UNIVERSIDADE FEDERAL DE PERNAMBUCO

Universidade Federal de Pernambuco
Centro de Tecnologia e Geociências
Departamento de Eletrônica e Sistemas

Graduação em Engenharia Eletrônica

Lais de Morais Coutinho Silva

Diagnóstico e predição do rendimento estudantil no ensino médio: uma estrutura analítica para interpretação e modelagem preditiva

Recife

2025

Lais de Morais Coutinho Silva

Diagnóstico e predição do rendimento estudantil no ensino médio: uma estrutura analítica para interpretação e modelagem preditiva

Trabalho de Conclusão apresentado ao Curso de Graduação em Engenharia Eletrônica, do Departamento de Eletrônica e Sistemas, da Universidade Federal de Pernambuco, como requisito parcial para obtenção do grau de Bacharel em Engenharia Eletrônica.

Orientador(a): Prof. Marcos Antonio Martins de Almeida, D.Sc

Recife
2025

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Silva, Lais de Morais Coutinho.
Diagnóstico e predição do rendimento estudantil no ensino médio: uma estrutura analítica para interpretação e modelagem preditiva / Lais de Morais Coutinho Silva. - Recife, 2025.
173 p. : il., tab.

Orientador(a): Marcos Antonio Martins de Almeida
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Tecnologia e Geociências, Engenharia Eletrônica - Bacharelado, 2025.
Inclui referências, apêndices.

1. Mineração de Dados Educacionais. 2. Aprendizado de Máquina. 3. Predição de Desempenho Escolar. 4. Metodologia e Processos de Ciência de Dados. I. Almeida, Marcos Antonio Martins de. (Orientação). II. Título.

000 CDD (22.ed.)

Lais de Morais Coutinho Silva

Diagnóstico e predição do rendimento estudantil no ensino médio: uma estrutura analítica para interpretação e modelagem preditiva

Trabalho de Conclusão apresentado ao Curso de Graduação em Engenharia Eletrônica, do Departamento de Eletrônica e Sistemas, da Universidade Federal de Pernambuco, como requisito parcial para obtenção do grau de Bacharel em Engenharia Eletrônica.

Aprovado em: 02/10/2025

Banca Examinadora

Prof. Marcos Antonio Martins de Almeida, D.Sc
Universidade Federal de Pernambuco

Prof. Fernanda Maria Ribeiro de Alencar, D.Sc.
Universidade Federal de Pernambuco

Prof. Patrícia Silva Lessa, D.Sc
Universidade Federal de Pernambuco

Se a educação sozinha não
transforma a sociedade, sem ela
tampouco a sociedade muda.

Paulo Freire

Resumo do Trabalho de Conclusão de Curso apresentado ao Departamento de Eletrônica e Sistemas, como parte dos requisitos necessários para a obtenção do grau de Bacharel em Engenharia Eletrônica(Eng.)

Diagnóstico e predição do rendimento estudantil no ensino médio: uma estrutura analítica para interpretação e modelagem preditiva

Lais de Morais Coutinho Silva

A mineração de dados educacionais aplica técnicas computacionais para extrair conhecimento de dados educacionais visando melhorar processos de ensino-aprendizagem. Este trabalho desenvolveu um *framework* analítico – conjunto modular e reutilizável de ferramentas – para diagnóstico e predição do rendimento estudantil no ensino médio, utilizando dados de 677 estudantes portugueses em português e matemática. A metodologia seguiu o processo CRISP-DM, desenvolvendo o índice PerfilScore para seleção de variáveis categóricas e implementando três estratégias complementares de seleção de atributos. Foram aplicados cinco algoritmos de classificação supervisionada (Regressão Logística, Support Vector Machine, Árvore de Decisão, Random Forest e AdaBoost) com otimização de hiperparâmetros via *grid search* e validação cruzada estratificada. Os modelos apresentaram melhor desempenho preditivo em português que em matemática, com o SVM demonstrando os melhores resultados. A análise identificou como principais fatores preditivos o histórico de reprovações, escolaridade dos pais e aspectos comportamentais como tempo livre e consumo de álcool. O *framework* analítico desenvolvido indica que sistemas preditivos baseados em aprendizado de máquina podem contribuir para a identificação precoce de estudantes em risco e auxiliar no desenvolvimento de intervenções educacionais mais direcionadas.

Palavras-chave: mineração de dados educacionais; aprendizado de máquina; predição de desempenho escolar; seleção de atributos; CRISP-DM

Abstract of Course Conclusion Work, presented to Departament of Eletronic and Systems, as a partial fulfillment of the requirements for the degree of Bachelor of Electronic Engineering (Eng.)

Diagnosis and Prediction of Student Performance in High School: An Analytical Framework for Interpretation and Predictive Modeling

Lais de Morais Coutinho Silva

Educational data mining applies computational techniques to extract knowledge from educational data aiming to improve teaching-learning processes. This work developed an analytical framework – a modular and reusable set of tools – for diagnosing and predicting high school student performance, using data from 677 Portuguese students in Portuguese language and mathematics courses. The methodology followed the CRISP-DM process, developing the PerfilScore index for categorical variable selection and implementing three complementary feature selection strategies. Five supervised classification algorithms were applied (Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, and AdaBoost) with hyperparameter optimization via grid search and stratified cross-validation. The models showed better predictive performance in Portuguese than in mathematics, with SVM demonstrating the best results. The analysis identified failure history, parental education, and behavioral aspects such as free time and alcohol consumption as the main predictive factors. The developed analytical framework indicates that machine learning-based predictive systems can contribute to the early identification of at-risk students and assist in developing more targeted educational interventions.

Keywords: educational data mining; machine learning; student performance prediction; feature selection; CRISP-DM

Lista de Figuras

Lista de Tabelas

Lista de Quadros

Lista de Abreviações

AED Análise Exploratória de Dados

AHP *Analytic Hierarchy Process*

AI Inteligência Artificial (*Artificial Intelligence*)

AIC Critério de Informação de Akaike (*Akaike Information Criterion*)

ANOVA Análise de Variância (*Analysis of Variance*)

AUC-PR Área sob a Curva *Precision-Recall*

AUC-ROC Área sob a Curva da Característica de Operação do Receptor (*Area Under the Receiver Operating Characteristic*)

BIC Critério de Informação Bayesiano (*Bayesian Information Criterion*)

BWM *Best-Worst Method*

CRISP-DM *Cross-Industry Standard Process for Data Mining*

CV Coeficiente de Variação

EAD Educação a Distância

EDM *Educational Data Mining*

EJA Educação de Jovens e Adultos

EWS *Early Warning Systems*

F1-Score Métrica harmônica entre precisão e sensibilidade (*recall*)

FN Falsos Negativos

FOMO *Fear of Missing Out*

FP Falsos Positivos

GPA *Grade Point Average*

IA Inteligência Artificial

INEP Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

IQR Intervalo Interquartil (*Interquartile Range*)

LA *Learning Analytics*

LIME *Local Interpretable Model-agnostic Explanations*

LMS *Learning Management Systems*

MCDA Análise Multicritério de Decisão (*Multi-Criteria Decision Analysis*)

ML *Machine Learning*

OECD Organização para a Cooperação e Desenvolvimento Econômico (*Organisation for Economic Co-operation and Development*)

PC1 Primeiro Componente Principal

PC2 Segundo Componente Principal

PCA Análise de Componentes Principais (*Principal Component Analysis*)

PISA Programa Internacional de Avaliação de Estudantes (*Programme for International Student Assessment*)

RBF Função de Base Radial (*Radial Basis Function*)

RMSE Erro Quadrático Médio (*Root Mean Square Error*)

SAMME *Stagewise Additive Modeling using a Multi-class Exponential loss*

SAMME.R *SAMME Real* (variante do algoritmo SAMME)

SHAP *SHapley Additive exPlanations*

SPP *Student Performance Prediction*

STEM *Science, Technology, Engineering, and Mathematics*

SVM *Support Vector Machine*

TN Verdadeiros Negativos

TP Verdadeiros Positivos

UCI *University of California, Irvine Machine Learning Repository*

VIF Fator de Inflação da Variância (*Variance Inflation Factor*)

XAI Inteligência Artificial Explicável (*Explainable Artificial Intelligence*)

Sumário

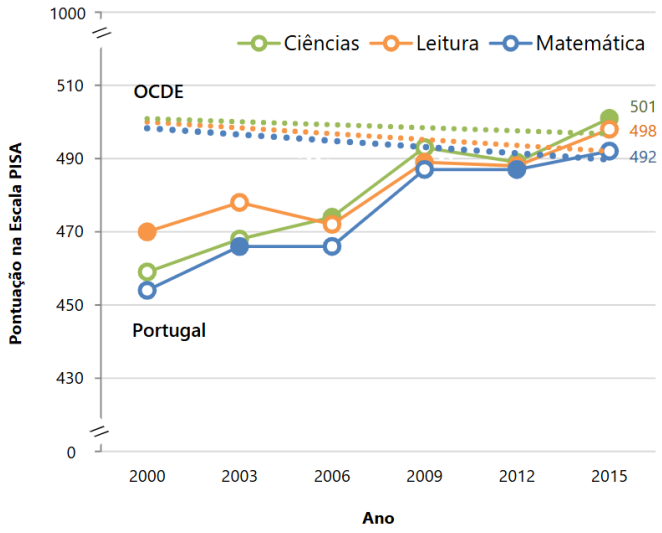
Capítulo 1

Introdução

Datificação é o processo de transformar aspectos do mundo anteriormente não quantificados em dados, tal fenômeno tem promovido uma mudança disruptiva na sociedade (IGUAL; SEGUÍ, 2017). A inteligência artificial e a ciência de dados foram amplamente incorporadas em diferentes domínios, desde tarefas administrativas até sistemas complexos. O acesso a grandes volumes de dados e os avanços no aprendizado de máquina impulsionaram progressos em áreas como visão computacional, sistemas de recomendação, diagnóstico médico e previsão financeira, impactando diversas disciplinas científicas e de engenharia (KOLLMANNSBERGER et al., 2021).

Seguindo essa tendência, a tomada de decisões institucionais tornou-se orientada por dados, inclusive no setor educacional, que adota técnicas como a mineração de dados e o aprendizado de máquina (BATISTA; FAGUNDES, 2023). A educação, por sua vez, exerce papel estratégico no desenvolvimento socioeconômico, ao contribuir para a produtividade, a redução das desigualdades e o crescimento sustentável. Compreender os fatores que afetam o desempenho estudantil é essencial para orientar políticas públicas e práticas pedagógicas eficazes.

Figura 1.1: Evolução das pontuações médias nos domínios do PISA em Portugal face às médias da OCDE.



Fonte: Instituto de Avaliação Educativa (IAVE) (2016).

Iniciativas como o PISA, programa internacional de avaliação de estudantes, conduzido pela OECD, evidenciam a importância de medir e comparar resultados educacionais, fornecendo insumos para melhorias nos sistemas de ensino. Portugal, conforme exemplificado na Figura 1.1, registrou avanços significativos entre 2000 e 2015 — período em que os dados utilizados neste estudo foram coletados —, o que permite investigar fatores críticos associados à aprovação ou reprovação escolar .

Nesse cenário, o presente estudo investiga a aplicação de modelos de aprendizado de máquina para prever a aprovação ou reprovação de estudantes portugueses do ensino médio, com base em dados demográficos, socioeconômicos e comportamentais nas disciplinas de português e matemática.

1.1 Justificativa

O presente estudo justifica-se pela sua relevância científica e social ao propor uma abordagem orientada por dados para compreender o desempenho escolar em disciplinas fundamentais — Português e Matemática. Do ponto de vista científico, busca-se aprofundar a discussão sobre o uso de métodos computacionais e estatísticos

na identificação dos fatores que influenciam o rendimento acadêmico. Do ponto de vista social, visa-se promover o desenvolvimento de soluções mais eficazes para os desafios educacionais, subsidiando políticas baseadas em evidências e intervenções pedagógicas direcionadas.

1.2 Objetivos

Este trabalho tem como objetivo investigar a aplicação de modelos de aprendizado de máquina e estratégias de seleção de atributos na classificação binária do desempenho estudantil — aprovação ou reprovação — com base em dados de estudantes do ensino médio nas disciplinas de Língua Portuguesa e Matemática. A proposta envolve modelagem preditiva e análise exploratória, com o intuito de identificar padrões relevantes entre as características dos estudantes e seus resultados acadêmicos.

Para alcançar esse objetivo, desenvolveu-se uma estrutura analítica baseado no processo CRISP-DM, integrando etapas de exploração, preparação, seleção de atributos e modelagem de dados. Essa arquitetura modular organiza o fluxo analítico de forma clara e sistemática, com potencial de replicação em diferentes contextos educacionais.

Assim, o estudo busca contribuir para a compreensão dos fatores que influenciam o rendimento acadêmico e subsidiar a formulação de intervenções educacionais mais eficazes.

1.2.1 Objetivos Específicos

1. Realizar a análise exploratória dos dados para identificar padrões, distribuições, correlações e tendências relacionadas ao desempenho dos estudantes.

2. Aplicar e avaliar técnicas de seleção de atributos, considerando a relevância das variáveis, os efeitos da multicolinearidade e a influência de diferentes estratégias na performance e interpretabilidade dos modelos.

3. Implementar e otimizar algoritmos de classificação supervisionada — incluindo Regressão Logística, Árvores de Decisão, *Random Forest*, *AdaBoost* e *Support Vector Machine* (SVM) — para prever a aprovação ou reprovação dos estudantes.

4. Avaliar e comparar o desempenho dos modelos com base em métricas adequadas a dados desbalanceados — F1-Score Macro, F1-Score da classe minoritária, *Recall* e AUC-ROC —, utilizando validação cruzada e considerando o impacto do balanceamento de classes e dos diferentes conjuntos de atributos.

5. Interpretar os resultados obtidos, identificando os fatores mais relevantes para o desempenho estudantil e discutindo suas implicações para a compreensão do fenômeno educacional no contexto analisado.

1.3 Organização do Trabalho

Este trabalho está estruturado em cinco capítulos, conforme descrito a seguir:

Capítulo 1 – Introdução: apresenta o tema, a justificativa da pesquisa, a contextualização do problema educacional e da base de dados, além da definição dos objetivos gerais e específicos.

Capítulo 2 – Fundamentação teórica: apresenta o panorama da pesquisa em mineração de dados educacionais e discute os principais conceitos relacionados, como aprendizado de máquina, algoritmos de classificação, métricas de avaliação, técnicas de pré-processamento, tratamento de dados desbalanceados e critérios para seleção de atributos.

Capítulo 3 – Procedimentos Metodológicos: descreve o *framework* analítico preditivo desenvolvido e os procedimentos metodológicos adotados com base no processo CRISP-DM, abrangendo desde a exploração dos dados até a modelagem preditiva e a avaliação de desempenho. Inclui o ambiente computacional, as etapas de preparação dos dados, as estratégias de seleção de atributos, os algoritmos aplicados e os critérios de comparação entre os modelos.

Capítulo 4 – Resultados e discussão: apresenta e analisa os resultados obtidos com a aplicação das técnicas de aprendizado de máquina, destacando o desempenho dos modelos sob diferentes configurações metodológicas e conjuntos de atributos.

Capítulo 5 – Conclusão: reúne as considerações finais com base nos objetivos propostos e discute as limitações do estudo, as dificuldades enfrentadas e sugestões para pesquisas futuras.

Capítulo 2

Fundamentação Teórica

Este capítulo apresenta os fundamentos teóricos que sustentam o presente trabalho, abordando os conceitos centrais da Mineração de Dados Educacionais e da Análise de Aprendizagem, bem como as técnicas estatísticas e computacionais aplicadas na análise e modelagem dos dados. A estrutura teórica fundamenta as decisões metodológicas adotadas e contextualiza os resultados obtidos no cenário científico contemporâneo.

2.1 Frentes de análise de dados educacionais

Apesar da ampla disponibilidade de dados gerados no contexto educacional, muitas instituições de ensino ainda enfrentam dificuldades em utilizá-los de forma estratégica, frequentemente conduzindo análises com atrasos significativos – o que compromete a eficácia de intervenções pedagógicas (CAMBRUZZI, 2014). Esse cenário é agravado pelo crescimento contínuo do volume e da complexidade dos dados disponíveis (ALALAWI et al., 2023; ROMERO; VENTURA, 2010), o que demanda abordagens analíticas mais robustas e integradas, capazes de revelar padrões relevantes e subsidiar a tomada de decisões educacionais (ALALAWI et al., 2023).

Para enfrentar esses desafios, emergiram dois campos interligados: a Mineração de Dados Educacionais (EDM) e o *Learning Analytics* (LA).

2.1.1 Conceitos fundamentais e evolução histórica

A EDM constitui um campo interdisciplinar voltado à aplicação de algoritmos estatísticos, técnicas de aprendizado de máquina e métodos de mineração de dados para responder a questões educacionais (AHMED, 2024; ROMERO; VENTURA, 2010). Seu objetivo central é extrair conhecimento a partir de grandes volumes de dados gerados em ambientes educacionais, contribuindo para a compreensão dos estudantes e dos contextos de aprendizagem (ALALAWI et al., 2023).

Em paralelo, o *Learning Analytics* (LA) é definido como o processo de medição, coleta, análise e interpretação de dados sobre os alunos e seus ambientes, com o intuito de entender e otimizar o aprendizado (OLIVEIRA et al., 2021). O campo tem se beneficiado diretamente da crescente digitalização do setor educacional.

A consolidação do *Learning Analytics* e da EDM como campos aplicados está diretamente relacionada à evolução da tecnologia educacional e das práticas institucionais de gestão de dados. A literatura aponta para um amadurecimento progressivo dessas áreas, cujos registros remontam à década de 1990, quando já se explorava o uso de algoritmos para prever o desempenho estudantil (ROMERO; VENTURA, 2010). Mais recentemente, a incorporação de técnicas de *Deep Learning* – que utilizam redes neurais profundas para reconhecer padrões complexos em dados – e o avanço das abordagens de Inteligência Artificial Explicável (*Explainable Artificial Intelligence* – XAI) têm ampliado o escopo e o refinamento desses métodos, consolidando seu potencial transformador na educação(GUNASEKARA; SAARELA, 2024).

2.1.2 Distinções e convergências

Embora frequentemente utilizados como sinônimos, EDM e LA apresentam enfoques distintos, ainda que convergentes em seus objetivos centrais (SHAFIQ et al., 2021). A EDM tende a concentrar-se nos aspectos metodológicos e técnicos, desenvolvendo algoritmos e abordagens analíticas para mineração de grandes volumes de dados educacionais (OLIVEIRA et al., 2021; SHAFIQ et al., 2021). Já o LA, por

sua vez, orienta-se mais fortemente para a tomada de decisão pedagógica e institucional, visando apoiar intervenções por meio da análise de contextos específicos de aprendizagem (SHAFIQ et al., 2021).

Apesar das distinções conceituais, observa-se uma tendência de convergência entre EDM e LA, especialmente, no que diz respeito à aplicação de métodos analíticos avançados para apoiar intervenções pedagógicas baseadas em dados. Ambos os campos têm como principais aplicações (SHAFIQ et al., 2021):

1. predição de desempenho;
2. identificação de estudantes em risco de evasão ou reprovação;
3. recomendação de recursos;
4. suporte à intervenção personalizada.

Diante desse panorama, observa-se que o enfoque deste trabalho se alinha conceitualmente à Mineração de Dados Educacionais (EDM), uma vez que privilegia a aplicação de algoritmos de aprendizado de máquina para a extração de padrões relevantes e predição do desempenho acadêmico.

2.2 Visão geral da modelagem preditiva em ambientes educacionais

Esta seção apresenta um panorama analítico das abordagens preditivas aplicadas à educação, com ênfase na previsão de desempenho estudantil — frequentemente referida na literatura como *Student Performance Prediction* (SPP). O objetivo é contextualizar os principais conceitos, técnicas e desafios associados ao uso de modelos preditivos em contextos educacionais, fornecendo uma base teórica para os capítulos subsequentes. Trata-se de uma revisão narrativa que privilegia a síntese conceitual em detrimento da exaustividade bibliográfica.

2.2.1 Algoritmos e técnicas predominantes

A SSP, enquanto subcampo da EDM, busca antecipar desfechos acadêmicos — como aprovação, reprovação ou evasão — com base em dados históricos, demográficos e comportamentais, entre outros. Mais do que explorar a acurácia de algoritmos, esta seção procura destacar também questões ligadas à interpretabilidade, às fontes de dados utilizadas e aos desafios metodológicos e éticos envolvidos.

Nas últimas décadas, têm-se observado uma expansão significativa no uso de algoritmos preditivos para análise de desempenho estudantil, abrangendo desde métodos estatísticos tradicionais até abordagens mais sofisticadas de aprendizado de máquina (ALBREIKI et al., 2021; IMRAN et al., 2019; ROMERO; VENTURA, 2010).

A literatura indica que não há um algoritmo universalmente superior, pois a eficácia de cada modelo está diretamente ligada às características do conjunto de dados e aos objetivos da análise (ALSARIERA et al., 2022; GUNASEKARA; SARELA, 2024). Nesse cenário, as estratégias de *ensemble*, que combinam múltiplos classificadores, surgem como uma alternativa para otimizar a performance geral (IMRAN et al., 2019). Contudo, essa otimização frequentemente ocorre em detrimento da interpretabilidade e ao custo de uma maior complexidade do modelo final (TONG; LI, 2025; YANG et al., 2024).

Além disso, embora a classificação seja a técnica mais recorrente, outras abordagens também são exploradas, como *clustering*, regras de associação e análise estatística, aplicadas ao agrupamento de perfis, descoberta de padrões e monitoramento de comportamentos de aprendizagem (ROMERO; VENTURA, 2010).

2.2.2 Fontes de dados e variáveis preditoras

As bases de dados mais utilizadas no contexto educacional incluem registros acadêmicos históricos, informações demográficas e *logs* de plataformas digitais de aprendizagem. Além disso, observa-se um movimento crescente em direção à incorporação de dados psicossociais e comportamentais, com o objetivo de tornar os modelos mais representativos da complexidade do processo de aprendizagem (ALA-

LAWI et al., 2023; ALBREIKI et al., 2021; SHAFIQ et al., 2021). Essa diversidade amplia o potencial analítico das predições e exige escolhas criteriosas quanto à seleção e ao tratamento das variáveis envolvidas.

De modo geral, os modelos preditivos baseiam-se em atributos provenientes de cinco grandes categorias:

- **Acadêmicos:** notas anteriores, histórico de reprovações, frequência às aulas;
- **Demográficos:** idade, gênero, localidade;
- **Socioeconômicos:** escolaridade dos pais, ocupação familiar, tipo de moradia;
- **Comportamentais:** tempo de estudo, hábitos extracurriculares, uso de tecnologias;
- **Psicossociais:** motivação, autorregulação, autoestima, apoio familiar.

Cada grupo de variáveis oferece uma perspectiva distinta sobre a trajetória do estudante, contribuindo não apenas para a acurácia do modelo, mas também para sua interpretabilidade. A integração dessas fontes favorece abordagens mais abrangentes e alinhadas à complexidade do processo educacional, desde que respeitadas as particularidades do contexto em que os dados são aplicados.

2.2.3 Desafios metodológicos

Apesar dos avanços técnicos e científicos em SSP, a predição do sucesso estudantil ainda enfrenta desafios metodológicos relevantes, que limitam tanto a eficácia quanto a aplicabilidade dos modelos desenvolvidos (ROMERO; VENTURA, 2024). Além das questões técnicas abordadas a seguir, destacam-se também desafios éticos centrais, como a transparência, a equidade algorítmica e o uso responsável de dados estudantis (JIN et al., 2024) — temas discutidos em maior detalhe na seção 2.8.

