, a incorporação da solução ao ecossistema do SUAP poderia ampliar sua utilização no cotidiano institucional, fornecendo subsídios para ações preventivas e para a tomada de decisão por gestores, coordenadores e equipes de apoio estudantil.

Do ponto de vista metodológico, estudos futuros podem explorar a utilização de outros algoritmos ou abordagens híbridas, incluindo modelos baseados em redes neurais ou técnicas de

ensemble mais complexas, visando avaliar possíveis ganhos de desempenho.

Por fim, recomenda-se a realização de estudos longitudinais, nos quais os estudantes sejam acompanhados ao longo do tempo, para avaliar o impacto real da utilização da ferramenta na redução da evasão, analisando se as intervenções orientadas pelas predições contribuem efetivamente para a permanência dos estudantes.

## REFERÊNCIAS

- BAKER, R.; YACEF, K. The state of educational data mining in 2009: A review and future visions. **Journal of Educational Data Mining**, v. 1, p. 3–17, 01 2009.
- BAKER, R. S.; INVENTADO, P. S. Educational data mining and learning analytics. In: \_\_\_\_\_. **Learning Analytics: From Research to Practice**. New York, NY: Springer New York, 2014. p. 61–75. ISBN 978-1-4614-3305-7. Disponível em: <[https://doi.org/10.1007/978-1-4614-3305-7\\_4](https://doi.org/10.1007/978-1-4614-3305-7_4)>.
- BARBOSA, D.; CABRAL, L.; DWAN, F.; FEITAS, E.; MELLO, R. Previsão da evasão escolar através da análise de dados e aprendizagem de máquina: Um estudo de caso. In: **II Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil**. Porto Alegre, RS, Brasil: SBC, 2023. p. 42–50. ISSN 0000-0000. Disponível em: <<https://sol.sbc.org.br/index.php/wapla/article/view/26127>>.
- BASTOS, N.; GOMES, L.; SILVEIRA, R.; OLIVEIRA, C. Mapeamento sistemático da mineração de dados educacionais no combate à evasão escolar no brasil. In: **Anais do IX Congresso sobre Tecnologias na Educação**. Porto Alegre, RS, Brasil: SBC, 2025. p. 177–187. ISSN 0000-0000. Disponível em: <<https://sol.sbc.org.br/index.php/ctrl/article/view/40480>>.
- BITENCOURT, W. A.; SILVA, D. M.; XAVIER, G. d. C. Pode a inteligência artificial apoiar ações contra evasão escolar universitária? **Ensaio: Avaliação e Políticas Públicas em Educação**, Fundação CESGRANRIO, v. 30, n. 116, p. 669–694, Jul 2022. ISSN 0104-4036. Disponível em: <<https://doi.org/10.1590/S0104-403620220003002854>>.
- BRAGA, L. **Introdução à Mineração de Dados - 2a edição: Edição ampliada e revisada**. e-papers, 2005. ISBN 9788576500353. Disponível em: <<https://books.google.com.br/books?id=0qcQbVsPxnkC>>.
- BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, oct 2001. ISSN 1573-0565. Disponível em: <<https://doi.org/10.1023/A:1010933404324>>.
- BRITO, B.; MELLO, R.; ALVES, G. Identificação de atributos relevantes na evasão no ensino superior público brasileiro. In: **XXXI Simpósio Brasileiro de Informática na Educação**. Porto Alegre, RS, Brasil: SBC, 2020. p. 1032–1041. ISSN 0000-0000. Disponível em: <<https://sol.sbc.org.br/index.php/sbie/article/view/12859>>.
- CAPES. **I Plano Nacional de Pós-Graduação (1975–1979)**. Brasília, 1975. Acesso em: 03 fev. 2026. Disponível em: <<https://www.gov.br/capes/pt-br/centrais-de-conteudo/i-pnpg-pdf>>.
- CAPES. **III Plano Nacional de Pós-Graduação (1986–1989)**. Brasília, 1986. Acesso em: 03 fev. 2026. Disponível em: <<https://www.gov.br/capes/pt-br/centrais-de-conteudo/iii-pnpg-pdf>>.
- CAPES. **V Plano Nacional de Pós-Graduação (2005–2010)**. Brasília, 2005. Acesso em: 03 fev. 2026. Disponível em: <<https://www.gov.br/capes/pt-br/centrais-de-conteudo/v-pnpg-pdf>>.
- CAPES. **Plano Nacional de Pós-Graduação 2011–2020**. Brasília, 2010. Acesso em: 03 fev. 2026. Disponível em: <<https://www.gov.br/capes/pt-br/centrais-de-conteudo/pnpg-2011-2020>>.
- CAPES. **Plano Nacional de Pós-Graduação (PNPG)**. 2022. Acesso em: 8 mar. 2026. Disponível em: <<https://www.gov.br/capes/pt-br/acesso-a-informacao/acoes-e-programas/plano-nacional-de-pos-graduacao-pnpg>>.

CAPES. **Plano Nacional de Pós-Graduação 2025–2029**. Brasília, 2025. Versão final. Acesso em: 03 fev. 2026. Disponível em: <[https://www.gov.br/capes/pt-br/centrais-de-conteudo/documentos/14072025\\_PNPG\\_20252029\\_FINALV3.pdf](https://www.gov.br/capes/pt-br/centrais-de-conteudo/documentos/14072025_PNPG_20252029_FINALV3.pdf)>.

CAPES. Plano Nacional de Pós-Graduação 2025-2029. **CAPES**, 2025.

CAPES, B. **Parecer CESU nº 977, de 3 de dezembro de 1965**. 1965. Acesso em: 19 jan. 2026. Disponível em: <<https://www.gov.br/capes/pt-br/centrais-de-conteudo/parecer-cesu-977-1965-pdf/view>>.

