Research, Society and Development, v. 14, n. 6, e6214649029, 2025
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v14i6.49029

# Machine learning: Uma aplicação para a prevenção da Evasão Escolar

## Machine learning: An application for School Dropout prevention

## Aprendizaje automático: Una aplicación para la prevención del Abandono Escolar

Recebido: 04/06/2025 | Revisado: 14/06/2025 | Aceitado: 15/06/2025 | Publicado: 17/06/2025

**Valberto Rômulo Feitosa Pereira**
ORCID: https://orcid.org/0000-0002-2541-4846
Instituto Federal de Educação, Ciência e Tecnologia do Ceará, Brasil
E-mail: valbertofeitosa@uol.com.br
**Mairton Cavalcante Romeu**
ORCID: https://orcid.org/0000-0001-5204-9031
Instituto Federal de Educação, Ciência e Tecnologia do Ceará, Brasil
E-mail: mairtoncavalcante@ifce.edu.br
**Nairys Costa de Freitas**
ORCID: https://orcid.org/0000-0002-0799-8489
Instituto Federal de Educação, Ciência e Tecnologia do Ceará, Brasil
E-mail: Nairys.freitas07@aluno.ifce.edu.br
**Vinicius Silva Pereira**
ORCID: https://orcid.org/0009-0002-7266-858X
Instituto Federal de Educação, Ciência e Tecnologia do Ceará, Brasil
E-mail: vinicius.pereira05@aluno.ifce.edu.br

**Resumo**
A retenção de estudantes em disciplinas de matemática perpetua-se como uma preocupação preeminente nas instituições educacionais, constituindo um desafio multifacetado, pois influencia diretamente a eficácia do processo pedagógico e, consequentemente, incita à evasão escolar. O objetivo do presente estudo é apresentar uma aplicação de aprendizagem de máquina na prevenção de evasão escolar. No escopo desta pesquisa, o intento consiste em discernir as probabilidades associadas à aprovação ou reprovação de alunos matriculados em disciplinas de matemática no primeiro semestre letivo, com base em registros históricos de desempenho acadêmico no Instituto Federal do Ceará (IFCE), campus de Fortaleza. A esse propósito, foi aplicado algoritmos de Aprendizagem de Máquina para indicar a probabilidade de um aluno ser aprovado ou reprovado em matemática no primeiro semestre letivo, gerando relatórios que auxiliem a tomada de decisão por parte das coordenações dos cursos em que essa matéria é lecionada. Serão utilizados os algoritmos Regressão Logística e Naive Bayes. Para isso, serão utilizadas bibliotecas disponíveis para a linguagem Python, a saber, Scikit-learn. O resultado esperado é de que os algoritmos de Aprendizagem de Máquina forneçam percepções na tomada de ações pedagógicas e gerenciais acadêmicas preventivas que visem à redução dos índices de reprovação e, consequentemente, a evasão escolar.
**Palavras-chave:** Retenção; Evasão; Aprendizado de máquina; *Naive bayes*; Regressão logística.

**Abstract**
The retention of students in mathematics subjects continues to be a major concern for educational institutions and is a multifaceted challenge, since it directly influences the effectiveness of the teaching process and, consequently, encourages school dropouts. The aim of this study is to present an application of machine learning to prevent school drop-outs. The aim of this research is to discern the probabilities associated with passing or failing students enrolled in mathematics subjects in the first semester of school, based on historical records of academic performance at the Federal Institute of Ceará (IFCE), Fortaleza campus. To this end, Machine Learning algorithms were applied to indicate the probability of a student passing or failing mathematics in the first semester, generating reports to help decision making by the coordinators of the courses in which this subject is taught. The Logistic Regression and Naive Bayes algorithms will be used. To do this, libraries available for the Python language will be used, namely Scikit-learn. The expected result is that the Machine Learning algorithms will provide insights into preventive academic pedagogical and managerial actions aimed at reducing failure rates and, consequently, school dropout.
**Keywords:** Retention; Dropout; Machine learning; Naive bayes; Logistic regression.

**Resumen**
La retención de alumnos en las asignaturas de matemática continúa siendo una gran preocupación para las instituciones de enseñanza y constituye un desafío multifacético, ya que influye directamente en la eficacia del proceso pedagógico y, consecuentemente, incita a la deserción escolar. El objetivo de esta investigación es discernir las probabilidades asociadas a la aprobación o reprobación de alumnos matriculados en asignaturas de matemáticas en

el primer semestre académico, a partir de los registros históricos de rendimiento académico en el Instituto Federal de Ceará (IFCE), campus Fortaleza. Para ello, se aplicaron algoritmos de Machine Learning para indicar la probabilidad de que un alumno apruebe o repruebe matemáticas en el primer semestre académico, generando informes que ayuden a la toma de decisiones por parte de los coordinadores de los cursos en los que se imparte esta asignatura. Se utilizarán los algoritmos de Regresión Logística y Naive Bayes. Para ello, se utilizarán librerías disponibles para el lenguaje Python, concretamente Scikit-learn. El resultado esperado es que los algoritmos de Aprendizaje Automático aporten ideas sobre acciones pedagógicas y de gestión académica preventivas encaminadas a reducir las tasas de fracaso y, en consecuencia, el abandono escolar.
**Palabras clave:** Retención; Abandono; Aprendizaje automático; Bayas ingenuas; Regresión logística.

## 1. Introdução

A educação profissionalizante no Brasil desempenha um papel de suma importância na inserção do aluno para o mercado de trabalho, além de contribuir para a formação de uma sociedade mais equânime, de acordo com o artigo "História da Educação Profissional no Brasil" (2021). Nesse contexto, com o objetivo de atender à demanda por ensino básico e profissionalizante, foi sancionada a Lei nº 11.892, de 29 de dezembro de 2008 (Brasil, 2008), que marcou o início dos Institutos Federais de Educação, Ciência e Tecnologia, democratizando o acesso a cursos técnicos, tecnólogos e cursos superiores para todo o país.

Entretanto, mesmo com a ampliação da oferta de ensino público federal, a retenção escolar ainda segue como um desafio para os educadores. A retenção escolar refere-se ao fenômeno em que alunos são obrigados a repetir um ano letivo devido ao desempenho insatisfatório em suas atividades acadêmicas. Esse fenômeno tem consequências no âmbito pessoal do aluno, podendo interferir no seu desenvolvimento pessoal e profissional, além de gerar prejuízos econômicos e organizacionais para as instituições de ensino (Felizardo, 2024).

Além disso, a retenção pode aumentar significativamente as chances de evasão escolar. Alunos que enfrentam repetição de ano frequentemente experimentam desmotivação, sentimento de fracasso e baixa autoestima, o que pode torná-los mais propensos a abandonar a escola (Mendoza Lira, 2023). Dessa forma, a retenção não apenas compromete o sucesso acadêmico imediato, mas também contribui para o aumento das taxas de evasão escolar, dificultando o alcance dos objetivos educacionais de formação integral e acesso ao mercado de trabalho.

Nesse contexto, a matemática frequentemente se destaca como um dos maiores desafios enfrentados pelos alunos, sendo fundamental para o desenvolvimento do pensamento lógico e para a construção de habilidades essenciais em áreas como física, química, economia e tecnologia. Dificuldades nessa disciplina geram desmotivação e insegurança, comprometendo não apenas o desempenho acadêmico imediato, mas também o sucesso em disciplinas correlatas. Quando a base matemática é mal estruturada, surgem lacunas no aprendizado que limitam as oportunidades de inserção no mercado de trabalho, especialmente em áreas técnico-científicas. Esse cenário contribui para um ciclo de baixo rendimento, retenção escolar e, em última instância, evasão, agravando desigualdades educacionais e profissionais.

Diante desses desafios, é necessário analisar soluções tecnológicas e inovadoras como o Machine Learning, que surge como uma alternativa promissora para evitar a evasão escolar. Machine Learning descreve a capacidade dos sistemas de aprender com dados de treinamento específicos de problemas para automatizar o processo de construção de modelos analíticos e resolver tarefas associadas (Janiesch, 2021). Para isso, o presente trabalho tem como objetivo apresentar uma aplicação de aprendizagem de máquina na prevenção de evasão escolar. No contexto escolar, essa tecnologia pode ser empregada para identificar padrões em grandes volumes de dados, ajudando a prever comportamentos de risco, como a propensão à evasão. Dessa forma, é possível propor intervenções mais assertivas e personalizadas, contribuindo para o fortalecimento da permanência e sucesso escolar.

## 2. Metodologia

Realizou-se uma pesquisa mista, em parte documental de fonte direta com coleta de dados de alunos em banco de dados do IFCE, em parte de natureza quantitativa (Pereira et al., 2018) e com emprego de estatística descritiva simples com uso de valores de classes de dados, média, desvio padrão, mediana, quartis (Shitsuka et al., 2014; Akamine & Yamamoto, 2009) e também de análise estátística (Bekman & Costa Neto, 2009) e com uso de machine learning (Burkov, 2019). quanti-qualitativa dos dados analisados (Gil, 2017; Pereira et al., 2018). Na parte quantitativa empregou-se estatística descritiva simples com uso de classes de dados, frequência absoluta e frequência relativa porcentual (Shitsuka et al., 2014; Akmanine & Yamamoto, 2009) e, machine learning (Burkov, 2019; Müller & Guido, 2017).