Natureza e variedade dos dados educacionais

O desempenho de modelos preditivos no contexto educacional está intimamente ligado à qualidade e à diversidade dos dados utilizados. Contudo, a ampla heteroge-

neidade das variáveis disponíveis — que vão desde informações acadêmicas formais até aspectos psicossociais — impõe desafios significativos à etapa de modelagem.

Um ponto crítico reside na complexidade dos atributos menos tangíveis, como engajamento e autorregulação, cuja mensuração demanda técnicas específicas e, muitas vezes, dados indiretos. Embora menos acessíveis, tais variáveis têm demonstrado alto potencial explicativo para o sucesso estudantil, especialmente quando combinadas com indicadores mais tradicionais, como notas e frequência.

Nesse cenário, a seleção criteriosa de atributos se torna decisiva para garantir modelos transparentes e interpretáveis (GUNASEKARA; SAARELA, 2024). Apesar do uso consolidado de dados históricos, demográficos e oriundos de ambientes virtuais de aprendizagem (LMS), variáveis de natureza comportamental, psicológica e associadas à atuação docente ainda são pouco exploradas na literatura especializada (ALALAWI et al., 2023; ALBREIKI et al., 2021; SHAFIQ et al., 2021).

Equilíbrio entre acurácia e interpretabilidade

Em SSP, a escolha do modelo preditivo envolve a ponderação entre a acurácia (capacidade de predições corretas) e interpretabilidade (facilidade de compreensão das predições)(ALALAWI et al., 2023; GUNASEKARA; SAARELA, 2024). Modelos complexos, como Redes Neurais Artificiais, tendem a apresentar um desempenho preditivo superior(GUNASEKARA; SAARELA, 2024; OLIVEIRA et al., 2021), mas são, em geral, obscuros do ponto de vista interpretativo, exigindo o uso de técnicas de explicação *post hoc* (GUNASEKARA; SAARELA, 2024). Em contraposição, modelos intrinsicamente explicáveis, como Árvores de Decisão, são frequentemente preferidos em contextos educacionais devido à sua transparência e facilidade de compreensão por usuários não técnicos (AHMED, 2024; ALALAWI et al., 2023; GUNASEKARA; SAARELA, 2024).

A literatura indica que não há um algoritmo universalmente superior, pois a eficácia de cada modelo está diretamente ligada às características do conjunto de dados e aos objetivos da análise (ALSARIERA et al., 2022). Nesse cenário, as estratégias de *ensemble*, que combinam múltiplos classificadores, surgem como uma

alternativa para otimizar a performance geral (IMRAN et al., 2019). Contudo, essa otimização frequentemente ocorre às custas da interpretabilidade e resulta em maior complexidade do modelo final (TONG; LI, 2025; YANG et al., 2024).

Lacunas em métricas de explicabilidade

Outra limitação, recorrente nos estudos de SSP, é a escassez de métricas específicas voltadas à avaliação da explicabilidade dos modelos (GUNASEKARA; SAARELA, 2024). Embora o debate sobre a importância de modelos transparentes tenha ganhado destaque, ainda são raros os estudos que adotam critérios sistemáticos para avaliar a qualidade das explicações fornecidas pelas técnicas empregadas (ALALAWI et al., 2023; GUNASEKARA; SAARELA, 2024).

As métricas mais utilizadas permanecem centradas no desempenho preditivo (AHMED, 2024; ALBREIKI et al., 2021; IMRAN et al., 2019; OLIVEIRA et al., 2021). Entre as principais estão acurácia, *recall*, precisão e área sob a curva ROC (AUC), cujas definições são apresentadas na Seção 2.7. Por outro lado, observa-se menor atenção às métricas de justiça (*fairness*), que avaliam se o modelo produz resultados equitativos entre diferentes grupos demográficos ou subgrupos de estudantes – por exemplo, paridade demográfica, igualdade de oportunidades e calibração por grupo (MEHRABI et al., 2022; TANG et al., 2023). Além disso, há ausência de padronização quanto à avaliação da clareza e da utilidade das explicações fornecidas aos usuários finais dos modelos, como professores e gestores educacionais (GUNASEKARA; SAARELA, 2024).

Preparação e validação de dados

A preparação adequada dos dados constitui uma etapa crítica para a construção de modelos preditivos robustos e confiáveis em contextos educacionais, uma vez que dados brutos frequentemente contêm inconsistências, valores ausentes e ruídos que podem comprometer a acurácia e a capacidade de generalização dos modelos. Assim, o pré-processamento envolve procedimentos como limpeza, categorização, normalização e seleção de atributos, sendo essencial para reduzir ruídos, uniformizar

escalas e estruturar o conjunto de dados para a modelagem (AHMED, 2024; IMRAN et al., 2019; NAFURI et al., 2022).

Entre os desafios recorrentes, destaca-se a alta dimensionalidade, caracterizada por um grande número de variáveis ou características disponíveis, a qual pode comprometer tanto a generalização quanto a interpretabilidade dos modelos. Para mitigar esse problema, a seleção criteriosa de atributos relevantes contribui para melhorar o desempenho preditivo, reduzir o risco de sobreajuste (*overfitting*) e facilitar a compreensão dos resultados gerados (ALALAWI et al., 2023; NAFURI et al., 2022).

Outro obstáculo frequente em bases educacionais é o desequilíbrio de classes. A predominância de estudantes aprovados em relação aos reprovados pode induzir vieses nos algoritmos, dificultando a identificação da classe minoritária — justamente aquela de maior interesse em muitos contextos (ALBREIKI et al., 2021; IMRAN et al., 2019; SHAFIQ et al., 2021). O tratamento adequado desse desbalanceamento requer estratégias como reamostragem, criação de instâncias sintéticas ou ponderação de classes, discutidas em mais detalhes na seção 2.7.

No que se refere à validação dos modelos — etapa metodológica essencial — é necessário cautela tanto na estimativa da performance real quanto na prevenção de sobreajustes (AULAKH et al., 2023). Apesar disso, a literatura evidencia um problema recorrente de generalização limitada, com estudos frequentemente baseados em amostras restritas a uma única instituição, curso ou região (ORDONEZ-AVILA et al., 2023). Essa limitação compromete a replicabilidade dos resultados e dificulta a aplicação dos modelos em contextos educacionais mais amplos (AHMED, 2024; AULAKH et al., 2023; OLIVEIRA et al., 2021; ORDONEZ-AVILA et al., 2023; ROMERO; VENTURA, 2010).

Da predição à ação

A maior parte dos estudos em EDM concentra-se na construção de modelos preditivos de alta acurácia, com pouca ênfase nas ações pedagógicas que poderiam ser desencadeadas com base nas previsões geradas (ALALAWI et al., 2025). No entanto, observa-se, na prática, uma lacuna entre predição e intervenção, a qual representa

um obstáculo à efetiva transformação dos resultados analíticos em melhorias concretas nos processos de ensino-aprendizagem (ALALAWI et al., 2023, 2025; CHEN et al., 2025).

Além disso, muitos modelos limitam-se a prever o que acontece, negligenciando a análise das causas que originam esses resultados. A compreensão das relações causais é crucial para o desenho de intervenções eficazes, permitindo que as instituições educacionais atuem não apenas de forma reativa, mas preventiva (ALALAWI et al., 2023; SILVA FILHO et al., 2023).

Padronização e acessibilidade

Na literatura, observa-se carência de ferramentas de EDM que sejam acessíveis a educadores e profissionais não especialistas(BAKER; YACEF, 2009; BATISTA; FAGUNDES, 2023). A ausência de interfaces intuitivas e interpretáveis limita a apropriação prática dos modelos nas instituições de ensino(BAKER; CARVALHO, 2010; ROMERO; VENTURA, 2010).

Ainda que a área tenha avançado consideravelmente, não há padronização consolidada em relação aos tipos de dados, tarefas preditivas e métricas de avaliação, o que dificulta a comparação entre estudos e compromete o progresso cumulativo da área(ALALAWI et al., 2025; ROMERO; VENTURA, 2010, 2024).

2.2.4 Tendências emergentes

O campo da SSP tem se transformado significativamente com o avanço de abordagens mais sofisticadas de *Machine Learning* e a crescente valorização da interpretabilidade e da ação pedagógica fundamentada em dados. Entre as principais tendências emergentes, destaca-se o fortalecimento da Inteligência Artificial Explicável (*Explainable Artificial Intelligence* – XAI), que busca oferecer maior transparência quanto ao funcionamento interno dos modelos, promovendo confiança na tomada de decisões educacionais automatizadas (GUNASEKARA; SAARELA, 2024). Modelos intrinsecamente interpretáveis, como Árvores de Decisão, têm sido favorecidos por sua capacidade de fornecer justificativas compreensíveis, enquanto modelos

mais robustos em desempenho preditivo — como Redes Neurais e *Support Vector Machines* (SVM) — requerem técnicas de explicação *post-hoc*, que justificam as previsões do modelo já realizadas, ressaltando a necessidade de equilibrar acurácia, explicabilidade e justiça algorítmica (GUNASEKARA; SAARELA, 2024).

Outra tendência relevante diz respeito à adoção de técnicas dinâmicas e de conjunto (*ensemble methods*) (AHMED, 2024; ALBREIKI et al., 2021), bem como ao uso de *hyperparameter tuning* para maximizar a acurácia dos modelos (AHMED, 2024). Estratégias híbridas, semi-supervisionadas e baseadas em *clustering* têm se mostrado promissoras para revelar padrões latentes em perfis de estudantes heterogêneos, embora ainda sejam pouco exploradas (SHAFIQ et al., 2021).

Há também um aprofundamento na utilização de fontes de dados temporais e comportamentais, como registros de atividades em plataformas de *e-learning*, ampliando as possibilidades de análise longitudinal e predição em tempo real (AHMED, 2024; IMRAN et al., 2019). Paralelamente, a integração de técnicas de Análise multicritério de decisão, como o *Analytic Hierarchy Process* (AHP) e o *Best-Worst Method* (BWM), tem sido incorporada em estudos voltados à prevenção da evasão escolar, permitindo a ponderação criteriosa de fatores preditivos (OLIVEIRA et al., 2021).

De forma alinhada, há um foco crescente na intervenção baseada em dados, especialmente por meio de *Early Warning Systems* (EWS), que permitem identificar precocemente estudantes em risco e acionar mecanismos personalizados de suporte. Evidências apontam que essas ações, quando bem calibradas, impactam positivamente as taxas de aprovação e de permanência (ALALAWI et al., 2023; ALBREIKI et al., 2021).

2.3 Fatores determinantes do desempenho escolar

O desempenho escolar resulta da interação complexa entre múltiplas dimensões do contexto social, familiar, comportamental e disciplinar. Esta seção examina essas diferentes dimensões, organizadas em três eixos analíticos: aspectos demográficos e

socioeconômicos, fatores comportamentais e psicossociais, e especificidades disciplinares. Tais fatores atuam de forma interdependente na configuração das oportunidades de aprendizagem (IMRAN et al., 2019; OLIVEIRA et al., 2021), do engajamento estudantil e, em última instância, do rendimento acadêmico.

2.3.1 Aspectos demográficos e socioeconômicos

Sob a perspectiva estrutural, destacam-se os fatores socioeconômicos, demográficos e de acesso a recursos educacionais. A escolaridade dos pais, por exemplo, figura como um preditor recorrente do desempenho escolar, refletindo o capital cultural disponível no ambiente familiar (ALBREIKI et al., 2021; ROSLAN; CHEN, 2022). De forma semelhante, a composição familiar e a condição socioeconômica moldam o suporte cotidiano à aprendizagem (NAFURI et al., 2022).

Variáveis como idade, gênero e localização geográfica também exercem influência, ainda que de forma contextual. Estudantes de áreas urbanas, por exemplo, tendem a ter maior acesso a infraestrutura escolar e a recursos pedagógicos diversificados (ROSLAN; CHEN, 2022). A presença de apoio institucional, aulas particulares e condições materiais adequadas pode intensificar ou atenuar os efeitos da origem social sobre o desempenho (NAFURI et al., 2022; NAWANG et al., 2022; OLIVEIRA et al., 2021).

Apesar dessas correlações, a literatura alerta para limitações éticas e riscos de viés ao empregar características demográficas como preditores em modelos analíticos, tema que será aprofundado na seção 2.8.

2.3.2 Fatores comportamentais e psicossociais

No plano processual, destaca-se o engajamento acadêmico como elemento central, integrando aspectos objetivos — como o tempo de estudo e a frequência às aulas — e subjetivos, como o envolvimento emocional com o aprendizado. Junto a ele, fatores de autorregulação, como persistência, autodisciplina e estratégias metacognitivas, ganham destaque na literatura por sua associação positiva ao desempenho, especi-

almente em disciplinas com alta exigência de prática constante (GUNASEKARA; SAARELA, 2024; OLIVEIRA et al., 2021; ROMERO; VENTURA, 2010; SHAFIQ et al., 2021).

Questões psicossociais, como suporte familiar, qualidade das relações interpessoais e estado de saúde, também impactam diretamente a trajetória escolar. Em contrapartida, o absenteísmo tem sido identificado de forma recorrente como um fator de risco, embora seus efeitos variem conforme a disciplina e o contexto institucional (ALBREIKI et al., 2021; OLIVEIRA et al., 2021; SHAFIQ et al., 2021).

2.3.3 Especificidades disciplinares

Evidências apontam que o desempenho estudantil é influenciado por características inerentes a cada disciplina(ALALAWI et al., 2023; ALRESHIDI, 2023; ROZGONJUK et al., 2020). As áreas do campo STEM, especialmente a Matemática, tendem a representar maiores desafios aos estudantes, em virtude da forte dependência de conhecimentos prévios, decorrente da natureza cumulativa dos conteúdos, e da presença de fatores emocionais como a ansiedade matemática (ALRESHIDI, 2023; ROZGONJUK et al., 2020).

A Matemática, em particular, destaca-se por exigir uma base conceitual sólida, o que a torna especialmente sensível a lacunas no percurso educacional anterior(ALRESHIDI, 2023). Além disso, fatores socioeconômicos e o apoio educacional familiar são amplamente reconhecidos como determinantes do desempenho acadêmico geral e da evasão escolar, reforçando desigualdades estruturais nos resultados educacionais (ALALAWI et al., 2023). Já outras dimensões, como o engajamento estudantil e o tempo de estudo, configuram variáveis processuais que também impactam diretamente o rendimento escolar, especialmente em disciplinas mais responsivas ao esforço individual (OLIVEIRA et al., 2021).

2.4 CRISP-DM como metodologia analítica

O CRISP-DM é um modelo de processo amplamente adotado para projetos de mineração de dados, desenvolvido em 1996 como uma abordagem independente de setor e consolidado como padrão de fato em mineração de dados(CHAPMAN et al., 2000; SCHRÖER et al., 2021). A metodologia é estruturada de forma hierárquica, com quatro níveis de abstração e seis fases interdependentes que seguem um fluxo iterativo e não-linear.

As fases do processo incluem:

- Compreensão do Negócio: traduz os objetivos organizacionais em termos analíticos;
- Compreensão dos Dados: etapa voltada à coleta, descrição e exploração inicial dos dados;
- Preparação dos Dados: abrange seleção, limpeza, transformação e formatação;
- Modelagem: é realizada construção e calibração dos modelos analíticos;
- Avaliação: verifica a adequação dos modelos aos objetivos iniciais;
- Implementação: responsável pela operacionalização das soluções desenvolvidas.

A natureza cíclica do modelo permite retornos a fases anteriores conforme novos insights emergem durante o processo analítico. Essa flexibilidade torna o CRISP-DM aplicável a diferentes tipos de tarefas, com adoção particularmente expressiva nos domínios da saúde e da educação (SCHRÖER et al., 2021).

No contexto da Mineração de Dados Educacionais (EDM), diversos estudos têm destacado a necessidade de adaptações nas etapas iniciais do processo — especialmente na compreensão do problema e na preparação dos dados — em função das particularidades semânticas e estruturais dos dados educacionais (ALBREIKI et al., 2021; ROMERO; VENTURA, 2010; SHAFIQ et al., 2021).

2.5 Seleção de atributos em dados educacionais

A seleção de atributos relevantes é uma etapa crítica em projetos de análise preditiva, impactando tanto a eficiência dos modelos quanto sua interpretabilidade. Os métodos de filtragem (*Filter Methods*) avaliam a relevância dos atributos com base em suas características intrínsecas, independentemente do algoritmo de aprendizado posterior .

Para dados categóricos, neste trabalho, duas abordagens principais se destacam na seleção de atributos:

- **Técnicas baseadas em entropia** que utilizam a entropia de *Shannon* e suas variações (ganho de informação, razão de ganho) para medir a capacidade de cada atributo reduzir a incerteza sobre a classe-alvo (LI et al., 2024).

- **Testes de independência**, como o teste Qui-quadrado e o V de *Cramér*, permitem avaliar associações estatisticamente significativas entre variáveis categóricas e a variável-alvo (MALIK et al., 2025).

Entropia como medida de diversidade informacional

A entropia de *Shannon* pode ser aplicada como uma métrica descritiva para avaliar a diversidade ou incerteza intrínseca na distribuição das categorias de variáveis qualitativas . A entropia é expressa pela Equação 2.1:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \tag{2.1}$$

onde $p(x_i)$ é a probabilidade associada à categoria i do atributo X. Para tornar comparáveis atributos com diferentes números de categorias, calcula-se a entropia relativa conforme a Equação 2.2:

$$H_{rel}(X) = \frac{H(X)}{\log_2 k} \tag{2.2}$$

em que k representa o número de categorias distintas do atributo. Quanto mais uniformemente distribuídas forem as categorias, maior será o valor da entropia,

indicando maior potencial informativo da variável.

2.5.1 Técnicas estatísticas para análise de associações

A análise de associações constitui uma das vertentes fundamentais da estatística aplicada à Mineração de Dados Educacionais, do ingês, do inglês, *Educational Data Mining*(EDM), englobando o que se categoriza como *relationship mining* (mineração de relacionamentos) para descobrir relações significativas entre variáveis que impactam o desempenho discente (BAKER; YACEF, 2009; MISHRA; SAHOO, 2016; ROMERO; VENTURA, 2010; ROY; FARID, 2024). Essas técnicas vão desde medidas descritivas básicas e coeficientes de correlação até testes estatísticos sofisticados que avaliam a independência entre variáveis ou diferenças entre grupos. No contexto educacional, a escolha adequada desses métodos depende das características dos dados (por exemplo, tipo de variável e distribuição) e do objetivo analítico pretendido.

Visão geral dos métodos

O arsenal metodológico para análise de associações em contextos educacionais é amplo e diversificado. Técnicas paramétricas tradicionais incluem estatística descritiva, correlação de Pearson, regressão linear e logística, e ANOVA. Além disso, métodos de redução dimensional, como a Análise de Componentes Principais (PCA) — amplamente utilizados em análise multivariada para simplificação de variáveis (OLIVEIRA et al., 2021) — auxiliam na redução de complexidade dos dados. Paralelamente, a mineração de regras de associação revela padrões do tipo “se-então” em grandes bases de dados educacionais (BAKER; YACEF, 2009).

As abordagens de aprendizado de máquina também integram este cenário, incluindo árvores de decisão pela sua interpretabilidade, algoritmos de agrupamento (como o K-means) para identificação de perfis estudantis e *Support Vector Machines* (SVM) para tarefas de classificação complexas (AHMED, 2024; ALALAWI et al., 2023; ROSLAN; CHEN, 2022). Além disso, métodos de reamostragem de dados são empregados para abordar problemas de desbalanceamento de classes (GRAMS, 2024), e até mesmo a análise de redes sociais é utilizada para explorar relações inter-

pessoais em ambientes educacionais(ALBREIKI et al., 2021; MALIK et al., 2025).

Foco em métodos não paramétricos

Dada a natureza frequentemente ordinal, categórica ou assimétrica das variáveis educacionais, destacam-se os métodos estatísticos não paramétricos. Esses métodos oferecem robustez analítica sem exigir pressupostos rígidos sobre a distribuição dos dados – uma vantagem crucial considerando que, na prática, dados educacionais reais raramente atendem às suposições de normalidade e homocedasticidade. Em outras palavras, os métodos não paramétricos permitem inferências estatísticas rigorosas mesmo quando não se cumprem os requisitos clássicos dos testes paramétricos.

Teste de *Kruskal-Wallis*

O teste de *Kruskal-Wallis* é amplamente utilizado para comparar distribuições entre três ou mais grupos independentes, funcionando como uma alternativa robusta à ANOVA quando os pressupostos paramétricos não são atendidos. Este teste permite avaliar se características contextuais — como o nível de escolaridade dos responsáveis, a faixa etária dos estudantes ou a modalidade de ensino — estão associadas a variações no desempenho acadêmico. A estatística H de *Kruskal-Wallis* compara as medianas dos grupos sob a hipótese nula de que não existem diferenças significativas entre eles. Essa abordagem é particularmente útil na análise de dados educacionais longitudinais ou de grupos com características heterogêneas, nos quais os dados podem violar as suposições de normalidade exigidas pelos testes paramétricos tradicionais.

Teste Qui-Quadrado de independência

Outro método amplamente adotado na análise de associações é o teste qui-quadrado (χ^2) de independência, aplicado para verificar relações entre variáveis qualitativas. Este teste permite averiguar, por exemplo, se há dependência estatística entre o histórico de reprovações de um aluno e o turno escolar que ele frequenta, ou entre o tipo de escola (pública/privada) e a aprovação em processos seletivos.

A versatilidade do qui-quadrado reside na capacidade de trabalhar com variáveis categóricas de qualquer número de categorias, fornecendo uma base estatística rigorosa para investigar relações entre fatores qualitativos no ambiente educacional. Na prática, o teste de qui-quadrado de *Pearson* verifica se as frequências observadas em uma tabela de contingência diferem significativamente das frequências esperadas sob independência, indicando se existe associação entre as variáveis (MALIK et al., 2025; MISHRA; SAHOO, 2016).

Coeficiente V de Cramer

Para complementar a análise do qui-quadrado, utiliza-se o coeficiente V de Cramer, que fornece uma medida padronizada da intensidade da associação entre variáveis categóricas. Esse coeficiente varia entre 0 e 1, em que valores próximos de 0 indicam associação fraca e valores próximos de 1 sugerem associação forte. Sua principal vantagem está na capacidade de comparar a força de associações mesmo ao trabalhar com tabelas de contingência de tamanhos diferentes ou com número desigual de categorias, permitindo uma interpretação mais aprofundada das relações identificadas. Em essência, o V de Cramer é derivado da estatística qui-quadrado e serve como medida de tamanho de efeito para os achados desse teste (MALIK et al., 2025).

Coeficiente de Correlação de Spearman

No que se refere à análise de correlação entre variáveis ordinais ou não normalmente distribuídas, o coeficiente de correlação de Spearman representa uma alternativa eficaz ao coeficiente de *Pearson*. Sua principal vantagem é captar relações monotônicas entre variáveis, mesmo na ausência de linearidade estrita. Isso é particularmente útil em contextos educacionais onde as variáveis estão em escalas subjetivas (como níveis de satisfação ou proficiência), apresentam distribuições assimétricas, ou são ordinais (como conceitos ou *rankings*). Diferentemente da correlação de *Pearson* – que pressupõe dados contínuos normalmente distribuídos e relacionamentos lineares – a correlação de *Spearman* baseia-se nos postos (*ranks*) dos dados e permanece

confiável mesmo quando tais pressupostos não são satisfeitos (MISHRA; SAHOO, 2016).