CARMO Êrica; GASPARINI, I.; OLIVEIRA, E. Identificação de trajetórias de aprendizagem em um curso de graduação e sua relação com a evasão escolar. In: **XXXIII Simpósio Brasileiro de Informática na Educação**. Porto Alegre, RS, Brasil: SBC, 2022. p. 323–333. ISSN 0000-0000. Disponível em: <<https://sol.sbc.org.br/index.php/sbie/article/view/22419>>.

CARNEIRO, M. G.; DUTRA, B. L.; PAIVA, J. G. S.; GABRIEL, P. H. R.; ARAÚJO, R. D. Educational data mining to support identification and prevention of academic retention and dropout: a case study in introductory programming. **Revista Brasileira de Informática na Educação**, v. 30, p. 379–395, set. 2022. Disponível em: <<https://journals-sol.sbc.org.br/index.php/rbie/article/view/2518>>.

CARVALHO, C.; MATTOS, J.; AGUIAR, M. Avaliação da interpretabilidade de modelos por meio da clusterização de explicações no contexto da predição de evasão no ensino superior. In: **XXXIV Simpósio Brasileiro de Informática na Educação**. Porto Alegre, RS, Brasil: SBC, 2023. p. 1191–1201. ISSN 0000-0000. Disponível em: <<https://sol.sbc.org.br/index.php/sbie/article/view/26746>>.

CECHINEL, C.; CAMARGO, S. da S. Mineração de dados educacionais: avaliação e interpretação de modelos de classificação. In: JAQUES, P. A.; SIQUEIRA, S.; BITTENCOURT, I.; PIMENTEL, M. (Ed.). **Metodologia de Pesquisa Científica em Informática na Educação: Abordagem Quantitativa**. Porto Alegre: Sociedade Brasileira de Computação, 2020. cap. 12.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. Smote: Synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research**, AI Access Foundation, v. 16, p. 321–357, jun. 2002. ISSN 1076-9757. Disponível em: <<http://dx.doi.org/10.1613/jair.953>>.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: Association for Computing Machinery, 2016. (KDD '16), p. 785–794. ISBN 9781450342322. Disponível em: <<https://doi.org/10.1145/2939672.2939785>>.

COLPO, M.; PRIMO, T.; AGUIAR, M. Predição da evasão estudantil: uma análise comparativa de diferentes representações de treino na aprendizagem de modelos genéricos. In: **XXXII Simpósio Brasileiro de Informática na Educação**. Porto Alegre, RS, Brasil: SBC, 2021. p. 873–884. ISSN 0000-0000. Disponível em: <<https://sol.sbc.org.br/index.php/sbie/article/view/18114>>.

COLPO, M.; PRIMO, T.; PERNAS, A.; CECHINEL, C. Mineração de dados educacionais na previsão de evasão: uma rsl sob a perspectiva do congresso brasileiro de informática na educação. In: **XXXI Simpósio Brasileiro de Informática na Educação**. Porto Alegre, RS, Brasil: SBC, 2020. p. 1102–1111. ISSN 0000-0000. Disponível em: <<https://sol.sbc.org.br/index.php/sbie/article/view/12866>>.

CORREIA, R.; MENDONÇA, H.; SILVA, C.; TOLEDO, D. Análise dos principais fatores que influenciam a evasão no ensino superior utilizando técnicas de mineração de dados educacionais. In: **XXXII Workshop sobre Educação em Computação**. Porto Alegre, RS, Brasil: SBC, 2024. p. 830–841. ISSN 2595-6175. Disponível em: <<https://sol.sbc.org.br/index.php/wei/article/view/29680>>.

CUNHA, L. A. Desenvolvimento desigual e combinado no ensino superior: estado e mercado. **Educação Amp; Sociedade**, v. 25, p. 795–817, 2004.

DE, N. S. L. M.; GOMES, J. B. F.; HOLANDA, M.; KOIKE, C. C.; COSTA, M. T. L. Study on computer science undergraduate students dropout at the university of brasilia. In: **2023 IEEE Frontiers in Education Conference (FIE)**. Piscataway, NJ: IEEE, 2023. p. 1–7.

DIAS, J. C.; SILVA, T. L. D.; JULIATTO, M. A.; PAIXÃO, A. N. D.; PRATA, D. N. School dropout in the federal network education of brazil: Is it an inherent individual attribute or it lies on setting conditions? In: **2023 International Symposium on Computers in Education (SIIE)**. Piscataway, NJ: IEEE, 2023. p. 1–10. ISBN 979-8-3503-2610-3.

FACELI, K.; LORENA, A. C.; GAMA, J.; ALMEIDA, T. A. d.; CARVALHO, A. C. P. d. L. F. d. **Inteligência artificial: uma abordagem de aprendizado de máquina**. Rio de Janeiro: LTC, 2021.

FALCÃO, A.; VILLWOCK, R.; MILOCA, S. Análise de dados pré-universidade para prever a evasão de alunos ingressantes em uma instituição de ensino superior. In: **XXXIV Simpósio Brasileiro de Informática na Educação**. Porto Alegre, RS, Brasil: SBC, 2023. p. 1293–1304. ISSN 0000-0000. Disponível em: <<https://sol.sbc.org.br/index.php/sbie/article/view/26755>>.

FAYYAD, U.; PIATETSKY-SHAPIO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v. 17, n. 3, p. 37, Mar. 1996. Disponível em: <<https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1230>>.

FERNÁNDEZ, A.; GARCÍA, S.; GALAR, M.; PRATI, R. C.; KRAWCZYK, B.; HERRERA, F. **Learning from Imbalanced Data Sets**. Cham: Springer, 2018. ISBN 978-3-319-98073-7.

FILHO, F.; VINUTO, T.; LEAL, B. Análise de classificadores para predição de evasão dos campi de uma instituição de ensino federal. In: **XXXI Simpósio Brasileiro de Informática na Educação**. Porto Alegre, RS, Brasil: SBC, 2020. p. 1132–1141. ISSN 0000-0000. Disponível em: <<https://sol.sbc.org.br/index.php/sbie/article/view/12869>>.