A metodologia adotada fundamenta-se na aplicação de algoritmos probabilísticos, seguindo uma abordagem sistemática para investigar a aplicação de técnicas avançadas de aprendizado de máquina as quais são instrumentalizados para prever o desfecho acadêmico dos estudantes, fornecendo visões pertinentes à gestão educacional e norteando a implementação de medidas preventivas em relação à disciplina de Matemática, com base em dados acadêmicos dos cursos técnicos integrados do Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE) - Campus Fortaleza. O processo analítico almeja contribuir para a formulação de estratégias pedagógicas e gerenciais embasadas em evidências, com vistas a mitigar os índices de reprovação e, por conseguinte, conter a evasão escolar. A construção do estudo foi organizada em múltiplas fases metodológicas, uma prática consolidada e bem fundamentada na pesquisa científica, especialmente nas áreas de aprendizado de máquina e análise de dados. Esta abordagem sistemática assegura uma organização clara e sequencial do processo de pesquisa, garantindo que cada fase seja meticulosamente planejada e executada. Conforme discutido por Creswell (2014), uma estrutura metodológica bem definida permite maior rigor e reprodutibilidade na pesquisa. Adicionalmente, Yin (2018) enfatiza que a segmentação do estudo em etapas distintas facilita a análise e a interpretação dos resultados. Além disso, tal estrutura facilita a replicação do estudo e a verificação dos resultados por parte de outros pesquisadores, contribuindo para a robustez e a transparência científica. O cerne do estudo é descrito a seguir:

**Acesso aos Dados**

O banco de dados é composto por informações dos alunos ingressos nos cursos técnicos integrados do Instituto Federal – IFCE – Campus Fortaleza, coletadas a partir do portal oficial (*https://qselecao.ifce.edu.br/lista_concursos.aspx*), abrangendo as seis modalidades de cursos técnicos integrados no período de 2018.2 a 2020.1. Inicialmente, essa coleta foi realizada manualmente. No entanto, no semestre de 2022.2, desenvolvemos um código em Python (Van Rossum et al., 2009) para automatizar o processo via *web scraping*, utilizando as bibliotecas *Selenium* e *BeautifulSoup*.

Os dados coletados incluem o curso escolhido, a reserva de vagas e a nota final no exame de seleção. Em um segundo momento, no semestre seguinte, a coleta da média final em Matemática foi realizada manualmente, por meio da busca individualizada do nome do aluno no sistema acadêmico do IFCE. Esses dados foram então integrados ao banco de dados original por meio da chave "nome do aluno". Para garantir a qualidade da informação coletada, foi realizada uma conferência visual dos registros, dado o tamanho relativamente pequeno do conjunto de dados. Embora esse procedimento não elimine completamente possíveis inconsistências, ele permitiu a identificação e correção de discrepâncias evidentes. Além disso, análises exploratórias foram conduzidas para avaliar padrões de distribuição das variáveis e minimizar erros nos registros utilizados na modelagem preditiva.

**O banco de dados está constituído pelas seguintes variáveis**

1. Nota_Matemática - Número real de 0 a 10. Representa a média na disciplina de Matemática no primeiro semestre.

2. Pontuação - Número real de 0 a 10. Representa a pontuação final no exame de seleção.
3. Reserva_vaga - String. Representa reserva de vaga. Pode ser uma das seguintes opções:
   a. Ampla Concorrência
   b. L1 - Candidatos com deficiência autodeclarados pretos, pardos ou indígenas, que tenham renda familiar bruta per capita igual ou inferior a 1,5 salário mínimo e que tenham cursado integralmente o ensino fundamental em escolas públicas.
   c. L2 - Candidatos autodeclarados pretos, pardos ou indígenas, que tenham renda familiar bruta per capita igual ou inferior a 1,5 salário mínimo e que tenham cursado integralmente o ensino fundamental em escolas públicas.
   d. L3 - Candidatos com deficiência que tenham renda familiar bruta per capita igual ou inferior a 1,5 salário mínimo e que tenham cursado integralmente o ensino fundamental em escolas públicas.
   e. L4 - Candidatos que tenham renda familiar bruta per capita igual ou inferior a 1,5 salário mínimo e que tenham cursado integralmente o ensino fundamental em escolas públicas.
   f. L5 - Candidatos com deficiência autodeclarados pretos, pardos ou indígenas, que independentemente da renda tenham cursado integralmente o ensino fundamental em escolas públicas.
   g. L6 - Candidatos autodeclarados pretos, pardos ou indígenas, que independentemente da renda tenham cursado integralmente o ensino fundamental em escolas públicas.
   h. L7 - Candidatos com deficiência que independentemente da renda tenham cursado integralmente o ensino fundamental em escolas públicas.
   i. L8 - Candidatos que independentemente da renda tenham cursado integralmente o ensino fundamental em escolas públicas.
4. Curso - String. Representa o curso técnico para o qual o aluno se inscreveu. Pode ser uma das seguintes opções:
   a. Técnico Integrado em Informática
   b. Técnico Integrado em Telecomunicações
   c. Técnico Integrado em Eletrotécnica
   d. Técnico Integrado em Química
   e. Técnico Integrado em Edificações
   f. Técnico Integrado em Mecânica Industrial
5. Semestre - String. Representa o semestre de ingresso. Pode ser uma das seguintes opções:
   a. 2018.1
   b. 2018.2
   c. 2019.1
   d. 2019.2
   e. 2020.1
   f. 2020.2
   g. 2021.1
   h. 2021.2
   i. 2022.1
   j. 2022.2

Inicialmente tínhamos 682 alunos, dos semestres entre 2018.2 até 2020.1, mas na coleta das médias em matemática, foram encontrados 468 alunos no sistema acadêmico da instituição, ou seja, 214 alunos não foram encontrados, um dado que

será investigado no futuro.

Os estudantes analisados foram divididos em duas classes: aprovado e reprovado em matemática no final do 1º semestre, sendo a aprovação o resultado de maior número. Essa discrepância de resultados pode levar ao direcionamento do modelo para a classe com mais amostras (aprovação), o que pode prejudicar o aprendizado em casos minoritários (reprovação). Dessa forma, o modelo pode concluir que todos foram aprovados e, consequentemente, acertar a maioria das previsões, mas não identificar quem seria reprovado. Para lidar com esse problema, foi utilizado neste artigo o algoritmo *Synthetic Minority Oversampling Technique* (SMOTE) a fim de balancear as classes, aumentando artificialmente a classe desfavorecida (Fernandéz et al, 2018).

**Pré-processamento dos Dados**

O pré-processamento dos dados desempenha um papel fundamental na análise de dados educacionais e em diversas outras áreas, sendo uma etapa crítica na preparação dos dados brutos antes da análise estatística e modelagem preditiva. Durante esse processo, são tratados problemas como valores incoerentes, ausentes, inconsistentes, duplicados, ruídos e imperfeições nos dados, garantindo maior qualidade e confiabilidade para os modelos preditivos (Faceli *et al*., 2021).

Para a realização dessas etapas, foram utilizadas bibliotecas amplamente reconhecidas na comunidade científica, incluindo Pandas, NumPy, Seaborn, Matplotlib, Scikit-learn e Excel, proporcionando um arcabouço computacional robusto para manipulação, visualização e análise dos dados (Mckinney, 2018).

Além disso, em conformidade com a Lei Geral de Proteção de Dados (LGPD) (Brasil, Lei nº 13.709/2018), foram removidas informações sensíveis, como nomes dos alunos, garantindo a anonimização dos dados e protegendo a identidade dos indivíduos. A LGPD estabelece diretrizes para o tratamento de dados pessoais, visando garantir a segurança, privacidade e uso ético das informações, conforme descrito no art. 12 da Lei nº 13.709/2018 (Brasil, 2018).

**Análise Exploratória dos Dados**

A análise exploratória de dados (EDA) é uma etapa essencial em estudos voltados para modelagem preditiva, pois permite compreender a estrutura dos dados e extrair informações relevantes para o desenvolvimento de modelos mais eficazes (Tukey, 1977).

No contexto deste estudo, a EDA foi aplicada a um conjunto de dados provenientes do processo seletivo para ingresso nos cursos técnicos integrados (ensino médio) do IFCE. O objetivo dessa análise foi garantir a qualidade dos dados antes da modelagem preditiva, seguindo as seguintes etapas:

Primeiramente, foi realizada a identificação e tratamento de dados ausentes, considerando que a presença de informações incompletas pode comprometer a precisão dos modelos (Géron, 2019). Em seguida, procedeu-se à verificação das escalas das variáveis, um fator relevante para normalização e padronização dos dados, garantindo que diferenças nas magnitudes das variáveis não influenciem negativamente os modelos estatísticos e de aprendizado de máquina (Hair, 2009).

