



INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PIAUÍ
CAMPUS TERESINA ZONA SUL
CURSO LICENCIATURA EM COMPUTAÇÃO

Cláudio Emanuel Dantas Campos

Aprendizado de Máquina: Modelo de Predição de Evasão Acadêmica

TERESINA

2026

Cláudio Emanuel Dantas Campos

Aprendizado de Máquina: Modelo de Predição de Evasão Acadêmica

Trabalho de Conclusão de Curso (artigo científico) apresentado como exigência parcial para obtenção do diploma do Curso de Licenciatura em Computação do Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Campus Teresina Zona Sul.

Orientador: Prof. Msc. Kelson Carvalho Santos.

TERESINA

2026

Aprendizado de Máquina: Modelo de Predição de Evasão Acadêmica

Cláudio Emanuel Dantas Campos¹

Kelson Carvalho Santos²

RESUMO

A evasão no ensino superior é um desafio persistente que impacta estudantes e instituições. Este trabalho tem como objetivo desenvolver um modelo preditivo baseado em aprendizagem de máquina capaz de identificar padrões associados ao abandono acadêmico a partir de dados socioeconômicos. A pesquisa caracteriza-se como um estudo quantitativo, utilizando dados históricos de alunos matriculados entre os anos de 2012 e 2024 em um curso de licenciatura. A metodologia envolveu etapas rigorosas de pré-processamento, abrangendo a imputação de valores faltantes, normalização, seleção de atributos por filtragem estatística e balanceamento de classes por meio de sobre amostragem sintética. Foram treinados e avaliados quatro algoritmos de classificação supervisionada. Os experimentos demonstraram que a preservação de dados associada à eliminação de variáveis irrelevantes potencializou significativamente o aprendizado dos modelos. O algoritmo baseado em florestas aleatórias obteve o melhor desempenho e equilíbrio geral, apresentando forte capacidade de generalização e consistência na identificação de discentes em situação de risco. Conclui-se que a abordagem metodológica adotada demonstrou-se tecnicamente viável, ainda que o desempenho preditivo obtido — limitado pelo reduzido volume de dados disponíveis — indiquem que os resultados atuais devam ser interpretados como uma prova de conceito, e não como um modelo pronto para aplicação direta na gestão institucional, consolidando-se, nesse estágio, como um ponto de partida para o desenvolvimento futuro de ferramentas de apoio à tomada de decisão voltadas à permanência estudantil.

Palavras-chave: aprendizagem de máquina; evasão acadêmica; predição.

ABSTRACT

Dropout in higher education is a persistent challenge that impacts students and institutions. This work aims to apply data mining techniques to develop a predictive model based on machine learning capable of identifying patterns associated with academic dropout using socioeconomic variables. The research is characterized as a quantitative study, using historical information of students enrolled between the years 2012 and 2024 in a teaching degree program. The methodology involved rigorous preprocessing steps, encompassing the imputation of missing values, normalization, feature selection through statistical filtering, and class balancing using synthetic

¹ Graduando em Licenciatura em Computação. IFPI(CATZS).E-mail: ceuniversidade@gmail.com.

² Doutorando em Ciência da Computação na Faculdade de Computação – FACOM da Universidade Federal de Uberlândia - UFU. Instituição de ensino. E-mail: kelson@ifpi.edu.br.

oversampling. Four supervised classification algorithms were trained and evaluated. The experiments demonstrated that information preservation associated with the elimination of irrelevant variables significantly enhanced the learning of the models. The algorithm based on random forests obtained the best performance and overall balance, presenting strong capacity for prediction, generalization and consistency in identifying students at risk. It is concluded that the adopted methodological approach proved technically viable, although the predictive performance obtained — still limited by the reduced volume of available data — indicates that the current results should be interpreted as a proof of concept rather than a model ready for direct application in institutional management, thereby constituting an initial step toward future decision-support tools aimed at student retention.

Keywords: machine learning; academic dropout; prediction.

Data de aprovação: 24/07/2026

LISTA DE ILUSTRAÇÕES

Figura 1- Fórmula da Acurácia	16
Figura 2- Fórmula da Precisão	17
Figura 3- Fórmula do recall	18
Figura 4- Fórmula do f1_score	18
Figura 5- Processo de criação de um modelo de Aprendizagem de Máquina	23
Figura 6- Desempenho de Algoritmos	35

LISTA DE TABELAS

Tabela 1- Relação em porcentagem de evasão e conclusão das turmas de cada ano	8
Tabela 2- Comparativo de características dos algoritmos de aprendizado de máquina	15
Tabela 3- Síntese dos trabalhos relacionados à predição de evasão no ensino superior	21
Tabela 4- Tipo, natureza e valores possíveis dos atributos utilizados	24
Tabela 5- Distribuição das classes antes e depois do balanceamento (conjunto de treino)	30
Tabela 6- Melhor resultado individual por cenário experimental	34

LISTA DE ABREVIações E SIGLAS

IFPI – Instituto Federal do Piauí

CATZS – Campus Teresina Zona Sul

IES – Instituições de Ensino Superior

PCA – *Principal Component Analysis*

SMOTE – *Synthetic Minority Over-sampling Technique*

UFPB – Universidade Federal da Paraíba

KNN – *K-Nearest Neighbors*

XGBoost – *eXtreme Gradient Boosting*

VP – Verdadeiro Positivo

VN – Verdadeiro Negativo

FP – Falso Positivo

FN – Falso Negativo

IFMG – Instituto Federal de Minas Gerais

EaD – Ensino a Distância

IFRS – Instituto Federal do Rio Grande do Sul

E1, 2, 3, ... – Experimento 1, 2, 3 ...

INEP – Instituto Nacional de Estudos e Pesquisas

SUMÁRIO

1 Introdução	8
2 Referencial Teórico	9
2.1 Evasão no Ensino Superior Brasileiro	9
2.2 Mineração de Dados Educacionais e Aprendizagem de Máquina	10
2.3 Técnicas de pré-processamento	11
2.3.1 Seleção de Atributos	11
2.3.2 Redução de Dimensionalidade - PCA	12
2.3.3 Balanceamento de Dados - SMOTE	13
2.4 Algoritmos de Classificação	13
2.5 Métricas de Avaliação	15
2.6 Ferramentas Utilizadas	18
3 Trabalhos Relacionados	19
4 Metodologia	21
4.1 Conjunto de Dados	22
4.2 Pré-processamento dos Dados	23
4.3 Divisão dos Dados	23
4.4 Algoritmos Utilizados - Treinamento do modelo	24
4.4.1 Técnicas de Otimização do Modelo	25
4.4.2 Técnicas de Balanceamento dos Dados	25
4.5 Avaliação do Desempenho do Modelo	26
5 Experimentos e Discussão de Resultados	26
5.1 Visão Geral dos Experimentos	27
5.2 Impacto do Tratamento e Valores Faltantes	27
5.3 Impacto da Técnica de Normalização	27
5.4 Impacto da Técnica de Seleção ou Redução de Atributos	28
5.5 Desempenho por Algoritmo	28
5.6 Limitações e Implicações	29
6 Considerações finais	30
7 Referências	33

1 Introdução

A evasão no ensino superior constitui um dos principais desafios enfrentados pelas instituições de educação superior no Brasil e em diversos países. Esse fenômeno caracteriza-se pela interrupção ou abandono do curso antes da sua conclusão, gerando impactos significativos tanto para os estudantes quanto para as instituições e para a sociedade. Entre as consequências da evasão estão o desperdício de recursos públicos e privados investidos na formação acadêmica, a redução da qualificação profissional da população e o comprometimento dos indicadores de desempenho institucional.

Segundo Carvalho e Oliveira (2014, p. 109), "por todo o Brasil as universidades apresentam alto índice de evasão nos setores de licenciatura (48% não chegam a se formar, todo ano 19,6% desistem do curso)".

Acentuando a situação, é possível notar, na tabela 1, uma certa semelhança com a situação de anos anteriores do curso de Licenciatura em Computação do Instituto Federal do Piauí, Campus Teresina Zona Sul (IFPI-CATZS).

Tabela 1- Relação em porcentagem de evasão e conclusão das turmas de cada ano

Ano de Ingresso	Matriculados	Evadidos	Concluídos	Ano de Conclusão	Previsão de Conclusão	Evadido %	Concluído %
2012	42	27	7	2016	-	64,29%	16,67%
2013	40	25	10	2017	-	62,50%	25,00%
2014	43	29	7	2018	-	67,44%	16,28%
2015	39	31	5	2019	-	79,49%	12,82%
2016	35	26	7	2020	-	74,29%	20,00%
2017	45	37	7	2021	-	82,22%	15,56%
2018	41	20	11	2022	-	48,78%	26,83%
2019	48	32	1	2023	-	66,67%	2,08%
2020	42	31	1	2024	-	73,81%	2,38%
2021	33	19	Em andamento	-	2025	-	-
2022	35	14	Em andamento	-	2026	-	-
2023	47	18	Em andamento	-	2027	-	-
2024	45	17	Em andamento	-	2028	-	-

Fonte: Adaptado - Controle Acadêmico - IFPI CATZS (2024).

Como é possível notar na tabela 1, há sempre uma taxa mais elevada de alunos evadidos do curso do que os alunos que o concluem. Pensando nesses

dados preocupantes, vê-se a necessidade da criação de uma estratégia para a mitigação desse problema.

Diante dessa problemática, este trabalho tem como objetivo a aplicação de técnicas de Aprendizagem de Máquina sobre dados socioeconômicos dos estudantes. Buscando-se, assim, identificar padrões associados à evasão acadêmica, fornecendo subsídios para a adoção de estratégias de intervenção e permanência estudantil. Auxiliando, assim, a gestão institucional do IFPI Campus Teresina Zona Sul (CATZS), com foco no curso de Licenciatura em Computação.