FREIRE, J.; LANDIM, F.; MORAES, L.; DELGADO, C.; PEDREIRA, C. Modelo para previsão precoce de abandono de uma disciplina de introdução à programação. In: **XXXII Workshop sobre Educação em Computação**. Porto Alegre, RS, Brasil: SBC, 2024. p. 635–645. ISSN 2595-6175. Disponível em: <<https://sol.sbc.org.br/index.php/wei/article/view/29663>>.

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. **Annals of Statistics**, v. 29, n. 5, p. 1189–1232, 2001.

GARCIA, L. L. da S.; LARA, D.; GOMES, R.; CAZELLA, S. Mineração de dados educacionais na predição do desempenho acadêmico: um prognóstico a partir do percurso curricular realizado. In: **XXXIII Simpósio Brasileiro de Informática na Educação**. Porto Alegre, RS, Brasil: SBC, 2022. p. 1077–1086. ISSN 0000-0000. Disponível em: <<https://sol.sbc.org.br/index.php/sbie/article/view/22483>>.

HAN, H.; WANG, W.-Y.; MAO, B.-H. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In: **Proceedings of the 2005 International Conference on Advances in Intelligent Computing - Volume Part I**. Berlin, Heidelberg: Springer-Verlag, 2005. (ICIC'05), p. 878–887. ISBN 3540282262. Disponível em: <[https://doi.org/10.1007/11538059\\_91](https://doi.org/10.1007/11538059_91)>.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. 3rd. ed. Waltham, MA: Morgan Kaufmann, 2012. ISBN 978-0123814791.

HASTIE, T.; FRIEDMAN, J.; TIBSHIRANI, R. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. Springer, 2006. (Springer series in statistics). Disponível em: <<https://books.google.com.br/books?id=EscrvgAACAAJ>>.

HE, H.; GARCIA, E. A. Learning from imbalanced data. **IEEE Transactions on Knowledge and Data Engineering**, v. 21, n. 9, p. 1263–1284, 2009.

IFCE. Instituto Federal de Educação, Ciência e Tecnologia do Ceará. **Plano de Dados Abertos (PDA): vigência maio 2024–maio 2026**. Fortaleza: IFCE, 2024. <<https://ifce.edu.br/dados-abertos-arquivos/pda-24-26.pdf>>. Aprovado pela Portaria n.º 6543/GABR/REITORIA, de 31 de agosto de 2023 (e subsequentes alterações) e elaborado por comissão instituída em novembro de 2023.

IFCE. Instituto Federal de Educação, Ciência e Tecnologia do Ceará. Pró-reitoria de Ensino - PROEN. **Plano estratégico para permanência e êxito dos estudantes do IFCE**. Fortaleza: IFCE, 2017. <<https://ifce.edu.br/proen/ensino/plano-de-permanencia-e-exito.pdf>>.

JESUS, H. O. de; RODRIGUEZ, L. C.; JUNIOR, A. d. O. C. Predição de evasão escolar na licenciatura em computação. **Revista Brasileira de Informática na Educação**, v. 29, p. 255–272, mar. 2021. Disponível em: <<https://journals-sol.sbc.org.br/index.php/rbie/article/view/2997>>.

JESUS, J. A. d.; GUSMÃO, R. P. d. Investigação da evasão estudantil por meio da mineração de dados e aprendizagem de máquina: Um mapeamento sistemático. **Revista Brasileira de Informática na Educação**, v. 32, p. 807–841, mar. 2024. Disponível em: <<https://journals-sol.sbc.org.br/index.php/rbie/article/view/3466>>.

JOLLIFFE, I. **Principal Component Analysis**. New York, NY: Springer-Verlag, 1986. (Springer series in statistics). ISBN 9780387962696. Disponível em: <<https://books.google.com.br/books?id=cN5UAAAAYAAJ>>.

JOLLIFFE, I. T. **Principal Component Analysis**. 2. ed. New York, NY: Springer, 2002. ISBN 9780387954424.

KANTORSKI, G.; MARTINS, R.; BALEJO, A.; FRICK, M. Mineração de dados educacionais para predição da evasão em cursos de graduação presenciais no ensino superior. In: **XXXIV Simpósio Brasileiro de Informática na Educação**. Porto Alegre, RS, Brasil: SBC, 2023. p. 1133–1142. ISSN 0000-0000. Disponível em: <<https://sol.sbc.org.br/index.php/sbie/article/view/26741>>.

KOHAVI, R.; JOHN, G. H. Wrappers for feature subset selection. **Artificial Intelligence**, v. 97, n. 1, p. 273–324, 1997. ISSN 0004-3702. Relevance. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S000437029700043X>>.

KRÜGER, J. G. C.; BRITTO, A. de S.; BARDDAL, J. P. An explainable machine learning approach for student dropout prediction. **Expert Systems with Applications**, v. 233, p. 120933, 2023. ISSN 0957-4174. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0957417423014355>>.

KUHN, M.; JOHNSON, K. **Applied Predictive Modeling**. Springer New York, 2013. (Springer-Link : Bücher). ISBN 9781461468493. Disponível em: <<https://books.google.com.br/books?id=xYRDAAAQBAJ>>.

LEITE, D.; FILHO, E.; OLIVEIRA, J. F. L.; CARNEIRO, R. E.; MACIEL, A. Early detection of students at risk of failure from a small dataset. In: **2021 International Conference on Advanced Learning Technologies (ICALT)**. Piscataway, NJ: IEEE, 2021. p. 42–46.

LIMA, L. M. D.; KROHLING, R. A. Discovering an aid policy to minimize student evasion using offline reinforcement learning. In: **2021 International Joint Conference on Neural Networks (IJCNN)**. IEEE, 2021. p. 1–8. Disponível em: <<http://dx.doi.org/10.1109/IJCNN52387.2021.9534159>>.