Além disso, foi verificada a tipificação correta das variáveis, pois a natureza de cada atributo determina a abordagem estatística e os algoritmos mais adequados para análise. Variáveis categóricas, por exemplo, requerem técnicas de codificação específicas, enquanto variáveis contínuas podem demandar normalização (Géron, 2019 & Bussab, 2010).

Por fim, a análise de tendências estatísticas foi conduzida por meio de médias descritivas e gráficos de dispersão, permitindo a identificação de padrões e relações entre as variáveis, elementos fundamentais para a construção de modelos preditivos robustos (Géron, 2019 & Bussab, 2010).

**Balanceamento de Classes**

O equilíbrio de classes desempenha um papel fundamental na modelagem de problemas de classificação, particularmente quando se encontram desigualdades significativas entre as classes. A presença de desarmonia pode implicar a eficácia dos modelos preditivos, resultando em viés e desempenho insatisfatório no reconhecimento da classe minoritária. De acordo com He e Garcia (2009), "o equilíbrio de classes é fundamental em problemas de classificação, pois a presença de classes desbalanceadas pode levar a modelos enviesados que não representam corretamente a distribuição dos dados."

Neste estudo, abordamos o desafio do desbalanceamento de classes na variável alvo, que representa a situação dos candidatos em relação à aprovação ou reprovação em matemática. A detecção inicial dessa discrepância motivou a aplicação da técnica de oversampling, especificamente o *Synthetic Minority Oversampling Technique* (SMOTE), uma perspectiva amplamente reconhecida na literatura por sua eficácia na geração de dados sintéticos para a classe minoritária, pois segundo (Chawla, Bowyer, Hall & Kegelmeyer 2002), "muitas vezes, os conjuntos de dados do mundo real são predominantemente composto de exemplos "normais" com apenas uma pequena porcentagem de exemplos "anormais" ou exemplos interessantes".

Ao empregar o SMOTE, tivemos em vista igualar o número de exemplos das classes majoritária e minoritária, promovendo deste modo uma distribuição mais equiponderante dos dados. Primeiramente, após dividir os dados em regressores (X_In) e variável alvo (y_In), foi reconhecida uma desproporção expressiva entre as classes da variável alvo. Posteriormente, aplicou-se o SMOTE com a estratégia de amostragem definida como “minority”, o que significa que a classe minoritária seria sobre-amostrada para corresponder à quantidade de observações da classe majoritária. Tal técnica contribui para a validade e confiabilidade dos resultados, certificando o compromisso com a integridade metodológica e a qualidade desse estudo.

**Seleção das Variáveis**

A seleção de variáveis foi realizada por meio de uma abordagem híbrida, combinando os métodos de Filtro e Envoltório para otimizar a escolha das características mais relevantes para o modelo preditivo (Tanimu *et al*., 2022).

Inicialmente, utilizamos o método de filtro, aplicando o algoritmo SelectKBest em conjunto com a Regra do Cotovelo. Segundo a documentação do Scikit-learn (2025), o SelectKBest é uma função da biblioteca Scikit-learn em Python que seleciona as k melhores variáveis com base em testes estatísticos univariados. Essa técnica é amplamente empregada para reduzir a dimensionalidade dos dados, mantendo as variáveis mais relevantes para o modelo.

A Regra do Cotovelo é frequentemente utilizada para determinar o número ideal de clusters em algoritmos de agrupamento, como o k-means. No entanto, esse conceito pode ser adaptado para a seleção de variáveis, permitindo identificar o ponto em que a adição de novas variáveis não contribui significativamente para a melhoria do modelo (Géron, 2019).

Após essa pré-seleção, aplicamos o método envoltório, testando diferentes subconjuntos de variáveis em modelos preditivos e analisando a acurácia de cada configuração. Essa etapa final possibilitou validar a escolha das variáveis não apenas com base em critérios estatísticos, mas também considerando o impacto direto no desempenho do modelo, caracterizando um processo envoltório.

Dessa forma, a estratégia híbrida adotada permitiu equilibrar eficiência computacional e qualidade preditiva, garantindo que o modelo fosse construído com um conjunto de variáveis otimizado e relevante (Tanimu *et al*., 2022).

**Normalização dos Dados em Aprendizado de Máquina**

A normalização dos dados é uma etapa essencial no pré-processamento de muitos algoritmos de aprendizado de máquina, garantindo que todas as variáveis possuam escalas comparáveis. Esse processo melhora a estabilidade numérica dos modelos e acelera a convergência durante o treinamento. Ele é especialmente importante em algoritmos baseados em distância,

como *K-Means*, *K-Nearest Neighbors* (KNN) e *Support Vector Machines* (SVM), onde escalas diferentes podem distorcer a análise de similaridade entre os dados (Géron, 2019 & Hair, 2009).

Existem várias técnicas de normalização, sendo as mais comuns:

a) *Min-Max Scaling:* Transforma os valores para um intervalo específico, geralmente entre 0 e 1, utilizando a fórmula:

$$X_{norm} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

b) *Z-Score* (Padronização): Transforma os dados para uma distribuição com média zero e desvio padrão unitário:

$$X_{standard} = \frac{X - \mu}{\sigma}$$

Onde μ é a média e σ é o desvio padrão da variável.

Segundo Géron (2019), a escolha do método de normalização depende do tipo de modelo e da distribuição dos dados. Enquanto a padronização (*Z-score*) é recomendada para dados com distribuição aproximadamente normal, a normalização *Min-Max* é mais adequada quando há restrições de intervalo, como em redes neurais que utilizam funções de ativação sigmoide.

Além disso, a normalização desempenha um papel fundamental no desempenho dos modelos de aprendizado de máquina, evitando que variáveis com magnitudes maiores influenciem desproporcionalmente o processo de aprendizado. Esse princípio também se aplica a algoritmos de clusterização, garantindo que todas as características tenham impacto equilibrado na formação dos grupos, resultando em uma segmentação mais precisa (Géron, 2019; Hair, 2009).

Portanto, a normalização dos dados é um passo fundamental para garantir um treinamento eficiente e uma melhor generalização dos modelos de aprendizado de máquina. Ao aplicar essas técnicas, é essencial considerar a natureza dos dados e os requisitos do algoritmo utilizado para obter o melhor desempenho possível (Géron, 2019).

**Treinamento e Avaliação dos Modelos**

No contexto deste estudo, a variável alvo (*target*) utilizada foi a Situação do Aluno (aprovado ou reprovado), caracterizando um problema de classificação supervisionada. Para abordar esse tipo de problema, selecionamos dois modelos amplamente utilizados: Regressão Logística e *Naive Bayes* (Morettin & Singer, 2021; Géron, 2019).

De acordo com Galdino (2018), a Regressão Logística é amplamente empregada em contextos que exigem modelos interpretáveis e robustos, sendo eficaz para prever probabilidades associadas a uma variável categórica binária. Sua principal vantagem reside na capacidade de interpretar os coeficientes das variáveis independentes, possibilitando a identificação dos fatores que mais influenciam a reprovação dos alunos. Batista (2015) reforça essa característica ao afirmar que "a regressão logística é uma técnica estatística que estima a probabilidade de ocorrência de um evento binário com base em um conjunto de variáveis explicativas".

O modelo Naive Bayes, por sua vez, fundamenta-se no Teorema de Bayes e assume independência condicional entre as variáveis preditoras. Segundo Sabry (2023), "o Naive Bayes é um método direto para construir classificadores, sendo eficaz

em problemas de classificação mesmo sob a forte suposição de independência entre as variáveis".

Os modelos foram treinados e avaliados utilizando um conjunto de dados previamente balanceados. A avaliação do desempenho considerou métricas como Acurácia Balanceada, Precisão, Revocação, F1-Score , Curva ROC e validação cruzada (Tanimu *et al*., 2022) (Géron 2019), permitindo uma análise abrangente da capacidade preditiva dos modelos. Essas métricas foram escolhidas por fornecerem diferentes perspectivas sobre o desempenho, desde a taxa global de acertos até a capacidade do modelo de distinguir corretamente alunos aprovados e reprovados. Além disso, no pré-processamento dos dados incluiu a separação entre os atributos descritivos (variáveis preditoras) e a variável alvo, que indicava a situação do aluno em matemática (aprovado ou reprovado), garantindo uma avaliação mais robusta e confiável (Géron 2019).

**Limitações da Metodologia**

Apesar dos resultados alcançados, algumas limitações metodológicas devem ser consideradas. O conjunto de dados analisado compreende 468 alunos, o que pode ser um número relativamente pequeno para garantir a generalização dos modelos preditivos em diferentes contextos acadêmicos. Além disso, os dados abrangem apenas o período de 2018.2 a 2020.1, o que pode não refletir mudanças estruturais ocorridas em anos subsequentes, como adaptações pedagógicas ou impactos externos significativos, incluindo a pandemia da COVID-19.