Para alcançar esse objetivo, foi utilizada uma base de dados previamente tratada contendo informações referentes às turmas do curso de Licenciatura em Computação do IFPI Campus Teresina Zona Sul (CATZS). Em seguida, foram aplicadas técnicas de pré-processamento e transformação dos dados, tornando-os adequados para utilização pelos algoritmos de Aprendizagem de Máquina avaliados neste estudo. O desenvolvimento do modelo foi realizado no ambiente Jupyter Notebook, utilizando a linguagem de programação Python.

2 Referencial Teórico

A construção desta proposta fundamenta-se em estudos científicos que subsidiam tanto a compreensão do fenômeno da evasão no ensino superior quanto a justificativa para a escolha das técnicas e ferramentas empregadas. Dessa forma, torna-se necessário compreender o contexto da evasão no ensino superior brasileiro, bem como os principais conceitos e etapas envolvidos no desenvolvimento de modelos de Aprendizagem de Máquina.

2.1 Evasão no Ensino Superior Brasileiro

A evasão no ensino superior brasileiro constitui um dos principais desafios enfrentados pelas Instituições de Ensino Superior (IES), representando perdas tanto para os estudantes quanto para as instituições e para a sociedade. Segundo Prestes e Fialho (2018), as políticas educacionais voltadas ao enfrentamento da evasão no ensino superior ainda são recentes e incipientes no Brasil, apesar dos impactos

significativos decorrentes desse fenômeno, o que reforça a necessidade de estudos e ações voltados à sua compreensão e mitigação.

Nesse contexto, a ampliação do acesso à educação superior somente se traduz em benefícios efetivos para a sociedade quando acompanhada por políticas que garantam a permanência dos estudantes até a conclusão do curso. Conforme Silva Filho et al. (2019), a expansão das vagas precisa estar associada a estratégias institucionais que favoreçam a permanência estudantil, reduzindo os índices de evasão e aumentando a efetividade das políticas públicas de educação superior.

Dessa forma, a identificação precoce de estudantes com maior probabilidade de evasão torna-se fundamental para que as instituições possam desenvolver ações preventivas e direcionadas, contribuindo para a permanência acadêmica e para a melhoria dos indicadores institucionais.

2.2 Mineração de Dados Educacionais e Aprendizagem de Máquina

Considerando a complexidade do fenômeno da evasão, técnicas de Mineração de Dados Educacionais associadas à Aprendizagem de Máquina têm se mostrado ferramentas promissoras para a construção de modelos preditivos capazes de identificar estudantes em situação de risco.

Segundo Santos e Pires Fernandes (2024), a aplicação de técnicas de Aprendizagem de Máquina busca identificar relações entre atributos acadêmicos e socioeconômicos dos estudantes, possibilitando a construção de modelos capazes de antecipar situações de evasão acadêmica e fornecer informações estratégicas para a tomada de decisão por parte da gestão institucional.

Entretanto, a evasão escolar caracteriza-se como um fenômeno multifatorial, envolvendo aspectos financeiros, sociais, psicológicos e acadêmicos, cujas influências variam conforme o contexto educacional (APOENA, 2023). Dessa maneira, torna-se necessária a utilização de modelos capazes de identificar padrões complexos presentes nos dados.

Estudos realizados em instituições públicas brasileiras demonstram a viabilidade dessa abordagem. Teodoro et al. (2020), por exemplo, aplicaram cinco algoritmos de Aprendizagem de Máquina utilizando dados do INEP, obtendo aproximadamente 80% de acerto com o algoritmo *Random Forest*. Os autores identificaram que características como idade e participação em atividades extracurriculares figuram entre os fatores mais relevantes para a predição da evasão.

No contexto da Licenciatura em Computação, Jesus et al. (2021) destacam que grande parte dos estudantes apresenta dificuldades nas disciplinas introdutórias de programação em decorrência da ausência de conhecimentos prévios, da dificuldade de resolução de problemas e do raciocínio lógico-matemático. Esses fatores tornam o histórico acadêmico uma importante fonte de informação para modelos de predição de evasão.

2.3 Técnicas de pré-processamento

As técnicas de pré-processamento consistem em etapas de preparação dos dados que visam melhorar o desempenho dos modelos de Aprendizagem de Máquina. Com base na literatura especializada, identificou-se a necessidade da aplicação das técnicas selecionadas neste trabalho, uma vez que elas contribuem para a melhoria da qualidade dos dados utilizados no treinamento dos algoritmos.

2.3.1 Seleção de Atributos

A seleção de atributos é o processo de identificação e escolha de um subconjunto de variáveis relevantes de um conjunto de dados para uso em modelos de Aprendizagem de Máquina. Segundo Segatto (2023), a seleção de atributos é importante por diversas razões: pode ajudar a melhorar o desempenho do modelo, reduzir sua complexidade e torná-lo mais fácil de interpretar e implementar.

Michelin (2021) destaca que variáveis irrelevantes ou parcialmente relevantes podem afetar negativamente o desempenho do modelo, especialmente em algoritmos lineares. Nesse contexto, a seleção de atributos automática permite

identificar quais recursos contribuem mais para a variável alvo, reduzindo o tempo de treinamento e eliminando ruído dos dados.

Neste trabalho foi utilizada a técnica de filtragem estatística por meio da função *SelectKBest* da biblioteca *Scikit-learn*, que seleciona as K variáveis com maior pontuação em um teste estatístico. Conforme Vasconcellos (2019), o funcionamento do *SelectKBest* é direto: informa-se à função que se deseja selecionar apenas as K melhores features do conjunto de dados com base em um teste estatístico, sendo utilizado o teste Qui-quadrado (χ^2) por ser adequado a variáveis categóricas codificadas numericamente, como os presentes na base de dados deste trabalho.

2.3.2 Redução de Dimensionalidade - PCA

A redução de dimensionalidade é um processo que tem por objetivo diminuir o número de variáveis de um conjunto de dados visando a menor perda de informações possível (GERMANOS; SILVA, 2020). Dentre as técnicas disponíveis, a Análise de Componentes Principais — do inglês *Principal Component Analysis* (PCA) — é uma das mais amplamente utilizadas na literatura de Aprendizagem de Máquina.

O PCA é um método estatístico que transforma um conjunto de dados de alta dimensionalidade em um novo conjunto de dimensões ortogonais, chamadas de componentes principais, escolhidas de maneira a maximizar a variância dos dados ao longo de cada componente. Dessa forma, os primeiros componentes capturam a maior parte da variabilidade dos dados originais, permitindo uma representação mais compacta e eficiente.

Conforme apontado por Francheto (2024), o PCA é utilizado para reduzir a dimensionalidade dos dados preservando ao máximo a variabilidade presente nos mesmos, o que ajuda a simplificar a análise e a visualização, facilitando a interpretação e permitindo que padrões subjacentes sejam mais facilmente identificados. No entanto, os componentes principais não correspondem diretamente às variáveis originais, o que pode dificultar a interpretação dos resultados, além de o PCA pressupor que as relações entre as variáveis são lineares, o que pode limitar sua aplicação em determinados contextos.

2.3.3 Balanceamento de Dados - SMOTE

Em problemas de classificação com bases de dados desbalanceadas — onde uma classe possui significativamente mais exemplos do que outra — os modelos de Aprendizagem de Máquina tendem a favorecer a classe majoritária, prejudicando a identificação da classe minoritária. Para contornar esse problema, foi utilizada a técnica SMOTE (*Synthetic Minority Over-sampling Technique*), proposta originalmente por Chawla et al. (2002).

O SMOTE realiza o sobre amostramento da classe minoritária criando exemplos sintéticos ao longo dos segmentos de linha que unem amostras minoritárias e seus K vizinhos mais próximos no espaço de features, em vez de simplesmente duplicar exemplos existentes. Essa abordagem é relevante pois, conforme destacado por Chawla et al. (2002), a criação sintética de registros evitará o *overfitting* que ocorreria com a simples duplicação de exemplos da classe minoritária, tornando a região de decisão do modelo mais generalizada.

2.4 Algoritmos de Classificação

Para a construção do modelo preditivo proposto neste trabalho, foram selecionados quatro algoritmos de classificação que, além de amplamente utilizados na literatura, apresentam resultados otimistas dentro da proposta da criação de um modelo preditivo de evasão acadêmica, todos os algoritmos apresentados possuem suas características únicas, características essas que colaboraram para a sua seleção e estão sendo apresentadas na tabela 2.

A *Decision Tree* é um algoritmo que constrói uma representação em formato de fluxograma, realizando divisões sucessivas dos dados com base em regras aprendidas durante o treinamento. Sua principal vantagem está na interpretação, permitindo a visualização das regras utilizadas para a classificação de cada instância. Oliveira e Medeiros (2024) aplicaram o *Decision Tree* em um modelo de predição de evasão na UFPB, obtendo acurácia superior a 87% na identificação de alunos com alta probabilidade de abandono.

O *Random Forest* é um algoritmo do tipo *ensemble*³ que combina múltiplas árvores de decisão para produzir um modelo mais robusto. Segundo Kantorski et al. (2023), sua ideia central é construir múltiplas árvores de decisão durante o treinamento e, para a predição, cada árvore vota no resultado, sendo cada amostra extraída com reposição do conjunto de dados original por meio da técnica de *bootstrap aggregating*⁴.

O *K-Nearest Neighbors* (KNN) é um algoritmo que classifica uma instância com base nas K instâncias mais similares presentes na base de treino, atribuindo a classe mais frequente entre os vizinhos identificados. Hortêncio Filho et al. (2025) utilizaram o KNN junto a outros classificadores para predição de evasão em um curso de Tecnologia da Informação de um Instituto Federal, utilizando variáveis demográficas e acadêmicas similares às adotadas no presente trabalho.