LINHARES, J.; VERAS, E.; OLIVEIRA, R.; GOMES, L.; SILVEIRA, R.; OLIVEIRA, C. de. Visualização de dados como ferramenta de apoio ao estudo analítico dos indicadores de permanência e desempenho acadêmico. In: **Anais da XIII Escola Regional de Computação do Ceará, Maranhão e Piauí**. Porto Alegre, RS, Brasil: SBC, 2025. p. 21–30. ISSN 0000-0000. Disponível em: <<https://sol.sbc.org.br/index.php/ercemapi/article/view/41240>>.

LITTLE, R.; RUBIN, D. **Statistical Analysis with Missing Data**. Wiley, 2019. (Wiley Series in Probability and Statistics). ISBN 9780470526798. Disponível em: <<https://books.google.com.br/books?id=BemMDwAAQBAJ>>.

MARQUES, L.; MARQUES, B.; ROCHA, R.; SILVA, L. e; CASTRO, A. de; QUEIROZ, P. G. Evasão acadêmica e suas causas em cursos de bacharelado em ciência da computação: Um estudo de caso na ufersa. In: **XXXI Simpósio Brasileiro de Informática na Educação**. Porto Alegre, RS, Brasil: SBC, 2020. p. 1042–1051. ISSN 0000-0000. Disponível em: <<https://sol.sbc.org.br/index.php/sbie/article/view/12860>>.

MEC. **Plataforma Nilo Peçanha**. 2026. Acesso em: 14 jan. 2026. Disponível em: <<https://www.gov.br/mec/pt-br/pnp>>.

MITCHELL, T. **Machine Learning**. McGraw-Hill Education, 1997. (McGraw-Hill international editions - computer science series). ISBN 9780070428072. Disponível em: <<https://books.google.com.br/books?id=xOGAngEACAAJ>>.

MOLNAR, C. **Interpretable Machine Learning**. Morrisville, NC: Lulu, 2020.

MORAIS, F.; MELO, A.; MOUTINHO, M.; FAGUNDES, R. Modelos de regressão aplicados na previsão da evasão escolar do ensino básico: uma revisão sistemática da literatura. In: **XXXII Simpósio Brasileiro de Informática na Educação**. Porto Alegre, RS, Brasil: SBC, 2021. p. 168–178. ISSN 0000-0000. Disponível em: <<https://sol.sbc.org.br/index.php/sbie/article/view/18053>>.

NASCIMENTO, P.; JUNIOR, A. S.; SCHULZ, C.; SANTOS, M.; MACIEL, A.; RODRIGUES, R.; NASCIMENTO, R.; ALENCAR, F. Análise dos impactos da gestão do tempo no desempenho acadêmico através da mineração de dados educacionais. In: **XXXII Simpósio Brasileiro de Informática na Educação**. Porto Alegre, RS, Brasil: SBC, 2021. p. 783–791. ISSN 0000-0000. Disponível em: <<https://sol.sbc.org.br/index.php/sbie/article/view/18106>>.

NOETZOLD, E.; PERTILE, S. de L. Análise e predição de evasão dos alunos de um curso de graduação em sistemas de informação por meio da mineração de dados educacionais. **RENOTE**, v. 19, n. 1, p. 351–360, jul. 2021. Disponível em: <<https://seer.ufrgs.br/index.php/renote/article/view/118525>>.

NÓBREGA, B. S. d.; MAIA, J. d. S.; FILHO, M. A. S.; JÚNIOR, E. E. A.; ALVES, M. B. Sistema para identificação e monitoramento de estudantes em risco de evasão: System for identifying and monitoring students at risk of circumventin. **Brazilian Journal of Development**, v. 8, n. 11, p. 74311–74323, Nov. 2022. Disponível em: <<https://ojs.brazilianjournals.com.br/ojs/index.php/BRJD/article/view/54518>>.

OLIVEIRA, J. L.; AMBRÓSIO, A. P.; SILVA, U.; BRANCHER, J.; FRANCO, J. J. Undergraduate students' effectiveness in an institution with high dropout index. In: **2020 IEEE Frontiers in Education Conference (FIE)**. Uppsala, Sweden: IEEE, 2020. p. 1–7.

OLIVEIRA, R.; MEDEIROS, F.; ALVES, K. Predição de evasão por meio de um instrumento sistemático de avaliação institucional. In: **II Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil**. Porto Alegre, RS, Brasil: SBC, 2023. p. 118–127. ISSN 0000-0000. Disponível em: <<https://sol.sbc.org.br/index.php/wapla/article/view/26135>>.

OLIVEIRA, R. d. S.; MEDEIROS, F. P. A. d. Modelo de predição de evasão escolar com base em dados de autoavaliação de cursos de graduação. **Revista Brasileira de Informática na Educação**, v. 32, p. 1–21, jan. 2024. Disponível em: <<https://journals-sol.sbc.org.br/index.php/rbie/article/view/3542>>.

PPGCC. Instituto Federal de Educação, Ciência e Tecnologia do Ceará. Programa de Pós-Graduação em Ciência da Computação. **Regimento PPGCC – IFCE v0.3**. 2017. <<https://ppgcc.ifce.edu.br/?wpdmp=regimento-ppgcc-ifce-v0-2>>. Criado em 08 jul. 2015; última atualização em 24 mar. 2017.

QUINLAN, J. **C4.5: Programs for Machine Learning**. Elsevier Science, 1993. (Morgan Kaufmann series in machine learning). ISBN 9781558602380. Disponível em: <<https://books.google.com.br/books?id=HExnepjbYroC>>.

RABELO, A.; ZÁRATE, L. A model for predicting dropout of higher education students. **Data Science and Management**, v. 8, 07 2024.

RAUFI, B.; LONGO, L. Comparing anova and powershap feature selection methods via shapley additive explanations of models of mental workload built with the theta and alpha eeg band ratios. **BioMedInformatics**, v. 4, n. 1, p. 853–876, 2024. ISSN 2673-7426. Disponível em: <<https://www.mdpi.com/2673-7426/4/1/48>>.