Integração e Aplicação Prática

Os métodos não paramétricos apresentados oferecem uma base teórica sólida para a análise de associações em ambientes educacionais reais, nos quais a complexidade dos dados e a heterogeneidade dos sujeitos impõem desafios aos métodos estritamente paramétricos. A escolha entre as diferentes abordagens deve considerar não apenas as características estatísticas dos dados, mas também a interpretabilidade dos resultados para os atores educacionais. A combinação estratégica desses métodos permite uma análise multifacetada: o teste de *Kruskal-Wallis* pode revelar diferenças entre grupos, o teste qui-quadrado pode confirmar associações categóricas, o V de Cramer quantifica a força dessas associações, e o coeficiente de Spearman evidencia relações ordinais subjacentes. Essa abordagem integrada fortalece a robustez das conclusões e oferece múltiplas perspectivas sobre os fenômenos educacionais investigados. Ao possibilitarem interpretações estatisticamente rigorosas sem comprometer a validade analítica, tais abordagens fortalecem a compreensão dos fatores associados ao sucesso ou fracasso escolar, contribuindo para a formulação de políticas educacionais baseadas em evidências e para o desenvolvimento de intervenções pedagógicas mais eficazes e personalizadas (ROY; FARID, 2024; SINGH; RATHI, 2016).

2.6 Algoritmos de Aprendizado de Máquina em EDM

A seleção adequada de algoritmos de aprendizado de máquina é fundamental para o sucesso das análises em EDM. Esta seção apresenta alguns dos principais algoritmos utilizados na área, abordando desde técnicas supervisionadas para classificação até métodos não-supervisionados para descoberta de padrões ocultos, discutindo suas características, vantagens e contextos de aplicação em ambientes educacionais.

2.6.1 Algoritmos supervisionados para classificação

Diversos algoritmos de aprendizado supervisionado são amplamente utilizados em EDM. A escolha dos algoritmos considera múltiplos fatores: interpretabilidade, especialmente importante em contextos educacionais onde compreender fatores é crucial; robustez, ou capacidade de lidar com *outliers* e dados ruidosos; complexidade computacional, adequação ao tamanho e características dos dados; e generalização, ou capacidade de desempenho em dados não vistos.

Regressão Logística

A Regressão Logística é um modelo estatístico amplamente utilizado para tarefas de classificação binária. A relação entre a probabilidade de ocorrência do evento e as variáveis explicativas é modelada pela função logística apresentada na Equação 2.3:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\cdots+\beta_k X_k)}} \tag{2.3}$$

Sua principal vantagem está na interpretabilidade dos coeficientes e na estimativa direta de probabilidades, características fundamentais em contextos educacionais onde a compreensão dos fatores influentes é tão importante quanto a acurácia preditiva.

Support Vector Machine (SVM)

O *Support Vector Machine* (SVM) é um algoritmo eficaz para tarefas de classificação, cuja estratégia central consiste em encontrar um hiperplano que maximize a margem de separação entre as classes. Quando os dados não são linearmente separáveis, utiliza funções *kernel* (linear, RBF) que transformam o espaço original para uma dimensão superior, onde a separação se torna possível.

A escolha do *kernel* e dos hiperparâmetros, como 'C' (penalização) e 'gamma' (no caso do RBF), influencia diretamente a flexibilidade e a capacidade de generalização do modelo. O SVM demonstra particular eficácia em problemas de alta dimensionalidade e quando há necessidade de separar classes com margens bem de-

finidas.

Árvores de Decisão e Métodos *Ensemble*

As Árvores de Decisão são modelos hierárquicos que particionam recursivamente o espaço de atributos com base em critérios como Ganho de Informação ou índice de Gini (uma medida de impureza que varia de 0, pureza total, a 0,5, máxima impureza em problemas binários). Sua interpretabilidade e facilidade de aplicação as tornam particularmente úteis em contextos educacionais, onde a compreensão das regras de decisão é fundamental.

Para mitigar limitações como o *overfitting*, são empregados métodos de conjunto (*ensemble*), que combinam múltiplos modelos para melhorar o desempenho preditivo. O *Random Forest* constrói múltiplas árvores com amostras e atributos aleatórios, reduzindo variância e prevenindo ajuste excessivo. A robustez é ampliada, pois ao agregar múltiplas árvores independentes, os erros individuais causados por *outliers* tendem a ser diluídos no resultado final. O *AdaBoost* combina iterativamente classificadores fracos, ajustando pesos para focar em exemplos difíceis. É especialmente útil em problemas binários, como a predição de aprovação/reprovação, demonstrando capacidade de melhorar progressivamente a classificação de casos complexos.

2.6.2 Exploração de padrões ocultos com PCA e clusterização

Análise de Componentes Principais (PCA)

A Análise de Componentes Principais é uma técnica de redução de dimensionalidade que transforma um conjunto de variáveis, possivelmente, correlacionadas em um novo conjunto de variáveis ortogonais chamadas componentes principais. Seu objetivo é capturar a maior variabilidade possível dos dados com um número reduzido de componentes, facilitando a visualização e a interpretação das estruturas subjacentes.

Em contextos educacionais, a PCA pode ser utilizada para identificar combinações de variáveis que melhor explicam as diferenças entre os estudantes, revelando dimensões latentes do desempenho acadêmico (JOLLIFFE, 2002). A técnica permite reduzir a complexidade dos modelos e evitar problemas de multicolinearidade, especialmente relevantes quando múltiplas variáveis socioeconômicas estão presentes.

Clusterização K-Means

A clusterização consiste na segmentação de um conjunto de dados em grupos homogêneos, de modo que instâncias dentro do mesmo grupo sejam mais similares entre si do que em relação a outros grupos. O algoritmo K-Means busca particionar dados em k *clusters* minimizando a função objetivo expressa na Equação 2.4:

$$\arg\min_{C} \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \tag{2.4}$$

onde C_i representa o conjunto de pontos pertencentes ao *cluster* i e μ_i é o centroide do *cluster*. A aplicação conjunta de PCA e K-Means permite identificar perfis latentes de desempenho estudantil, revelando grupos com características semelhantes que podem orientar intervenções pedagógicas personalizadas.

2.7 Métricas de Avaliação e Validação de Modelos

A escolha apropriada de métricas de avaliação é fundamental para validar modelos preditivos em contextos educacionais. Esta seção apresenta as principais métricas validação utilizadas e suas aplicações para dados desbalanceados.

2.7.1 Matriz de confusão e métricas derivadas

Tabela 2.1: Estrutura da matriz de confusão para classificação binária

		Predição	
		Positivo	Negativo
Real	Positivo	TP	FN
	Negativo	FP	TN

Fonte: Adaptada da literatura.

A matriz de confusão é uma ferramenta fundamental para visualizar o desempenho de modelos de classificação binária, comparando as classificações realizadas pelo modelo com os rótulos reais. Para problemas de classificação binária, a matriz possui dimensão 2×2, onde linhas representam as classes reais e colunas representam as predições do modelo. A Tabela 2.1 apresenta sua estrutura, que permite identificar quatro tipos de resultados:

- **Verdadeiros Positivos (TP)**: Casos positivos corretamente identificados pelo modelo;
- **Falsos Positivos (FP)**: Casos negativos incorretamente classificados como positivos (erro tipo I);
- **Falsos Negativos (FN)**: Casos positivos incorretamente classificados como negativos (erro tipo II);
- **Verdadeiros Negativos (TN)**: Casos negativos corretamente identificados pelo modelo.

A partir da matriz de confusão, derivam-se as principais métricas de avaliação:

- **Acurácia**: Proporção de classificações corretas considerando todas as classes.

$$Acurácia = \frac{TP + TN}{TP + FP + FN + TN} \tag{2.5}$$

- **Precisão**: Proporção de acertos entre todas as previsões positivas realizadas pelo modelo.

$$Precisão = \frac{TP}{TP + FP} \tag{2.6}$$

- **Sensibilidade (*Recall*)**: Capacidade do modelo de identificar corretamente os exemplos da classe positiva.

$$Sensibilidade = \frac{TP}{TP + FN} \tag{2.7}$$

- F_1**-*Score***: Média harmônica entre precisão e sensibilidade, especialmente útil em cenários com desbalanceamento entre classes.

$$F_1 = \frac{2 \cdot Precisão \cdot Sensibilidade}{Precisão + Sensibilidade} \tag{2.8}$$

2.7.2 Métricas para dados desbalanceados

Em contextos educacionais, o desbalanceamento de classes é frequente, especialmente quando se analisa aprovação/reprovação. Métricas específicas para esses cenários incluem:

- **AUC-ROC**: Área sob a curva *Receiver Operating Characteristic*, que avalia a capacidade discriminativa do modelo em diferentes limiares de classificação.
- **AUC-PR**: Área sob a curva *Precision-Recall*, especialmente relevante em cenários com classes desbalanceadas, por fornecer uma avaliação mais sensível ao desempenho da classe minoritária — que pode ser mascarado pela AUC-ROC, mesmo quando o modelo apresenta baixa performance nesse grupo.

A AUC-PR é particularmente valiosa em contextos educacionais onde identificar corretamente estudantes em risco (classe minoritária) é mais crítico do que a acurácia geral do modelo.

2.7.3 Estratégias de tratamento de desbalanceamento

O desbalanceamento de classes é um desafio metodológico central em contextos educacionais, especialmente quando se analisa aprovação/reprovação. Diversas estratégias podem ser empregadas para lidar com esse problema:

- **Ponderação de classes**: Parâmetro **'class_weight'** aplicado diretamente nos algoritmos, ajustando a importância relativa das classes durante o treinamento. Em bibliotecas como o **'scikit-learn'**, algoritmos como Regressão Logística e SVM já oferecem suporte nativo a essa funcionalidade.

- **Validação cruzada estratificada**: Preservação de proporções de classes em todos os subconjuntos de validação, assegurando avaliação representativa e mantendo a distribuição das classes proporcional em cada subconjunto.

O diagnóstico de *overfitting* emprega estratégias para identificar sobreajuste, incluindo comparação entre desempenho em treinamento e validação. Diferenças superiores a 10% entre essas métricas podem indicar sobreajuste, embora esse limite seja apenas um parâmetro empírico e não uma regra absoluta, devendo ser analisado de acordo com o contexto do modelo e dos dados.

A escolha da estratégia de balanceamento deve considerar as características específicas dos dados e os objetivos da modelagem, priorizando a capacidade de generalização sobre a performance em dados de treinamento.

2.8 Considerações Éticas e Limitações

A aplicação de técnicas de EDM para predição em contextos escolares transcende aspectos puramente técnicos, exigindo reflexão crítica sobre suas implicações éticas e limitações metodológicas. Esta seção examina os principais desafios éticos no uso de dados educacionais, as limitações estruturais que afetam a validade dos resultados e os *frameworks* que devem orientar o desenvolvimento responsável desses sistemas.

2.8.1 Desafios éticos em EDM

A aplicação de técnicas de EDM envolve considerações éticas fundamentais que impactam o desenvolvimento e implementação de modelos preditivos. O uso de dados estudantis apresenta implicações éticas significativas, especialmente quando se trata de informações sensíveis e pessoais. A transparência quanto ao uso dos dados, o consentimento informado e a proteção à privacidade dos estudantes são princípios que devem nortear qualquer aplicação de EDM (SLADE; PRINSLOO, 2013).

A equidade algorítmica – que diz respeito à garantia de que modelos preditivos não produzam resultados sistematicamente desfavoráveis para grupos específicos – representa uma preocupação central, exigindo o desenvolvimento de algoritmos mais justos e a detecção de instâncias de injustiça algorítmica, especialmente importante em contextos com dados desbalanceados (KIZILCEC et al., 2020). Além disso, a equidade nas previsões e na explicação dos resultados para diferentes subgrupos da população estudantil é uma preocupação crescente. Modelos que perpetuam ou ampliam desigualdades podem comprometer a legitimidade ética das soluções propostas (OCHOA et al., 2017).

A transparência e explicabilidade emergem como requisitos fundamentais, demandando *frameworks* abrangentes que contemplem dimensões fundamentais da explicabilidade em IA educacional (WANG et al., 2021). A priorização de modelos interpretáveis torna-se essencial quando o objetivo inclui compreensão dos fatores influentes, não apenas acurácia preditiva.

Mesmo utilizando bases de dados já anonimizadas e disponibilizadas publicamente, mantém-se o compromisso com análises responsáveis e interpretações contextualizadas (SLADE; PRINSLOO, 2013). A proteção da privacidade dos dados educacionais, que frequentemente incluem informações pessoais sensíveis, é fundamental para evitar prejuízos aos estudantes e garantir conformidade com normas de privacidade e segurança.

2.8.2 Limitações metodológicas

Diversas limitações estruturais devem ser consideradas na interpretação dos resultados de pesquisas em EDM. A qualidade dos dados educacionais frequentemente é comprometida por incompletude, vieses ou estruturação inadequada, impactando a eficácia dos modelos preditivos (GUNASEKARA; SAARELA, 2024; LIAO, 2022; NAWANG et al., 2022). A natureza observacional dos dados limita inferências causais diretas, restringindo a interpretação dos resultados a associações estatísticas.

A necessidade de validação em diferentes contextos e populações é crucial, uma vez que modelos desenvolvidos em contextos específicos podem não se aplicar universalmente. Elementos como clima escolar, práticas pedagógicas específicas e dinâmicas institucionais podem influenciar resultados observados sem serem capturados nos dados disponíveis.

Uso controverso de variáveis demográficas como preditores

O uso de variáveis demográficas (raça, gênero, status socioeconômico) como preditores diretos em modelos educacionais é considerada uma questão particularmente controversa na literatura de EDM (GARDNER et al., 2019). Embora a inclusão dessas variáveis possa potencialmente melhorar o desempenho preditivo dos modelos, sua utilização apresenta riscos significativos que merecem consideração cuidadosa.

A principal limitação reside na redução da capacidade de ação (*actionability*): variáveis demográficas são inerentemente não manipuláveis pelas instituições educacionais, limitando severamente a utilidade prática das predições para intervenções individuais (GARDNER et al., 2019). Adicionalmente, existe o risco de reforçar vieses existentes nos tomadores de decisão, quando educadores podem interpretar predições baseadas em demografia como justificativas para atribuir resultados a fatores imutáveis dos estudantes, obscurecendo causas sistêmicas que requerem intervenção (KIZILCEC et al., 2020).

A literatura contemporânea recomenda que variáveis demográficas sejam prioritariamente utilizadas para auditorias de justiça (*fairness audits*) e análises de disparidades em nível de grupo, informando intervenções sistêmicas em vez de de-

cisões individuais (GARDNER et al., 2019). Esta abordagem permite identificar desigualdades estruturais sem comprometer a interpretabilidade ética dos modelos preditivos.

O desequilíbrio amostral entre diferentes grupos (escolas, regiões, perfis socioeconômicos) pode comprometer a generalização dos achados, limitando a aplicabilidade dos modelos a contextos similares aos da amostra original. A ausência de dados longitudinais também limita a compreensão das trajetórias de desenvolvimento acadêmico e dos efeitos de longo prazo das intervenções.

2.8.3 Frameworks éticos e boas práticas

O desenvolvimento responsável de sistemas de EDM requer a adoção de *frameworks* éticos robustos - conjuntos estruturados de princípios e diretrizes que orientam decisões e práticas - para todas as etapas do processo analítico (SLADE; PRINSLOO, 2013). Estes *frameworks* devem contemplar:

- **Consentimento informado:** mesmo quando utilizando dados secundários, é fundamental assegurar que a coleta original respeitou princípios de consentimento;
- **Minimização de viés:** implementação de estratégias para identificar e mitigar vieses algorítmicos que possam perpetuar desigualdades educacionais;
- **Transparência algorítmica:** disponibilização de informações claras sobre metodologias, limitações e potenciais impactos dos modelos desenvolvidos;
- **Responsabilidade social:** consideração dos potenciais impactos sociais das predições e recomendações geradas pelos modelos.

Capítulo 3

Procedimentos Metodológicos

Este capítulo apresenta a abordagem metodológica adotada para o desenvolvimento da estrutura analítica preditiva proposta neste trabalho, referenciada ao longo do texto como *framework* analítico, cujo objetivo central é investigar fatores associados ao desempenho escolar de estudantes do ensino médio a partir de dados reais. As etapas foram organizadas segundo os princípios da metodologia CRISP-DM (*Cross-Industry Standard Process for Data Mining*), conforme discutido no Capítulo 2.

3.1 Estrutura preditiva para desempenho escolar

Neste estudo foi desenvolvido um *framework* analítico preditivo para prever o desempenho escolar de estudantes. A estrutura é composta por cinco componentes que adaptam a metodologia CRISP-DM às características dos dados educacionais, considerando questões como o desbalanceamento entre categorias de desempenho e a necessidade de modelos interpretáveis. Para demonstrar a aplicabilidade e validar a abordagem proposta, o *framework* foi implementado utilizando o conjunto de dados *Student Performance* (CORTEZ; SILVA, 2008a), amplamente utilizado na literatura como referência para problemas de predição educacional. Os detalhes técnicos dos módulos programáticos estão disponíveis nos Apêndices.

3.1.1 Arquitetura geral e princípios norteadores

O *framework* organiza o processo de análise em cinco componentes interligados, que funcionam de forma iterativa e permitem ajustes contínuos ao longo das etapas. As adaptações metodológicas implementadas buscam responder aos desafios específicos do contexto educacional, tais como:

- **Desbalanceamento estrutural nas taxas de aprovação**: estudantes aprovados representam a classe majoritária, exigindo técnicas de ponderação e métricas sensíveis à classe minoritária (reprovados);
- **Necessidade de análises comparativas entre domínios curriculares**: o desempenho pode variar significativamente entre disciplinas (português vs. matemática), demandando análises integradas que identifiquem padrões específicos e transversais;
- **Relevância de variáveis contextuais além do histórico acadêmico**: fatores como apoio familiar, tempo de estudo, consumo de álcool, qualidade de relacionamentos e saúde podem influenciar o desempenho, mesmo quando controladas as notas anteriores;
- **Demanda por predições interpretáveis e acionáveis**: os modelos devem não apenas prever a aprovação, mas também identificar fatores modificáveis (como frequência e tempo de estudo) que orientem intervenções pedagógicas, ao contrário de atributos imutáveis como gênero ou escola de origem.

O *framework* estrutura-se em cinco componentes, correspondentes às fases do CRISP-DM com adaptações específicas ao estudo realizado:

1. **Analise contextual**: busca Compreensão do panorama educacional no contexto temporal e geográfico da base estudada.
2. **Compreensão dos Dados**: incorpora técnicas de análises individuais e comparativas para as disciplinas de português e matemática. Nesta etapa, propõe-se o **índice de perfil composto (PerfilScore)** para seleção de atributos

categóricos com base em diversidade informacional e impacto discriminativo. Adicionalmente, realiza-se a identificação de perfis estudantis latentes via PCA e clusterização.

3. **Preparação dos Dados**: propõe **três estratégias complementares** de seleção de atributos: (i) baseada em evidências exploratórias, (ii) orientada por regressão linear múltipla com controle de multicolinearidade, e (iii) fundamentada em testes estatísticos inferenciais. Essa abordagem multinível permite avaliar o impacto de diferentes conjuntos de preditores na capacidade de generalização dos modelos.

4. **Modelagem**: implementa a comparação de cinco algoritmos (Regressão Logística, SVM, Árvore de Decisão, Random Forest e AdaBoost) com otimização via *grid search* e tratamento integrado de desbalanceamento por ponderação de classes. Cada algoritmo é avaliado em versões base e otimizada.

5. **Avaliação**: incorpora validação cruzada estratificada (5-fold), diagnóstico de sobreajuste mediante comparação entre conjuntos de teste e validação, e análise visual consolidada com curvas ROC, Precision-Recall e matrizes de confusão. Utiliza métricas sensíveis ao desbalanceamento, com atenção especial à classe minoritária (reprovados).

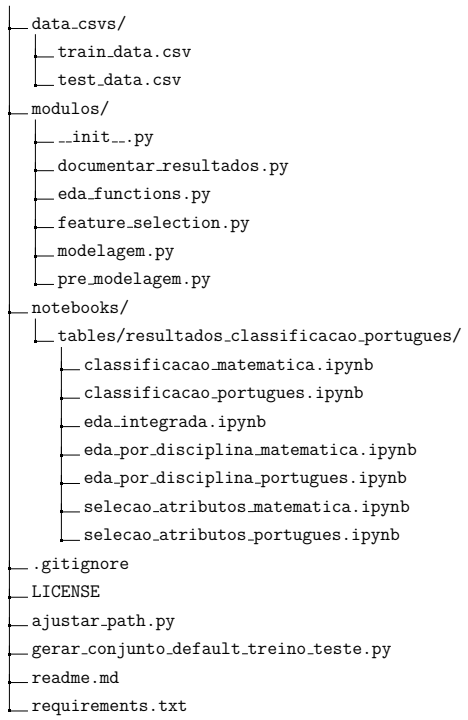
Quadro 3.1: Síntese do framework analítico: fases, ferramentas e procedimentos

Fase (CRISP-DM)	**Ferramentas e scripts**	**Procedimentos e objetivos principais**
Entendimento do Negócio (Análise Contextual)	Revisão bibliográfica; análise documental; definição do problema preditivo	Contextualização do problema no domínio educacional; definição dos objetivos da modelagem preditiva e das variáveis de interesse com base em estudos prévios e nas características do conjunto de dados.
Entendimento dos Dados	'pandas', 'numpy', 'seaborn', 'matplotlib', 'scipy'; 'eda_functions.py'	Exploração inicial do conjunto de dados; análise de distribuições, correlações, valores ausentes, outliers e aspectos estatísticos relevantes. Apoio à formulação de hipóteses e à seleção preliminar de atributos.
Preparação dos Dados	'pandas', 'scikit-learn', 'imblearn', 'scipy'; 'pre_modelagem.py', 'feature_selection.py'	Codificação de variáveis, tratamento de multicolinearidade e seleção de atributos relevantes para os modelos preditivos.
Modelagem	'scikit-learn'; 'modelagem.py'	Treinamento e comparação de classificadores binários com e sem otimização de hiperparâmetros ('GridSearchCV'); validação cruzada e análise de desempenho preditivo.
Avaliação	'scikit-learn', 'matplotlib', 'seaborn'; 'modelagem.py'	Análise de desempenho dos modelos com base em métricas como acurácia, F_1 e AUC; identificação de sobreajuste e geração de visualizações para apoio à interpretação dos resultados.

Fonte: Elaborado pela autora (2025).

3.2 Ambiente computacional e ferramentas utilizadas

Figura 3.1: Estrutura de diretórios do projeto.



Fonte: Elaborado pela autora (2025).

A análise e a modelagem dos dados foram conduzidas predominantemente em ambiente *Python versão 3.9.0*, selecionado por sua flexibilidade e amplo ecossistema de bibliotecas para ciência de dados (MCKINNEY, 2018). As bibliotecas utilizadas incluíram:

- **'pandas'**, para manipulação e análise tabular;
- **'scikit-learn'**, para pré-processamento, modelagem e validação cruzada;
- **'imblearn'**, para tratamento de desbalanceamento com técnicas de reamostragem;
- **'seaborn'** e **'matplotlib'**, para visualização de dados;
- **'scipy'** e **'statsmodels'**, para testes estatísticos e regressão.

A arquitetura do projeto foi modularizada em scripts Python organizados em diretórios específicos (Figura 3.1), permitindo reuso, manutenibilidade e rastreabilidade das análises. A estrutura organizacional incluiu um diretório **'data_csvs/'** para armazenamento dos conjuntos de dados, um diretório **'modulos/'** contendo os scripts modulares, e um diretório **'notebooks/'** com análises exploratórias e documentação dos resultados.