O *eXtreme Gradient Boosting* (XGBoost) é um algoritmo do tipo *ensemble* fundamentado no conceito de aumento de gradiente (*gradient boosting*) aplicado em árvores de decisão. Segundo Chen e Guestrin (2016), sua arquitetura foi desenvolvida visando o máximo desempenho computacional e controle de *overfitting*, operando de maneira iterativa, onde cada nova árvore é construída para corrigir os erros residuais da árvore anterior. Em contextos acadêmicos, Primão (2022) aplicou o XGBoost para a predição de evasão no ensino superior em um estudo realizado no Instituto Federal de Santa Catarina (IFSC). A pesquisadora constatou que o XGBoost obteve capacidade de generalização e métricas de desempenho (como o F1-Score) superiores às dos modelos de *baseline* testados, consolidando-se como uma ferramenta altamente eficaz e escalável na identificação de discentes em risco de abandono.

³ *Ensemble* (ou comitê de máquinas) é uma abordagem em Aprendizagem de Máquina que combina as predições de múltiplos modelos individuais (chamados de aprendizes base ou fracos) para formar um modelo único mais robusto e preciso. Essa técnica busca compensar as fraquezas de um algoritmo isolado utilizando os pontos fortes dos demais, resultando em predições com maior estabilidade, melhor capacidade de generalização e menor margem de erro.

⁴ *Bootstrap Aggregating* (ou *Bagging*) é uma técnica estatística que consiste na criação de múltiplos subconjuntos de dados a partir do conjunto de treino original, gerados por amostragem aleatória com reposição (*bootstrap*). Um modelo independente é treinado para cada subconjunto e as previsões finais são combinadas (agregadas) por votação ou média, reduzindo a variância geral e minimizando o risco de *overfitting*.

Tabela 2- Comparativo de características dos algoritmos de aprendizado de máquina

Algoritmos	Mecanismos de Funcionamento	Características
<i>Decision Tree</i>	Realiza divisões sucessivas e hierárquicas dos dados a partir de regras "Se-Então", criando uma estrutura de fluxograma.	• Alta Interpretação dos dados • Facilidade em entendimento de padrões
<i>Random Forest</i>	Cria um comitê (<i>ensemble</i>) de múltiplas <i>Decision Tree</i> usando amostras aleatórias (<i>bootstrap</i>) e define a classe final por votação majoritária.	• Robustez • Mais chance de memorizar o padrão correto
<i>K-Nearest Neighbors(KNN)</i>	Classifica um novo dado calculando a distância matemática dele para os "K" vizinhos mais próximos na base de treino, adotando a classe da maioria.	• Simplicidade • Não possui um parâmetro matemático fixo, ajudando na interpretação dos dados
<i>eXtreme Gradient Boosting (XGBoost)</i>	Constrói um comitê sequencial de <i>Decision Tree</i> , onde cada nova árvore é focada exclusivamente em corrigir os erros residuais cometidos pelas árvores anteriores.	• Alto poder de processamento dos dados • Focada em corrigir os erros, aumenta as chances de precisão e a entender padrões complexos

Fonte: Elaborada pelo autor (2026).

2.5 Métricas de Avaliação

A avaliação de modelos preditivos de evasão requer atenção especial à escolha das métricas. Conforme Oliveira e Medeiros (2024), métricas como acurácia, precisão, *recall* e f1-score fornecem uma visão abrangente do desempenho preditivo, sendo essenciais para embasar decisões institucionais e a implementação

de políticas eficazes contra a evasão. Para mensurar a eficácia dos modelos desenvolvidos neste estudo, a validação baseou-se na extração dos valores de Verdadeiros Positivos (VP), Falsos Positivos (FP), Verdadeiros Negativos (VN) e Falsos Negativos (FN), utilizando essas quatro métricas: acurácia, precisão, *recall* e *f1-score*, cujas definições são fundamentadas em Faceli et al. (2021) e na documentação do Scikit-learn (2026) que é a biblioteca utilizada para aplicar as métricas.

A acurácia é a métrica mais elementar de avaliação, calculando a proporção geral de previsões corretas realizadas pelo modelo em relação ao total de amostras analisadas. No contexto de predição acadêmica, ela fornece uma visão panorâmica sobre o nível de acerto global do algoritmo. Contudo, seu uso isolado pode ser enganoso em bases de dados desbalanceadas, pois um modelo que classifica todos os alunos na classe majoritária obterá uma alta acurácia, mesmo errando todos os casos da classe minoritária. O cálculo da acurácia é representado na Figura 1.

Figura 1- Fórmula da Acurácia

$$\text{Acurácia} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}}$$

Fonte: Elaborado pelo autor com base em Faceli et al. (2021).

A precisão avalia a proporção de classificações positivas corretas em relação ao total de previsões positivas feitas pelo algoritmo. Em outras palavras, essa métrica responde à seguinte questão: de todos os discentes que o modelo apontou como possíveis evadidos, quantos realmente evadiram? Para a gestão institucional, uma precisão alta colabora ao evitar o desperdício de recursos e esforços pedagógicos com "falsos alarmes", garantindo que as intervenções sejam direcionadas aos alunos corretos. A representação do cálculo da precisão pode ser observada na Figura 2.

Figura 2- Fórmula da Precisão

$$\text{Precisão} = \frac{VP}{VP + FP}$$

Fonte: Elaborado pelo autor com base em Faceli et al. (2021).

O *recall* (também conhecido como Revocação ou Sensibilidade) mensura a proporção de instâncias positivas reais que foram identificadas corretamente pelo modelo. Para este trabalho, o *recall* assume o papel de métrica principal e mais crítica, pois avalia a capacidade do algoritmo de encontrar todos os alunos que efetivamente irão evadir. Em um sistema de retenção estudantil, o custo de um Falso Negativo (não identificar um discente em risco e perdê-lo) é significativamente maior do que o custo de um Falso Positivo. Essa escolha metodológica fundamenta-se em um argumento de custo assimétrico do erro: em um sistema de retenção estudantil, o Falso Positivo — isto é, um discente sinalizado como em risco que, de fato, não evadirá — implica, no pior cenário, o direcionamento de uma ação preventiva desnecessária, cujo custo institucional é relativamente baixo. O Falso Negativo, por sua vez, representa a perda de um discente em situação real de risco sem que a instituição tenha tido a oportunidade de intervir, configurando um custo consideravelmente mais elevado do ponto de vista da gestão institucional. Dessa forma, ainda que um *recall* elevado possa implicar em um maior número de falsos positivos, esse é um custo aceitável frente ao objetivo institucional de não deixar nenhum discente em risco sem possibilidade de intervenção. Cabe ressaltar que o *recall* reportado neste trabalho corresponde à média macro entre as duas classes (conforme detalhado na Seção 4.5), de modo a garantir peso equivalente à identificação de discentes evadidos e concluintes. O método de cálculo do *recall* é ilustrado na Figura 3.

Figura 3- Fórmula do recall

$$\text{Recall} = \frac{VP}{VP + FN}$$

Fonte: Elaborado pelo autor com base em Faceli et al. (2021).

O *f1_score* consiste na média harmônica entre a precisão e o *recall*, oferecendo uma avaliação unificada que penaliza extremos. Essa métrica é excepcionalmente útil para a análise do modelo deste estudo, pois a base de dados original apresenta desproporcionalidade entre evadidos e concluintes. O *f1-score* garante que o classificador possua um desempenho equilibrado, evitando que um modelo seja considerado excelente apenas por maximizar o *recall* à custa de uma precisão ínfima (ou vice-versa). A fórmula do *f1-score* está demonstrada na Figura 4.

Figura 4- Fórmula do *f1_score*

$$\text{F1_Score} = \frac{2 * (\text{Precisão} + \text{Recall})}{(\text{Precisão} + \text{Recall})}$$

Fonte: Elaborado pelo autor com base em Faceli et al. (2021).

2.6 Ferramentas Utilizadas

A implementação de modelos preditivos voltados à Mineração de Dados Educacionais requer a escolha de um ecossistema tecnológico que alia robustez computacional, flexibilidade e ampla aceitação na comunidade científica. Diante dessas necessidades, o desenvolvimento do presente trabalho foi realizado utilizando a linguagem de programação Python, a biblioteca *Scikit-learn* e o ambiente interativo *Jupyter Notebook*.

A escolha da linguagem Python fundamenta-se na sua consolidação como o padrão *de facto* para projetos de ciência de dados e inteligência artificial. Conforme destaca McKinney (2018), o ecossistema Python sobressai-se pela simplicidade de

sua sintaxe, o que reduz o tempo de codificação e facilita a manutenção dos modelos. Ademais, a linguagem possui uma vasta e madura gama de bibliotecas nativas e de terceiros voltadas ao tratamento, manipulação e modelagem estatística de dados, conferindo alta escalabilidade ao processo de desenvolvimento desde o pré-processamento até a execução dos classificadores.

Para a aplicação prática das etapas de Aprendizagem de Máquina, utilizou-se a biblioteca *Scikit-learn*. Conforme apontado por Pedregosa et al. (2011), o *Scikit-learn* é uma ferramenta consolidada para modelagem preditiva em Python, sendo amplamente adotada em pesquisas acadêmicas devido à eficiência de seus algoritmos, à padronização de suas APIs de desenvolvimento (como o uso de *Pipelines*) e ao suporte nativo para tarefas de validação cruzada, normalização e avaliação de desempenho.

A condução dos experimentos e a documentação do código foram realizadas por meio da plataforma *Jupyter Notebook*. Essa ferramenta foi escolhida por sua arquitetura baseada em células executáveis de forma independente e assíncrona. Segundo Kluyver et al. (2016), o *Jupyter Notebook* promove um ambiente altamente interativo e visual que favorece a análise exploratória rápida de dados e o diagnóstico ágil do comportamento dos algoritmos. Ele permite que o pesquisador visualize imediatamente os resultados e tabelas de saída logo abaixo do bloco de código correspondente, além de viabilizar a integração direta de documentações textuais em formato *Markdown*, garantindo a reprodutibilidade e a clareza metodológica do experimento científico.