Outra limitação importante está relacionada à coleta e qualidade dos dados. Embora parte do processo tenha sido automatizada por meio de *web scraping*, a extração de informações pode estar sujeita a erros, inconsistências ou lacunas em atributos relevantes. Em particular, os dados socioeconômicos dos alunos não estavam preenchidos para a maioria dos casos, o que impediu sua utilização na análise. Essa ausência pode comprometer a compreensão mais ampla dos fatores que influenciam o desempenho acadêmico, uma vez que aspectos como condições financeiras, acesso a recursos educacionais e contexto familiar podem exercer um papel determinante.

Além disso, o estudo se baseia em modelos específicos, como Regressão Logística e Naive Bayes, o que pode limitar a captura de padrões mais complexos nos dados. Outro desafio é a possível autocorrelação, pois alunos de um mesmo curso e semestre tendem a compartilhar características semelhantes, o que pode impactar os resultados das análises. Para superar essas limitações, estudos futuros podem explorar técnicas de aprendizado profundo, ampliar o período analisado e incorporar novas variáveis para aumentar a robustez e a capacidade preditiva dos modelos.

## 3. Resultados e Discussões

**Pré-processamento dos Dados**

O pré-processamento dos dados foi realizado em várias etapas para garantir a qualidade, consistência e estruturação adequada do conjunto de dados antes da modelagem preditiva. Inicialmente, parte do tratamento foi feito manualmente no Excel, incluindo a eliminação de alunos com dados faltantes ou que não concluíram o semestre, além da verificação de inconsistências e ruídos nos registros. Essa verificação demandou um trabalho minucioso e manual, analisando os dados individualmente para garantir sua integridade.

Após essa etapa inicial, os dados foram carregados no ambiente Python a partir de um arquivo Excel, utilizando a biblioteca Pandas, o que possibilitou a estruturação e manipulação eficiente do conjunto de dados. A partir desse ponto, foram aplicadas diversas técnicas de organização e preparação dos dados, conforme descrito a seguir:

**1. Reordenamento e Conversão de Tipos de Dados**

- As colunas foram reorganizadas para manter uma estrutura padronizada.

- As variáveis numéricas foram verificadas novamente, garantindo que estivessem no tipo adequado (float ou int).
- As variáveis categóricas foram analisadas, assegurando que estavam corretamente representadas como strings, permitindo sua posterior codificação.

**2. Codificação de Variáveis Categóricas**

Para que os algoritmos de aprendizado de máquina pudessem interpretar corretamente as informações categóricas, foi aplicada a técnica de one-hot encoding (Géron, 2019). Esse processo converteu variáveis como Reserva de Vaga e Curso em variáveis binárias, garantindo que os modelos conseguissem processá-las corretamente.

A conversão foi realizada utilizando a função pd.get_dummies(), conforme ilustrado abaixo:

Integrado1 = pd.get_dummies(Integrado, columns=['Reserva_vaga', 'Curso'], drop_first=True)

O parâmetro drop_first=True foi utilizado para evitar multicolinearidade, reduzindo a dimensionalidade do conjunto de dados sem perda de informação.

**3. Criação da Variável-Alvo**

Para transformar a variável Nota_Matemática em um problema de classificação binária, os alunos foram categorizados da seguinte forma:

- Aprovado (1): alunos com nota ≥ 6,0.
- Reprovado (0): alunos com nota < 6,0.

A conversão foi feita com a função pd.cut():

Integrado1['Situação_Matemática'] = pd.cut(Integrado1.Nota_Matemática, bins=[-1, 5.9, 10], labels=[0, 1])

Essa abordagem permitiu que os algoritmos pudessem lidar com a variável-alvo de forma estruturada, facilitando a modelagem preditiva.

**4. Resultado do Processo**

O resultado desse pré-processamento foi um conjunto de dados limpo, estruturado e pronto para ser utilizado na modelagem preditiva, garantindo que os algoritmos operassem de maneira eficiente e confiável (Faceli *et al*., 2021).

**Análises exploratória dos dados**

Após carregar a base de dados com as variáveis: 'Semestre', 'Nota_Matemática','Pontuação', 'Reserva_vaga', 'Curso', verificamos se tinha dados faltantes e se as variáveis estavam caracterizadas corretamente, usando o comando NomedaBase.info(). Assim, a Tabela 1 a seguir apresenta a estrutura do conjunto de dados.

**Tabela 1 -** Estrutura do Conjunto de Dados.

| Coluna | Contagem Não Nula | Tipo de Dado |
|---|---|---|
| Semestre | 468 non-null | object |
| Nota_Matemática | 468 non-null | float64 |
| Pontuação | 468 non-null | float64 |
| Reserva_vaga | 468 non-null | object |
| Curso | 468 non-null | object |

Fonte: Dados coletados do estudo.

Em seguida fizemos a estatística descritiva da variável Pontuação por Reserva_vaga, Nota_Matemática por Reserva_vaga, Pontuação por Curso, Nota_Matemática por Curso, Pontuação por Semestre e Nota_Matemáticapor Semestre a fim de detectar alguma anomalia. Feito isto, as Tabelas 2 e 3 apresentam a estatística descritiva por categoria de reserva de vagas.

**Tabela 2 -** Estatística Descritiva por Categoria de Reserva de Vaga.

| Reserva de Vaga | N (count) | Média (mean) | Desvio Padrão (std) | Mínimo (min) | 1° Quartil (25%) | Mediana (50%) | 3° Quartil (75%) | Máximo (max) |
|---|---|---|---|---|---|---|---|---|
| AC | 273.0 | 7.916484 | 0.706946 | 6.16 | 7.3700 | 8.00 | 8.440 | 9.50 |
| L2 | 4.0 | 5.432500 | 0.365183 | 5.16 | 5.2425 | 5.30 | 5.490 | 5.97 |
| L2 | 81.0 | 6.669877 | 1.040162 | 3.47 | 6.0200 | 6.78 | 7.470 | 9.07 |
| L3 | 1.0 | 6.250000 | NaN | 6.25 | 6.2500 | 6.25 | 6.250 | 6.25 |
| L4 | 19.0 | 6.701579 | 1.604779 | 1.22 | 6.0850 | 6.96 | 7.680 | 8.81 |
| L6 | 75.0 | 6.422400 | 1.074192 | 2.21 | 5.7300 | 6.59 | 7.235 | 8.76 |
| L7 | 1.0 | 0.590000 | NaN | 0.59 | 0.5900 | 0.59 | 0.590 | 0.59 |
| L8 | 14.0 | 6.639286 | 0.754631 | 5.68 | 5.9825 | 6.50 | 7.330 | 7.76 |

Fonte: Dados coletados do estudo.

**Tabela 3 -** Estatística Descritivas por Categoria de Reserva de Vaga

| Reserva de Vaga | N (count) | Média (mean) | Desvio Padrão (std) | Mínimo (min) | 1° Quartil (25%) | Mediana (50%) | 3° Quartil (75%) | Máximo (max) |
|---|---|---|---|---|---|---|---|---|
| AC | 273.0 | 7.598535 | 2.195784 | 0.0 | 6.600 | 8.10 | 9.30 | 10.0 |
| L2 | 4.0 | 3.025000 | 4.050000 | 0.0 | 1.125 | 1.55 | 3.45 | 9.0 |
| L2 | 81.0 | 6.492593 | 2.741431 | 0.0 | 5.400 | 7.20 | 8.60 | 10.0 |
| L3 | 1.0 | 1.000000 | NaN | 1.0 | 1.000 | 1.00 | 1.00 | 1.0 |
| L4 | 19.0 | 6.063158 | 3.048900 | 0.0 | 5.050 | 6.50 | 8.10 | 9.8 |
| L6 | 75.0 | 5.626667 | 2.687978 | 0.0 | 5.000 | 6.20 | 7.40 | 10.0 |
| L7 | 1.0 | 2.700000 | NaN | 2.7 | 2.700 | 2.70 | 2.70 | 2.7 |
| L8 | 14.0 | 5.300000 | 3.117198 | 1.2 | 2.750 | 6.00 | 7.15 | 10.0 |

Fonte: Dados coletados do estudo.

A Tabela 4, a seguir apresenta as estatísticas descritivas dos cursos técnicos integrados em Edificações, Eletrotécnica, Informática, Mecânica Industrial, Química e Telecomunicações no campus Fortaleza.

**Tabela 4** - Estatísticas Descritivas por Curso Técnico Integrado - Campus Fortaleza

| Curso Tecnico Integrado | N (count) | Média (mean) | Desvio Padrão (std) | Mínimo (min) | 1° Quartil (25%) | Mediana (50%) | 3° Quartil (75%) | Máximo (max) |
|---|---|---|---|---|---|---|---|---|
| Edificações | 75.0 | 7.484800 | 1.030415 | 4.83 | 6.8050 | 7.560 | 8.1850 | 9.35 |
| Eletrotécnica | 77.0 | 7.268442 | 1.094961 | 4.57 | 6.5200 | 7.410 | 8.1400 | 9.33 |
| Informática | 84.0 | 7.561548 | 1.231914 | 1.22 | 6.9575 | 7.755 | 8.3825 | 9.21 |
| Mecânica Industrial | 74.0 | 6.855000 | 1.180351 | 3.47 | 6.0750 | 7.030 | 7.8050 | 8.70 |
| Química | 77.0 | 7.549221 | 1.222406 | 0.59 | 7.0100 | 7.670 | 8.3100 | 9.50 |
| Telecomunicações | 81.0 | 7.249753 | 1.059272 | 2.21 | 6.8000 | 7.410 | 8.0000 | 8.99 |

Fonte: Dados coletados do estudo.