Os módulos foram estruturados de forma a se alinharem, ainda que de maneira não estritamente delimitada, às etapas do CRISP-DM, conforme descrito a seguir:

- 'pre_modelagem.py': importação, renomeação de atributos, tradução de categorias, criação da variável-alvo binária ('aprovacao'), imputação de dados e escalonamento.

- 'eda_functions.py': funções para análise exploratória, como visualizações personalizadas, mapas de calor, análise de *outliers*, detecção de padrões extremos e comparação entre disciplinas.

- 'feature_selection.py': seleção de atributos com base em correlações, testes estatísticos, verificação de multicolinearidade (VIF) e regressões lineares explicativas.

- 'modelagem.py': treinamento, avaliação e comparação de classificadores binários, com ou sem otimização via *'GridSearchCV'*, incluindo geração de métricas, curvas ROC/PR e diagnóstico de sobreajuste.

- 'documentar_resultados.py': suporte à exportação de gráficos e tabelas em formatos apropriados para uso em documentos científicos (LaTeX).

As análises foram conduzidas em *Jupyter notebooks* distintos para cada etapa (análise exploratória integrada, análise por disciplina, seleção de atributos e classificação), organizados no subdiretório 'tables/resultados_classificacao_portugues/', o que permitiu a rastreabilidade e a replicação dos experimentos. Os conjuntos de treino e teste foram gerados previamente por meio do *script* 'gerar_conjunto_default_treino_teste.py', com divisão estratificada por aprovação ('test_size=0.3', 'ran-

dom_state=42'), e salvos em arquivos 'CSV' separados para garantir a integridade e reprodutibilidade do processo. Arquivos auxiliares como 'ajustar_path.py' foram implementados para facilitar o gerenciamento de caminhos relativos entre módulos e notebooks, assegurando portabilidade do código.

3.3 Análise contextual

A análise contextual, correspondente à etapa de entendimento do negócio, teve como objetivo mapear o fenômeno educacional em estudo e alinhar as metas do projeto com as possibilidades analíticas oferecidas pelos dados disponíveis.

O recorte temporal da base analisada — o ano letivo de 2005 a 2006 — coincide com o ciclo do PISA de 2008, conduzido pela Organização para a Cooperação e Desenvolvimento Econômico (OCDE). Naquele momento, Portugal apresentava médias de desempenho estatisticamente inferiores à média da OCDE nas três áreas avaliadas (leitura, matemática e ciência), sendo a matemática o domínio de menor pontuação (466, abaixo da média geral dos países avaliados, de 498) (OECD, 2007).

Além disso, o relatório evidenciava a forte influência do contexto socioeconômico sobre o rendimento escolar dos estudantes portugueses, apontando para desigualdades estruturais persistentes no sistema educacional.

É neste cenário educacional que se insere a presente investigação, utilizando a base de dados pública *Student Performance*, originalmente coletada e analisada por Cortez e Silva (2008). Ao empregar esta base de referência, o presente estudo estabelece um diálogo direto com o trabalho seminal, aproveitando o contexto educacional pré-existente para validar a aplicabilidade e a eficácia do *framework* preditivo proposto. O objetivo é demonstrar que a nova estrutura analítica (com foco nas estratégias multinível de seleção de atributos) oferece um avanço na capacidade de identificar e predizer o risco de insucesso escolar. O conjunto consolidado abrange dados de 677 estudantes únicos de duas escolas públicas portuguesas, matriculados nas disciplinas de Português e Matemática.

Com base nessa contextualização, e com a intenção de definir o fluxo de análise

e o framework, foram formuladas as seguintes perguntas de pesquisa que orientaram os métodos do estudo:

- Quais atributos estão mais associados à aprovação ou reprovação dos estudantes?
- Existem padrões distintos entre o desempenho em português e matemática?
- Quais variáveis contextuais têm maior poder explicativo sobre o rendimento discente?
- É possível predizer a aprovação final sem utilizar as notas parciais como variáveis preditoras?
- Como a escolha dos atributos influencia a acurácia e a interpretabilidade dos modelos?
- Há indícios de assimetrias estruturais entre as disciplinas que impactam a performance dos alunos?

Essas questões guiaram a organização das etapas subsequentes do processo analítico estruturado no *framework*, definindo os critérios de preparação, seleção de atributos, modelagem e avaliação preditiva. A próxima seção detalha o entendimento dos dados, suas características e os procedimentos exploratórios realizados.

3.4 Entendimento dos dados

A etapa teve como finalidade compreender a estrutura e a qualidade dos dados, identificando padrões relevantes, possíveis inconsistências e relações iniciais entre os atributos e as variáveis-alvo, de modo a orientar a preparação e a modelagem preditiva subsequente.

Para isso foram realizadas duas análises exploratórias complementares:

1. Análise segmentada por disciplina: abordagem voltada à compreensão específica do desempenho nas áreas de português e matemática;

2. Análise integrada: empregada com enfoque nos estudantes matriculados em ambas as disciplinas, permitindo comparações diretas entre os respectivos desempenhos escolares.

As subseções a seguir apresentam o detalhamento das etapas de aquisição dos dados, caracterização dos atributos disponíveis e a análise exploratória realizada.

3.4.1 Aquisição e descrição dos dados

Os dados utilizados nesta pesquisa foram obtidos a partir da base de dados "*Student Performance*", disponibilizada no repositório *UCI Machine Learning Repository* (CORTEZ; SILVA, 2008a). A base original encontra-se segmentada em dois arquivos distintos: um referente aos alunos matriculados na disciplina de português (649 instâncias) e outro à disciplina de matemática (395 instâncias). Ambos os conjuntos apresentam 33 atributos que descrevem características demográficas, socioeconômicas, escolares e comportamentais dos estudantes.

A estrutura inicial dos dados foi inspecionada com a função 'info' da biblioteca 'pandas', a qual confirmou a ausência de valores faltantes e a tipagem dos atributos: 16 variáveis numéricas do tipo 'int64' e 17 variáveis categóricas do tipo 'object'. A partir dessas bases brutas, aplicaram-se rotinas de padronização implementadas no módulo 'pre_modelagem.py', com as seguintes etapas principais:

- Tradução dos nomes dos atributos e dos valores categóricos para o português, com base na documentação da base original e na semântica dos campos (vide Tabela A.1);
- Padronização de nomes de colunas e uniformização de categorias equivalentes;
- Criação da variável-alvo binária 'aprovacao', que indica se o estudante foi aprovado na disciplina, com base no critério de nota final maior ou igual a 10 pontos. Esse limiar segue o critério de aprovação adotado no conjunto de dados original, conforme descrito por Cortez e Silva (CORTEZ; SILVA, 2008b), e disponibilizado pela *UCI Machine Learning Repository* (CORTEZ; SILVA, 2008a).

Para estruturar a análise, os atributos foram classificados em três categorias principais: variáveis quantitativas (discretas ou contínuas), categóricas nominais e categóricas ordinais. As variáveis quantitativas incluem contagens e medidas com significado matemático, como notas, número de faltas e tempo de estudo. As variáveis categóricas nominais representam categorias sem ordenação implícita — como sexo, escola, profissão dos pais e tipo de moradia —, enquanto as categóricas ordinais apresentam uma hierarquia natural, como escolaridade dos responsáveis, percepção de saúde e tempo de deslocamento até a escola. Essa classificação orientou a seleção dos testes estatísticos e a construção das visualizações exploratórias.

Além da natureza estatística dos atributos, considerou-se o comportamento específico de algumas variáveis, em especial as notas finais. Apesar de apresentarem valores inteiros, foram tratadas como variáveis contínuas devido ao seu amplo intervalo (0 a 20) e ao comportamento numérico progressivo, que permite análises mais detalhadas. Ainda assim, para avaliar associações envolvendo essas variáveis, optou-se pelo coeficiente de correlação de *Spearman*, reconhecido por sua robustez frente a distribuições assimétricas, ausência de normalidade e presença de *outliers* (DE WINTER et al., 2016; SHAQIRI et al., 2023; XU et al., 2013).

3.4.2 Análise exploratória dos dados

A Análise Exploratória dos Dados (AED) teve papel central na compreensão inicial das bases e na identificação de padrões relevantes para a modelagem preditiva. Foram examinadas características estruturais, demográficas, socioeconômicas e comportamentais dos estudantes, com destaque para o desempenho escolar nas disciplinas de português e matemática.

A seguir, apresentam-se os aspectos metodológicos das análises realizadas, que fundamentam a modelagem preditiva e promovem insights relevantes para a compreensão dos dados.

Inspeção e validação da base de dados

Inicialmente, buscou-se garantir a adequação e integridade da base de dados por meio de inspeção que confirmou a ausência de duplicatas e valores faltantes, assegurando a confiabilidade das análises subsequentes. Para preservar a validade estatística dos modelos preditivos, realizou-se a divisão dos conjuntos de dados em amostras de treinamento e teste via amostragem estratificada, utilizando a variável binária *aprovação* como critério. Esse procedimento visou manter as proporções originais de estudantes aprovados e reprovados em ambos os subconjuntos.

Com o intuito de assegurar a representatividade amostral e permitir análises contextuais, averiguou se a estratificação foi estendida aos principais atributos demográficos — incluindo escola, localização residencial, gênero e nível de escolaridade dos responsáveis. Essa decisão metodológica possibilitou não apenas confirmar a preservação proporcional do conjunto original, mas também identificar diferenças contextuais entre subgrupos populacionais, aspecto relevante para a compreensão de fatores que influenciam o desempenho acadêmico.

Organização das variáveis e análises descritivas iniciais

As análises univariadas incluíram a observação de medidas de tendência central, dispersão e assimetria, além de testes de normalidade (*Shapiro-Wilk*). Para facilitar a interpretação das distribuições, foram utilizadas visualizações como histogramas e boxplots, as quais também permitiram identificar valores atípicos com base nos limites do intervalo interquartil (IQR), o que possibilitou mapear a variabilidade das variáveis contínuas e discretas, oferecendo uma visão inicial sobre a presença de possíveis extremos.

As análises bivariadas, por sua vez, exploraram associações entre notas, faltas e demais atributos quantitativos, com base no coeficiente de *Spearman*. Essa etapa foi complementada por gráficos de dispersão, regressões lineares simples e mapas de calor, permitindo identificar padrões de correlação e variações de desempenho entre grupos. A variável 'faltas', em especial, foi segmentada em faixas com base no IQR, viabilizando a análise comparativa do rendimento escolar em função da frequência.

A decisão de manter os valores atípicos nas análises subsequentes baseou-se tanto nas ferramentas estatísticas utilizadas quanto nas observações extraídas da própria base. Considerou-se que esses casos, apesar de extremos, apresentavam consistência interna e relevância interpretativa no contexto educacional, uma vez que os *outliers* podem representar estudantes com trajetórias atípicas — como situações de evasão iminente ou desempenho excepcional — cuja exclusão poderia comprometer a compreensão da heterogeneidade do conjunto analisado e a identificação de padrões críticos de risco ou excelência.

Análise de variáveis categóricas e índice de perfil composto

As variáveis qualitativas foram analisadas com base em frequências absolutas e relativas, entropia de *Shannon*, medidas de diversidade e taxas de aprovação por categoria. Os atributos foram agrupados tematicamente em: perfil demográfico e estrutural (gênero, moradia, tipo de escola), contexto familiar e socioeconômico (profissão e escolaridade dos pais, interesse no ensino superior), apoio institucional (atividades extracurriculares, apoio escolar) e estilo de vida (hábitos de consumo, relacionamento, saúde).

Para ranquear as variáveis nominais em termos de potencial informativo, foi desenvolvido um *índice de perfil composto*, baseado principalmente em dois critérios centrais: entropia relativa (diversidade interna) e variação média de desempenho entre categorias ('gap'). O índice buscou sistematizar a priorização de variáveis com maior potencial discriminativo, levando em consideração simultaneamente a diversidade das categorias e a magnitude das diferenças de desempenho associadas.

Proposta metodológica: Índice de Perfil Composto ('PerfilScore').

Com base nos fundamentos teóricos apresentados, propõe-se neste trabalho uma metodologia específica para a avaliação e seleção de variáveis categóricas, denominada Índice de Perfil Composto ('PerfilScore').

Este método foi desenvolvido como critério para identificar atributos simultaneamente informativos e discriminativos, combinando duas dimensões complementares:

- a *diversidade informacional*, capturada pela entropia relativa do atributo;

- o *impacto sobre a variável de interesse*, representado pela variação ('gap') entre as médias de desempenho nas diferentes categorias do atributo.

Embora ambos os componentes tenham respaldo na literatura estatística e práticas analíticas (vide Seção 2.5), a combinação ponderada desses critérios em um único escore configurou-se como uma estratégia metodológica particular deste trabalho. Inicialmente foi proposta uma abordagem mais rigorosa, contudo, durante o processo analítico, identificou-se que variáveis com alta diferença média no desempenho não deveriam ser descartadas apenas devido à diversidade interna moderada. Assim, optou-se por flexibilizar o limiar mínimo da entropia relativa para $0,4$, incluindo também um alerta específico para variáveis com alta dispersão. Essa decisão visou garantir uma análise mais criteriosa e contextualizada.

O cálculo do 'PerfilScore' envolveu etapas sequenciais de avaliação da entropia, mensuração do 'gap' de desempenho, aplicação criteriosa de filtros e normalização para a composição final do índice. O processo detalhado de implementação encontra-se descrito no Apêndice B.2, acompanhado de trechos selecionados do código-fonte.

Identificação de grupos extremos de desempenho

Com o propósito de identificar estudantes com desempenho particularmente alto ou baixo, foi aplicada uma estratégia baseada em quantis adaptativos da variável 'nota_final'. Para operacionalizar essa análise, desenvolveu-se a função 'identificar_extremos_comparaveis', que define limiares de corte considerando simultaneamente a magnitude da diferença e o equilíbrio no tamanho dos grupos — conforme a implementação detalhada no Apêndice B.3. Na sequência, foi conduzida uma análise qualitativa dos atributos mais prevalentes em cada grupo, a fim de explorar interpretações relevantes sobre os fatores associados aos extremos de rendimento escolar.

Comparação integrada entre disciplinas

Com foco nos estudantes matriculados em ambas as disciplinas, foi realizada uma AED integrada, por meio da junção das bases com auxílio de um identificador composto, conforme descrito na documentação da base (CORTEZ; SILVA, 2008a). Essa

etapa permitiu comparar diretamente os desempenhos em português e matemática, revelando padrões e assimetrias. Foram destacados casos como:

- estudantes com reprovação exclusiva em uma das disciplinas;
- alunos com discrepância acentuada de desempenho (diferença de nota superior a 4 pontos);
- diferenças marcantes de frequência (mais de 10 faltas de diferença entre disciplinas);
- coincidência de desempenho (mesma nota final em ambas as disciplinas).

As evidências observadas sugerem que o componente curricular pode exercer influência distinta sobre o rendimento dos estudantes, o que reforçou a formulação de hipóteses testadas nas etapas seguintes de análise e modelagem. Os principais resultados dessa etapa estão organizados no Capítulo 4, em alinhamento com os objetivos do estudo.

3.4.3 Análise orientada por regressão

Com o objetivo de identificar fatores contextuais associados ao desempenho intermediário dos estudantes, foi conduzida uma análise de regressão linear considerando a variável 'nota2' como dependente. Inicialmente, empregou-se a 'nota1' como único preditor, de forma a mensurar sua relação direta com a 'nota2'. Em seguida, incorporaram-se variáveis escolares, socioeconômicas e comportamentais, de modo a avaliar a estabilidade do coeficiente da 'nota1' na presença de outros preditores.

A utilização dessas variáveis justifica-se pela forte correlação observada entre 'nota1', 'nota2' e a 'nota_final', o que indica seu papel como indicadores intermediários do desempenho acadêmico global. Essa abordagem possibilita compreender como fatores adicionais influenciam a trajetória escolar, fornecendo indícios sobre a progressão do rendimento ao longo do tempo.

Ainda que esta análise não tenha sido utilizada como critério direto para a seleção de atributos nos modelos de classificação binária, ela permitiu explorar de forma

detalhada as relações contextuais entre preditores e desempenho parcial. Para isso, foram testadas diferentes estratégias de otimização do modelo, incluindo:

1. avaliação da significância estatística dos coeficientes por meio de valores-p;
2. diagnóstico de multicolinearidade para detecção e tratamento de variáveis redundantes;
3. aplicação de procedimentos de seleção *stepwise*, guiados pelos critérios de informação de Akaike (AIC) e bayesiano (BIC).

Tais etapas visaram compreender a contribuição individual de cada atributo para a variação observada na 'nota2'.

3.4.4 Perfis estudantis latentes: redução de dimensionalidade e clusterização

Como extensão da análise exploratória, foi conduzida uma abordagem não supervisionada com o objetivo de identificar padrões latentes de agrupamento entre os estudantes, de modo a explorar possíveis perfis associados ao desempenho escolar. Com esse objetivo, foram combinadas técnicas de redução de dimensionalidade, via Análise de Componentes Principais (*Principal Component Analysis* – PCA), e de clusterização, por meio do algoritmo *K-means*.

A PCA foi aplicada sem prévia redução do número de atributos, transformando o espaço original em componentes ortogonais que maximizam a variância explicada. A decisão de reter dois componentes principais baseou-se na análise conjunta de três critérios:

1. método do cotovelo aplicado à inércia;
2. índice de silhueta;
3. proporção acumulada da variância explicada.

Para interpretar a estrutura latente, calcularam-se os *loadings* (correlações entre variáveis originais e componentes principais), identificando as variáveis com maior

magnitude nos dois primeiros componentes. As dez variáveis mais influentes foram representadas em gráfico de barras, destacando atributos que exerceram maior peso na diferenciação dos perfis.

As duas componentes principais obtidas foram incorporadas ao conjunto de dados como variáveis 'PCA1' e 'PCA2', servindo como base para a aplicação do *K-means*. A escolha do número de agrupamentos foi fundamentada nos mesmos critérios empregados na seleção de componentes, priorizando a convergência entre o ponto de inflexão do método do cotovelo e o maior valor do índice de silhueta. O modelo final adotou dois *clusters*, posteriormente incorporados à base como rótulos categóricos.

A caracterização dos grupos foi conduzida por meio de visualizações no plano bidimensional das componentes principais, com diferenciação de cores por *cluster*. Boxplots permitiram comparar a distribuição da *nota_final* entre os grupos, enquanto gráficos de barras evidenciaram diferenças na taxa de aprovação. Complementarmente, realizaram-se análises descritivas das variáveis escolares, socioeconômicas e comportamentais em cada *cluster*, permitindo delinear perfis estudantis latentes e avaliar associações entre características observadas e desempenho acadêmico.

3.5 Preparação dos dados

A etapa de preparação dos dados compreende o conjunto de procedimentos aplicados sobre os dados brutos para adequá-los à modelagem preditiva supervisionada. Conforme o ciclo CRISP-DM, esta fase visa garantir a integridade, consistência e relevância dos atributos utilizados nos modelos, sendo dividida neste trabalho em três subetapas: pré-processamento, balanceamento e seleção de atributos.

3.5.1 Pré-processamento e preparação para modelagem

Em consonância com a natureza flexível e iterativa do processo CRISP-DM, algumas etapas técnicas — como a divisão estratificada em conjuntos de treino e teste — foram mencionadas anteriormente, ainda no contexto do entendimento e análise exploratória dos dados. Tal antecipação foi necessária para assegurar a

integridade analítica, evitando vazamento de informação, especialmente em análises que poderiam influenciar a seleção de atributos.

Assim, embora a divisão dos dados seja tradicionalmente parte da preparação para modelagem, neste trabalho ela foi planejada e executada de forma antecipada, para contextualizar adequadamente a AED empregada.

Todas as etapas subsequentes de pré-processamento foram conduzidas exclusivamente sobre os conjuntos estratificados, garantindo consistência e reprodutibilidade ao longo do processo analítico.

Para fins de análise preditiva, os dados passaram por uma série de transformações conduzidas pelas funções 'preparar_dados' e 'preparar_treino_e_teste', detalhadas no Apêndice A. Tais funções foram responsáveis por:

- Mapeamento de variáveis ordinais e binárias (como 'apoio_escolar', 'interesse_ensino_superior' e 'tamanho_familia');
- Codificação de variáveis nominais por 'One-Hot Encoding';
- Remoção opcional das colunas de notas ('nota1', 'nota2' e 'nota_final') para evitar vazamento de informação;
- Imputação de valores ausentes com a média;
- Escalonamento opcional das variáveis numéricas por padronização *z-score* (via 'StandardScaler');
- Separação do conjunto em preditores (X) e variável-alvo (y).

As variáveis categóricas binárias foram codificadas via 'Label Encoding' (0 e 1), enquanto as ordinais mantiveram sua ordem original. As nominais foram transformadas por 'One-Hot Encoding' para representar adequadamente categorias não ordinais. Quando habilitada, aplicou-se normalização *z-score* às variáveis quantitativas e ordinais.

3.5.2 Diagnóstico de multicolinearidade entre preditores

Como etapa complementar da preparação dos dados, foi realizado um diagnóstico de multicolinearidade entre as variáveis contextuais por meio da análise conjunta do Fator de Inflação da Variância (VIF — *Variance Inflation Factor*) e dos coeficientes de correlação entre os preditores. O objetivo foi identificar possíveis redundâncias estruturais que pudessem comprometer a estabilidade das análises posteriores. Para isso, adotou-se como referência valores de VIF acima de 5 para alerta e acima de 10 para criticidade, enquanto para o coeficiente de correlação de Pearson estabeleceu-se um limite flexível de $|r| > 0,6$, considerando a natureza dos dados e a importância de preservar variáveis potencialmente relevantes. Valores elevados de VIF e correlações fortes foram utilizados como alerta para revisão da composição dos conjuntos de atributos considerados, sobretudo nas etapas de regressão e avaliação estatística.

3.5.3 Seleção de atributos

Para a escolha de subconjuntos relevantes de atributos para a modelagem da classificação binária, foram adotadas três estratégias complementares. Cada uma delas foi concebida para capturar dimensões distintas dos dados: abordagens exploratórias permitem identificar padrões iniciais, enquanto técnicas estatísticas e inferenciais oferecem rigor analítico e controle de variáveis. Essa triangulação visa ampliar a robustez da análise e favorecer a construção de modelos mais informativos e interpretáveis.

Os valores de corte, níveis de significância e demais critérios numéricos utilizados nos procedimentos descritos nesta subseção serão apresentados no Capítulo 4.

Primeira estratégia: seleção baseada na Análise Exploratória de Dados

A primeira estratégia fundamentou-se em evidências obtidas por meio da Análise Exploratória de Dados (AED), conduzida de forma segmentada por disciplina. A seleção dos atributos considerou relações observadas com a 'nota_final', diferenças de desempenho entre categorias e a exclusão de variáveis altamente correlacionadas

ou conceitualmente redundantes. Essa abordagem buscou reduzir atributos pouco discriminatórios e priorizar variáveis de maior potencial analítico.

Segunda estratégia: seleção por regressão linear múltipla com variáveis contextuais

A segunda estratégia consistiu na aplicação de um modelo de regressão linear múltipla, tendo a 'nota_final' como variável dependente e utilizando exclusivamente variáveis contextuais — excluindo-se notas intermediárias como 'nota1' e 'nota2', a fim de evitar vazamento de informação na modelagem da 'aprovacao'.