3 Trabalhos Relacionados

A revisão de trabalhos correlatos permite contextualizar a pesquisa frente aos avanços já obtidos na literatura sobre evasão estudantil, evidenciando as lacunas que o presente estudo busca preencher.

Na literatura recente, Bitencourt, Silva e Xavier (2022) investigaram se a inteligência artificial poderia apoiar ações contra a evasão escolar universitária no Instituto Federal de Minas Gerais (IFMG). Os autores utilizaram algoritmos de

Aprendizagem de Máquina aplicados a variáveis demográficas, acadêmicas e familiares, propondo um sistema de detecção e acompanhamento do risco de evasão focado em auxiliar intervenções precoces por parte da gestão institucional.

Avaliando a modalidade de Educação a Distância (EaD), Juliani (2025) desenvolveu um modelo baseado no algoritmo de Árvore de Decisão para integrar as causas exógenas e endógenas da evasão. O estudo demonstrou a viabilidade de utilizar esse modelo preditivo para gerar percepções estratégicas de retenção de estudantes, destacando a interpretabilidade das árvores de decisão para entender os fatores críticos de abandono.

Em uma abordagem voltada especificamente para o contexto dos Institutos Federais, Silvestri (2023) aplicou técnicas de Mineração de Dados Educacionais para classificar alunos em risco de evasão, explorando características socioeconômicas e acadêmicas de discentes do IFRS. Os resultados evidenciaram que variáveis associadas ao rendimento acadêmico inicial funcionam como fortes indicadores de risco.

Relacionando-se com parágrafos anteriores, trabalhos voltados para a vertente de análise de evasão acadêmica, utilizando de ferramentas de aprendizagem de máquina e algoritmos, tendem para a utilização de algoritmos baseados em árvores. Assim como é possível se ver no trabalho Teodoro et al. (2020), que por sua vez, utilizou de uma base de dados fornecida pelo INEP no qual possui uma gama maior de dados, além de contextos diversos sobre a educação, contribuindo para o melhor treinamento e desempenho do modelo. Em sua pesquisa identificou o algoritmo *Random Forest* como mais robusto para a sua proposta, com isso conseguiu de resultado um desempenho de cerca de 80% envolvendo métricas de acurácia e com isso concluindo que, dentro da sua base de dados, os atributos idade e atividades extracurriculares possuíram destaque positivo no desempenho do seu modelo

Diferenciando-se das pesquisas supracitadas, a presente proposta contribui para a literatura ao avaliar, de forma comparativa, múltiplos classificadores (KNN, *Random Forest*, *Decision Tree* e *XGBoost*) focados na realidade do curso de Licenciatura em Computação do IFPI. Além disso, este estudo busca inovar ao mensurar o impacto direto de diferentes técnicas de pré-processamento — como a

Seleção de Atributos (*SelectKBest*) e o balanceamento sintético via SMOTE — visando otimizar o controle de *overfitting* em bases de dados institucionais desbalanceadas.

Por fim para melhorar a compreensão das diferenças, deste e os demais trabalhos descritos, será apresentado na tabela 3, as características de cada trabalho

Tabela 3- Síntese dos trabalhos relacionados à predição de evasão no ensino superior

Autores	Base de Dados/ Contexto	Algoritmos/Técnica	Principais Resultados e Contribuições
Teodoro et al. (2020)	Dados do INEP	<i>Random Forest</i> e outros 4 algoritmos de Aprendizagem de Máquina.	Obtenção de aproximadamente 80% de acerto. Identificou que a idade e atividades extracurriculares são fatores altamente relevantes na evasão.
Silva (2022)	Dados do Instituto Federal de Minas Gerais (IFMG).	Comitês de Máquinas (<i>Ensemble</i>) e <i>Random Forest</i> .	Demonstrou que a inteligência artificial apoia o monitoramento precoce de discentes, apresentando superioridade no desempenho preditivo.
Primão (2022)	Dados do Instituto Federal de Santa Catarina (IFSC).	<i>XGBoost</i> e modelos de <i>baseline</i> .	O <i>XGBoost</i> alcançou alta capacidade de generalização e <i>F1-Score</i> superior aos demais modelos, provando ser eficaz na classificação de risco.
Oliveira e Medeiros (2024)	Dados de estudantes da Universidade Federal da Paraíba (UFPB).	<i>Decision Tree</i> (Árvore de Decisão).	Atingiu acurácia superior a 87%. Destacou a alta interpretabilidade das regras de decisão para

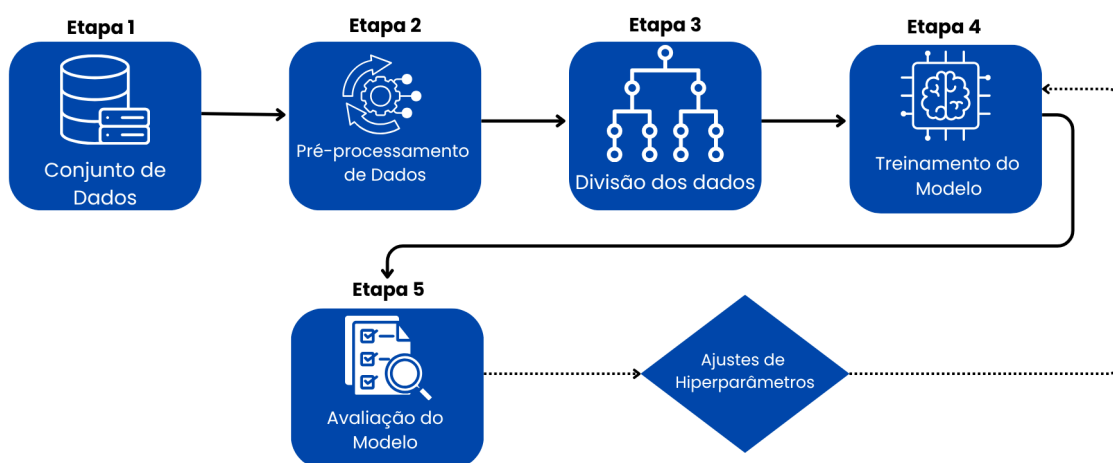
			subsidiar as ações de gestão acadêmica.
Correia et al. (2024)	Mineração de dados em instituição de ensino superior.	Técnicas de Mineração de Dados Educacionais e seleção de atributos.	Alcançou até 92% de acurácia utilizando menos de 20% das características originais, provando o impacto da seleção de atributos.
Hortêncio Filho et al. (2025)	Curso de Tecnologia da Informação em Instituto Federal.	<i>K-Nearest Neighbors</i> (KNN) e outros classificadores.	Evidenciou que atributos demográficos e socioeconômicos são preditores eficazes para o mapeamento da evasão no eixo tecnológico.
Este Trabalho (2026)	Licenciatura em Computação do IFPI CATZS.	<i>Random Forest, KNN, Decision tree, XGBoost, Seleção de Atributos, StandardScaler e SMOTE.</i>	Desenvolveu um modelo preditivo com dados socioeconômicos locais. Identificou gargalos de dados exíguos e propôs o uso balanceado de <i>Recall</i> para gestão de risco.

Fonte: Elaborada pelo autor (2026).

4 Metodologia

O presente trabalho caracteriza-se como uma pesquisa aplicada de natureza quantitativa, tendo como base o conjunto de dados socioeconômicos dos discentes do curso de Licenciatura em Computação do Instituto Federal do Piauí, Campus Teresina Zona Sul (IFPI CATZS), compreendendo o período de 2012 a 2024. A metodologia utilizada neste trabalho, seguirá os seguintes passo a passo demonstrado na figura 5.

Figura 5- Processo de criação de um modelo de Aprendizagem de Máquina



Fonte: Elaborado pelo autor (2026).

A metodologia adotada está organizada em etapas sequenciais, conforme ilustrado na Figura 5, contemplando desde a preparação dos dados até a avaliação dos modelos de Aprendizagem de Máquina.

4.1 Conjunto de Dados

O conjunto de dados utilizado neste trabalho é resultado de uma pesquisa feita paralelamente também envolvendo o curso de licenciatura em computação do IFPI CATZS, sendo composta por informações socioeconômicas coletadas junto à instituição. As variáveis presentes na base contemplam características pessoais, socioeconômicas e acadêmicas dos discentes, tais como: sexo, idade, cor/raça, estado civil, bairro e cidade de residência, zona de residência, profissão, renda mensal familiar, participação econômica na família, modalidade de ingresso via cotas, formação no nível médio, acesso a computador, meio de transporte utilizado para o IFPI, utilização de transporte escolar público, tipo de residência, estado civil dos pais, portador de necessidades especiais, se cursou outro nível superior, além da situação final do discente no curso — se evadiu ou não — que corresponde à

variável alvo codificada de forma binária, sendo 0 correspondente à evasão do curso e 1 a sua conclusão. Os dados abrangem as turmas ingressantes entre os anos de 2012 e 2024, totalizando o histórico completo disponível do curso até o referido período.

Com vistas a uma caracterização mais precisa do problema investigado, o Tabela 4 apresenta o tipo e a natureza de cada atributo utilizado, bem como seus respectivos valores possíveis. Predominam na base variáveis de natureza categórica nominal, refletindo o caráter majoritariamente socioeconômico e demográfico do conjunto de dados, o que justifica, adiante, a adoção do teste estatístico Qui-quadrado (χ^2) na etapa de seleção de atributos, por ser adequado a variáveis dessa natureza.