RODRIGUES, H.; MORAES, L.; SANTIAGO, E.; CAMPOS, J.; JÚNIOR, E. G.; WANDERLEY, G.; GARCIA, A.; MELLO, C.; ALVARES, R.; SANTOS, R. Predicting student dropout on the information systems undergraduate program of unirio using decision trees. In: **XXXII Workshop sobre Educação em Computação**. Porto Alegre, RS, Brasil: SBC, 2024. p. 588–598. ISSN 2595-6175. Disponível em: <<https://sol.sbc.org.br/index.php/wei/article/view/29659>>.

ROMERO, C.; VENTURA, S. Educational data mining: A review of the state of the art. **IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)**, v. 40, n. 6, p. 601–618, 2010.

SANTOS, C. H.; MARTINS, S.; PLASTINO, A. É possível prever evasão com base apenas no desempenho acadêmico? In: **XXXII Simpósio Brasileiro de Informática na Educação**. Porto Alegre, RS, Brasil: SBC, 2021. p. 792–802. ISSN 0000-0000. Disponível em: <<https://sol.sbc.org.br/index.php/sbie/article/view/18107>>.

SANTOS, G.; SOUZA, A.; MANTOVANI, R.; CRUZ, R.; CORDEIRO, T.; SOUZA, F. An exploratory analysis on gender-related dropout students in distance learning higher education using machine learning. In: **Proceedings of the 20th Brazilian Symposium on Information Systems**. New York, NY, USA: Association for Computing Machinery, 2024. (SBSI '24). ISBN 9798400709968. Disponível em: <<https://doi.org/10.1145/3658271.3658323>>.

SANTOS, J.; SOUSA, J. D.; MELLO, R.; CRISTINO, C.; ALVES, G. Um modelo para análise do impacto da retenção e evasão no ensino superior utilizando cadeias de markov absorventes. In: **XXXII Simpósio Brasileiro de Informática na Educação**. Porto Alegre, RS, Brasil: SBC, 2021. p. 813–823. ISSN 0000-0000. Disponível em: <<https://sol.sbc.org.br/index.php/sbie/article/view/18109>>.

SANTOS, J. K. Magalhães dos; ROCHA, H. O. da; OKAMURA, E. M. R.; DIAS, V. A. A.; MELO, L. H. Pessoa de; VIANA, G. B. O.; RODRIGUES, V. C.; SILVA, D. A. da. A review of ia use in education analysis. In: **2023 Workshop on Communication Networks and Power Systems (WCNPS)**. Brasilia, Brazil: IEEE, 2023. p. 1–7.

SANTOS, V.; SARAIVA, D.; OLIVEIRA, C. Uma análise de trabalhos de mineração de dados educacionais no contexto da evasão escolar. In: **Anais do XXXII Simpósio Brasileiro de Informática na Educação**. Porto Alegre, RS, Brasil: SBC, 2021. p. 1196–1210. ISSN 0000-0000. Disponível em: <<https://sol.sbc.org.br/index.php/sbie/article/view/18142>>.

SARAIVA, D.; PEREIRA, S.; BRAGA, R.; OLIVEIRA, C. Análise de agrupamentos para caracterização de indicadores de evasão. In: **Anais do XXIX Workshop sobre Educação em Computação**. Porto Alegre, RS, Brasil: SBC, 2021. p. 238–247. ISSN 2595-6175. Disponível em: <<https://sol.sbc.org.br/index.php/wei/article/view/15915>>.

SARAIVA, D.; PEREIRA, S.; BRAGA, R.; OLIVEIRA, C. Análise de agrupamentos para caracterização de indicadores de evasão. In: **XXIX Workshop sobre Educação em Computação**. Porto Alegre, RS, Brasil: SBC, 2021. p. 238–247. ISSN 2595-6175. Disponível em: <<https://sol.sbc.org.br/index.php/wei/article/view/15915>>.

SCHOEFFEL, P.; RAMOS, V.; WAZLAWICK, R. A method to predict at-risk students in introductory computing courses based on motivation. In: **Anais Estendidos do IX Congresso Brasileiro de Informática na Educação**. Porto Alegre, RS, Brasil: SBC, 2020. p. 41–41. ISSN 0000-0000. Disponível em: <[https://sol.sbc.org.br/index.php/cbie\\_estendido/article/view/13023](https://sol.sbc.org.br/index.php/cbie_estendido/article/view/13023)>.

SILVA, J.; ROMAN, N. Predicting dropout in higher education: a systematic review. In: **Anais do XXXII Simpósio Brasileiro de Informática na Educação**. Porto Alegre, RS, Brasil: SBC, 2021. p. 1107–1117. ISSN 0000-0000. Disponível em: <<https://sol.sbc.org.br/index.php/sbie/article/view/18134>>.

SILVA, J. C. L. Definição de modelos de aprendizado de máquina para predição de evasão de alunos do curso técnico. **Refas - Revista Fatec Zona Sul**, v. 9, n. 3, p. 1–12, fev. 2023. Disponível em: <<https://www.revistarefas.com.br/RevFATECZS/article/view/438>>.

SILVEIRA, R. F.; HOLANDA, M.; RAMOS, G. N.; VICTORINO, M.; SILVA, D. D. Analysis of student performance and social-economic data in introductory computer science courses at the university of brasília. In: **2022 IEEE Frontiers in Education Conference (FIE)**. Uppsala, Sweden: IEEE, 2022. p. 1–8.

SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. **Information Processing Management**, v. 45, n. 4, p. 427–437, 2009. ISSN 0306-4573. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0306457309000259>>.

SONNENSTRAHL, T. S.; BERNARDI, G.; PERTILE, S. Análise de interações do ambiente virtual de aprendizagem para predição de evasão em cursos no ensino a distância. **EaD em Foco**, v. 11, n. 1, jul. 2021. Disponível em: <<https://eademfoco.cecierj.edu.br/index.php/Revista/article/view/1463>>.

SOUSA, R.; FACHINI-GOMES, J.; HOLANDA, M.; LEÃO, M. Um estudo da evasão no curso de licenciatura em computação da universidade de brasília. In: **XXXII Workshop sobre Educação em Computação**. Porto Alegre, RS, Brasil: SBC, 2024. p. 715–725. ISSN 2595-6175. Disponível em: <<https://sol.sbc.org.br/index.php/wei/article/view/29670>>.