Assim, as Tabelas 5 a 7 apresentam as estatísticas descritivas da nota em Matemática por semestre.

**Tabela 5 -** Estatísticas Descritivas da Nota em Matemática por Curso Técnico Integrado - Campus Fortaleza.

| Curso Tecnico Integrado | N (count) | Média (mean) | Desvio Padrão (std) | Mínimo (min) | 1° Quartil (25%) | Mediana (50%) | 3° Quartil (75%) | Máximo (max) |
|---|---|---|---|---|---|---|---|---|
| Edificações | 75.0 | 6.553333 | 2.255104 | 0.0 | 5.550 | 6.80 | 7.950 | 10.0 |
| Eletrotécnica | 77.0 | 7.590909 | 2.175533 | 0.0 | 7.000 | 8.10 | 9.200 | 10.0 |
| Informática | 84.0 | 6.388095 | 3.137546 | 0.0 | 5.000 | 7.45 | 8.800 | 10.0 |
| Mecânica Industrial | 74.0 | 5.941892 | 3.143378 | 0.0 | 3.125 | 6.60 | 8.675 | 10.0 |
| Química | 77.0 | 7.963636 | 2.030745 | 0.2 | 6.800 | 8.60 | 9.500 | 10.0 |
| Telecomunicações | 81.0 | 6.938272 | 2.262055 | 0.0 | 6.300 | 7.40 | 8.600 | 10.0 |

Fonte: Dados coletados do estudo.

**Tabela 6 -** Estatísticas Descritivas da Nota em Matemática por Semestre.

| Semestre | N (count) | Média (mean) | Desvio Padrão (std) | Mínimo (min) | 1° Quartil (25%) | Mediana (50%) | 3° Quartil (75%) | Máximo (max) |
|---|---|---|---|---|---|---|---|---|
| 2018.2 | 103.0 | 6.617476 | 1.272604 | 0.59 | 6.0500 | 6.850 | 7.350 | 8.48 |
| 2019.1 | 110.0 | 7.067909 | 1.058272 | 1.22 | 6.5225 | 7.155 | 7.835 | 8.85 |
| 2019.2 | 118.0 | 7.339322 | 0.985214 | 4.49 | 6.8400 | 7.560 | 8.010 | 8.99 |
| 2020.1 | 137.0 | 8.079416 | 0.816043 | 5.93 | 7.6200 | 8.210 | 8.660 | 9.50 |

Fonte: Dados coletados do estudo.

**Tabela 7 -** Estatísticas Descritivas da Nota em Matemática por Semestre.

| Semestre | N (count) | Média (mean) | Desvio Padrão (std) | Mínimo (min) | 1° Quartil (25%) | Mediana (50%) | 3° Quartil (75%) | Máximo (max) |
|---|---|---|---|---|---|---|---|---|
| 2018.2 | 103.0 | 6.088350 | 3.037585 | 0.00 | 4.700 | 6.90 | 8.400 | 10.00 |
| 2019.1 | 110.0 | 6.730909 | 2.461772 | 0.00 | 5.825 | 7.10 | 8.700 | 10.00 |
| 2019.2 | 118.0 | 6.774576 | 2.662310 | 0.00 | 6.000 | 7.25 | 8.975 | 10.00 |
| 2020.1 | 137.0 | 7.741606 | 2.133784 | 0.00 | 7.300 | 8.30 | 9.000 | 10.00 |

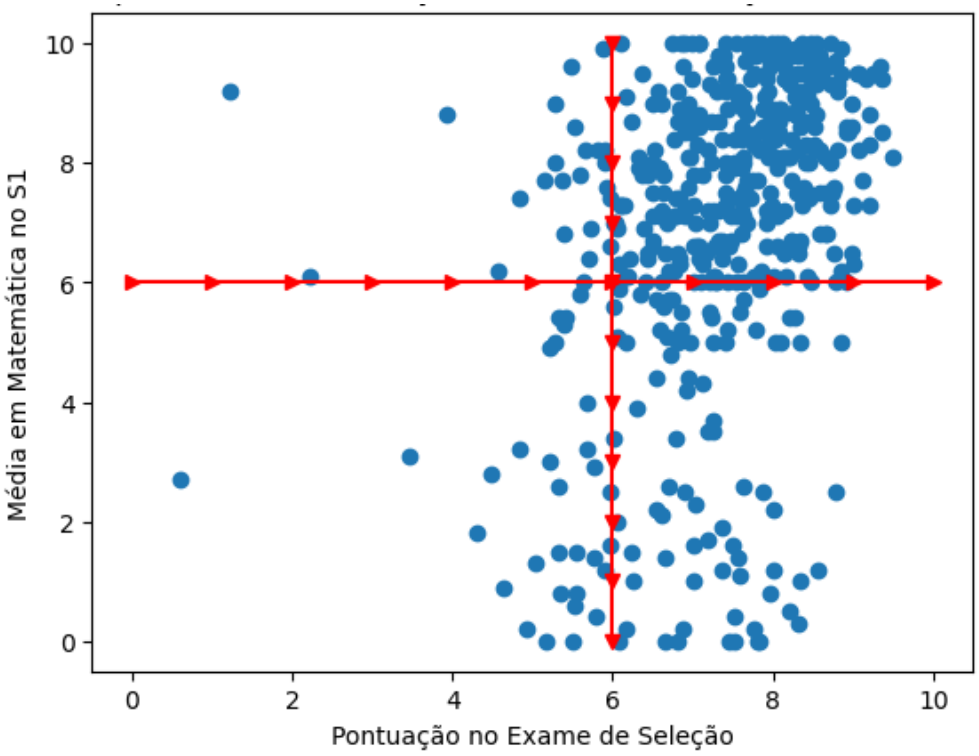
Fonte: Dados coletados do estudo.

O conjunto de dados analisado, composto por 468 registros completos, abrange variáveis categóricas (Semestre, Reserva de Vaga, Curso) e numéricas (Nota de Matemática, Pontuação). A análise descritiva revelou diferenças significativas entre as categorias, mas não indicou a presença de anomalias evidentes, como valores extremamente discrepantes ou inconsistências nos dados. Em relação à Reserva de Vaga, candidatos da Ampla Concorrência apresentaram, em média, melhores desempenhos tanto na pontuação geral (7.91) quanto na nota de Matemática (7.59), com menor dispersão. Algumas categorias, como L3 e L7, possuem poucos registros e valores ausentes no desvio padrão, limitando conclusões mais robustas, mas sem indicar valores inesperados. No recorte por Curso Técnico Integrado, cursos como Informática e Química destacaram-se com as maiores médias, enquanto Edificações teve o menor desempenho médio em Matemática (6.55). A variabilidade das notas sugere diferenças no perfil dos alunos e na abordagem pedagógica entre os cursos, mas sem indicar padrões incomuns.

Ao analisar os semestres, observou-se uma tendência de crescimento nas médias da pontuação e da nota de Matemática ao longo do tempo, chegando a 8.07 e 7.74, respectivamente, no semestre 2020.1. A redução na dispersão sugere maior homogeneidade no desempenho, possivelmente refletindo melhorias na preparação dos alunos ou ajustes nos critérios avaliativos. No entanto, não foram detectadas anomalias evidentes, como outliers extremos ou falhas na escala dos dados, indicando que o conjunto de dados é estatisticamente consistente e adequado para análises mais aprofundadas. Assim, os resultados reforçam a influência da categoria de ingresso e do curso no desempenho dos alunos, mas sem apontar distorções inesperadas nos dados coletados.

Assim, foi plotado o Gráfico de Dispersão entre Pontuação no Exame de Seleção e Nota em Matemática, conforme a Figura 1 a seguir:

**Figura 1 -** Gráfico de Dispersão entre Pontuação no Exame de Seleção e Média em Matemática.




Fonte: Dados coletados do estudo.

O gráfico de dispersão apresenta a relação entre a Pontuação no Exame de Seleção e a Média em Matemática no primeiro semestre (S1). Cada ponto representa um estudante, e observa-se que há uma concentração significativa de pontos na região superior direita, indicando que muitos alunos com alta pontuação no exame também obtiveram boas médias em Matemática no S1.