Tabela 4- Tipo, natureza e valores possíveis dos atributos utilizados

Atributo	Tipo	Valores possíveis
Sexo	Categórica nominal binária	Masculino / Feminino
Idade	Numérica	16 ~ 56 anos
Cor/Raça	Categórica nominal	Branco, Negro, Pardo, Não Declarado
Cotista	Categórica binária	Sim ou Não
Estado Civil	Categórica nominal	Casados, Solteiros, Viúvos e Divorciados
Bairro/ Cidade de Residência	Categórica nominal	Categorizado por Zona Norte, Sul, Leste ou Centro

Zona de Residência	Categórica nominal binária	Rural ou Urbana
Profissão	Categórica nominal binária	Trabalha ou Não Trabalha
Renda Mensal Familiar	Categórica nominal	2.000 ~ 20.000
Participação econômica na família	Categórica nominal	Principal participação, parcialmente participante, pouco participante ou não participante
Modalidade de Ingresso	Categórica nominal	Vestibular da Instituição, ENEM, transferência de outro curso
Formação de Nível médio	Categórica nominal	Instituição pública, particular, EJA ou Ensino Domiciliar
Acesso ao Computador	Categórica nominal binária	Sim ou Não
Meio de transporte até o IFPI	Categórica nominal	Carro, Moto, Ônibus, Bicicleta ou Outro
Uso de Transporte escolar público	Categórica nominal binária	Sim ou Não

Tipo de Residência	Categórica nominal	Casa própria, apartamento próprio, casa alugada ou apartamento alugado
Estado Civil dos Pais	Categórica nominal	Casados, Divorciados ou Viúvo(a)
Portador de Necessidades Especiais	Categórica nominal binária	Sim ou Não
Cursou Outro Nível Superior	Categórica nominal binária	Sim ou Não
Situação Final (<i>Target</i>)	Categórica binária	Evadido ou Não Evadido

Fonte: Elaborado pelo autor (2026).

4.2 Pré-processamento dos Dados

Por se tratar de uma base de dados que passou por etapas iniciais de coleta e codificação anteriores a este trabalho, foram realizadas etapas complementares de pré-processamento para adequação ao modelo preditivo. As variáveis categóricas presentes na base foram previamente codificadas em valores numéricos, viabilizando sua interpretação pelos algoritmos de Aprendizagem de Máquina.

A aplicação de técnicas de normalização justifica-se pela heterogeneidade de escalas presentes entre as variáveis do conjunto de dados — que combina, por exemplo, atributos numéricos como a idade, atributos monetários como a renda familiar e atributos binários codificados como 0 e 1. Essa disparidade de magnitude

pode comprometer o desempenho de algoritmos sensíveis à distância entre observações, como o *K-Nearest Neighbors*, uma vez que atributos de maior amplitude numérica tendem a dominar o cálculo da distância euclidiana, mesmo quando não apresentam maior relevância preditiva. Cabe destacar, entretanto, que algoritmos baseados em estruturas de árvore, como o *Decision Tree* e o *Random Forest*, realizam divisões por limiar de forma independente para cada atributo, sendo, portanto, pouco sensíveis à escala das variáveis. O *XGBoost*, embora também fundamentado em estruturas de árvore, pode se beneficiar indiretamente da padronização quando combinada aos seus mecanismos de regularização. Dessa forma, o principal ganho esperado com a normalização recai sobre o desempenho do KNN, sendo sua aplicação aos demais algoritmos mantidos por consistência metodológica ao longo do pipeline experimental comparativo. Como técnica de normalização, foi aplicado o *MinMaxScaler*, que escala os valores numéricos para o intervalo entre 0 e 1, eliminando a disparidade de escala entre as variáveis e garantindo que nenhuma característica exerça influência desproporcional sobre o modelo em razão da magnitude de seus valores.

Assim como também, foi aplicada a padronização com a técnica *StandardScaler*, que determina os valores numéricos por meio do intervalo de um desvio padrão no qual o valor varia entre -1 e 1, aumentando assim o detalhamento dos valores dos atributos.

Desta forma o benefício proporcionado por essas técnicas ajuda nos desempenhos dos algoritmos escolhidos neste trabalho, pois evita que os atributos de maior magnitude numérica dominem o treinamento.

4.3 Divisão dos Dados

A divisão entre os conjuntos de treino e teste foi realizada de forma temporal, respeitando a ordem cronológica dos dados. Os registros compreendidos entre 2012 e 2024 foram destinados ao treinamento e teste dos modelos — treinado com informações do passado é aplicado a situações futuras.

Devido à necessidade metodológica de particionamento dos dados, os registros referentes às turmas de 2012 a 2024 foram divididos utilizando a proporção de 70% para treinamento e 30% para teste. A escolha por essa taxa de separação baseou-se no volume absoluto de instâncias disponíveis na base, buscando um

ponto de equilíbrio. Dessa forma, assegurou-se uma quantidade de informações adequada para o treinamento robusto dos classificadores, preservando simultaneamente um volume amostral representativo para a avaliação imparcial do poder de generalização do modelo.

Com essa determinação, faz-se necessário dizer que, para uma análise detalhada de desempenho, foram testadas técnicas diferentes de análise de dados a fim de proporcionar o melhor desempenho para o modelo, sendo elas a seleção de atributos, redução de dimensão

4.4 Algoritmos Utilizados - Treinamento do modelo

Para a construção do modelo preditivo, foram selecionados quatro algoritmos supervisionados de classificação amplamente utilizados na literatura de Aprendizagem de Máquina:

A *Decision Tree* (Árvore de Decisão) é um algoritmo que constrói um modelo em formato hierárquico, realizando divisões sucessivas dos dados com base em regras aprendidas durante o treinamento. Sua principal vantagem reside na interpretabilidade, permitindo a visualização clara do caminho lógico utilizado para a classificação das instâncias.

O *Random Forest* é um algoritmo *ensemble* que combina múltiplas árvores de decisão, treinadas a partir de amostras aleatórias do conjunto de dados original. A predição final ocorre por meio de uma votação majoritária entre as árvores. Essa abordagem tende a produzir modelos mais robustos e menos suscetíveis ao *overfitting*⁵ quando comparada a uma árvore isolada.

O *K-Nearest Neighbors* (KNN) é um algoritmo de aprendizado baseado em instâncias que classifica um novo dado fundamentando-se na sua similaridade com as amostras de treinamento. Ele identifica as K instâncias mais próximas

⁵ *Overfitting* (ou sobreajuste) ocorre quando o modelo de aprendizagem de máquina se ajusta excessivamente aos dados de treinamento, memorizando o ruído e as oscilações aleatórias em vez de aprender o padrão subjacente. Isso resulta em um desempenho excelente nos dados de treino, mas em uma baixa capacidade de generalização ao lidar com dados novos (dados de teste).

matematicamente e atribui à nova amostra a classe mais frequente entre esses vizinhos.

O *XGBoost* (*eXtreme Gradient Boosting*) é um algoritmo baseado em comitês de árvores de decisão que utiliza o princípio de impulsionamento de gradiente (*gradient boosting*). Ele opera de forma sequencial, construindo iterativamente novas árvores focadas em corrigir os erros residuais cometidos pelas árvores anteriores. Sua arquitetura é voltada para o alto desempenho computacional, velocidade de processamento e eficiência no controle da generalização.

Todos os classificadores foram implementados em linguagem Python, dentro do ambiente *Jupyter Notebook*, utilizando a biblioteca *Scikit-learn* e a biblioteca nativa do *XGBoost*. Os algoritmos foram encapsulados em um *Pipeline*, estruturando o fluxo de execução para automatizar as etapas de normalização e treinamento de forma padronizada e sequencial.

4.4.1 Técnicas de Otimização do Modelo

Com o objetivo de avaliar o impacto de diferentes abordagens de pré-processamento no desempenho do modelo preditivo, foram testadas duas técnicas complementares em conjunto com os algoritmos de classificação. A Seleção de Atributos via *SelectKBest* foi aplicada com o intuito de identificar as variáveis socioeconômicas com maior relevância estatística para a predição, eliminando características redundantes ou irrelevantes que poderiam introduzir ruído ao modelo. A Redução de Dimensionalidade via PCA foi testada como alternativa, buscando verificar se a compressão das características em componentes principais produziria ganhos de desempenho. Os três cenários — sem técnica, com Seleção de Atributos e com PCA — foram avaliados comparativamente por meio das métricas de acurácia, *Recall* e F1-Score, permitindo uma escolha metodologicamente embasada da abordagem mais adequada ao problema de predição de evasão estudantil.

A escolha da técnica *SelectKBest*, associada ao teste estatístico Qui-quadrado (χ^2), fundamenta-se na natureza predominantemente categórica das variáveis presentes na base de dados, conforme detalhado na Seção 4.1, sendo o χ^2 estatisticamente adequado à avaliação de associação entre variáveis

categóricas codificadas numericamente e a variável alvo (VASCONCELLOS, 2019). Já a técnica de Redução de Dimensionalidade via PCA foi incorporada ao desenho experimental como abordagem alternativa e complementar, permitindo comparar uma estratégia de seleção — que preserva a interpretabilidade das variáveis originais — a uma estratégia de transformação, que compacta os atributos em componentes principais de mais difícil interpretação semântica, conforme discutido na Seção 2.3.2.

Após a aplicação do SelectKBest, os atributos selecionados são `estudante_profissional`, `estado_civil`, `meio_de_transporte`, `oriundo_escola`, `atualmente_reside`, `acesso_ao_computador`, `renda_mensal`, `participacao_economica`, `cota` e `cursou_outro_nivel_superior`, foram os que obtiveram maior pontuação estatística e, portanto, foram mantidos no subconjunto utilizado para o treinamento dos modelos, ao passo que os demais atributos foram descartados por apresentarem menor associação estatística com a variável alvo.

4.4.2 Técnicas de Balanceamento dos Dados

Após uma avaliação do conjunto de dados disponibilizados, notou-se uma grande diferença no que diz respeito a quantidade de alunos evadidos, que supera, sobre a quantidade de alunos concluintes do curso, essa discrepância nos dados pode ocasionar em um mau treinamento do modelo ocasionando um problema de falta de entendimento, por parte do modelo, de padrões, dessa forma, tornando-se um modelo inutilizável. Portanto para a mitigação desse problema, aplicou-se a técnica SMOTE nos dados de treino do modelo.