O objetivo foi identificar atributos com associação estatisticamente significativa ao desempenho final, controlando simultaneamente múltiplos fatores escolares, socioeconômicos e comportamentais. Após a estimação do modelo, foi conduzida uma análise de multicolinearidade por meio do Fator de Inflação da Variância (VIF) para eliminar atributos redundantes.

Terceira estratégia: seleção por testes estatísticos inferenciais

Na terceira estratégia, variáveis categóricas ordinais foram avaliadas quanto à associação com a 'nota_final' por meio de dois testes não paramétricos complementares:

- Correlação de *Spearman*, para identificar relações monotônicas;
- Teste de *Kruskal-Wallis*, para verificar diferenças significativas entre grupos.

Já para variáveis nominais, foi utilizado o teste de independência do Qui-Quadrado, complementado pelo cálculo do V de *Cramér* como medida padronizada da intensidade da associação.

Independentemente da estratégia empregada, a variável 'faltas' foi incluída no conjunto final de atributos. Sua utilização não se deu apenas pelo potencial explicativo observado nos dados, mas também pelo reconhecimento consolidado na literatura como um fator associado ao desempenho escolar.

3.5.4 Considerações finais sobre a preparação dos dados

As estratégias adotadas, combinando métodos exploratórios, estatísticos e inferenciais, geraram conjuntos complementares de atributos. Essa diversidade foi essencial para avaliar o impacto das escolhas na modelagem e na interpretação dos fatores relacionados ao desempenho escolar. A aplicação prática dessas estratégias resultou em conjuntos preparados e integrados ao processo de modelagem descrito na Seção 3.6.

3.6 Modelagem

A modelagem, etapa central do processo CRISP-DM, envolveu a construção e avaliação de modelos supervisionados para prever a aprovação dos estudantes, com base em diferentes combinações de atributos e algoritmos de aprendizado de máquina.

Foram selecionados cinco algoritmos amplamente consolidados na literatura e comumente aplicados em contextos de EDM: Regressão Logística, SVM, Árvore de Decisão, *Random Forest* e *AdaBoost*. A escolha desses modelos contemplou a diversidade de fundamentos teóricos — desde classificadores lineares até métodos baseados em margens e ensembles — e levou em consideração aspectos práticos como interpretabilidade, robustez frente a ruídos e capacidade de adaptação a dados desbalanceados. Tais algoritmos destacam-se por sua eficácia em tarefas de classificação binária no campo educacional, especialmente quando combinados a técnicas de pré-processamento e seleção de atributos (BAKER; CARVALHO, 2010).

Tais modelos foram selecionados, também, por apresentarem flexibilidade estrutural e potencial de ajuste fino por meio de múltiplos hiperparâmetros. Com base nessa seleção de classificadores, estruturou-se o componente de modelagem do *framework*, cuja implementação foi centralizada na função 'avaliar_classificadores_binarios_otimizados', desenvolvida no módulo 'modelagem.py', a qual consolidou as etapas de treinamento, avaliação, otimização e documentação dos modelos (ver Apêndice D).

É sabido que em contextos de aprovação escolar há um desbalanceamento entre a proporção de aprovados e reprovados. Para lidar com tal aspecto, inerente ao problema de predição de aprovação escolar, utilizou-se o parâmetro 'class_weight', configurado diretamente nos algoritmos compatíveis. Essa abordagem permitiu compensar a desproporção entre classes de forma integrada ao processo de treinamento.

Para avaliar a qualidade dos resultados gerados as abordagens abrangeram tanto métricas convencionais (acurácia, precisão, *recall*, F_1-Score, AUC-ROC) quanto visualizações diagnósticas (curvas ROC, curvas *Precision-Recall* e matrizes de confusão). A análise comparativa entre os modelos base e otimizados foi complementada pelo diagnóstico de sobreajuste, utilizando a função 'verificar_overfitting', que quantificou a discrepância entre as métricas obtidas na validação cruzada e no teste final. Essa análise foi fundamental para aferir a estabilidade e a generalização dos modelos ajustados.

Por fim, a função 'comparar_resultados_classificacao' sintetizou visualmente os desempenhos relativos dos diferentes algoritmos, evidenciando padrões de desempenho e consistência. Os resultados dessas análises estão organizados no Capítulo 4, onde são discutidos à luz dos objetivos preditivos e das implicações educacionais do estudo..

3.6.1 Espaços de Busca e Justificativas dos Hiperparâmetros

A seleção dos melhores hiperparâmetros foi realizada por meio de *grid search*, que explora sistematicamente todas as combinações possíveis. Essa abordagem foi viável devido à compatibilidade com os recursos computacionais disponíveis.

Antes da aplicação do *'grid search'*, definiram-se os espaços de busca considerando simplicidade estrutural, diversidade de estratégias e viabilidade computacional. Para cada algoritmo, estabeleceram-se intervalos ajustáveis que abrangeram desde configurações conservadoras até variantes mais complexas, permitindo avaliar diferentes níveis de capacidade de ajuste e generalização.

Em bases de dados educacionais, a variável de aprovação escolar frequentemente apresenta distribuição desbalanceada, com predominância de casos positivos. Para

reduzir o impacto desse viés no processo de aprendizado dos modelos, foram aplicadas técnicas de ponderação de classes nos algoritmos compatíveis.

Para classificadores baseados em árvores, como Árvores de Decisão e *Random Forest*, variaram-se critérios de impureza ('gini', 'entropy'), profundidade máxima ('max_depth') e parâmetros mínimos para divisão e folhas. Para a *Random Forest*, também foram testadas diferentes estratégias de amostragem de atributos ('max_features') e números de estimadores ('n_estimators').

No *AdaBoost*, além da taxa de aprendizado e do número de estimadores, testaram-se as variantes 'SAMME' e 'SAMME.R', bem como diferentes configurações para o estimador base — árvores com distintas profundidades e pesos de classe. A escolha por árvores rasas ('max_depth = 1' ou '3') está alinhada à literatura, que recomenda classificadores fracos para maximizar os ganhos incrementais do *boosting* (HASTIE et al., 2009).

A Regressão Logística foi configurada com penalidade 'l2', solucionador 'lbfgs' e variações do hiperparâmetro 'C', que controla o grau de regularização. A penalização 'l2' foi mantida fixa por sua estabilidade numérica e robustez em contextos com múltiplos preditores correlacionados, sendo recomendada para o solucionador adotado (PEDREGOSA et al., 2011).

Para o SVM, combinaram-se *kernels* linear e RBF com diferentes valores de 'C' e 'gamma', além da ponderação por classe via 'class_weight'. Essa combinação busca equilibrar interpretabilidade (kernel linear) e capacidade de modelar relações não lineares (kernel RBF) (CORTES; VAPNIK, 1995).

Quadro 3.2: Espaços de hiperparâmetros por modelo e estratégia de balanceamento

Algoritmo	**Hiperparâmetros**	**Valores testados**
Árvore de Decisão	criterion	gini, entropy
	max_depth	1, 3, 5, 10, None
	min_samples_split	2, 5, 10
	min_samples_leaf	1, 2, 4
	class_weight	None, balanced, {0:3, 1:1}, {0:5, 1:1}
Random Forest	n_estimators	50, 100, 300
	max_depth	None, 5, 10
	min_samples_split	2, 3, 5
	min_samples_leaf	1, 2, 3
	max_features	None, sqrt, log2
	class_weight	None, balanced, {0:3, 1:1}, {0:5, 1:1}
AdaBoost	n_estimators	50, 100, 200, 300
	learning_rate	0.001, 0.01, 0.1, 1.0
	algorithm	SAMME, SAMME.R
	estimator	Árvores com max_depth = 1 ou 3 e diferentes class_weight
Regressão Logística	penalty	l2
	C	0.01, 0.1, 1, 10
	solver	lbfgs
	class_weight	None, balanced, {0:3, 1:1}, {0:5, 1:1}
SVM	kernel	linear, rbf
	C	0.1, 1, 10, 100
	gamma	scale, auto, 0.1
	class_weight	None, balanced, {0:3, 1:1}, {0:5, 1:1}

Fonte: Elaborado pela autora.

Os espaços de busca e as configurações adotadas para cada modelo estão apresentados no Quadro 3.2. O delineamento desses espaços foi conduzido de modo a equilibrar robustez preditiva, interpretabilidade e viabilidade computacional. Para lidar com o desbalanceamento da variável de interesse, empregou-se validação cruzada estratificada e ponderação por classe nos algoritmos compatíveis, alinhando o processo de ajuste às práticas recomendadas para problemas dessa natureza no

contexto educacional.

3.7 Avaliação dos modelos

A etapa de Avaliação teve como objetivo verificar a qualidade, estabilidade e capacidade de generalização dos modelos preditivos construídos, assegurando que os resultados estivessem alinhados aos objetivos definidos na fase de entendimento do problema e aos requisitos operacionais de um sistema de apoio à decisão educacional.

As avaliações foram conduzidas a partir de dois conjuntos distintos: o conjunto de teste independente, reservado desde o início do processo analítico, e os subconjuntos derivados do treino utilizados na validação cruzada durante a otimização de hiperparâmetros. Essa estratégia garantiu a imparcialidade dos resultados finais e ofereceu uma estimativa robusta de desempenho em novos dados.

A validação cruzada seguiu o esquema *Stratified K-Fold*, com cinco partições estratificadas para preservar a proporção entre as classes em cada divisão — prática recomendada para cenários com desbalanceamento da variável-alvo (GRAMS, 2024). A estratificação mitiga o risco de partições com escassez de amostras da classe minoritária (estudantes reprovados), fator crítico para a avaliação confiável da capacidade de generalização dos modelos.

A escolha de K = 5 considerou o porte moderado da base de dados, o desbalanceamento entre as classes e o custo computacional das etapas de otimização e validação. Embora valores como K = 10 sejam frequentemente utilizados na literatura (CUNHA, 2019) por reduzirem o viés do estimador, essa configuração pode levar a partições com amostras minoritárias insuficientes e maior variância em conjuntos menores. Assim, cinco partições equilibraram robustez estatística, estabilidade dos resultados e viabilidade computacional.

Com base nessa estrutura de validação, adotaram-se métricas que combinam medidas globais, como acurácia e AUC-ROC, com indicadores sensíveis ao desbalanceamento, como o F_1-Score (média macro e direcionado à classe minoritária), além de precisão e *recall* por classe. Apesar de a classe majoritária (aprovados)

ter sido o foco da predição, o desempenho sobre a classe minoritária (reprovados) recebeu atenção especial, dada sua relevância prática no contexto educacional.

Além das métricas pontuais, foram geradas curvas ROC, curvas *Precision-Recall* e matrizes de confusão para cada modelo, em versões base e otimizadas. As curvas ROC permitiram observar o desempenho dos classificadores sob diferentes limiares, considerando ambas as classes. As curvas *Precision-Recall* foram calculadas com base na classe majoritária (aprovados) e utilizadas como apoio visual para complementar a análise dos modelos. As matrizes de confusão permitiram uma análise detalhada dos tipos de erro mais frequentes, auxiliando na compreensão prática da utilidade dos modelos em contextos educacionais reais.

Diagnóstico de sobreajuste

A identificação de possíveis casos de sobreajuste (*overfitting*) foi realizada por meio de uma função específica, descrita no Apêndice D, que compara os valores obtidos no conjunto de teste com as médias da validação cruzada para cada métrica, calculando a diferença percentual relativa entre ambos. Adotou-se como critério que diferenças inferiores a 10% entre os desempenhos no teste e na validação cruzada seriam consideradas aceitáveis. Diferenças acima desse limiar, quando favoráveis ao conjunto de teste, foram interpretadas como indicativas de sobreajuste; quando favoráveis à validação cruzada, como possíveis sinais de instabilidade ou subajuste.

Essa abordagem baseia-se na premissa de que a validação cruzada, ainda que estratificada, não produz partições totalmente independentes, podendo subestimar a taxa real de erro, especialmente em contextos de desbalanceamento severo ou alta complexidade do modelo (CUNHA, 2019; GRAMS, 2024). Para ampliar a confiabilidade da avaliação da capacidade de generalização, incorporou-se a comparação com um conjunto de teste independente — previamente separado e não utilizado durante as etapas de treinamento ou validação.

A discussão detalhada dos resultados — incluindo a comparação entre modelos, a influência das estratégias de seleção de atributos e as implicações práticas no contexto educacional — está sistematizada no Capítulo 4 deste trabalho.

3.8 Considerações Metodológicas

Esta seção apresenta os aspectos metodológicos centrais deste trabalho, abordando decisões analíticas, éticas e tecnológicas que permearam a pesquisa. Discutem-se o uso responsável de variáveis demográficas, a aplicação de ferramentas de Inteligência Artificial no processo textual e a validação do índice 'PerfilScore' para seleção de atributos categóricos. O objetivo é assegurar transparência metodológica, rigor científico e reflexão crítica sobre os procedimentos adotados.

3.8.1 Considerações éticas sobre variáveis demográficas

Esta pesquisa reconhece o debate atual na literatura de *Educational Data Mining* sobre as implicações éticas do uso de variáveis demográficas em modelos preditivos. Conforme (BAKER et al., 2023), a utilização desses atributos como preditores pode representar uma prática questionável sob a perspectiva da justiça algorítmica.

Neste estudo, variáveis demográficas foram empregadas prioritariamente na análise exploratória e caracterização da população estudada, preservando o foco em variáveis acionáveis — como tempo de estudo, frequência e desempenho anterior — nas etapas de seleção de atributos. Quando incluídas em modelos preditivos, essas variáveis foram selecionadas com base em critérios estatísticos específicos e acompanhadas de justificativas pedagógicas explícitas.

Reconhece-se, contudo, que sua presença nos dados de treinamento pode contribuir para o reforço de vieses estruturais, reduzir o grau de *actionability* (capacidade de ação) das predições e perpetuar desigualdades pré-existentes no contexto educacional.

3.8.2 Considerações sobre o uso de ferramentas de Inteligência Artificial

Durante o desenvolvimento deste trabalho, ferramentas de Inteligência Artificial — especificamente *ChatGPT* e *Gemini AI* — foram utilizadas como assistentes no processo de elaboração e revisão textual. Todo o conteúdo gerado foi posteriormente

revisado, adaptado e validado pela autora, assegurando rigor metodológico e integridade ética.

3.8.3 Considerações sobre a validação do 'PerfilScore'

O 'PerfilScore' foi concebido como um critério heurístico de apoio à análise exploratória e à seleção de atributos categóricos. Sua validação formal, por meio de comparação sistemática com outros métodos ou métricas padronizadas de seleção de atributos, não foi escopo deste trabalho.

Buscou-se, contudo, uma verificação indireta de sua utilidade prática ao observar se os atributos priorizados pelo critério correspondiam àqueles destacados por outras estratégias analíticas (correlação, regressão e testes de hipótese). Complementarmente, sua aplicação contribuiu para a construção de modelos preditivos com desempenho consistente, oferecendo indícios preliminares de sua viabilidade como instrumento exploratório.

Portanto, o 'PerfilScore' deve ser entendido como uma proposta metodológica em fase inicial, com potencial de aplicação prática, mas que requer validações adicionais em estudos futuros para aferição mais assertiva de sua eficácia. Detalhes sobre sua construção e cálculo encontram-se no Apêndice B.2.

Capítulo 4

Resultados e Discussões

Este capítulo apresenta e discute os resultados obtidos a partir da metodologia descrita no Capítulo 3, seguindo as etapas do *pipeline* analítico: análise exploratória, seleção de atributos e avaliação dos modelos preditivos para português e matemática.

A discussão articula-se com o referencial teórico em Mineração de Dados Educacionais (EDM) e *Learning Analytics*, buscando responder às perguntas de pesquisa. Para facilitar a leitura, utilizou-se padronização cromática por disciplina –português: azul e vermelho; matemática: verde e laranja – reforçando visualmente a distinção entre os domínios analisados.

4.1 Análise exploratória dos dados segmentada por disciplina

Esta seção apresenta as principais análises exploratórias conduzidas sobre os dados, focalizando a caracterização das variáveis, identificação de padrões relacionais e construção de perfis estudantis, de modo a subsidiar as etapas posteriores de modelagem preditiva.

4.1.1 Avaliação preliminar da base de dados e validação da estratificação

A análise preliminar das bases indicou que ambas possuem boa cobertura e integridade, sem registros duplicados ou valores ausentes. As bases foram segmentadas em conjuntos de treinamento e teste com estratificação da variável *aprovação*, preservando as proporções originais de aprovados e reprovados e assegurando a representatividade das amostras.

A estratificação da variável *aprovação* preservou a proporção de aprovados nos conjuntos de treino e teste:

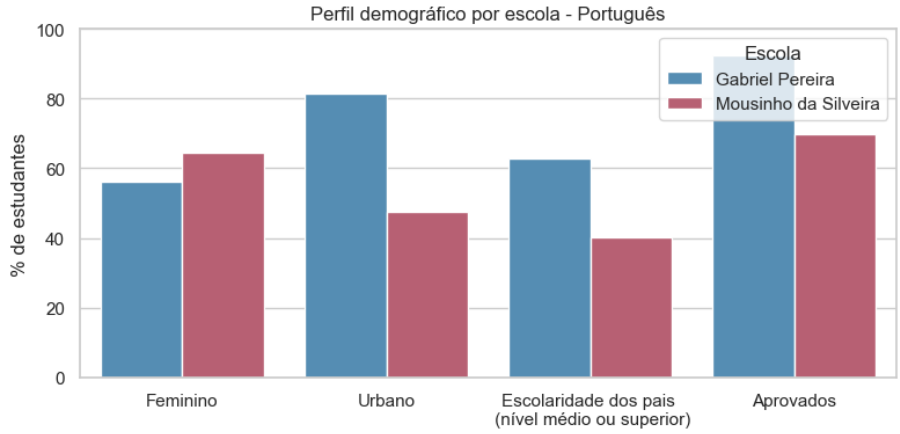
- **Português**: 649 registros (85 % aprovados); treino com 454 observações.
- **Matemática**: 395 registros (67 % aprovados); treino com 276 observações.

Do ponto de vista técnico, a maior dimensão da base de português representa uma vantagem para o desenvolvimento de modelos preditivos mais robustos, proporcionando maior diversidade de padrões e maior capacidade de generalização. Em contrapartida, a base de matemática, mais reduzida, pode apresentar menor potencial para capturar variabilidade e padrões menos frequentes, o que pode impactar negativamente o desempenho dos modelos.

Sob uma perspectiva qualitativa, em matemática o percentual de reprovações é mais que o dobro daquele observado em português. O que indica um desempenho mais comprometido na disciplina e sugere a presença de desafios pedagógicos específicos, que serão retomados ao longo da análise.

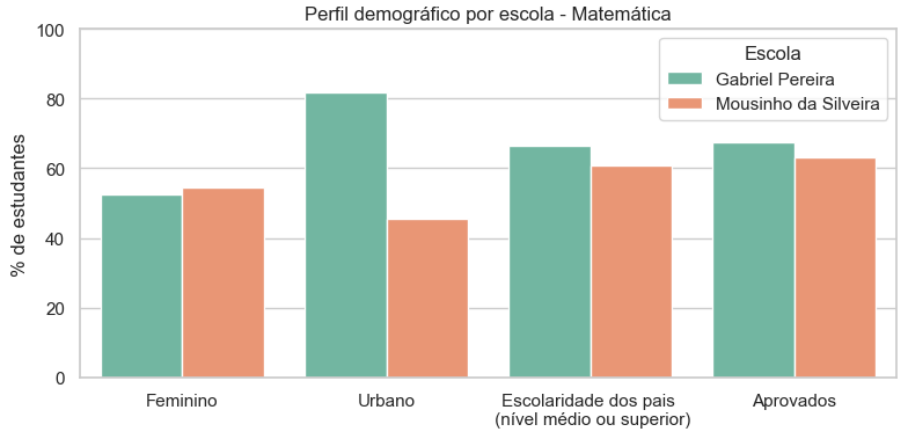
Para as seções subsequentes, as análises consideram a base de dados consolidada por disciplina, sem segmentação por escola. A segmentação entre Gabriel Pereira e Mousinho da Silveira foi previamente examinada (conforme Figuras 4.1 e 4.2) para compreensão do contexto sociodemográfico, mas os quadros estatísticos e os modelos apresentados a seguir resumem informações agregadas para cada disciplina, permitindo uma visão geral do comportamento dos alunos em português e matemática.

Figura 4.1: Indicadores sociodemográficos e aprovação em português segmentados por escola(Gabriel Pereira e Mousinho da Silveira).



Fonte: Elaborado pela autora, com base nos dados analisados (2025).

Figura 4.2: Indicadores sociodemográficos e aprovação em matemática segmentados por escola(Gabriel Pereira e Mousinho da Silveira).



Fonte: Elaborado pela autora, com base nos dados analisados (2025).

A maioria dos alunos é do sexo feminino e mora em zona urbana. No entanto, as Figuras 4.1 e 4.2 mostram que, considerando a variável escola, o Colégio Mousinho da Silveira tem perfil mais rural e feminino, enquanto o Gabriel Pereira apresenta maior equilíbrio entre zona de residência e gênero — padrão menos evidente na base de matemática.

As imagens também expõem diferenças na escolaridade dos responsáveis: Gabriel Pereira concentra mais pais com ensino médio e superior, enquanto Mousinho da Silveira registra predominância de níveis mais baixos. Em português, a taxa de

aprovação tende a acompanhar esse cenário.

Implicações para a modelagem preditiva

Os resultados preliminares contribuíram para a escolha dos algoritmos empregados na modelagem preditiva. Optou-se por classificadores com diferentes graus de robustez e capacidade de generalização, incluindo regressão logística, máquinas de vetor de suporte (SVM), florestas aleatórias (RF), *AdaBoost* e árvores de decisão.

A maior variabilidade observada na base de matemática evidenciou a necessidade de modelos mais sensíveis à dispersão dos dados e, possivelmente, de estratégias específicas de balanceamento. No caso da SVM, foram ajustados os hiperparâmetros C e *gamma* para reduzir os efeitos da escala e melhorar a adaptação aos padrões identificados na análise exploratória.

Além disso, os achados reforçam a existência de contextos socioeconômicos distintos entre as escolas e confirmam o alinhamento estrutural entre as bases completas e os conjuntos de treinamento, assegurando condições adequadas para a construção e validação dos modelos preditivos.

4.1.2 Exploração das variáveis quantitativas e padrões de outliers

A análise das variáveis quantitativas expôs padrões contrastantes de desempenho, frequência e perfil etário entre os estudantes nas disciplinas de português e matemática. As Tabelas 4.1 e 4.2 sintetizam os principais indicadores estatísticos das bases, servindo de base para as interpretações subsequentes.

Tabela 4.1: Resumo estatístico das variáveis quantitativas — português

Variável	Média	Mediana	Desvio Padrão	CV	Shapiro-Wilk (p)
Idade	16,71	17	1,23	0,07	$< 0{,}001$
Faltas	3,86	2	4,74	1,23	$< 0{,}001$
Nota 1	11,46	11	2,73	0,24	$< 0{,}001$
Nota 2	11,56	11	2,89	0,25	$< 0{,}001$
Nota final	11,95	12	3,13	0,26	$< 0{,}001$

Fonte: Elaborado pela autora, com base nos dados analisados (2025).

Tabela 4.2: Resumo estatístico das variáveis quantitativas — matemática

Variável	Média	Mediana	Desvio Padrão	CV	Shapiro-Wilk (p)
Idade	16,67	17	1,24	0,07	$< 0{,}001$
Faltas	5,81	3	7,82	1,35	$< 0{,}001$
Nota 1	10,88	11	3,32	0,31	$< 0{,}001$
Nota 2	10,76	11	3,71	0,35	$< 0{,}001$
Nota final	10,43	11	4,58	0,44	$< 0{,}001$

Fonte: Elaborado pela autora, com base nos dados analisados (2025).