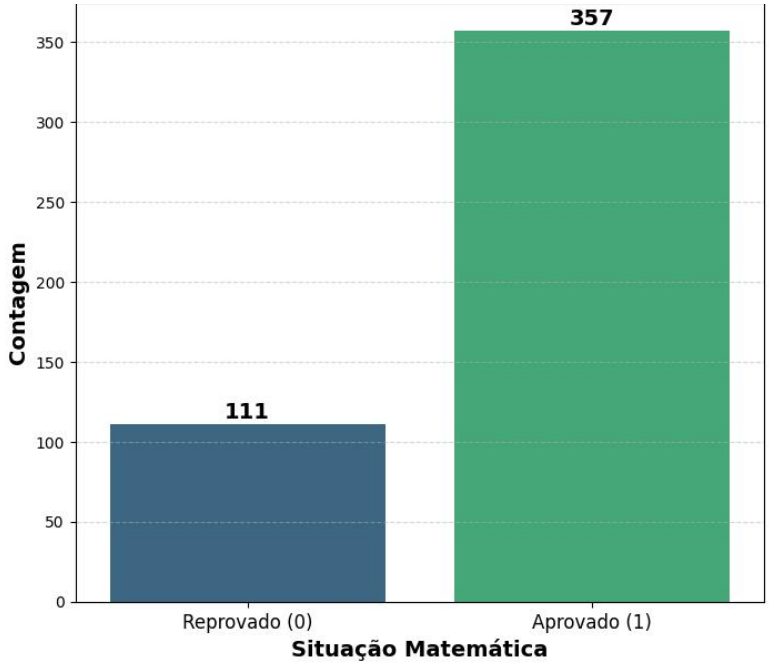
Os eixos vermelhos traçam as médias das duas variáveis, permitindo visualizar a distribuição dos dados em relação a esses valores centrais. Nota-se que a maior parte dos estudantes está acima da média da pontuação no exame e da média da disciplina, mas há uma dispersão considerável abaixo desses valores.

Além disso, percebe-se que, embora exista uma tendência de correlação positiva entre as variáveis, há alunos com baixa pontuação no exame que alcançaram boas médias em Matemática e vice-versa. Esse padrão pode sugerir que, apesar da pontuação no exame ser um indicador inicial de desempenho, outros fatores influenciam o rendimento acadêmico ao longo do curso.

**Balanceamento das Classes**

Os conjuntos de dados iniciais apresentavam um desequilíbrio entre as classes. Para mitigar esse problema, aplicou-se a técnica de oversampling utilizando SMOTE (Synthetic Minority Over-sampling Technique), conforme o Gráfico 2.

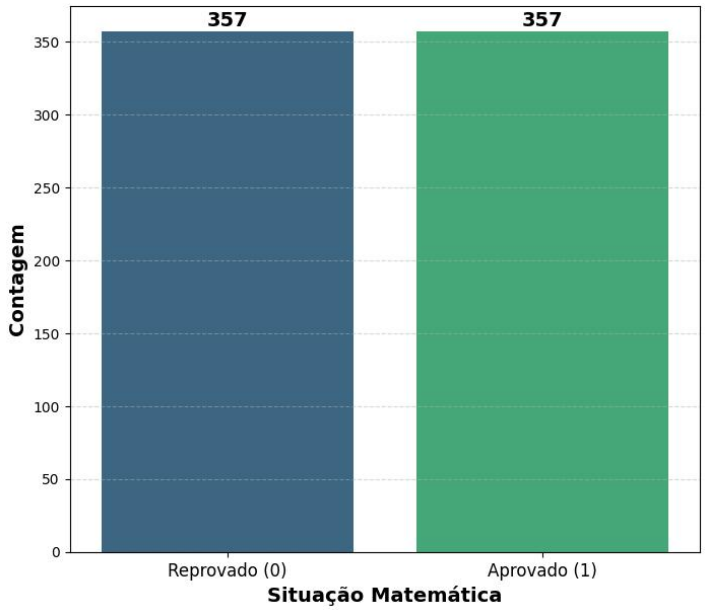
**Figura 2 -** Distribuição das Classes de Situação Matemática.




Fonte: Dados coletados do estudo.

Após o balanceamento, obteve-se os seguintes resultados no Gráfico exposto na Figura 3.

**Figura 3 -** Distribuição das Classes de Situação Matemática após o balanceamento.




Fonte: Dados coletados do estudo.

**Seleção das Variáveis**

Ao aplicarmos o algoritmo SelectKBest ao conjunto de variáveis preditoras, obtivemos a seguinte tabela de escores, na qual cada variável foi ranqueada de acordo com sua relevância para o modelo. *(Inserir Imagem 1 - Tabela de Scores das Variáveis)*, conforme a Tabela 8.

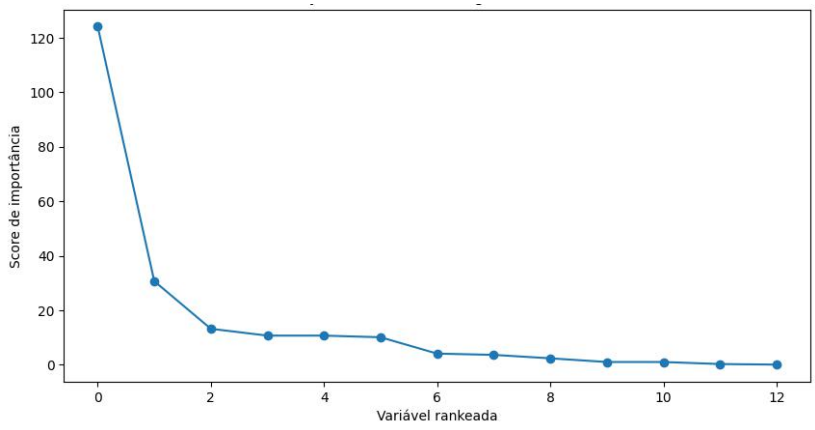
**Tabela 8 -** Escores das variáveis.

| Variável | Score |
|---|---|
| Pontuação | 124.20 |
| Curso_Técnico_Integrado_em_Química_Campus_Fortaleza | 30.69 |
| Curso_Técnico_Integrado_em_Eletrotécnica_Campus_Fortaleza | 13.20 |
| Reserva_vaga_L6 | 10.69 |
| Curso_Técnico_Integrado_em_Mecânica_Industrial | 10.69 |
| Curso_Técnico_Integrado_em_Telecomunicações | 10.08 |
| Reserva_vaga_L2 | 4.09 |
| Reserva_vaga_L1 | 3.61 |
| Curso_Técnico_Integrado_em_Informática_Campus_Fortaleza | 2.35 |
| Reserva_vaga_L3 | 1.00 |
| Reserva_vaga_L7 | 1.00 |
| Reserva_vaga_L8 | 0.26 |
| Reserva_vaga_L4 | 0.04 |

Fonte: Dados coletados do estudo.

Em seguida, plotamos o gráfico abaixo e, com base na Regra do Cotovelo, identificamos o ponto de inflexão e selecionamos as variáveis com escores acima de 10, garantindo que apenas as mais relevantes fossem mantidas. Essa abordagem está alinhada com os limiares de scores gerados pelo algoritmo SelectKBest, onde foram analisados os pontos de corte nos valores 5, 10 e 20. *(Inserir Imagem 2 - Gráfico da Regra do Cotovelo)*.

**Figura 4 -** Seleção de Variáveis (Regra do Cotovelo).




Fonte: Dados coletados do estudo.

A seleção das variáveis mais influentes se alinha com os padrões observados na análise estatística descritiva e nos gráficos exploratórios previamente apresentados. A Pontuação no Exame de Seleção se destaca como a variável de maior impacto, reforçando sua forte correlação com o desempenho acadêmico, conforme evidenciado pelo gráfico de dispersão, no qual alunos com pontuação mais alta tendem a obter melhores médias em Matemática. *(Inserir Imagem 3 - Gráfico de*

*Dispersão entre Pontuação e Média em Matemática)*.

Além disso, os cursos técnicos integrados, especialmente Química, Eletrotécnica e Mecânica Industrial, apresentaram altos escores, indicando que a formação escolhida influencia diretamente o desempenho dos estudantes. Esse achado é consistente com as estatísticas descritivas, que mostraram que os alunos desses cursos tendem a ter médias mais altas. No que se refere às categorias de reserva de vaga, observa-se que algumas delas possuem escores relevantes, o que sugere que a forma de ingresso pode impactar o desempenho acadêmico. As análises anteriores mostraram diferenças estatísticas entre os grupos de ampla concorrência e cotistas, justificando a manutenção dessas variáveis no modelo preditivo.

Para validar a escolha do limiar mais adequado na seleção das variáveis, testamos diferentes valores (1, 5, 10 e 20) e avaliamos a acurácia dos modelos Logistic Regression e Naive Bayes. Os resultados indicaram que o limiar 10.0 garantiu um equilíbrio entre simplificação do modelo e manutenção de um desempenho consistente, uma vez que apresentou valores estáveis de acurácia, próximos aos melhores resultados observados. *(Inserir Imagem 4 - Tabela de Limiar, Modelo e Acurácia)*.

Portanto, a seleção das variáveis utilizando a Regra do Cotovelo e o critério de corte do *SelectKBest* se alinha às observações feitas na estatística descritiva e na performance dos modelos testados, garantindo que o modelo final retenha apenas os fatores que efetivamente influenciam os resultados acadêmicos, evitando ruídos e contribuindo para maior precisão nas previsões.

A seguir, a Tabela 9 é de Limiar, modelo e acurácia.