A fim de dimensionar o grau de desbalanceamento enfrentado e o impacto da técnica de balanceamento aplicada, o Tabela 5 apresenta a distribuição das classes "evadido" e "concluído" no conjunto de treino antes e depois da aplicação do SMOTE, para os dois cenários de tratamento de dados faltantes considerados neste trabalho.

Tabela 5- Distribuição das classes antes e depois do balanceamento (conjunto de treino)

Cenário	Classe “evadido” (0)	Classe “concluído” (1)	Total	Proporção Evadido:não evadido
Antes do SMOTE – Dados Reais	163	33	196	~ 5 : 1
Antes do SMOTE – Imputação por Moda	444	60	504	37 : 5
Depois do SMOTE (ambos os cenários)	310	310	620	1 : 1

Fonte: Elaborada pelo próprio autor (2026)

4.5 Avaliação do Desempenho do Modelo

A avaliação dos modelos foi realizada por meio de métricas amplamente adotadas na literatura para problemas de classificação binária. A acurácia representa a proporção geral de classificações corretas realizadas pelo modelo. A precisão mensura, dentre os discentes classificados como evadidos, quantos de fato evadiram. O *Recall*, considerado a métrica principal deste trabalho, foi calculado por meio da média macro entre as duas classes, garantindo peso equivalente à correta identificação tanto de discentes evadidos quanto de concluintes — minimizando os casos em que alunos em situação de risco passam despercebidos pela instituição. O F1-Score corresponde à média harmônica entre precisão e *Recall*, oferecendo uma visão equilibrada do desempenho do modelo.

Para garantir a confiabilidade estatística dos resultados, foi aplicada a técnica de validação cruzada *K-Fold* com cinco partições (cv=5), que divide os dados em cinco subconjuntos e realiza cinco rodadas de treinamento e avaliação, utilizando

uma partição diferente como conjunto de teste em cada rodada. Os resultados são expressos pela média e desvio padrão do *Recall* obtido nas cinco rodadas.

5 Experimentos e Discussão de Resultados

Esta seção descreve a metodologia experimental adotada e a subsequente análise de desempenho dos modelos preditivos. Para garantir uma avaliação abrangente, os testes foram estruturados a partir de seis categorias de conjuntos de dados, subdivididos entre bases compostas estritamente por valores reais e bases submetidas à imputação de valores sintéticos. Esses conjuntos foram aplicados sob três abordagens distintas de pré-processamento: ausência de técnica adicional, Seleção de Atributos e Redução de Dimensionalidade (PCA), gerando 18 cenários experimentais. Por fim, cada um dos 18 conjuntos de dados foi submetido ao treinamento com quatro algoritmos de classificação diferentes, culminando em um total de 72 experimentos conduzidos e analisados no escopo deste trabalho.

5.1 Visão Geral dos Experimentos

A execução da extensa bateria de testes proporcionou uma observação panorâmica sobre o comportamento dos classificadores. Ao compilar os resultados, constatou-se que o desempenho dos modelos variou significativamente de acordo com o critério observado, evidenciando que nenhuma combinação metodológica isolada alcançou superioridade absoluta em todas as métricas simultaneamente. Esse achado demonstra que não existe uma solução analítica trivial para o problema abordado, reforçando a pertinência de uma investigação comparativa exaustiva. Dessa forma, a análise ampla conduzida neste trabalho mostrou-se indispensável para mapear as vantagens e limitações de cada técnica, garantindo uma tomada de decisão fundamentada para a seleção do modelo final.

Cumprе esclarecer que o *overfitting* apresentado ao longo desta seção e no Apêndice A foi calculado como a diferença percentual entre o desempenho do modelo no conjunto de treino e no conjunto de teste (*overfitting* = acurácia treino – acurácia teste). Valores positivos e elevados indicam que o modelo memorizou padrões específicos do conjunto de treinamento sem generalizar adequadamente

para dados não vistos, ao passo que valores próximos de zero, ou mesmo negativos, indicam boa capacidade de generalização, podendo sugerir, inclusive, um comportamento levemente conservador do modelo durante o treinamento.

5.2 Impacto do Tratamento e Valores Faltantes

A comparação entre os *datasets* com valores reais — onde as linhas com dados faltantes foram removidas — e os *datasets* com imputação pela moda revelou uma tendência consistente: os conjuntos com imputação pela moda apresentaram desempenho superior na maioria dos experimentos. Esse resultado era esperado, uma vez que a imputação preserva um maior número de amostras — 504 registros contra aproximadamente 196 nos *datasets* com remoção de linhas — oferecendo ao modelo uma base de aprendizado mais ampla, fator especialmente relevante em bases de dados de tamanho reduzido como a utilizada neste trabalho.

5.3 Impacto da Técnica de Normalização

A comparação entre o *StandardScaler* e o *MinMaxScaler* demonstrou diferenças sutis entre as duas abordagens. O *StandardScaler* apresentou *Recall* médio de 55.5% ao longo de todos os experimentos, enquanto o *MinMaxScaler* obteve 56.4% — uma diferença marginal que sugere que, para este conjunto de dados, a escolha da técnica de normalização não foi o fator determinante no desempenho dos modelos. Contudo, o *StandardScaler* mostrou-se mais adequado em experimentos com Seleção de Atributos, onde produziu o melhor resultado geral do trabalho.

5.4 Impacto da Técnica de Seleção ou Redução de Atributos

A análise comparativa entre os três cenários — sem técnica, com Seleção de Atributos e com Redução de Dimensionalidade — revelou resultados relevantes para a escolha metodológica mais adequada ao problema.

A Seleção de Atributos via *SelectKBest* apresentou o melhor desempenho médio em termos de *Recall* (57.4%) e o menor *overfitting* médio entre as três abordagens (15.09%), evidenciando que a eliminação de features irrelevantes

contribuiu tanto para o desempenho quanto para a capacidade de generalização dos modelos.

A abordagem sem nenhuma técnica adicional apresentou *Recall* médio de 56.7% e *overfitting* médio de 16.58% — resultados ligeiramente inferiores à Seleção de Atributos, indicando que a presença de atributos irrelevantes introduz ruído nos modelos.

A Redução de Dimensionalidade via PCA apresentou consistentemente os piores resultados, com *Recall* médio de 53.5% e *overfitting* médio de 23.55% — o mais elevado entre as três abordagens. Esse comportamento confirma que a transformação das features em componentes abstratos prejudica a capacidade dos modelos de identificar padrões relevantes em bases de dados socioeconômicas, além de amplificar o *overfitting* pela perda de informação interpretável.

Para além das médias apresentadas, na Tabela 6 sintetiza a melhor combinação individual obtida em cada um dos três cenários de pré-processamento avaliados, permitindo identificar o teto de desempenho alcançado por cada abordagem.

Tabela 6- Melhor resultado individual por cenário experimental

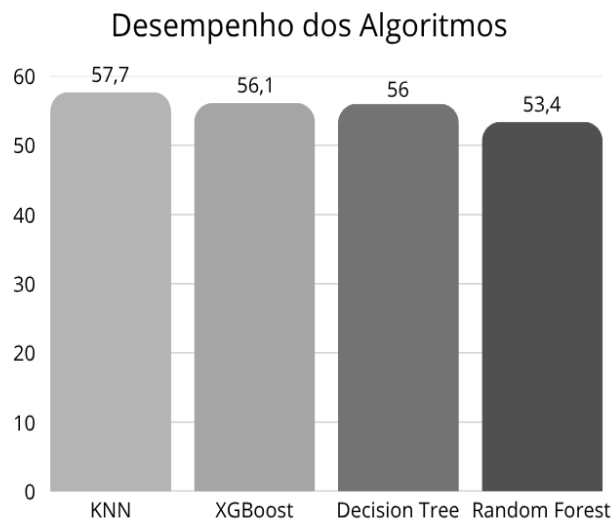
Cenário	Melhor combinação (conjunto de dados + algoritmo)	<i>Recall</i>	<i>Overfitting</i>
Sem técnica Adicional	Experimento 3 (Dados reais + StandardScaler + LabelEncoder) — KNN	70%	31,15%
Seleção de Atributos (SelectKBest)	Experimento 8 (Imputação por moda + StandardScaler + Seleção de Atributos) — Random Forest	72%	6,03%
PCA	Experimento 14 (Imputação por moda + StandardScaler + PCA) — KNN	69%	22,99%

Fonte: Elaborada pelo próprio autor (2026)

5.5 Desempenho por Algoritmo

A análise do desempenho médio dos quatro algoritmos ao longo de todos os experimentos (conforme tabela detalhada no Apêndice A) revelou os seguintes resultados, ilustrados na Figura 6 em termos de porcentagem de recall macro:

Figura 6- Desempenho de Algoritmos



Fonte: Elaborado pelo autor (2026).

O KNN apresentou o maior *recall* médio entre os algoritmos avaliados, porém com alta variabilidade entre os experimentos — registrando desde 36% no E13 até 70% no E3 — e *overfitting* elevado em vários cenários, chegando a 31.61% no DS1. O *XGBoost* demonstrou as maiores acurácias absolutas em vários experimentos, chegando a 86.84% nos datasets E10 e E12, porém com *Recall* macro frequentemente abaixo dos demais algoritmos, indicando dificuldade em identificar corretamente a classe minoritária. O Random Forest apresentou o melhor equilíbrio entre *Recall* e controle de *overfitting*, sendo responsável pelo melhor resultado geral

do trabalho. A Decision Tree registrou o menor *Recall* médio, porém com casos pontuais de excelente controle de *overfitting* — como no E10, onde a diferença entre treino e teste foi de apenas -3.75%.