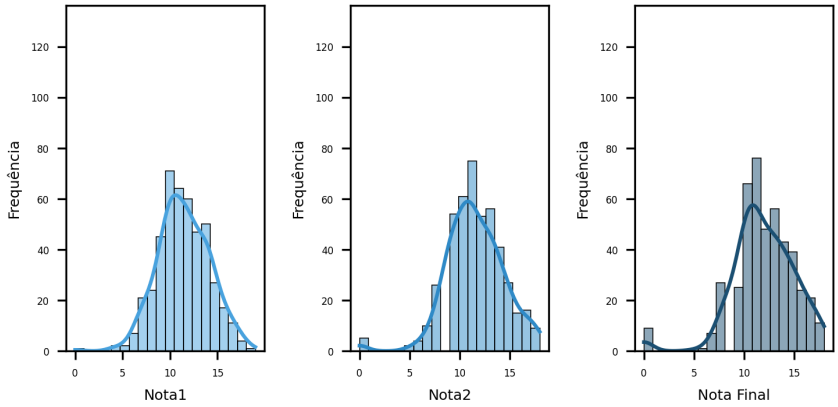
Perfil etário: As duas bases apresentam perfil etário semelhante, com média em torno de 16,7 anos e baixo coeficiente de variação (CV), indicando relativa homogeneidade. A maioria dos estudantes situa-se na faixa etária esperada para o ensino médio, embora os valores máximos — 22 anos em português e 21 em matemática — sugiram trajetórias escolares não lineares, possivelmente marcadas por reprovações ou interrupções.

Padrões de frequência: No que se refere às faltas, percebe-se um comportamento mais heterogêneo, sobretudo em matemática, que apresenta média de 5,81 faltas — cerca de 50% superior à de português (3,86) — e maior coeficiente de variação (1,35 contra 1,23), reforçando a maior dispersão nos padrões de frequência. Essa discrepância pode refletir níveis distintos de engajamento ou percepção de dificuldade, alinhando-se a evidências de maior evasão ou desmotivação em matemática (ALRESHIDI, 2023; ROZGONJUK et al., 2020).

Desempenho acadêmico: O desempenho acadêmico também reforça esse contraste entre as disciplinas. Em matemática, a média das notas finais é de 10,43, inferior à de português (11,95), com diferença de aproximadamente 1,5 ponto. A dispersão também é mais acentuada — o coeficiente de variação da nota final é de 0,44, comparado a 0,26 em português —, indicando uma distribuição mais assimétrica e com maior concentração de casos de baixo rendimento.

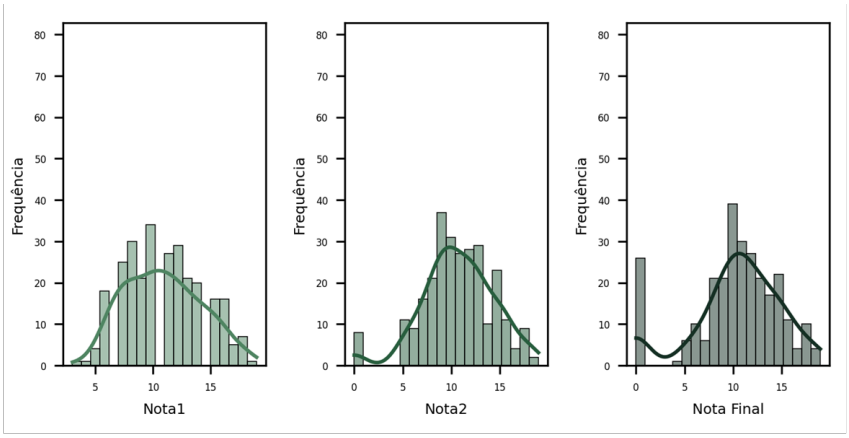
As Figuras 4.3 e 4.4 evidenciam visualmente padrões como caudas alongadas e concentração em torno do limiar de aprovação.

Figura 4.3: Distribuição das notas — português



Fonte: Elaborado pela autora, com base nos dados analisados (2025).

Figura 4.4: Distribuição das notas — matemática



Fonte: Elaborado pela autora, com base nos dados analisados (2025).

Os gráficos demonstram que, em ambas as disciplinas, as distribuições de no-

tas apresentam padrões assimétricos, embora com características distintas. Em português, observa-se uma leve assimetria à esquerda em todas as avaliações, com concentração próxima ao limiar de aprovação, sugerindo um comportamento de manutenção ou leve melhora ao longo do ano letivo. Em matemática, além de maior dispersão e variação entre as avaliações, as distribuições revelam maior irregularidade e presença mais visível de notas em faixas muito baixas.

Implicações estatísticas: A assimetria nas distribuições, evidente nos histogramas, confirma a não normalidade das variáveis de nota — evidência corroborada pelos testes de Shapiro-Wilk ($p < 0{,}001$), que indicaram rejeição sistemática da hipótese de normalidade para todas as variáveis analisadas. Esses resultados justificam a adoção de abordagens analíticas robustas, capazes de lidar com assimetrias acentuadas e presença de outliers.

Diante disso, optou-se por métodos estatísticos não paramétricos nas análises inferenciais. Para variáveis ordinais, foram utilizados os testes de *Spearman* e Kruskal-Wallis; para variáveis nominais, os testes de qui-quadrado de independência e o coeficiente V de Cramer.

Faixas por agrupamento e aprovação

A análise do desempenho acadêmico com base em agrupamentos empíricos de ausência e idade revelou tendências importantes. A categorização — construída a partir de quartis e limites de *outliers* — permitiu estruturar os dados em perfis comparáveis entre português e matemática. As principais tendências foram sintetizadas no Quadro 4.1.

Quadro 4.1: Resumo comparativo por faixas de ausência, idade e aprovação

Dimensão	**português**	**matemática**
Faixa de ausência	Regular (0–6), moderada (7–15), elevada (>15)	Regular (0–8), moderada (9–19), elevada (>19)
Tendência de desempenho	Redução gradual das notas conforme aumento das ausências; impacto mais pronunciado entre reprovados	Padrão semelhante, mas com maior dispersão e presença de *outliers* entre aprovados com ausência elevada
Diferença entre aprovados e reprovados	Gap consistente: variação de 4 a 5 pontos em todas as faixas; reforça validade preditiva da nota final	Diferença mantida, porém com maior sobreposição entre grupos em faixas intermediárias de ausência.
Perfil etário	Predominância de estudantes com idades inferiores a 19 anos; média de 16,6 anos; faixas mais velhas associadas a maior reprovação	Padrão equivalente, mas mais estudantes acima de 18 anos entre reprovados e com ausência elevada

Fonte: Elaborado pela autora, com base nos dados analisados (2025).

De modo geral, os padrões confirmam a associação entre maior número de faltas, idade mais avançada e menores níveis de desempenho, especialmente em português. Observou-se que o aumento das ausências reduz gradualmente a média final, com impacto mais acentuado entre os reprovados. A segmentação etária também evidenciou maior risco de reprovação entre estudantes mais velhos, sobretudo aqueles com frequência irregular. Esses achados reforçam o potencial discriminativo dessas variáveis e justificam sua permanência nas etapas seguintes de análise.

Identificação e caracterização de outliers

A identificação de valores atípicos foi conduzida com base no critério do Intervalo Interquartil (IQR), estabelecendo limites superiores e inferiores para a classificação de casos extremos nas variáveis faltas e nota final. Foram identificados 9 *outliers* para a variável *nota final* em português e 26 em matemática. No caso das *faltas*, os casos extremos somaram 17 em português e 12 em matemática. Embora o número de outliers seja pequeno, sua análise permite explorar perfis críticos com potencial valor explicativo, tanto para fins analíticos quanto para modelagem preditiva.

Outliers em nota final O Quadro 4.2 revela que os casos de desempenho extremamente baixo não estão distribuídos de forma homogênea entre as instituições. Em português, a maioria concentra-se na escola Mousinho da Silveira, enquanto em matemática predominam na Gabriel Pereira, sugerindo possíveis fragilidades curriculares ou dificuldades institucionais localizadas — ainda que não se possa estabelecer generalizações.

Quadro 4.2: Perfil predominante dos *outliers* com nota final extremamente baixa

Dimensão	Perfil predominante (português)	Perfil predominante (matemática)
Escola	Mousinho da Silveira (88,89%)	Gabriel Pereira (88,46%)
Residência	Zona Rural (66,67%)	Zona Urbana (84,62%)
Apoio educacional	Baixo: sem apoio escolar e familiar (>77%)	Baixo: sem apoio escolar (96,15%)
Família	Estruturas maiores (>3 membros), pais casados e mãe como responsável legal	Estruturas maiores (>3 membros)
Estilo de vida	Pouco tempo de estudo, mas atividades extracurriculares presentes	Frequente participação em atividades extracurriculares
Escolaridade parental	Baixa a média, especialmente do pai (nível 2 predominante)	Predomínio de nível 2 (34,62%)
Expectativas	Interesse no ensino superior (66,67%)	Interesse no ensino superior (80,77%)
Idade	$\geq$ 18 anos	$\geq$ 18 anos
Consumo de álcool	Baixo a moderado: dias úteis majoritariamente baixo; fim de semana mais disperso	Baixo a moderado: predominância de consumo ocasional
Assiduidade	O grupo apresentou zero faltas	Faltas variáveis, mas grupo não homogêneo

Fonte: Elaborado pela autora, com base nos dados analisados (2025).

Entre os perfis identificados, observa-se predominância de estudantes com 18 anos ou mais, inseridos em famílias numerosas, com baixa escolaridade parental (sobretudo do pai) e acesso limitado a apoio educacional, tanto escolar quanto familiar. Apesar dessas limitações, a maioria declara interesse no ensino superior — especialmente em matemática (80,77%) — indicando um descompasso entre expectativas

acadêmicas e condições objetivas de permanência e sucesso escolar.

Chama atenção, no caso de português, a ocorrência de baixo desempenho mesmo entre estudantes com zero faltas, o que pode levantar hipóteses sobre a insuficiência da mera presença física em sala como garantia de aprendizagem em contextos vulneráveis. Já em matemática, as faltas aparecem de forma mais variada, sugerindo uma possível relação mais complexa entre assiduidade, currículo e suporte institucional — hipótese que exigiria investigação mais aprofundada.

***Outliers* em faltas** Os casos analisados concentram-se integralmente no Colégio Gabriel Pereira, em ambas as disciplinas, sugerindo um problema localizado de assiduidade. Predominam estudantes do sexo feminino, residentes em área urbana e com pouco apoio escolar. Em português, a maioria não participa de atividades extracurriculares (64,71%), enquanto em matemática 75% participam, indicando um possível engajamento seletivo que não se reflete na frequência às aulas.

Vê-se ainda o alto interesse pelo ensino superior (acima de 75%) e o histórico escolar sem reprovações na maioria dos casos. Esse perfil — de alta abstenção — sugere que a evasão parcial não está necessariamente ligada ao desinteresse acadêmico ou à repetência, apontando para a complexidade dos fatores envolvidos.

Quadro 4.3: Perfil predominante dos *outliers* com número elevado de faltas

Dimensão	**Perfil predominante (português)**	**Perfil predominante (matemática)**
Escola	Gabriel Pereira (100%)	Gabriel Pereira (100%)
Residência	Urbana (82,35%)	Urbana (91,67%)
Gênero	Feminino (64,71%)	Feminino predominante (75%)
Apoio escolar	Ausente em 100%	Ausente em 91,67%
Aulas particulares	Ausentes em 100%	Presentes em 75%
Atividades extracurriculares	Ausentes em 64,71%	Presentes em 75%
Interesse em ensino superior	Presente em 76,47%	Presente em 91,67%
Tempo de transporte	-	Menor que 15 minutos (66,67%)
Histórico escolar	Sem reprovações (70,6%)	Sem reprovações (50%)

Fonte: Elaborado pela autora, com base nos dados analisados (2025).

Os achados apresentados reforçam que, apesar de representarem casos extremos e envolverem amostras reduzidas, os *outliers* identificados oferecem indícios relevantes sobre perfis de risco frequentemente invisíveis em análises centradas na média. Sua manutenção no conjunto de dados visa ampliar a representatividade de realidades educacionais periféricas e enriquecer o processo de modelagem preditiva. Ressalta-se, contudo, que as hipóteses levantadas nesta seção devem ser interpretadas com cautela, dado o caráter descritivo da abordagem e a limitação amostral envolvida.

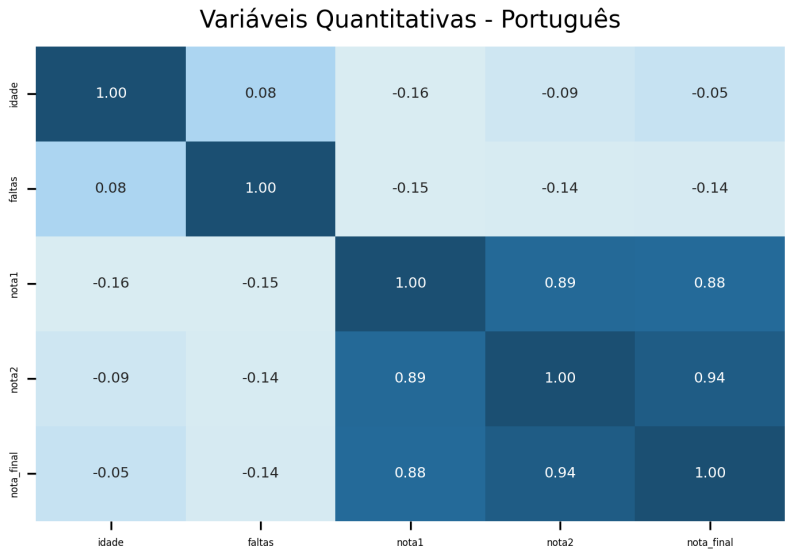
4.1.3 Relações entre fatores e desempenho escolar

A investigação das correlações de *Spearman* entre as variáveis quantitativas e ordinais buscou identificar padrões de associação relevantes ao desempenho estudantil. A seguir, são apresentados os principais coeficientes observados para português e matemática, permitindo uma leitura comparativa das relações entre atributos e possíveis influências sobre os resultados finais.

Relações entre medidas quantitativas e desempenho

As Figuras 4.5 e 4.6 apresentam, respectivamente, as matrizes de correlação entre as variáveis quantitativas das bases de português e matemática, permitindo a visualização conjunta dos padrões de associação discutidos.

Figura 4.5: Correlação de *Spearman* — variáveis quantitativas — português

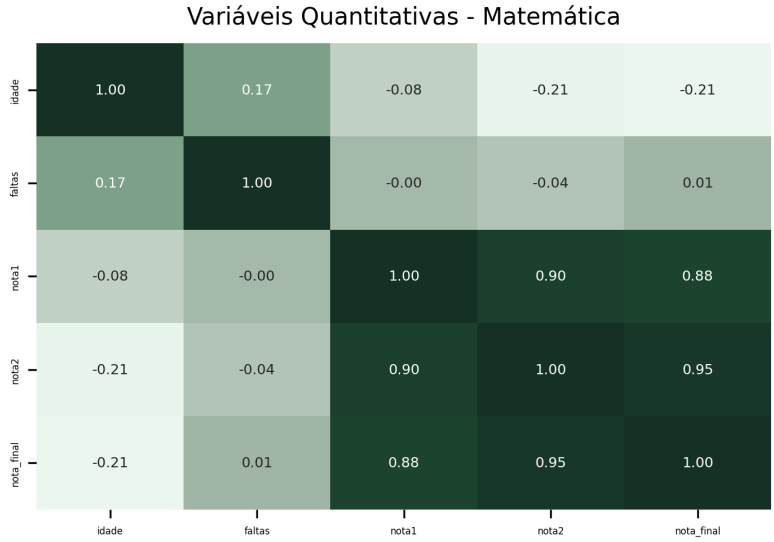


Fonte: Elaborado pela autora, com base nos dados analisados (2025).

Ao se analisarem as associações entre as notas, observam-se correlações elevadas entre as diferentes avaliações em ambas as disciplinas, o que sinaliza uma forte consistência no desempenho dos estudantes ao longo do ano letivo. Avaliando a correlação das parciais com as notas finais apresentam associação alta com a segunda avaliação — $\rho = 0{,}94$ em português e $\rho = 0{,}95$ em matemática — enquanto a correlação entre a primeira e a segunda avaliação, embora ligeiramente inferior, permanece robusta ($\rho = 0{,}89$ e $\rho = 0{,}90$, respectivamente). Tal padrão sugere uma possível progressão ao longo do período letivo, com maior alinhamento dos estudantes aos conteúdos e critérios avaliativos nas etapas finais.

No que se refere à frequência escolar — constantemente considerada um indicativo relevante de engajamento — observaram-se padrões de correlação pouco expressivos com o desempenho final. As faltas apresentaram correlações fracas em ambas as disciplinas. Em português, identificou-se uma associação negativa e modesta ($\rho \approx -0{,}14$), sugerindo uma leve tendência de queda no rendimento à medida

Figura 4.6: Correlação de *Spearman* — variáveis quantitativas — matemática

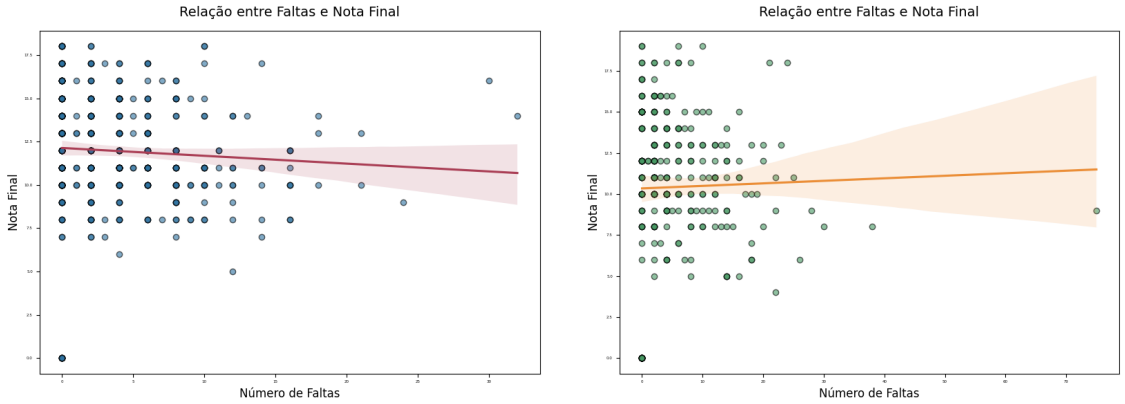


Fonte: Elaborado pela autora, com base nos dados analisados (2025).

que o número de faltas aumenta. Em matemática, por outro lado, a correlação foi praticamente nula ($\rho \approx 0{,}01$), indicando que, nesse caso, a assiduidade teve impacto desprezível sobre as notas finais.

Esses resultados dialogam com as hipóteses discutidas anteriormente nesta subseção, segundo as quais a simples presença em sala de aula não garante, por si só, a aprendizagem — especialmente em contextos de maior vulnerabilidade, como o observado no Colégio Gabriel Pereira. No conjunto total, verificou-se que o número de faltas não apresenta correlação forte com as notas finais, reforçando a ideia de que a assiduidade, embora relevante, não atua isoladamente sobre o desempenho. Em matemática, o padrão de dispersão mais acentuado e a ausência de correlação significativa sugerem uma relação mais complexa, possivelmente mediada por fatores curriculares ou institucionais. Essas tendências estão representadas na Figura 4.7, que apresenta os gráficos de dispersão entre número de faltas e nota final, com linhas de tendência linear para ambas as disciplinas.

Figura 4.7: Relação entre faltas e nota final



(a) português (b) matemática

Fonte: Elaborado pela autora, com base nos dados analisados (2025).

A variável idade também apresentou correlações discretas, embora com distinções relevantes entre as disciplinas. Em português, verificaram-se correlações negativas fracas entre idade e desempenho, sinalizando que estudantes mais velhos tendem a apresentar ligeiramente menores notas finais. A associação com o número de faltas, nesse caso, foi positiva, porém quase nula ($\rho \approx 0{,}08$). Em matemática, os padrões foram mais pronunciados: a idade apresentou correlações negativas com a nota final e a segunda avaliação ($\rho \approx -0{,}21$), além de uma correlação positiva com as faltas ($\rho \approx 0{,}16$).

Embora esses dados indiquem uma tendência estatística, é essencial interpretar tais padrões com cautela. A maior idade pode refletir trajetórias escolares irregulares — como reprovações, interrupções temporárias ou retorno tardio à escola —, frequentemente associadas a contextos de vulnerabilidade. Portanto, mais do que um fator causal, a idade pode funcionar como marcador de desigualdades educacionais acumuladas. Essa perspectiva reforça a necessidade de considerar princípios de justiça (*fairness*) e equidade na análise de desempenho escolar, especialmente ao utilizar variáveis demográficas em processos de modelagem preditiva.

Impacto de variáveis ordinais no rendimento dos alunos

A Tabela 4.3 apresenta as maiores correlações de *Spearman* entre variáveis ordinais, evidenciando padrões comportamentais e contextuais consistentes entre as bases de português e matemática. A associação mais forte foi observada entre os níveis de escolaridade dos pais (ρ = 0,65 e 0,66), sugerindo certa homogeneidade educacional dentro dos núcleos familiares. Da mesma forma, os consumos de álcool durante a semana e aos fins de semana apresentaram correlações elevadas ($\rho > 0{,}60$), indicando padrões de comportamento de risco relativamente estáveis entre os estudantes.

Tabela 4.3: Maiores correlações de *Spearman* entre variáveis ordinais

Variável 1	**Variável 2**	**português**	**matemática**
escolaridade_pai	escolaridade_mae	0,65	0,66
álcool (fim de semana)	álcool (dias úteis)	0,60	0,62
freq. saídas	tempo livre	0,37	0,29
álcool (fim de semana)	freq. saídas	0,33	0,41
álcool (dias úteis)	freq. saídas	0,23	0,26
álcool (dias úteis)	tempo livre	—	0,21
reprovações	escolaridade_mae	-0,20	-0,21
reprovações	escolaridade_pai	—	-0,21
álcool (fim de semana)	tempo de estudo	-0,24	-0,23
álcool (dias úteis)	tempo de estudo	-0,21	-0,24
tempo de transporte	escolaridade_mae	-0,26	—
tempo de transporte	escolaridade_pai	-0,20	—

Fonte: Elaborado pela autora, com base nos dados analisados (2025).

Outras correlações relevantes emergem entre variáveis relacionadas ao cotidiano dos estudantes. Tempo livre e frequência de saídas se correlacionam positivamente, sugerindo que jovens com maior autonomia social tendem a dedicar mais tempo a atividades de lazer. Já o tempo de estudo apresentou correlação negativa com o consumo de álcool — tanto durante a semana quanto nos fins de semana —, reforçando a hipótese de que determinados hábitos prejudiciais ao bem-estar comprometem o engajamento escolar. O histórico de reprovações, por sua vez, associou-se negativa-

mente à escolaridade dos pais, sugerindo que famílias com menor capital educacional enfrentam maiores dificuldades para sustentar trajetórias escolares estáveis.