**Tabela 9 -** Tabela de Limiar, Modelo e Acurácia.

| Limiar | Modelo | Acurácia |
|---|---|---|
| 1.0 | Logistic Regression | 0.664336 |
| 1.0 | Naive Bayes | 0.622378 |
| 5.0 | Naive Bayes | 0.657343 |
| 5.0 | Logistic Regression | 0.650350 |
| 10.0 | Naive Bayes | 0.657343 |
| 10.0 | Logistic Regression | 0.650350 |
| 20.0 | Naive Bayes | 0.643357 |
| 20.0 | Logistic Regression | 0.615385 |

Fonte: Dados coletados do estudo.

**Normalização dos Dados em Aprendizado de Máquina**

Após a seleção das variáveis, realizamos a normalização dos dados utilizando a função StandardScaler(), que transforma os valores para uma distribuição com média zero e desvio padrão unitário, garantindo que todas as variáveis possuam escalas comparáveis. Esse procedimento é essencial para algoritmos que são sensíveis à magnitude das variáveis, como Regressão Logística e KNN (Géron, 2019).

**Treinamento e Avaliação dos Modelos**

Os modelos foram treinados utilizando as seis variáveis selecionadas, conforme descrito na seção Seleção das Variáveis. Para a construção dos modelos preditivos, foram empregados os algoritmos Regressão Logística (LogisticRegression) e Naive Bayes Gaussiano (GaussianNB), ambos disponíveis na biblioteca Scikit-learn (Pedregosa *et al*., 2011). Esses modelos são amplamente utilizados para tarefas de classificação supervisionada, sendo referenciados em diversos estudos acadêmicos e literaturas especializadas (Morettin & Singer, 2021; Géron, 2019).

Para garantir a reprodutibilidade e avaliar adequadamente a generalização dos modelos, os dados foram divididos em 80% para treinamento e 20% para teste (Géron, 2019; Scikit-learn, 2024). O particionamento foi realizado sem validação cruzada, utilizando a função train_test_split da biblioteca Scikit-learn (Scikit-learn, 2024), conforme mostrado a seguir:

X_treinamento_over, X_teste_over, y_treinamento_over, y_teste_over = train_test_split(X_over, y_over, test_size=0.20, random_state=0)

As dimensões dos conjuntos resultantes foram verificadas com:

X_treinamento_over.shape, X_teste_over.shape

Essa divisão permitiu avaliar a capacidade dos modelos de generalizar para novos conjuntos de dados, assegurando uma comparação justa entre as abordagens testadas (Géron, 2019; Scikit-learn, 2024).

No processo de validação cruzada, utilizamos a função cross_val_score, conforme o código abaixo:

scores = cross_val_score(Modelo, X_selected_scaled, y_over, cv=10)

Essa abordagem permite avaliar o desempenho do modelo em diferentes subconjuntos de dados, reduzindo o risco de overfitting e garantindo uma avaliação mais robusta da capacidade preditiva (Scikit-learn, 2024; Géron, 2019).

Após o treinamento, os modelos foram avaliados no conjunto de teste, onde analisamos métricas essenciais para a avaliação do desempenho preditivo, incluindo Precisão (Precision), Recall, F1-Score, Acurácia, Acurácia na Validação Cruzada e Curva ROC.

As métricas Precisão, Recall, F1-Score e Acurácia foram obtidas por meio da matriz de confusão, utilizando o seguinte procedimento:

plot_confusion_matrix(y_teste_over, y_pred1_test, "Modelo - Teste")

Já a Curva ROC, utilizada para avaliar a capacidade discriminativa dos modelos, foi calculada com:

fpr, tpr, thresholds = roc_curve(y_treinamento_over, y_scores)

roc_auc = auc(fpr, tpr)

A validação dessas métricas possibilitou uma análise detalhada do equilíbrio entre falsos positivos e falsos negativos, além da capacidade do modelo de generalizar para novos conjuntos de dados. Todo o processo foi realizado com o suporte da biblioteca Scikit-learn, que disponibiliza ferramentas robustas para avaliação de modelos de aprendizado de máquina (Scikit-learn, 2024; Géron, 2019).

Os resultados detalhados dessas métricas estão apresentados na Tabela 10 a seguir.

**Tabela 10 -** Resultados das métricas para avaliação dos modelos de aprendizagem de máquina.

| Métrica | Logistic Regression | Naive Bayes |
|---|---|---|
| Precision (Classe 0) | 0.78 | 0.74 |
| Precision (Classe 1) | 0.68 | 0.67 |
| Recall (Classe 0) | 0.67 | 0.67 |
| Recall (Classe 1) | 0.79 | 0.74 |
| F1-Score (Classe 0) | 0.72 | 0.74 |
| F1-Score (Classe 1) | 0.73 | 0.70 |
| Acurácia | 0.73 | 0.70 |
| Macro Avg | 0.73 | 0.70 |
| Weighted Avg | 0.73 | 0.70 |
| Média da Acurácia na Validação Cruzada | 0.68 | 0.65 |
| Curva ROC | 0.77 | 0.66 |

Fonte: Dados coletados do estudo.

A análise dos resultados apresentados na Tabela 10 demonstra que a Regressão Logística superou o Naive Bayes em diversos aspectos, apresentando maior eficiência na previsão de alunos em risco de reprovação. Para avaliar o desempenho dos modelos, analisamos métricas essenciais, como Precisão (Precision), Recall, F1-Score, Acurácia, Validação Cruzada e Curva ROC, conforme detalhado a seguir.

**Precisão e Recall**

A Regressão Logística apresentou maior precisão para a Classe 0 (aprovações), atingindo um valor de 0,78, enquanto o Naive Bayes obteve 0,74. Isso indica que a Regressão Logística gera menos falsos positivos, ou seja, tem menor probabilidade de prever erroneamente a aprovação de alunos que deveriam ser reprovados.

Além disso, ao analisarmos o recall para a Classe 1 (reprovações), a Regressão Logística novamente se destaca, alcançando um valor de 0,79, enquanto o Naive Bayes registrou 0,74. Esse resultado indica que a Regressão Logística é mais sensível e consegue identificar corretamente mais alunos com risco de reprovação.

**F1-Score e Acurácia**

O F1-Score, que representa o equilíbrio entre precisão e recall, também favorece a Regressão Logística, que apresentou valores superiores em ambas as classes. Isso reforça a consistência do modelo, garantindo que a taxa de erro seja reduzida em diferentes cenários.

Além disso, ao avaliarmos a Acurácia Geral, observamos que a Regressão Logística atingiu 0,73, enquanto o Naive Bayes alcançou 0,70. Esse resultado confirma que a Regressão Logística possui uma maior taxa de acertos globais, tornando-a uma opção mais confiável para prever corretamente se um aluno será aprovado ou reprovado.

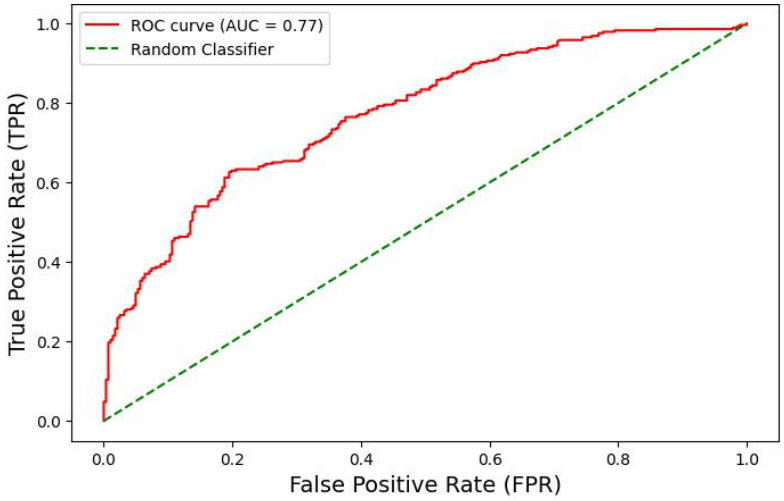
**Validação Cruzada**

A validação cruzada revelou que a Regressão Logística apresentou maior estabilidade, com um desempenho médio de 0,68, enquanto o Naive Bayes obteve 0,65. Esse resultado sugere que a Regressão Logística tem menor variação de desempenho entre diferentes subconjuntos de dados, garantindo maior generalização quando aplicada a novos conjuntos de dados.

**Curva ROC e Discriminação do Modelo**

A Curva ROC, que avalia a capacidade do modelo de distinguir corretamente entre alunos aprovados e reprovados, mostrou uma diferença significativa entre os modelos. A Regressão Logística apresentou um valor de 0,77, Figura 5 , enquanto o Naive Bayes obteve 0,66, Figura 6.