SOUZA, A. L.; BRAGA, A. Uma análise dos algoritmos de classificação com base na evasão dos estudantes dos cursos técnicos integrados ao ensino médio do campus ceres do if goiano. In: **XXXII Simpósio Brasileiro de Informática na Educação**. Porto Alegre, RS, Brasil: SBC, 2021. p. 1276–1285. ISSN 0000-0000. Disponível em: <<https://sol.sbc.org.br/index.php/sbie/article/view/18149>>.

SOUZA, J.; KOMATI, K.; ANDRADE, J. Análise de sobrevivência: um estudo de caso em um curso de sistemas de informação. In: **XXX Workshop sobre Educação em Computação**. Porto Alegre, RS, Brasil: SBC, 2022. p. 392–403. ISSN 2595-6175. Disponível em: <<https://sol.sbc.org.br/index.php/wei/article/view/20847>>.

SOUZA, V. F. d.; SANTOS, T. C. B. d. Processo de mineração de dados educacionais aplicado na previsão do desempenho de alunos: Uma comparação entre as técnicas de aprendizagem de máquina e aprendizagem profunda. **Revista Brasileira de Informática na Educação**, v. 29, p. 519–546, jun. 2021. Disponível em: <<https://journals-sol.sbc.org.br/index.php/rbie/article/view/2975>>.

SUTTON, R.; BARTO, A. **Reinforcement Learning, second edition: An Introduction**. MIT Press, 2018. (Adaptive Computation and Machine Learning series). ISBN 9780262039246. Disponível em: <<https://books.google.com.br/books?id=sWV0DwAAQBAJ>>.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. 2nd. ed. New York, NY: Pearson, 2018. ISBN 978-0133128901.

TEODORO, L.; FERREIRA, A.; KAPPEL, M. Graduai – sistema com aprendizagem de máquina para avaliação de risco de evasão. In: **Anais Estendidos do XII Congresso Brasileiro de Informática na Educação**. Porto Alegre, RS, Brasil: SBC, 2023. p. 84–95. ISSN 0000-0000. Disponível em: <[https://sol.sbc.org.br/index.php/cbie\\_estendido/article/view/27485](https://sol.sbc.org.br/index.php/cbie_estendido/article/view/27485)>.

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, p. 838–863, nov. 2020. Disponível em: <<https://journals-sol.sbc.org.br/index.php/rbie/article/view/3691>>.

TINTO, V. **Completing College: Rethinking Institutional Action**. University of Chicago Press, 2012. ISBN 9780226804545. Disponível em: <[https://books.google.com.br/books?id=r9\\_kMkIPzikC](https://books.google.com.br/books?id=r9_kMkIPzikC)>.

VELHO, L. Evasão na pós brasileira: Uma crise em formação? **Jornal da Universidade Estadual de Campinas**, 2005.

VIANA, F.; SANTANA, A.; RABÊLO, R. Avaliação de classificadores para predição de evasão no ensino superior utilizando janela semestral. In: **XXXIII Simpósio Brasileiro de Informática na Educação**. Porto Alegre, RS, Brasil: SBC, 2022. p. 908–919. ISSN 0000-0000. Disponível em: <<https://sol.sbc.org.br/index.php/sbie/article/view/22469>>.

VILLAR, A.; ANDRADE, C. R. V. de. Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study. **Discover Artificial Intelligence**, v. 4, n. 1, p. 2, Jan 2024.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining: Practical Machine Learning Tools and Techniques**. San Francisco, CA: Morgan Kaufmann, 2016.

**APÊNDICE A – ATRIBUTOS DERIVADOS**

Tabela 12 – Atributos derivados.

| <b>Nº</b> | <b>Atributo</b>          |
|-----------|--------------------------|
| 1         | Idade                    |
| 2         | Anos_Apos_Graduacao      |
| 3         | n_matriculas_pessoa      |
| 4         | primeira_matricula       |
| 5         | reingresso               |
| 6         | n_matriculas_maior_que_1 |

**APÊNDICE B – ATRIBUTOS ELIMINADOS DA FONTE DE DADOS DO PPGCC POR CONTEREM TODAS AS LINHAS COM VALORES INTEGRALMENTE AUSENTES**

Tabela 13 – Atributos eliminados da fonte de dados do PPGCC por conterem todas as linhas com valores integralmente ausentes.

| Nº | Atributo               | Nº | Atributo                    |
|----|------------------------|----|-----------------------------|
| 1  | Cod_Responsavel        | 15 | __Tratamento                |
| 2  | Renda_Per_Capita_Inep  | 16 | Dt_Limite_Interesse_Emprego |
| 3  | Necessidade_Fisica     | 17 | Id_Grupo_Etnico             |
| 4  | Necessidade_Visual     | 18 | Desc_Grupo_Etnico           |
| 5  | Necessidade_Auditiva   | 19 | Documento_Estrangeiro       |
| 6  | Necessidade_Mental     | 20 | Numero_Cns                  |
| 7  | Necessidade_Multipla   | 21 | DESC_EAD_POLO               |
| 8  | Outras_Necessidades    | 22 | DT_COLACAO_GRAU             |
| 9  | Superdotado            | 23 | BOLSA_FAMILIA               |
| 10 | Condutas_Tipicas       | 24 | BOLSA_ESCOLA                |
| 11 | Sindrome_Down          | 25 | Desc_Renda_Per_Capita_INEP  |
| 12 | necessidades_especiais | 26 | clNecessidades              |
| 13 | Cod_Distrito           | 27 | clApoios_Sociais            |
| 14 | __Observacoes          |    |                             |

**APÊNDICE C – ATRIBUTOS ELIMINADOS DA FONTE DE DADOS DO PPGCC  
POR CONTEREM O MESMO VALOR EM TODAS AS LINHAS**

Tabela 14 – Atributos eliminados da fonte de dados do PPGCC por conterem o mesmo valor em todas as linhas.