5.6 Limitações e Implicações

Apesar do rigor metodológico empregado e dos resultados promissores alcançados, o desenvolvimento desta pesquisa enfrentou óbices estruturais complexos relacionados diretamente à qualidade e à quantidade da base de dados disponível, os quais impuseram limitações reflexas ao desempenho final dos modelos preditivos.

O principal fator limitante consistiu no tamanho exíguo do conjunto de dados coletado para o curso de Licenciatura em Computação ao longo do recorte temporal analisado. O volume total de amostras reais mostrou-se numericamente reduzido para atender às demandas de alta parametrização que os algoritmos de Aprendizagem de Máquina exigem para mapear relações multifatoriais complexas. Essa escassez quantitativa de registros restringiu a capacidade dos classificadores de extraírem correlações estatísticas mais profundas e generalizáveis.

Somado a isso, a base de dados original apresentou um severo índice de desproporcionalidade entre a classe majoritária e a minoritária. Embora técnicas de balanceamento sintético via *SMOTE* tenham sido aplicadas com o intuito de mitigar o desequilíbrio durante o treinamento, a geração de dados artificiais por aproximação matemática não foi capaz de substituir plenamente a riqueza e a diversidade biográfica de dados reais que seriam necessários para o correto entendimento dos padrões de evasão.

Como consequência direta dessas barreiras de dados, os modelos construídos herdaram falhas de sensibilidade e inconsistências pontuais (conforme observado nas oscilações de desempenho nos cenários experimentais), evidenciando que os classificadores encontraram um teto metodológico limitante. Sob a configuração atual da base de dados institucional, tornou-se inviável conduzir o modelo ao seu ápice de desenvolvimento técnico e capacidade preditiva ideal, ressaltando que a eficácia da inteligência artificial aplicada à gestão educacional é

estritamente dependente da robustez dos ecossistemas de dados armazenados pelas instituições.

6 Considerações finais

Este trabalho teve como objetivo central investigar a aplicação de técnicas de Aprendizagem de Máquina para a predição da evasão acadêmica no curso de Licenciatura em Computação do Instituto Federal do Piauí, Campus Teresina Zona Sul (IFPI CATZS). Diante dos altos índices de abandono observados historicamente no curso, buscou-se desenvolver um modelo preditivo baseado em dados socioeconômicos e acadêmicos que pudesse auxiliar a gestão institucional na identificação precoce de discentes em situação de risco, fornecendo subsídios para a criação de políticas de permanência estudantil.

Para alcançar esse objetivo, foi conduzida uma extensa bateria de testes, totalizando 72 combinações entre diferentes algoritmos, técnicas de normalização, tratamento de dados faltantes e seleção de atributos. Essa abordagem metodológica ampla evidenciou que a predição de evasão é um problema complexo e que nenhuma técnica isolada apresenta superioridade absoluta, ressaltando a importância do processo de experimentação e ajuste fino (*tuning*).

No que diz respeito ao pré-processamento dos dados, constatou-se que a preservação do volume de informações é crucial. A imputação de valores faltantes pela moda demonstrou resultados superiores à simples exclusão de registros, garantindo uma base de aprendizado mais robusta para os algoritmos. Em relação à seleção e redução de características, a técnica de filtragem estatística *SelectKBest* apresentou o melhor desempenho geral, maximizando a métrica de *Recall* e reduzindo o *overfitting*. Em contrapartida, a Análise de Componentes Principais (PCA) obteve os piores resultados, indicando que a transformação de características socioeconômicas em componentes abstratos prejudica a capacidade do modelo de identificar os padrões associados à evasão.

A avaliação dos algoritmos de classificação revelou comportamentos distintos. Embora o *XGBoost* tenha alcançado as maiores taxas de acurácia global,

apresentou dificuldades na identificação correta da classe minoritária (alunos evadidos). O *K-Nearest Neighbors* (KNN) obteve o maior *Recall* médio, porém demonstrou instabilidade e níveis elevados de *overfitting* em diversos cenários. Conclui-se, portanto, que o algoritmo *Random Forest* apresentou o melhor equilíbrio geral para este problema específico, aliando uma boa capacidade de generalização (controle de *overfitting*) a um desempenho consistente na métrica de *Recall* – a mais importante para o contexto do estudo, uma vez que o custo de não identificar um aluno em risco é superior ao de falsos alarmes.

É importante ressaltar, contudo, que o recall médio obtido ao longo dos experimentos — próximo de 50% a 58% a depender do cenário — situa-se ainda distante de um patamar ideal de aplicação prática, o que impõe cautela na interpretação dos resultados como um indicativo de excelência preditiva. A relevância deste trabalho reside, portanto, menos no desempenho absoluto alcançado pelos modelos e mais na validação da abordagem metodológica comparativa proposta — envolvendo múltiplos algoritmos, técnicas de pré-processamento e estratégias de balanceamento —, a qual demonstrou ser tecnicamente aplicável ao contexto institucional do IFPI CATZS. O teto de desempenho observado decorre primariamente do volume reduzido de dados disponíveis (entre 196 e 504 registros), conforme discutido na Seção 5.6, e não de uma limitação intrínseca da abordagem, o que reforça a importância dos trabalhos futuros sugeridos para a ampliação e o enriquecimento da base de dados institucional.

Por fim, conclui-se que os objetivos propostos por este trabalho foram plenamente alcançados. Isso se justifica porque os resultados confirmam a viabilidade técnica da utilização da Mineração de Dados Educacionais como ferramenta de apoio à tomada de decisão no IFPI CATZS, entregando um modelo validado (*Random Forest*) capaz de identificar discentes em situação de risco. Como trabalhos futuros, sugere-se a ampliação da base de dados com a inclusão do histórico de notas dos estudantes nos primeiros semestres, o que pode aumentar o poder preditivo dos modelos, além do desenvolvimento de uma interface que integre o modelo ao sistema de controle acadêmico da instituição, facilitando o acesso e a utilização pelos gestores educacionais.

7 Referências

JESUS, Hiago et al. Predição de Evasão Escolar na Licenciatura em Computação. *Revista Brasileira de Informática na Educação*, v. 29, p. 255–273, 2021.

KANTORSKI, Gustavo Zanini et al. Mineração de Dados Educacionais para Predição da Evasão em Cursos de Graduação Presenciais no Ensino Superior. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO (SBIE), 34., 2023, Passo Fundo/RS. Anais [...]. Porto Alegre: SBC, 2023. p. 1133-1142.

OLIVEIRA, R. dos S.; MEDEIROS, F. P. A. de. 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, 2024.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

SANTOS, K. dos; PIRES FERNANDES, E. A Previsão de Evasão em Cursos de Graduação Utilizando Machine Learning: Uma Análise com Alunos de uma Instituição de Ensino Superior particular. *Caderno Virtual*, v. 1, n. 58, 2024.

TEODORO, W. et al. 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. 838–857, 2020.

HORTÊNCIO FILHO, F. W. B. et al. Análise de Classificadores para Predição de Evasão em um Curso Superior de Tecnologia da Informação de um Instituto Federal. In: *COMPUTER ON THE BEACH*, v. 16, p. 146-153. Univali, 2025.

CHAWLA, N. V. et al. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, v. 16, p. 321–357, 2002.

GERMANOS, L. V.; SILVA, T. Redução de Dimensionalidade. *Medium*, 2020. Disponível em: <https://luvi01.medium.com>. Acesso em: 2025.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

BITENCOURT, Wanderci Alves; SILVA, Diego Mello; XAVIER, Gláucia do Carmo. Pode a inteligência artificial apoiar ações contra evasão escolar universitária?. *Ensaio: Avaliação e Políticas Públicas em Educação*, Rio de Janeiro, v. 30, n. 116, p. 669-694, jul./set. 2022.

JULIANI, J. P. Predição de evasão em cursos EAD: um modelo baseado em árvore de decisão que integra causas exógenas e endógenas. *EaD em Foco*, Rio de Janeiro, v. 15, n. 1, e2504, 2025.

SILVESTRI, Felipe. Mineração de dados educacionais: uma análise sobre os preditores de evasão no ensino superior. 2023. Trabalho de Conclusão de Curso (Tecnologia em Análise e Desenvolvimento de Sistemas) – Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul, Ibirubá, 2023.

CHEN, Tianqi; GUESTRIN, Carlos. XGBoost: a scalable tree boosting system. *In: PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING*, 22., 2016, São Francisco. Nova York: ACM, 2016. p. 785-794.

PRIMÃO, Aline Pacheco. Uso de algoritmos de *machine learning* para prever a evasão escolar no ensino superior: um estudo no Instituto Federal de Santa Catarina. 2022. 95 f. Dissertação (Mestrado em Administração Universitária) – Centro Socioeconômico, Universidade Federal de Santa Catarina, Florianópolis, 2022.

FACELI, Katti; LORENA, Ana Carolina; GAMA, João; CARVALHO, André C. P. L. F. de. Inteligência artificial: uma abordagem de aprendizado de máquina. 2. ed. Rio de Janeiro: LTC, 2021.

SCIKIT-LEARN DEVELOPERS. Metrics and scoring: quantifying the quality of predictions. Scikit-learn 1.3 Documentation. Disponível em: https://scikit-learn.org/stable/modules/model_evaluation.html. Acesso em: 15 jul. 2026.

KLUYVER, Thomas et al. Jupyter Notebooks – a publishing format for reproducible computational workflows. *In: LOIZIDES, Fernando; SCHMIDT, Birgit (ed.). Positioning professional research and science in the world of today*. Amsterdã: IOS Press, 2016. p. 87-90.

MCKINNEY, Wes. Python para análise de dados: tratamento, processamento e análise de dados com pandas, NumPy e IPython. Tradução de Lúcia A. Kinoshita. São Paulo: Novatec, 2018.

PEDREGOSA, Fabian et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, [S. l.], v. 12, p. 2825-2830, 2011.