No que se refere ao desempenho acadêmico, as variáveis com maior magnitude de correlação foram o número de reprovações anteriores e o tempo de estudo, ambas com sinais esperados. As reprovações se associaram negativamente às notas finais ($\rho = -0{,}46$ em português e $-0{,}36$ em matemática). A escolaridade dos pais também mostrou associação positiva com o desempenho, embora com menor intensidade.

Esses dados estão sintetizados na Tabela 4.4, que apresenta as principais correlações entre atributos ordinais e desempenho nas disciplinas analisadas.

Tabela 4.4: Principais correlações de *Spearman* entre atributos categóricos e desempenho acadêmico

Variável ordinal	**português**			**matemática**		
	Faltas	Nota 1	Nota final	Faltas	Nota 1	Nota final
Álcool (dias úteis)	0,18	-0,20	-0,20	0,11	-0,10	-0,14
Álcool (fim de semana)	0,16	-0,14	-0,16	0,23	-0,13	-0,15
Escolaridade da mãe	-0,02	0,26	0,26	0,12	0,21	0,23
Escolaridade do pai	0,08	0,25	0,25	-0,01	0,23	0,21
Frequência de saídas	0,10	-0,09	-0,12	0,15	-0,16	-0,17
Relação familiar	-0,10	0,07	0,09	-0,13	0,02	0,07
Reprovações anteriores	0,12	-0,44	-0,46	0,12	-0,32	-0,36
Saúde	-0,05	-0,05	-0,10	-0,01	-0,06	-0,04
Tempo de estudo	-0,09	0,28	0,26	-0,12	0,12	0,05
Tempo livre	-0,05	-0,08	-0,12	0,04	0,05	0,02
Tempo de transporte	0,03	-0,17	-0,13	-0,00	-0,07	-0,12

Fonte: Elaborado pela autora, com base nas matrizes de correlação de *Spearman* (2025).

Um aspecto que merece atenção específica é o comportamento da variável tempo de estudo, que, embora positiva nas duas disciplinas, apresentou correlações bem mais expressivas em português ($\rho = 0{,}28$ com a Nota 1 e 0,26 com a Nota Final) do que em matemática ($\rho = 0{,}12$ e 0,05, respectivamente). Esse contraste sugere possíveis diferenças na natureza das aprendizagens envolvidas. Em português, o es-

tudo autônomo parece estar mais diretamente associado à melhora no desempenho, enquanto em matemática o efeito é substancialmente menor — o que pode indicar maior dependência de mediação pedagógica, tutoria estruturada ou resolução assistida de problemas. Essa assimetria pode estar relacionada à complexidade cumulativa dos conteúdos matemáticos e à necessidade de acompanhamento sistemático para superar lacunas conceituais, especialmente entre estudantes com histórico de dificuldades.

Por fim, os dados também revelam o papel dos fatores comportamentais sobre o desempenho e a assiduidade. O consumo de álcool e a alta frequência de saídas mantêm correlações negativas com as notas e positivas com o número de faltas, sugerindo um possível comprometimento da rotina escolar e do rendimento acadêmico. Embora essas correlações sejam de baixa magnitude, os padrões observados são consistentes com a literatura sobre comportamentos de risco e vulnerabilidade escolar.

A exposição dos resultados está organizada em blocos temáticos, visando facilitar a comparação e a interpretação:

1. perfil demográfico e estrutural;
2. contexto familiar e socioeconômico;
3. apoio institucional e recursos educacionais;
4. estilo de vida e hábitos comportamentais;
5. condições acadêmicas e comportamento estudantil.

Distribuições e taxas de aprovação por categoria

As Tabelas 4.5 a 4.9 apresentam, antes dos comentários interpretativos, as taxas de aprovação por categoria, organizadas tematicamente para comparação entre disciplinas.

Tabela 4.5: Taxa de Aprovação por Categoria – Perfil Demográfico e Estrutural

Variável	**Categoria**	**Português (%)**	**Matemática (%)**
Gênero	Mulher	85,77	65,54
Gênero	Homem	82,89	68,75
Endereço	Urbano	87,62	69,27
Endereço	Rural	77,70	58,62
Status parental	Juntos	85,06	65,85
Status parental	Separados	81,36	76,67
Tamanho da família	Mais de 3 membros	84,33	65,28
Tamanho da família	Até 3 membros	85,19	71,08

Fonte: Elaborado pela autora, com base nos dados analisados (2025).

Perfil demográfico e estrutural As diferenças por gênero mostram-se pequenas, com vantagens discretas das mulheres em Português e dos homens em Matemática, sugerindo que o gênero isoladamente exerce efeito limitado sobre o desempenho. Em contraste, a diferença urbano–rural é substancial, com aproximadamente 10 pontos percentuais de vantagem das áreas urbanas em ambas as disciplinas, sugerindo desigualdades estruturais no acesso a recursos educacionais. Em relação ao status parental, a maior taxa de aprovação observada entre estudantes com pais separados em Matemática pode refletir a influência de fatores mediadores não mensurados. No entanto, essa interpretação deve ser feita com cautela, pois pode haver desbalanceamento amostral entre os grupos analisados.

Tabela 4.6: Taxa de Aprovação por Categoria – Contexto Familiar e Socioeconômico

Variável	Categoria	Português (%)	Matemática (%)
Escolaridade da mãe	0 - Nenhuma escolaridade	100,00	100,00
Escolaridade da mãe	1 - Ensino fundamental (até 4º ano)	74,74	57,14
Escolaridade da mãe	2 - Ensino fundamental (5º ao 9º ano)	86,15	64,00
Escolaridade da mãe	3 - Ensino médio	82,18	65,67
Escolaridade da mãe	4 - Ensino superior	92,00	74,73
Escolaridade do pai	0 - Nenhuma escolaridade	100,00	100,00
Escolaridade do pai	1 - Ensino fundamental	72,65	54,72
Escolaridade do pai	2 - Ensino fundamental (5º ao 9º ano)	84,67	66,28
Escolaridade do pai	3 - Ensino médio	93,55	67,16
Escolaridade do pai	4 - Ensino superior	89,89	76,81
Profissão da mãe	Outra profissão	84,36	61,39
Profissão da mãe	Serviços	83,51	70,83
Profissão da mãe	Dona de casa	80,22	73,53
Profissão da mãe	Professor(a)	94,23	66,67
Profissão da mãe	Área da saúde	85,71	70,83
Profissão do pai	Outra profissão	83,67	69,80
Profissão do pai	Serviços	83,58	58,23
Profissão do pai	Dono de casa	83,33	72,73
Profissão do pai	Professor(a)	95,83	78,95
Profissão do pai	Área da saúde	93,33	66,67
Responsável legal	Mãe	82,92	68,21
Responsável legal	Pai	90,20	69,49
Responsável legal	Outro responsável	83,33	50,00

Fonte: Elaborado pela autora, com base nos dados analisados (2025).

As taxas de aprovação de 100% observadas entre estudantes cujos pais não possuem escolaridade formal resultam de amostras extremamente reduzidas — cerca de 1 caso em Matemática e de 3 a 5 em Português — e, portanto, não são representativas. Excluindo-se esses casos extremos, observa-se uma tendência geral de aumento das taxas de aprovação à medida que cresce a escolaridade dos pais, em consonância com a teoria do capital cultural de *Bourdieu* (EPSJV, 2009). As diferenças entre os níveis mais baixos e mais altos de escolaridade superam 20 pontos percentuais, sendo mais acentuadas em Matemática. No que se refere à profissão dos pais, lares com pai professor apresentam desempenho consistentemente elevado em ambas as disciplinas. Para mães professoras, o efeito é particularmente marcante em Português, mas menos uniforme em Matemática. As diferenças registradas entre responsáveis legais (pai versus mãe) podem refletir dinâmicas distintas de engajamento doméstico.

Tabela 4.7: Taxa de Aprovação por Categoria – Apoio Institucional

Variável	**Categoria**	**Português (%)**	**Matemática (%)**
Apoio escolar	Sim	85,71	52,78
Apoio escolar	Não	84,44	69,17
Atividades extracurriculares	Sim	86,18	65,71
Atividades extracurriculares	Não	83,12	68,38
Aulas particulares	Sim	86,36	70,80
Aulas particulares	Não	84,49	63,31
Acesso à internet	Sim	86,57	67,67
Acesso à internet	Não	77,88	63,64
Frequentou creche	Sim	84,44	66,22
Frequentou creche	Não	85,11	70,37
Escola	Gabriel Pereira	92,64	68,16
Escola	Mousinho da Silveira	69,03	58,06

Fonte: Elaborado pela autora, com base nos dados analisados (2025).

Apoio institucional e recursos educacionais O apoio escolar, entendido como suporte extra oferecido pela instituição, apresentou associação negativa com a taxa de aprovação, possivelmente por ser implementado de forma reativa, após a identi-

ficação de dificuldades. O acesso à internet, que pode servir de suporte acadêmico, teve efeito positivo, embora moderado. As aulas particulares associaram-se a melhores resultados, sobretudo em Matemática, enquanto as atividades extracurriculares mostraram efeito pouco consistente entre as disciplinas. A variável referente a frequência à creche apresentou impacto reduzido. Por fim, observam-se diferenças relevantes entre escolas, que podem refletir tanto características institucionais quanto perfis distintos de estudantes, incluindo possíveis vínculos com contextos urbanos ou rurais.

Tabela 4.8: Taxa de Aprovação por Categoria – Estilo de Vida e Hábitos

Variável	**Categoria**	**Português (%)**	**Matemática (%)**
Álcool (dias úteis)	1 - muito baixo	87,04	70,56
Álcool (dias úteis)	2 - baixo	79,22	60,00
Álcool (dias úteis)	3 - moderado	80,00	58,82
Álcool (dias úteis)	4 - alto	77,78	37,50
Álcool (dias úteis)	5 - muito alto	71,43	75,00
Álcool (fim de semana)	1 - muito baixo	87,57	68,22
Álcool (fim de semana)	2 - baixo	87,96	71,19
Álcool (fim de semana)	3 - moderado	79,76	72,73
Álcool (fim de semana)	4 - alto	80,77	56,76
Álcool (fim de semana)	5 - muito alto	75,76	50,00
Frequência de saídas	1 - muito baixa	75,00	70,59
Frequência de saídas	2 - baixa	88,50	75,64
Frequência de saídas	3 - moderada	90,34	74,07
Frequência de saídas	4 - alta	82,47	52,38
Frequência de saídas	5 - muito alta	73,13	56,76
Relacionamento romântico	Sim	81,29	59,78
Relacionamento romântico	Não	86,57	70,65
Reprovações	sem reprovações	91,05	74,77
Reprovações	1 reprovação	58,33	51,35
Reprovações	2 reprovações	42,86	9,09
Reprovações	3 ou mais reprovações	33,33	20,00

Fonte: Elaborado pela autora, com base nos dados analisados (2025).

Estilo de vida e hábitos comportamentais Observa-se uma tendência gradativa em que maior consumo de álcool está associado a menores taxas de aprovação, especialmente nas faixas intermediárias e altas; as variações nas extremidades indicam amostras reduzidas. A frequência de saídas apresenta um pico de aprovação em níveis moderados, caindo nas categorias mais baixas e mais altas, sugerindo possível efeito de equilíbrio entre socialização e estudo. Estar em relacionamento romântico associa-se a menor aprovação, possivelmente por competição de tempo e atenção. Por fim, o histórico de reprovações anteriores mostra efeito negativo acentuado e monotônico, evidenciando trajetórias acadêmicas de risco que podem se beneficiar de intervenção precoce.

Tabela 4.9: Taxa de Aprovação por Categoria – Interesse e Rotina Escolar

Variável	**Categoria**	**Português (%)**	**Matemática (%)**
Interesse ensino superior	Sim	87,78	68,32
Interesse ensino superior	Não	55,56	42,86
Motivo escolha escola	Curso específico	80,31	61,46
Motivo escolha escola	Próximo de casa	90,27	68,42
Motivo escolha escola	Reputação da escola	91,26	69,23
Motivo escolha escola	Outro motivo	73,33	76,92
Tempo de estudo (semanal)	1 – Menos de 2 horas	75,51	63,89
Tempo de estudo (semanal)	2 – De 2 a 5 horas	86,64	67,57
Tempo de estudo (semanal)	3 – De 5 a 10 horas	94,29	72,50
Tempo de estudo (semanal)	4 – Mais de 10 horas	95,00	62,50
Tempo livre	1 - muito baixo	85,71	53,85
Tempo livre	2 - baixo	89,87	69,57
Tempo livre	3 - moderado	85,89	66,99
Tempo livre	4 - alto	84,73	67,09
Tempo livre	5 - muito alto	69,57	68,57
Tempo transporte	1 – Menos de 15 minutos	86,38	69,73
Tempo transporte	2 – De 15 a 30 minutos	84,72	63,64
Tempo transporte	3 – De 30 minutos a 1 hora	72,50	54,55
Tempo transporte	4 – Mais de 1 hora	84,62	33,33

Fonte: Elaborado pela autora, com base nos dados analisados (2025).

Condições acadêmicas e comportamento estudantil O interesse em cursar o ensino superior associa-se a taxas de aprovação substancialmente mais altas em ambas as disciplinas, sugerindo papel relevante da motivação acadêmica. O motivo da escolha da escola também se relaciona ao desempenho: os melhores resultados são observados entre estudantes que optaram pela instituição por reputação ou proximidade, enquanto as menores taxas aparecem no grupo que indicou "outro motivo". Quanto ao tempo de estudo semanal, há ganhos expressivos até a faixa de 5 a 10 horas, seguidos de retornos decrescentes nas categorias mais altas, especialmente em Matemática, o que indica a importância da qualidade e eficiência do estudo. O tempo livre apresenta associação pouco consistente, sem padrão claro entre disciplinas. Já o tempo de transporte mostra relação negativa com o desempenho à medida que a distância aumenta, possivelmente em função de fadiga ou menor disponibilidade para atividades acadêmicas.

4.1.4 Análise exploratória orientada por regressão

Com o objetivo de identificar fatores associados ao desempenho intermediário (*nota2*), foi realizada análise exploratória com modelos de regressão linear. Testaram-se quatro abordagens:

1. modelo simples, apenas com a *nota1* como preditor;
2. modelo completo, com todas as variáveis disponíveis;
3. modelo com as 15 variáveis de menor p-valor;
4. modelo *stepwise*, guiado pelos critérios AIC e BIC.

A Tabela 4.10 compara o desempenho dessas abordagens para Português e Matemática. Os modelos *stepwise* apresentaram melhor equilíbrio entre ajuste e parcimônia, explicando 75,1% e 79,3% da variação em *nota2*, respectivamente, com menos de 11 variáveis significativas. Em ambos os casos, a *nota1* foi o preditor mais forte ($\beta = 0{,}854$ em Português e $\beta = 0{,}950$ em Matemática), reforçando seu papel dominante.

Tabela 4.10: Comparação de desempenho dos modelos de regressão para predição da segunda avaliação

Disciplina	Modelo	R^2 Ajustado	AIC	BIC	RMSE	Nº Variáveis	Signif. ($p < 0{,}05$)
Português	Simples	0,732	1654,63	1662,87	1,49	1	1
	Completo	0,739	1680,19	1849,03	1,47	40	5
	Top 15 p-valores	0,749	1638,44	1704,33	1,45	15	6
	Stepwise	**0,751**	**1630,14**	**1671,32**	**1,44**	**9**	**7**
Matemática	Simples	0,760	1115,47	1122,71	1,82	1	1
	Completo	0,782	1124,61	1273,05	1,73	40	3
	Top 15 p-valores	0,790	1091,25	1149,18	1,70	15	3
	Stepwise	**0,793**	**1083,77**	**1123,59**	**1,69**	**10**	**6**

Fonte: Elaborado pela autora, com base nos dados analisados (2025).

A análise dos coeficientes nos modelos *stepwise* revelou perfis distintos de influência por disciplina (Quadro 4.4). Em Português, predominam fatores psicossociais, como qualidade das relações familiares ($\beta = 0{,}200$; $p = 0{,}005$) e saúde ($\beta = -0{,}118$; $p = 0{,}013$). Já em Matemática, sobressaem aspectos estruturais e socioeconômicos, como residência urbana ($\beta = 0{,}562$; $p = 0{,}029$) e ocupação dos pais.

Quadro 4.4: Fatores explicativos significativos nos modelos *stepwise* de regressão

Variável	Português (β, p)	Matemática (β, p)
Primeira nota	0,854, $p < 0{,}001$	0,950, $p < 0{,}001$
Reprovações anteriores	$-0{,}454$, $p < 0{,}001$	$-0{,}347$, $p = 0{,}033$
Idade	0,186, $p = 0{,}002$	$-0{,}326$, $p < 0{,}001$
Relação familiar	0,200, $p = 0{,}005$	—
Saúde	$-0{,}118$, $p = 0{,}013$	—
Relacionamento romântico	$-0{,}341$, $p = 0{,}017$	—
Profissão do pai (professor)	0,618, $p = 0{,}045$	—
Endereço urbano	—	0,562, $p = 0{,}029$
Profissão do pai (serviços)	—	0,508, $p = 0{,}031$
Profissão da mãe (outra)	—	0,475, $p = 0{,}040$

Fonte: Elaborado pela autora, com base nos dados analisados (2025).

As direções opostas do efeito da idade merecem destaque: positivo em português,

sugerindo ganho de maturidade linguística, e negativo em matemática, possivelmente refletindo defasagens acumuladas. Os resultados reforçam que as disciplinas respondem a conjuntos distintos de fatores contextuais, exigindo estratégias pedagógicas específicas.

4.1.5 Perfis estudantis latentes

A etapa final da análise exploratória buscou identificar padrões latentes nos dados por meio da combinação de Análise de Componentes Principais (PCA) e clusterização via K-Means, permitindo reduzir a dimensionalidade das variáveis e destacar eixos latentes relacionados ao desempenho escolar.

Preparação e interpretação dos componentes

Foram removidas as variáveis diretamente associadas ao desempenho (*nota1*, *nota2* e *nota_final*), padronizando-se as quantitativas via *z-score*. Em ambas as disciplinas, os dois primeiros componentes explicaram cerca de 25% da variância (Português: 25,36%; Matemática: 26,17%).

O **PC1** representou um eixo de *risco acadêmico e comportamental*, com pesos positivos para consumo de álcool, frequência de saídas e reprovações, e negativos para tempo de estudo e escolaridade parental. Já o **PC2** refletiu *capital educacional e barreiras estruturais*, separando alunos com pais mais escolarizados, menor tempo de transporte e menos reprovações daqueles em contextos mais adversos.

Clusterização e identificação de perfis estudantis

Para identificar o número ótimo de *clusters*, empregou-se o algoritmo *k-means* precedido pela padronização das variáveis. A definição do número de grupos foi realizada mediante varredura sistemática, testando valores de k de 2 a 8, com base na análise conjunta da inércia (método do cotovelo) e do índice de silhueta. Ambos os critérios de avaliação convergiram consistentemente para k=2, nas duas disciplinas analisadas.

Em **Português**, o *Cluster* 0 (72,5% dos estudantes) caracterizou-se por um perfil estruturado, apresentando predominância do gênero feminino, maior dedicação ao tempo de estudo, menor consumo de álcool, maior escolaridade parental e menor incidência de reprovações. Em contraposição, o *Cluster* 1 (27,5%) concentrou estudantes em situação de vulnerabilidade acadêmica, caracterizados por maior idade média, maior frequência de reprovações, menor engajamento escolar e comportamentos de risco mais acentuados.

Em **Matemática**, observou-se uma inversão na distribuição dos perfis: o grupo em situação de vulnerabilidade (*Cluster* 0 – 33,3%) concentrou estudantes com maior número de faltas, menor taxa de aprovação histórica e maior defasagem idade-série. Por sua vez, o *Cluster* 1 (66,7%) apresentou características mais estruturadas, incluindo predominância feminina, maior escolaridade parental, menor consumo de álcool e maior dedicação aos estudos.

A seguir, o Quadro 4.5 sintetiza os principais achados da análise, apresentando uma visão comparativa dos componentes principais, perfis estudantis e fatores associados ao desempenho em cada disciplina. A sistematização permite visualizar tanto as convergências quanto as particularidades de cada área do conhecimento.

Quadro 4.5: Síntese dos componentes principais e perfis estudantis por disciplina

Elemento	Português	Matemática
PC1	Risco acadêmico e comportamental: maior consumo de álcool, saídas frequentes, histórico de reprovações; menor tempo dedicado ao estudo e menor escolaridade parental.	Risco acadêmico e comportamental: consumo de álcool, saídas noturnas, reprovações, idade elevada; menor tempo de estudo e baixa escolaridade parental.
PC2	Capital educacional e barreiras estruturais: maior escolaridade dos pais, menor tempo de deslocamento e reduzido histórico de reprovações.	Capital educacional e barreiras estruturais: pais com maior escolarização, menor tempo de transporte e idade adequada à série; barreiras associadas à menor estrutura familiar de apoio.
Perfil 0	Estruturado (72,5%): predominância feminina, maior tempo dedicado ao estudo, menor consumo de álcool, alta escolaridade parental, baixa incidência de reprovações.	Vulnerabilidade acadêmica (33,3%): predominância masculina, maior absenteísmo, menor taxa histórica de aprovação, maior defasagem idade-série.
Perfil 1	Risco acadêmico (27,5%): predominância masculina, maior idade média, histórico de reprovações, menor engajamento escolar, maior consumo de álcool.	Estruturado (66,7%): predominância feminina, maior escolaridade parental, menor consumo de álcool, maior tempo dedicado ao estudo, melhor assiduidade.
Fatores de proteção	Alta escolaridade parental, tempo de estudo elevado, baixa frequência de consumo de álcool, assiduidade escolar, apoio familiar estruturado.	Alta escolaridade parental, baixa frequência de consumo de álcool, assiduidade escolar, adequação idade-série, apoio ao estudo domiciliar.
Fatores de risco	Histórico de reprovações, consumo de álcool, baixa escolaridade parental, menor engajamento escolar, defasagem idade-série.	Reprovações anteriores, consumo de álcool, tempo de transporte elevado, baixa escolaridade parental, absenteísmo escolar.

Fonte: Elaborado pela autora, com base nos resultados do PCA e *K-Means* (2025).

A análise comparativa evidencia a existência de padrões universais de proteção e risco — destacando-se o papel fundamental da escolaridade parental e do tempo dedicado ao estudo — bem como especificidades disciplinares relevantes. Matemática demonstrou maior sensibilidade a barreiras estruturais (como tempo de deslocamento e defasagem idade-série), enquanto o desempenho em Português mostrou-se mais fortemente associado ao engajamento individual, aos hábitos de estudo e às práticas de letramento familiar. Essa distinção sugere a necessidade de estratégias

diferenciadas de intervenção pedagógica, considerando as características específicas de cada área do conhecimento e os perfis de vulnerabilidade identificados.

4.1.6 Análise integrada entre disciplinas

Para comparar fatores universais e específicos por disciplina, realizou-se análise integrada considerando apenas estudantes com registros completos em português e matemática, totalizando 366 casos.

Desempenho comparativo

A Tabela 4.11 mostra que o desempenho médio foi superior em português (diferença de 2,02 pontos), com menor variabilidade em relação à matemática, indicando maior heterogeneidade nesta última.

Tabela 4.11: Comparativo de desempenho entre disciplinas na base integrada

Variável	Média	Mediana	Desvio Padrão	CV	Mínimo	Máximo
Nota final português	12,57	13,0	2,94	0,23	0,0	19,0
Nota final matemática	10,55	11,0	4,53	0,43	0,0	20,0
Faltas português	3,55	2,0	4,70	1,32	0,0	32,0
Faltas matemática	5,43	4,0	7,69	1,42	0,0	75,0
Idade	16,57	17,0	1,18	0,07	15,0	22,0

Fonte: Elaborado pela autora, com base nos dados analisados (2025).