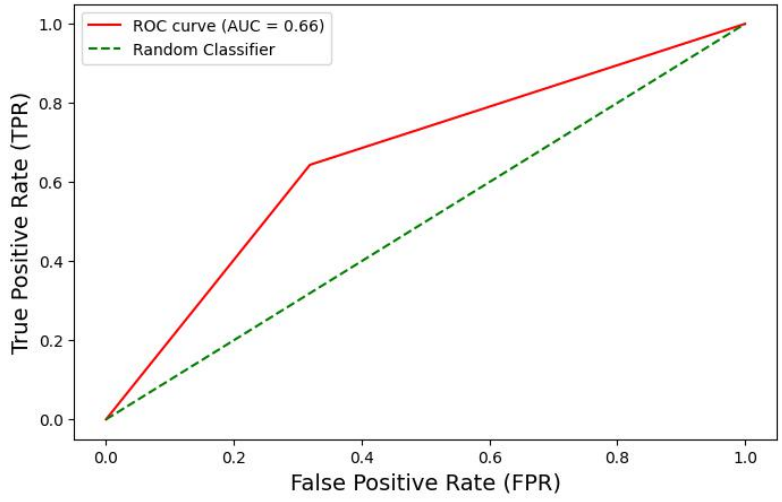
**Figura 5 -** Curva ROC Regressão Logística.




Fonte: Dados coletados do estudo.

A Figura 6, a seguir, apresenta a curva ROC Naive Bayes.

**Figura 6 -** Curva ROC Naive Bayes




Fonte: Dados coletados do estudo.

Esse resultado evidencia que a Regressão Logística tem melhor discriminação entre as classes, sendo mais eficaz na separação entre alunos com alto risco de reprovação e aqueles que possuem maior probabilidade de aprovação.

## 4. Conclusão

Neste estudo, exploramos a aplicação de técnicas de Machine Learning para prever a retenção de alunos em disciplinas de matemática no Instituto Federal do Ceará (IFCE), com o intuito de mitigar a evasão escolar. Utilizamos Regressão Logística e Naive Bayes, modelos amplamente empregados em problemas de classificação, para identificar padrões que possam auxiliar a gestão acadêmica na formulação de estratégias preventivas.

Um dos destaques deste trabalho foi a seleção de variáveis, essencial para garantir a eficiência e interpretabilidade dos modelos. Para esse fim, combinamos o método SelectKBest com a Regra do Cotovelo (*Elbow Method*), permitindo identificar o número ótimo de variáveis preditoras. Os resultados demonstraram que essa abordagem não apenas reduziu a dimensionalidade do conjunto de dados, mas também manteve ou melhorou a acurácia e a capacidade de generalização dos modelos. Essa metodologia se mostrou eficaz na escolha das variáveis mais relevantes, evitando o uso excessivo de atributos irrelevantes e reduzindo a complexidade computacional.

Além disso, a normalização dos dados e a aplicação da técnica SMOTE para balanceamento das classes foram fundamentais para aprimorar o desempenho dos modelos, minimizando viéses e garantindo previsões mais equilibradas. A avaliação das métricas, incluindo F1-score, matriz de confusão e ROC-AUC, reforçou que a Regressão Logística apresentou um desempenho superior ao Naive Bayes, sendo mais eficaz na identificação de alunos em risco de reprovação.

Apesar dos avanços obtidos, algumas limitações devem ser consideradas, como o tamanho da amostra e a impossibilidade de incluir fatores externos que podem influenciar o desempenho dos alunos, como aspectos socioeconômicos e metodologias de ensino. Para estudos futuros, recomenda-se a ampliação do conjunto de dados e a experimentação com redes neurais artificiais e modelos mais avançados de aprendizado profundo, buscando aprimorar ainda mais a precisão preditiva.

Por fim, os resultados deste trabalho demonstram que o uso de Machine Learning na educação pode ser um aliado valioso na identificação precoce de alunos em risco de reprovação, permitindo intervenções pedagógicas mais assertivas. A adoção de métodos como SelectKBest aliado à Regra do Cotovelo contribui para a construção de modelos mais eficientes, facilitando a tomada de decisão e promovendo um ensino mais inclusivo e adaptado às necessidades dos estudantes

## Referências

Akamine, C. T. & Yamamoto, R. K. (2009). *Estudo dirigido: estatística descritiva.* (3ed). Editora Érica.

Batista, A.S. (2015). *Regressão Logística: Uma introdução ao modelo estatístico - Exemplo de aplicação ao Revolving Credit*. Vida Economica Editorial. Recuperado de https://books.google.com.br/books?id=EtAsCgAAQBAJ.

Brasil. (2018). *Lei nº 13.709, de 14 de agosto de 2018*. Dispõe sobre a proteção de dados pessoais e altera a Lei nº 12.965, de 23 de abril de 2014 (Marco Civil da Internet).Brasília, DF: Presidência da República. Recuperado de: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm.

Bussab, Wilton de O.; Morettin, Pedro A. (2010). Estatística básica. In: *Estatística básica*. p. xvi, 540-xvi, 540.

Burkov, A. (2019). *The Hundred-Page Machine Learning Boo*k. Ed. Andriy Burkov. ISBN-13: 978-1999579500.

Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research (JAIR), 16,* 321-357. https://doi.org/10.1613/jair.953

Costa, N. ., Nascimento, M. ., Nascimento, A. C. ., Azevedo, C. ., & Suela, M. . (2024). Classificadores Híbridos Baseados no Naive Bayes Avaliados em Diferentes Níveis de Dependência Entre Variáveis. *Enciclopedia Biosfera*, *21*(47), 47-61. https://conhecer.org.br/ojs/index.php/biosfera/article/view/5749

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches.* (4th ed.). Sage Publications.

Faceli, K., Lorena, A. C., Gama, J., & Carvalho, A. C. P. de L. F. de. (2011). I*nteligência artificial: uma abordagem de aprendizado de máquina.* Rio de Janeiro: LTC.

Felizardo, L. F., Gualberto, D. R., & Carrano, D. P. (2024). Redes neurais artificiais em ambientes educacionais: business intelligence aplicado à pedagogia. *Caderno Pedagógico*, *21*(6), e5210. https://doi.org/10.54033/cadpedv21n6-258

Fernández, A., García, S., Herrera, F., Chawla N. V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research*, 61, 863-905.

Galdino, M. V. (2018). *Modelos probabilísticos e não probabilísticos de classificação binária para pacientes com ou sem demência como auxílio na prática clínica em geriatria* [Tese de Doutorado, Universidade Estadual Paulista]. Repositório UNESP.

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media.

Gil, A. C. (2017). *Como elaborar projetos de pesquisa*. (6ed). Editora Atlas.

Hair, Joseph F. et al. (2009). *Análise multivariada de dados*. Bookman editora.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263-1284. https://doi.org/10.1109/TKDE.2008.239

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley.

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets, 31*, 685–695. https://doi.org/10.1007/s12525-021-00475-2.

Mckinney, Wes. *Python para análise de dados: Tratamento de dados com Pandas, NumPy e IPython*. Novatec Editora, 2018.

Melgaço da Silva, L., & Ciasca, M. (2021). história da educação profissional no Brasil: do período colonial ao Governo Michel Temer (1500-2018). *Educação Profissional E Tecnológica Em Revista*, *5*(1), 73-101. https://doi.org/10.36524/profept.v5i1.677

Mendoza Lira, Michelle, Quiroz Muñoz, Javiera, Muñoz Pérez, Daniela, Contreras Pérez, Camila, & Ballesta Acevedo, Emilio. (2023). Conceituações de evasão e retenção escolar segundo uma escola primária no Chile. *Cadernos de Pesquisa Educacional* , *14* (2), e206. Epub 1 de dezembro de 2023. https://doi.org/10.18861/cied.2023.14.2.3368.

Montgomery, D. C., & Runger, G. C. (2010). *Applied Statistics and Probability for Engineers*. John Wiley & Sons.

Moore, D. S., McCabe, G. P., & Craig, B. A. (2014). *Introduction to the Practice of Statistics* (8th ed.). W.H. Freeman and Company.

Morettin, Pedro A.; Singer, Julio M. (2021). *Estatıstica e Ciência de Dados*. Texto Preliminar, IME-USP.

Müller, A. C. & Guido, S. (2017). *Introduction of machine leaning with phyton*. Published by O'Reilly Media, Inc. https://www.nrigroupindia.com/e-book/Introduction%20to%20Machine%20Learning%20with%20Python%20(%20PDFDrive.com%20)-min.pdf.

Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830, 2011. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html.

Pereira A. S. et al. (2018). Metodologia da pesquisa científica. [free e-book]. Editora da UAB/NTE/UFSM.

Sabry, F. (2023). *Naive Bayes Classifier: Fundamentals and Applications* [Série Artificial Intelligence]. One Billion Knowledgeable.

Scikit-Learn. (2024). *Sklearn.feature_selection.SelectKBest*. https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html.

Scikit-Learn. (2024). *Scikit-learn: Machine Learning in Python*. https://scikit-learn.org/stable/.

Shitsuka et al. (2014). *Matemática fundamental para a tecnologia*. Editora Érica.

Tanimu, Jesse Jeremiah et al. A machine learning method for classification of cervical cancer. *Electronics*, 11(3), 463, 2022.
Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.

Yin, R. K. (2018). *Case study research and applications: Design and methods*. (6th ed.). Sage Publications.