Apêndice A-Resultados dos Experimentos dos 4 Modelos

Experimento	Modelo	Accuracy (Teste)	Recall (Macro)	Precision (Macro)	F1-Score (Macro)	Overfitting (Treino - Teste)
1 - Reais + StandardScaler	Decision Tree	76.27%	54.00%	55.00%	54.00%	14.08%
1 - Reais + StandardScaler	Random Forest	69.49%	50.00%	50.00%	50.00%	20.42%
1 - Reais + StandardScaler	KNN	50.85%	54.00%	53.00%	46.00%	31.61%
1 - Reais + StandardScaler	XGB	74.58%	49.00%	48.00%	48.00%	25.42%
2 - Moda + StandardScaler	Decision Tree	52.63%	54.00%	52.00%	44.00%	19.79%
2 - Moda + StandardScaler	Random Forest	73.68%	63.00%	57.00%	57.00%	11.96%
2 - Moda + StandardScaler	KNN	61.84%	66.00%	57.00%	52.00%	22.67%
2 - Moda + StandardScaler	XGB	86.18%	58.00%	63.00%	60.00%	13.34%
3 - Reais + StandardScaler + LabelEncoder	Decision Tree	69.49%	62.00%	58.00%	58.00%	17.79%
3 - Reais + StandardScaler + LabelEncoder	Random Forest	76.27%	54.00%	55.00%	54.00%	14.08%
3 - Reais + StandardScaler + LabelEncoder	KNN	62.71%	70.00%	61.00%	57.00%	31.15%
3 - Reais + StandardScaler + LabelEncoder	XGB	77.97%	55.00%	57.00%	55.00%	22.03%
4 - Moda + StandardScaler + LabelEncoder	Decision Tree	79.61%	55.00%	54.00%	54.00%	2.17%
4 - Moda + StandardScaler + LabelEncoder	Random Forest	82.89%	59.00%	59.00%	59.00%	5.17%
4 - Moda + StandardScaler + LabelEncoder	KNN	71.71%	55.00%	53.00%	52.00%	21.19%
4 - Moda + StandardScaler + LabelEncoder	XGB	84.87%	55.00%	58.00%	56.00%	14.97%
5 - Reais + MinMaxScaler + LabelEncoder	Decision Tree	59.32%	64.00%	58.00%	53.00%	25.33%
5 - Reais + MinMaxScaler + LabelEncoder	Random Forest	77.97%	55.00%	57.00%	55.00%	13.70%
5 - Reais + MinMaxScaler + LabelEncoder	KNN	66.10%	60.00%	56.00%	55.00%	24.69%

5 - Reais + MinMaxScaler + LabelEncoder	XGB	74.58%	53.00%	53.00%	53.00%	25.42%
6 - Moda + MinMaxScaler + LabelEncoder	Decision Tree	73.03%	56.00%	53.00%	53.00%	10.36%
6 - Moda + MinMaxScaler + LabelEncoder	Random Forest	81.58%	56.00%	56.00%	56.00%	6.32%
6 - Moda + MinMaxScaler + LabelEncoder	KNN	73.68%	51.00%	51.00%	51.00%	19.06%
6 - Moda + MinMaxScaler + LabelEncoder	XGB	84.87%	55.00%	58.00%	56.00%	15.13%
7 - Reais + StandardScaler + Sel. Atributos	Decision Tree	72.88%	52.00%	52.00%	52.00%	13.96%
7 - Reais + StandardScaler + Sel. Atributos	Random Forest	69.49%	50.00%	50.00%	50.00%	20.86%
7 - Reais + StandardScaler + Sel. Atributos	KNN	54.24%	53.00%	51.00%	47.00%	29.97%
7 - Reais + StandardScaler + Sel. Atributos	XGB	76.27%	58.00%	58.00%	58.00%	21.54%
8 - Moda + StandardScaler + Sel. Atributos	Decision Tree	67.76%	65.00%	57.00%	55.00%	12.56%
8 - Moda + StandardScaler + Sel. Atributos	Random Forest	80.26%	72.00%	63.00%	65.00%	6.03%
8 - Moda + StandardScaler + Sel. Atributos	KNN	73.68%	63.00%	57.00%	57.00%	11.16%
8 - Moda + StandardScaler + Sel. Atributos	XGB	84.21%	55.00%	57.00%	56.00%	10.14%
9 - Reais + StandardScaler + LabelEncoder (V2)	Decision Tree	64.41%	43.00%	44.00%	43.00%	27.26%
9 - Reais + StandardScaler + LabelEncoder (V2)	Random Forest	72.88%	60.00%	58.00%	58.00%	19.66%
9 - Reais + StandardScaler + LabelEncoder (V2)	KNN	62.71%	66.00%	59.00%	56.00%	18.87%
9 - Reais + StandardScaler + LabelEncoder (V2)	XGB	79.66%	68.00%	66.00%	66.00%	16.39%
10 - Moda + StandardScaler + LabelEncoder (V2)	Decision Tree	85.53%	49.00%	44.00%	46.00%	-3.75%

10 - Moda + StandardScaler + LabelEncoder (V2)	Random Forest	84.21%	62.00%	62.00%	62.00%	5.14%
10 - Moda + StandardScaler + LabelEncoder (V2)	KNN	73.03%	51.00%	51.00%	50.00%	9.71%
10 - Moda + StandardScaler + LabelEncoder (V2)	XGB	86.84%	52.00%	57.00%	51.00%	9.45%
11 - Reais + MinMaxScaler + LabelEncoder + Sel.	Decision Tree	66.10%	52.00%	51.00%	51.00%	21.62%
11 - Reais + MinMaxScaler + LabelEncoder + Sel.	Random Forest	76.27%	62.00%	60.00%	61.00%	15.40%
11 - Reais + MinMaxScaler + LabelEncoder + Sel.	KNN	59.32%	60.00%	55.00%	52.00%	24.45%
11 - Reais + MinMaxScaler + LabelEncoder + Sel.	XGB	77.97%	67.00%	64.00%	65.00%	18.52%
12 - Moda + MinMaxScaler + LabelEncoder + Sel.	Decision Tree	76.97%	56.00%	54.00%	54.00%	9.00%
12 - Moda + MinMaxScaler + LabelEncoder + Sel.	Random Forest	82.24%	54.00%	55.00%	54.00%	5.34%
12 - Moda + MinMaxScaler + LabelEncoder + Sel.	KNN	71.71%	48.00%	49.00%	48.00%	11.35%
12 - Moda + MinMaxScaler + LabelEncoder + Sel.	XGB	86.84%	52.00%	57.00%	51.00%	9.61%
13 - Reais + StandardScaler + Red. Dimensão	Decision Tree	50.00%	36.00%	41.00%	37.00%	31.54%
13 - Reais + StandardScaler + Red. Dimensão	Random Forest	62.50%	38.00%	39.00%	38.00%	26.35%
13 - Reais + StandardScaler + Red. Dimensão	KNN	50.00%	36.00%	41.00%	37.00%	30.00%
13 - Reais + StandardScaler + Red. Dimensão	XGB	75.00%	45.00%	41.00%	43.00%	25.00%
14 - Moda + StandardScaler + Red. Dimensão	Decision Tree	45.54%	47.00%	49.00%	39.00%	27.70%

14 - Moda + StandardScaler + Red. Dimensão	Random Forest	63.37%	54.00%	52.00%	49.00%	23.25%
14 - Moda + StandardScaler + Red. Dimensão	KNN	58.42%	69.00%	58.00%	51.00%	22.99%
14 - Moda + StandardScaler + Red. Dimensão	XGB	80.20%	56.00%	56.00%	56.00%	19.52%
15 - Reais + StandardScaler + LabelEncoder + Red.	Decision Tree	52.50%	43.00%	46.00%	42.00%	24.81%
15 - Reais + StandardScaler + LabelEncoder + Red.	Random Forest	70.00%	54.00%	53.00%	53.00%	24.23%
15 - Reais + StandardScaler + LabelEncoder + Red.	KNN	60.00%	65.00%	58.00%	54.00%	19.23%
15 - Reais + StandardScaler + LabelEncoder + Red.	XGB	72.50%	55.00%	55.00%	55.00%	27.50%
16 - Moda + StandardScaler + LabelEncoder + Red.	Decision Tree	44.55%	58.00%	53.00%	40.00%	27.56%
16 - Moda + StandardScaler + LabelEncoder + Red.	Random Forest	66.34%	59.00%	54.00%	52.00%	17.46%
16 - Moda + StandardScaler + LabelEncoder + Red.	KNN	63.37%	61.00%	55.00%	51.00%	15.93%
16 - Moda + StandardScaler + LabelEncoder + Red.	XGB	85.15%	56.00%	59.00%	56.00%	14.85%
17 - Reais + MinMaxScaler + LabelEncoder + Red.	Decision Tree	44.55%	58.00%	53.00%	40.00%	27.56%
17 - Reais + MinMaxScaler + LabelEncoder + Red.	Random Forest	66.34%	59.00%	54.00%	52.00%	17.46%
17 - Reais + MinMaxScaler + LabelEncoder + Red.	KNN	63.37%	61.00%	55.00%	51.00%	15.93%
17 - Reais + MinMaxScaler + LabelEncoder + Red.	XGB	85.15%	56.00%	59.00%	56.00%	14.85%
18 - Moda + MinMaxScaler +	Decision Tree	54.46%	53.00%	51.00%	44.00%	23.43%

LabelEncoder + Red.						
18 - Moda + MinMaxScaler + LabelEncoder + Red.	Random Forest	63.37%	47.00%	48.00%	45.00%	25.08%
18 - Moda + MinMaxScaler + LabelEncoder + Red.	KNN	69.31%	61.00%	55.00%	54.00%	14.77%
18 - Moda + MinMaxScaler + LabelEncoder + Red.	XGB	81.19%	60.00%	59.00%	59.00%	18.81%