| <b>Nº</b> | <b>Atributo</b>                 | <b>Valor</b>                                |
|-----------|---------------------------------|---------------------------------------------|
| 1         | Curso_Estrutura_Curso           | 57                                          |
| 2         | Cod_curso                       | 3881                                        |
| 3         | Cod_instituicao                 | 1                                           |
| 4         | Desc_Instituicao                | Campus Fortaleza                            |
| 5         | Sigla_Curso                     | 1803                                        |
| 6         | desc_curso                      | Mestrado Acadêmico em Ciência da Computação |
| 7         | Emancipado                      | 0                                           |
| 8         | Regime_Internato                | 0                                           |
| 9         | Tipo_Concomitancia              | 0                                           |
| 10        | Ch_Eletiva_Prevista             | 0                                           |
| 11        | Ch_Eletiva_Cumprida             | 0                                           |
| 12        | Ch_Eletiva_Cursada              | 0                                           |
| 13        | Desc_Regime_Internato           | Não Internos                                |
| 14        | Nivel_Ensino                    | Pós-graduação                               |
| 15        | Desc_Modalidade_Curso           | Mestrado                                    |
| 16        | Qtd_Periodos                    | 1                                           |
| 17        | Carga_Hor_Complementar_Cumprida | 0                                           |
| 18        | Carga_Hor_Prevista              | 0                                           |
| 19        | F_Coeficiente_Rendimento        | 0                                           |
| 20        | CH_Obrigatoria_Cursada          | 0                                           |

**APÊNDICE D – ATRIBUTOS ELIMINADOS DA FONTE DE DADOS DO PPGCC  
POR BAIXO PERCENTUAL DE PREENCHIMENTO**

Tabela 15 – Atributos removidos da fonte de dados do PPGCC por baixo percentual de preenchimento.

| Nº | Atributo                            | % de Preenchimento |
|----|-------------------------------------|--------------------|
| 1  | Ano_Conclusao_1_Grau                | 14,51%             |
| 2  | Ano_Conclusao_Pos_Graduacao         | 0,39%              |
| 3  | Area_Procedencia_Escola_Origem      | 27,06%             |
| 4  | Cod_Escola_1_Grau                   | 17,65%             |
| 5  | Cod_Escola_Pos_Graduacao            | 0,78%              |
| 6  | Cod_Estado_Civil_Pais               | 37,65%             |
| 7  | Cod_Grau_Instrucao_Mae              | 45,49%             |
| 8  | Cod_Grau_Instrucao_Pai              | 44,71%             |
| 9  | Cod_turma                           | 18,04%             |
| 10 | Csm                                 | 37,65%             |
| 11 | Desc_Area_Procedencia_Escola_Origem | 27,06%             |
| 12 | Desc_Curso_2_Grau                   | 4,31%              |
| 13 | Desc_Curso_Pos_Graduacao            | 1,18%              |
| 14 | Desc_Escola_1_Grau                  | 17,65%             |
| 15 | Desc_Escola_Pos_Graduacao           | 0,78%              |
| 16 | Desc_Estado_Civil_Pais              | 37,65%             |
| 17 | Desc_Renda_Per_Capita               | 37,65%             |
| 18 | Desc_Renda_Per_Capita_PNP           | 30,59%             |
| 19 | Desc_Renda_Per_Capita_SIG           | 19,61%             |
| 20 | Desc_renda_familiar                 | 47,06%             |
| 21 | Grau_Parentesco_Responsavel         | 0,78%              |
| 22 | N_Filhos                            | 24,71%             |
| 23 | N_Membros_Familia                   | 9,41%              |
| 24 | Ocupacao_Responsavel                | 0,39%              |
| 25 | Percentual_Frequencia               | 30,98%             |
| 26 | Renda_Familiar                      | 47,06%             |
| 27 | Renda_Familiar_1                    | 47,06%             |
| 28 | Renda_Per_Capita                    | 37,65%             |

| Nº | Atributo             | % de Preenchimento |
|----|----------------------|--------------------|
| 29 | Renda_Per_Capita_PNP | 30,59%             |
| 30 | Renda_Per_Capita_Sig | 19,61%             |
| 31 | Reside               | 24,31%             |
| 32 | Tipo_Responsavel     | 0,39%              |
| 33 | Tipo_Sanguineo       | 18,43%             |
| 34 | clAtividades_Extras  | 0,78%              |
| 35 | cod_turma_Atual      | 18,04%             |

**APÊNDICE E – ATRIBUTOS ELIMINADOS DA FONTE DE DADOS DO PPGCC  
POR SEREM DUPLICADOS**

Tabela 16 – Grupos de atributos com indícios de duplicidade na base de dados do PPGCC.

| <b>Grupo</b> | <b>Descrição Semântica</b> | <b>Atributos Identificados</b>                      |
|--------------|----------------------------|-----------------------------------------------------|
| G1           | Identificação do estudante | Cod_Pessoa<br>Cod_Pessoa_1                          |
| G2           | Conclusão do Ensino Médio  | __Dt_Conclusao_Ensino_Medio<br>Ano_Conclusao_2_Grau |
| G3           | Tipo de escola de origem   | Tipo_escola_Origem<br>Tipo_Escola_Origem_1          |
| G4           | Sexo do estudante          | sexo<br>Sexo_1                                      |
| G5           | Código da cidade           | Cod_cidade<br>Cod_Cidade_1                          |
| G6           | Informações de matrícula   | dt_matricula<br>DT_Rematricula                      |
| G7           | Indicadores de reingresso  | reingresso<br>n_matriculas_maior_que_1              |

**APÊNDICE F – ATRIBUTOS ELIMINADOS DA FONTE DE DADOS DO PPGCC  
POR SEREM DIRETAMENTE RELACIONADOS AO ALVO**

Tabela 17 – Atributos eliminados da fonte de dados do PPGCC por serem diretamente relacionadas ao alvo.

| <b>Nº</b> | <b>Atributo</b>                |
|-----------|--------------------------------|
| 1         | Desc_Sit_Matricula             |
| 2         | Sit_Matricula                  |
| 3         | Situacao_Ultimo_Periodo_Letivo |
| 4         | DESC_SIT_MATRICULA_PERIODO     |
| 5         | __DIPLOMA_PASTA                |
| 6         | __DIPLOMA_N_REGISTRO           |
| 7         | __DIPLOMA_N_LIVRO_N_FOLHA      |
| 8         | N_Pasta                        |
| 9         | __Dt_Conclusao_Curso           |
| 10        | Ultimo_Evento_Matricula        |

**APÊNDICE G – ATRIBUTOS ELIMINADOS DA FONTE DE DADOS DO PPGCC  
POR SEREM REFERENTES A EVENTOS OCORRIDOS APÓS O INGRESSO DO  
ESTUDANTE NO CURSO**

Tabela 18 – Atributos eliminados da fonte de dados do PPGCC por serem referentes  
a eventos ocorridos após o ingresso do estudante no curso.

| <b>Grupo</b> | <b>Categoria Funcional</b>                   | <b>Variáveis Removidas</b>                                                                                                                                    |
|--------------|----------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
| R1           | Variáveis acadêmicas pós-ingresso            | Ultimo_Periodo_Letivo_Presente,<br>Ultima_Aula_Presente,<br>Ano_Let_Atual,<br>Periodo_Let_Atual,<br>Periodo_Atual,<br>Coeficiente_Rendimento                  |
| R2           | Variáveis de identificação codificada        | Cod_Pessoa,<br>cod_matricula,<br>Cod_Aluno,<br>Cod_Escola_2_Grau,<br>Cod_cidade,<br>Cod_Nacionalidade,<br>Cod_Naturalidade,<br>Cod_Turno,<br>Cod_Estado_Civil |
| R3           | Variáveis temporais relacionadas ao ingresso | ano_letivo_ini,<br>dt_matricula,<br>Dt_Cadastro,<br>Desc_Forma_Ingresso_Matricula,<br>clPeriodo_Let_ini                                                       |

## APÊNDICE H – ATRIBUTOS DA FONTE DE DADOS DO PPGCC APÓS LIMPEZA

Tabela 19 – Atributos e completude dos registros.

| Nº | Atributo                 | Total de Registros Não Nulos | % de Não Nulos | Valor                                          |
|----|--------------------------|------------------------------|----------------|------------------------------------------------|
| 1  | Matriz_Estrutura_Curso   | 255                          | 100,00%        | 20                                             |
| 2  | Periodo_letivo_ini       | 255                          | 100,00%        | 2                                              |
| 3  | Dt_Nascimento            | 255                          | 100,00%        | 1984-07-31<br>00:00:00                         |
| 4  | sexo                     | 255                          | 100,00%        | M                                              |
| 5  | Desc_Cor                 | 255                          | 100,00%        | Parda                                          |
| 6  | Desc_Turno               | 255                          | 100,00%        | Integral                                       |
| 7  | Desc_Turno_Ini           | 255                          | 100,00%        | Integral                                       |
| 8  | Idade                    | 255                          | 100,00%        | 30                                             |
| 9  | n_matriculas_pessoa      | 255                          | 100,00%        | 1                                              |
| 10 | primeira_matricula       | 255                          | 100,00%        | 1                                              |
| 11 | reingresso               | 255                          | 100,00%        | 0                                              |
| 12 | n_matriculas_maior_que_1 | 255                          | 100,00%        | 0                                              |
| 13 | DESC_NACIONALIDADE       | 243                          | 95,29%         | Brasileiro                                     |
| 14 | Desc_Naturalidade        | 238                          | 93,33%         | Fortaleza - CE                                 |
| 15 | Texto_cidade             | 237                          | 92,94%         | Fortaleza - CE                                 |
| 16 | Pai_Falecido             | 235                          | 92,16%         | 0.0                                            |
| 17 | Mae_Falecida             | 234                          | 91,76%         | 0.0                                            |
| 18 | Data_Titulo_Ele          | 233                          | 91,37%         | 19/02/2002                                     |
| 19 | Estado_Titulo_Ele        | 230                          | 90,20%         | CE                                             |
| 20 | Desc_Estado_Civil        | 227                          | 89,02%         | SOLTEIRO(A)                                    |
| 21 | Desc_Cota                | 225                          | 88,24%         | Não possui                                     |
| 22 | Cod_Grau_Instrucao       | 220                          | 86,27%         | GG                                             |
| 23 | Status                   | 210                          | 82,35%         | 1.0                                            |
| 24 | Tipo_escola_Origem       | 174                          | 68,24%         | P                                              |
| 25 | Desc_Tipo_Escola_Origem  | 174                          | 68,24%         | Privada                                        |
| 26 | Ano_Conclusao_Graduacao  | 165                          | 64,71%         | 2023.0                                         |
| 27 | Desc_Curso_Graduacao     | 165                          | 64,71%         | Bacharelado em<br>Engenharia de<br>Mecatrônica |

| Nº | Atributo                    | Total de Registros Não Nulos | % de Não Nulos |                                                                             |
|----|-----------------------------|------------------------------|----------------|-----------------------------------------------------------------------------|
| 28 | Anos_Apos_Graduacao         | 165                          | 64,71%         | 1.0                                                                         |
| 29 | Profissao                   | 165                          | 64,71%         | Bombeiro Militar                                                            |
| 30 | Desc_Escola_Graduacao       | 160                          | 62,75%         | INSTITUTO<br>FEDERAL DE<br>EDUCAÇÃO,<br>CIÊNCIA E<br>TECNOLOGIA<br>DO CEARÁ |
| 31 | Desc_Escola_2_Grau          | 155                          | 60,78%         | 7 DE SETEM-<br>BRO COLEGIO                                                  |
| 32 | __Dt_Conclusao_Ensino_Medio | 148                          | 58,04%         | 2001.0                                                                      |