Correlações interdisciplinares

As notas finais apresentaram correlação positiva moderada ($\rho = 0{,}56$), sugerindo relação entre desempenhos, mas com desafios específicos em matemática. Variáveis comportamentais mostraram correlações altas entre disciplinas, enquanto as acadêmicas foram mais disciplina-dependentes.

Casos extremos e perfis contrastantes

Matemática concentrou mais casos críticos: 32 alunos (8,7%) com nota zero, contra 5 (1,4%) em português, e 26,8% de reprovação exclusiva, frente a 1,9% na

outra disciplina. Diferenças superiores a 4 pontos nas notas finais ocorreram em 77 casos, sendo 70 (19,1%) favoráveis a português e 7 (1,9%) a matemática.

Fatores de impacto

Os fatores que influenciam o desempenho nas duas disciplinas apresentaram características distintas. Em português, destacaram-se variáveis relacionadas ao engajamento individual do estudante: tempo de estudo, escolaridade paterna e consumo de álcool nos fins de semana, sugerindo que hábitos pessoais e dedicação aos estudos exercem maior influência. Já em matemática, sobressaíram-se indicadores do capital educacional familiar, como a escolaridade dos pais e a profissão materna, evidenciando uma dependência mais acentuada do contexto socioeconômico familiar. Além disso, matemática apresentou variabilidade 87% maior que português, com mais casos de desempenho extremo e maior sensibilidade aos fatores socioeconômicos.

Esses achados convergem com os perfis latentes identificados nas análises anteriores e apontam para uma distinção importante: enquanto o bom desempenho em português associa-se predominantemente ao esforço e engajamento individuais do estudante, o desempenho em matemática mostra-se mais dependente das oportunidades educacionais e do suporte proporcionados pelo ambiente familiar.

4.2 Seleção de atributos

A seleção de atributos constitui etapa fundamental no desenvolvimento de modelos preditivos, visando identificar subconjuntos de variáveis que maximizem o poder discriminativo enquanto reduzem a complexidade computacional e o risco de sobreajuste. Neste trabalho, foram implementadas três estratégias complementares de seleção, fundamentadas em diferentes critérios estatísticos e exploratórios, conforme descrito na metodologia.

As estratégias adotadas buscaram equilibrar rigor estatístico com interpretabilidade prática, considerando as especificidades identificadas na análise exploratória para cada disciplina. A validação da efetividade das seleções foi posteriormente ava-

liada mediante comparação de desempenho dos modelos preditivos construídos com diferentes subconjuntos de atributos.

4.2.1 Estratégia baseada na AED

A primeira estratégia de seleção fundamentou-se integralmente nos achados da análise exploratória de dados, priorizando variáveis que demonstraram associações consistentes com o desempenho acadêmico através de múltiplas perspectivas analíticas. Esta abordagem combinou evidências descritivas, correlacionais e de diferenciação entre grupos, conforme os critérios estabelecidos na metodologia.

Critérios de seleção aplicados

A seleção de atributos foi orientada pela convergência de múltiplas evidências obtidas na análise exploratória, combinando critérios estatísticos, empíricos e teóricos. Foram considerados, entre outros, coeficientes de correlação de *Spearman* com a variável *nota_final* superiores a $|\rho| > 0{,}15$, diferenças de médias acima de 1,5 pontos entre categorias e variações relevantes nas taxas de aprovação — especialmente aquelas superiores a 15 pontos percentuais entre grupos.

Complementarmente, foi incorporado o desempenho das variáveis no índice de perfil composto, priorizando aquelas classificadas com impacto ”forte”ou ”moderado”. A relevância teórica dos atributos também foi levada em conta, com base em evidências consolidadas na literatura de Mineração de Dados Educacionais, garantindo a seleção de variáveis com significado interpretativo e valor explicativo.

Essa abordagem multifacetada permitiu eliminar variáveis altamente correlacionadas ou conceitualmente redundantes, assegurando maior parcimônia ao conjunto final de atributos. O resultado foi um conjunto robusto e enxuto de preditores, sensível às especificidades das disciplinas analisadas e alinhado às exigências dos modelos preditivos subsequentes.

Atributos selecionados

Tabela 4.12: Atributos selecionados para modelagem preditiva com base na AED segmentada por disciplina — português

Tipo	Variável	Justificativa
Quantitativa	Faltas	Indicador direto de envolvimento escolar com associação negativa consistente
	Idade	Relacionada à defasagem escolar e trajetórias não lineares
Categórica — Ordinal	Reprovações	Variável mais discriminante; diferenças extremas nas taxas de aprovação
	Tempo de estudo	Forte relação direta com aprovação, sem sobreposição com outras variáveis
	Apoio escolar	Identifica padrões ocultos de dificuldade (relação inversa)
	Apoio familiar	Potencial moderador de desempenho e ambiente de suporte
	Escolaridade da mãe	Indicador de capital educacional; preferida à escolaridade paterna
	Álcool (dias úteis)	Indicador de risco acadêmico; maior impacto que consumo em fins de semana
	Frequência de saídas	Indicador comportamental complementar ao consumo de álcool
	Relação familiar	Ambiente emocional e doméstico do estudante
Categórica — Nominal	Interesse ensino superior	Indicador motivacional com alto poder discriminatório
	Escola	Contexto institucional com variação significativa entre unidades
	Acesso à internet	Infraestrutura tecnológica
	Gênero	Controlador sociodemográfico básico
	Atividades extracurriculares	Organização temporal e estímulo ao desenvolvimento

Fonte: Elaborado pela autora (2025).

Tabela 4.13: Atributos selecionados para modelagem preditiva com base na AED segmentada por disciplina — matemática

Tipo	**Variável**	**Justificativa**
Quantitativa	Faltas	Indicador de envolvimento escolar; presença marcante entre casos críticos
	Idade	Indicador para defasagem escolar; independência em relação às demais variáveis
Categórica — Ordinal	Reprovações	Maior impacto discriminativo (variação de até 47 p.p. na taxa de aprovação entre categorias)
	Tempo de estudo	Associação direta com aprovação; relação clara com dedicação acadêmica
	Escolaridade da mãe	Indicador de capital educacional; maior variação e consistência
	Álcool (dias úteis)	Risco comportamental; maior correlação com desempenho
	Frequência de saídas	Comportamento social com impacto visível em baixo desempenho
	Tempo de transporte	Redutor de tempo de estudo; associado a casos de nota zero
	Apoio escolar	Associada a padrões de dificuldade (relação inversa)
	Relação familiar	Ambiente socioafetivo com variações entre aprovados/-reprovados
Categórica — Nominal	Responsável legal	Configuração familiar complementar a outras variáveis familiares
Categórica — Binária	Interesse ensino superior	Forte correlação com notas altas e motivação acadêmica
	Acesso à internet	Infraestrutura de aprendizagem e recursos tecnológicos

Fonte: Elaborado pela autora, com base na análise exploratória de dados (2025).

A partir da convergência das evidências exploratórias, foram selecionados conjuntos de atributos diferenciados por disciplina: 14 variáveis para português e 13 para matemática. Esta diferença quantitativa reflete a maior complexidade estrutural e menor número de fatores estatisticamente relevantes identificados em matemática durante a análise exploratória.

A seleção priorizou variáveis com forte impacto estatístico, ausência de redundância e cobertura multidimensional dos fatores associados ao desempenho acadêmico. As Tabelas 4.12 e 4.13 apresentam os conjuntos finais de variáveis, organizados por tipo

e acompanhados das respectivas justificativas.

Os conjuntos selecionados abrangem adequadamente os principais domínios identificados na análise exploratória: aspectos individuais (idade, faltas, gênero), familiares (escolaridade materna, apoio familiar, relação familiar), institucionais (escola, apoio escolar), comportamentais (consumo de álcool, frequência de saídas) e motivacionais (interesse no ensino superior, tempo de estudo), compondo perfis multidimensionais sensíveis às especificidades de cada disciplina e às desigualdades educacionais identificadas.

4.2.2 Estratégia por regressão linear múltipla

Com o objetivo de reforçar a seleção de atributos com base em associações estatisticamente significativas com o desempenho(nota final), foram ajustados modelos de regressão linear múltipla para cada disciplina, utilizando a nota final como variável dependente e excluindo-se as notas intermediárias para evitar vazamentos. A análise concentrou-se nas variáveis contextuais, priorizando aquelas com significância estatística e baixa multicolinearidade.

Critérios de seleção aplicados

Os critérios de seleção adotados foram:

- Variáveis com $p - valor < 0{,}05$ foram selecionadas automaticamente;
- Variáveis com $0{,}05 \leq p \leq 0{,}10$ foram consideradas relevantes de forma condicional, desde que apresentassem reforço teórico ou evidência em outras análises;
- Variáveis diretamente relacionadas às notas (*nota1*, *nota2*) foram propositalmente excluídas para evitar viés na modelagem da aprovação.

Resultados da modelagem regressiva

Os modelos finais apresentaram capacidade explicativa satisfatória, com R^2 ajustado de 0,41 em português e 0,39 em matemática. A aplicação dos critérios de signi-

ficância resultou na identificação de conjuntos diferenciados de variáveis preditoras, conforme apresentado nas Tabelas 4.14 e 4.15.

Tabela 4.14: Variáveis selecionadas por regressão linear múltipla — português

Variável	Coeficiente	*p*-valor	Justificativa
Escola (Mousinho da Silveira)	−1,46	0,000	Alta significância institucional
Tempo de estudo	0,41	0,014	Associação positiva com dedicação
Interesse ensino superior	1,68	0,000	Consistente com expectativas
Reprovações	−1,49	0,000	Forte impacto negativo
Apoio escolar	−1,61	0,000	Indicador de dificuldade (padrão inverso)
Saúde	−0,21	0,020	Relação negativa com desempenho
Idade	0,21	0,089	Mantida por reforço estatístico
Relação familiar	0,22	0,093	Relevância psicossocial
Relacionamento romântico	−0,51	0,055	Indicador comportamental significativo

Fonte: Elaborado pela autora, com base nos dados analisados (2025).

Tabela 4.15: Variáveis selecionadas por regressão linear múltipla — matemática

Variável	Coeficiente	*p*-valor	Justificativa
Reprovações	−2,03	0,000	Forte impacto negativo sobre desempenho
Gênero (Mulher)	−1,46	0,016	Diferença significativa entre gêneros
Idade	−0,64	0,021	Relação negativa com desempenho
Apoio familiar	−1,20	0,036	Indicador do ambiente doméstico
Tamanho da família	−1,22	0,035	Impacto estrutural na vida escolar
Profissão da mãe (Professor(a))	−2,60	0,042	Relevância estatística e interpretativa
Apoio escolar	−1,42	0,074	Significância marginal; mantida por contexto
Interesse ensino superior	2,20	0,081	Valor educativo; coerente com expectativa
Álcool (dias úteis)	−0,71	0,090	Possível interferência comportamental

Fonte: Elaborado pela autora, com base nos dados analisados (2025).

Atributos selecionados

A regressão linear múltipla resultou na seleção de conjuntos diferenciados por disciplina:

Português (9 variáveis): Escola, Tempo de estudo, Interesse no ensino superior, Reprovações, Apoio escolar, Saúde, Idade, Relação familiar, Relacionamento romântico.

Matemática (9 variáveis): Reprovações, Gênero, Idade, Apoio familiar, Tamanho da família, Profissão da mãe, Apoio escolar, Interesse no ensino superior, Álcool (dias úteis).

A diferença qualitativa entre os conjuntos reflete a maior sensibilidade do desempenho em português a fatores psicossociais (saúde, relação familiar, relacionamento romântico), enquanto matemática responde primariamente a variáveis estruturais e demográficas (gênero, tamanho da família, profissão da mãe).

Análise de multicolinearidade

A avaliação de multicolinearidade foi conduzida mediante análise do Fator de Inflação da Variância (VIF) e correlações de Pearson entre pares de variáveis. Os resultados revelaram a presença de multicolinearidade substancial no conjunto completo de variáveis, justificando a necessidade de seleção criteriosa de atributos.

Principais problemas identificados:

- **Correlações elevadas:** Escolaridade dos pais (ρ = 0,65 em matemática; ρ = 0,65 em português), consumo de álcool (ρ = 0,63 em matemática; ρ = 0,62 em português) e profissões dos pais (ρ = 0,69 em matemática; ρ = 0,72 em português);
- **VIF críticos:** Múltiplas variáveis apresentaram VIF ¿ 10, com destaque para idade (VIF = 90,3 em português; VIF = 25, em matemática) e escolaridade materna (VIF = 24,1 em ambas);
- **Redundância conceitual:** Variáveis como escolaridade dos pais e profissões parentais capturam dimensões sobrepostas do capital socioeconômico familiar.

Estratégias de mitigação adotadas: A seleção por regressão linear múltipla priorizou variáveis com menor VIF dentre aquelas estatisticamente significativas, resultando em conjuntos mais parcimoniosos. Para as variáveis mantidas nos modelos finais, os valores de VIF foram substancialmente reduzidos, embora algumas

correlações residuais permaneçam dentro de limites aceitáveis para modelagem preditiva.

Esta análise reforçou a importância da seleção criteriosa de atributos, orientando as escolhas metodológicas nas etapas subsequentes de modelagem para evitar instabilidade estatística e preservar a interpretabilidade dos modelos.

Implicações para a modelagem preditiva

A convergência parcial entre as estratégias de seleção baseada na AED e por regressão linear múltipla reforçou a robustez de variáveis como reprovações, interesse no ensino superior e apoio escolar. As especificidades identificadas na seleção por regressão — particularmente a relevância de fatores psicossociais em português e estruturais em matemática — complementaram adequadamente a perspectiva da seleção baseada na AED, fornecendo subsídios para a construção de modelos preditivos diferenciados por disciplina.

4.2.3 Estratégia por testes estatísticos inferenciais

A terceira estratégia de seleção baseou-se na aplicação de testes estatísticos inferenciais, com o objetivo de identificar variáveis significativamente associadas ao desempenho escolar. A análise diferenciou variáveis ordinais e nominais, adotando critérios combinados de significância estatística e magnitude da associação.

Critérios de seleção aplicados

Para variáveis ordinais: Correlação de *Spearman* com a nota final ($|\rho| \geq 0{,}15$ e $p < 0{,}05$) e teste de *Kruskal-wallis* para diferenças entre grupos.

Para variáveis nominais: Teste qui-quadrado de independência com a variável aprovação ($p < 0{,}05$) e V de Cramér como medida de associação ($V \geq 0{,}10$).

Resultados para variáveis ordinais

Tabela 4.16: Variáveis ordinais selecionadas por correlação de *Spearman*

Variável	Português		Matemática	
	ρ	***p*-valor**	ρ	***p*-valor**
Reprovações	−0,457	$7{,}80 \times 10^{-25}$	−0,364	$4{,}82 \times 10^{-10}$
Escolaridade da mãe	0,263	$1{,}28 \times 10^{-8}$	0,227	0,01
Tempo de estudo	0,263	$1{,}30 \times 10^{-8}$	—	—
Escolaridade do pai	0,246	$1{,}06 \times 10^{-7}$	0,209	0,01
Álcool (dias úteis)	−0,198	$2{,}16 \times 10^{-5}$	—	—
Álcool (fim de semana)	−0,157	0,001	−0,145	0,016
Tempo de transporte	−0,126	0,007	—	—
Frequência de saídas	−0,118	0,012	−0,169	0,005
Tempo livre	−0,116	0,013	—	—
Saúde	−0,095	0,042	—	—
Relação familiar	0,093	0,047	—	—

Fonte: Elaborado pela autora, com base nos dados analisados (2025).

Resultados para variáveis nominais

Tabela 4.17: Variáveis nominais selecionadas por qui-quadrado e V de Cramér

Variável	Português		Matemática	
	***p*-valor**	**V de Cramér**	***p*-valor**	**V de Cramér**
Escola	$9{,}90 \times 10^{-11}$	0,304	—	—
Interesse ensino superior	$4{,}68 \times 10^{-8}$	0,256	—	—
Motivo escolha escola	0,004	0,172	—	—
Endereço	0,011	0,120	—	—

Fonte: Elaborado pela autora, com base nos dados analisados (2025).

Atributos selecionados por disciplina

Português (16 variáveis): 11 variáveis ordinais significativas, 4 variáveis nominais com associação adequada, e a variável faltas mantida por relevância pedagógica.

Matemática (6 variáveis): 5 variáveis ordinais significativas, nenhuma variável nominal atendeu aos critérios estabelecidos, e a variável faltas mantida por relevância pedagógica.

A diferença acentuada entre as disciplinas (16 vs 6 variáveis) corrobora os achados da análise exploratória sobre a maior complexidade estrutural e menor número de fatores estatisticamente relevantes em matemática, requerendo seleções mais focalizadas em determinantes estruturais básicos.

Implicações metodológicas

A estratégia inferencial evidenciou que português apresenta maior sensibilidade a uma gama diversificada de fatores contextuais, comportamentais e institucionais, enquanto matemática responde a um conjunto mais restrito de variáveis, predominantemente relacionadas ao capital educacional familiar e a comportamentos de risco específicos.

A inclusão da variável faltas por justificativa pedagógica em ambas as disciplinas reconhece sua relevância educacional estabelecida, mesmo quando os testes estatísticos não capturam completamente sua influência sobre o desempenho final.

4.2.4 Análise comparativa das estratégias

Quadro 4.6: Síntese comparativa das estratégias de seleção de atributos

Estratégia	português	matemática	Características
AED	15 variáveis	13 variáveis	Abrangência multidimensional; cobertura de domínios explicativos
Regressão Linear	9 variáveis	9 variáveis	Parcimônia; significância estatística; especificidade disciplinar
Testes Inferenciais	16 variáveis	6 variáveis	Rigor estatístico; força de associação; contraste disciplinar acentuado

Fonte: Elaborado pela autora, com base nos dados analisados (2025).

A comparação entre as três estratégias de seleção identificou padrões consistentes e especificidades disciplinares importantes, conforme sintetizado no Quadro 4.6.

A convergência de variáveis-chave (reprovações, escolaridade parental, interesse no ensino superior, escola) em múltiplas estratégias reforçou sua relevância preditiva. Em contrapartida, as diferenças quantitativas entre as disciplinas — particularmente evidente na estratégia inferencial — corroboraram os achados da análise exploratória sobre a maior complexidade estrutural e menor número de fatores estatisticamente relevantes em matemática.

Implicações para a modelagem preditiva

A implementação das três estratégias de seleção permitiu a criação de subconjuntos de atributos complementares, cada um capturando dimensões específicas dos determinantes do desempenho acadêmico. A estratégia baseada na AED privilegiou abrangência temática e cobertura multidimensional, a seleção por regressão linear múltipla priorizou parcimônia e significância estatística, enquanto a abordagem inferencial enfatizou força de associação e rigor estatístico.

Esta diversidade metodológica forneceu base robusta para a avaliação comparativa de modelos preditivos, permitindo investigar tanto a influência da amplitude dos preditores quanto o impacto da especificidade disciplinar na capacidade de generalização dos classificadores desenvolvidos.

4.3 Modelagem preditiva

A etapa de modelagem preditiva concentrou-se na construção e avaliação de modelos supervisionados para classificação binária da aprovação escolar, utilizando cinco algoritmos consolidados na literatura de Mineração de Dados Educacionais: Regressão Logística, Árvore de Decisão, *Random Forest*, *AdaBoost* e SVM. A implementação seguiu um protocolo sistemático contemplando avaliação com hiperparâmetros padrão, otimização via *grid search*, tratamento de desbalanceamento e diagnóstico de estabilidade.

Os modelos foram desenvolvidos separadamente para cada disciplina, utilizando os três conjuntos de atributos selecionados na etapa anterior (baseado na AED, regressão linear múltipla e testes inferenciais), além de um conjunto controle com todas as variáveis disponíveis. Esta abordagem permitiu avaliar tanto o impacto das estratégias de seleção quanto a especificidade dos padrões preditivos entre português e matemática.

4.3.1 Otimização de hiperparâmetros

A otimização dos hiperparâmetros foi conduzida conforme os espaços de busca definidos na Seção 3.6.1, utilizando a técnica de *'GridSearchCV'* com validação cruzada estratificada em 5 *folds*. O objetivo foi maximizar o F_1-Score Macro, métrica adequada ao contexto de classes desbalanceadas, visando melhorar a capacidade preditiva dos modelos sem comprometer sua generalização.

Resultados da otimização

O processo de otimização produziu melhorias diferenciadas por algoritmo e disciplina. Em português, as melhorias no F_1-Score Macro variaram entre 5% (Regressão Logística) e 25% (*AdaBoost*), evidenciando alta responsividade dos algoritmos *ensemble*. Em matemática, os ganhos foram mais modestos, com melhorias entre 2% (SVM) e 15% (Árvore de Decisão), refletindo as limitações estruturais desta disciplina para modelagem preditiva.

O SVM demonstrou comportamento contrastante entre as disciplinas: alta responsividade em português, beneficiando-se especialmente do *kernel* RBF, enquanto em matemática o *kernel* linear mostrou-se mais efetivo, porém com ganhos limitados. As árvores de decisão apresentaram melhorias consistentes em ambas as disciplinas, confirmando sua adaptabilidade a diferentes contextos de dados e distribuições de classes.

Os *ensembles* (*Random Forest* e *AdaBoost*) tiveram bom desempenho em alguns cenários, porém, o SVM foi o mais consistente em português, onde a maior complexidade dos padrões beneficiou-se de abordagens de combinação de modelos. Em

matemática, contudo, a preferência por modelos mais simples ficou evidente, com árvores individuais frequentemente superando seus equivalentes *ensemble*.

4.3.2 Tratamento do desbalanceamento

O tratamento do desbalanceamento foi implementado via parâmetro *class_weight*, estratégia que se mostrou mais robusta para preservar a distribuição original dos dados e manter a capacidade de generalização dos modelos.

Impacto diferenciado por disciplina

O impacto da técnica de balanceamento de classes (*class_weight*) no desempenho dos modelos é ilustrado na Tabela 4.18, que compara as métricas de Português e Matemática. O balanceamento busca mitigar o desequilíbrio entre a classe majoritária (Aprovados) e minoritária (Reprovados). As colunas 'Média' e 'Melhor' auxiliam na análise, representando a performance média do conjunto de modelos e o resultado ótimo (Melhor) alcançado por um modelo individual, respectivamente. Os resultados mostram uma tendência de aumento no *Recall* (Reprovados) após o balanceamento, evidenciando uma melhoria na capacidade de detecção de risco.

Tabela 4.18: Impacto do balanceamento via *class_weight* por disciplina

Disciplina	**Métrica**	**Sem Balanceamento**		**Com Balanceamento**	
		Média	**Melhor**	**Média**	**Melhor**
português	F_1-Score Macro	0,612	0,673	0,651	0,732
	Recall (Reprovados)	0,287	0,333	0,521	0,833
	Precisão (Reprovados)	0,523	0,588	0,461	0,531
matemática	F_1-Score Macro	0,609	0,665	0,634	0,710
	Recall (Reprovados)	0,401	0,487	0,498	0,641
	Precisão (Reprovados)	0,565	0,667	0,536	0,595

Fonte: Elaborado pela autora, com base nos dados analisados (2025).

Efetividade diferenciada por disciplina

O balanceamento demonstrou efetividade superior em português, promovendo melhorias sistemáticas no *recall* da classe minoritária (reprovados), com o melhor modelo atingindo 83,3% de sensibilidade. Esta melhoria é particularmente relevante para sistemas de alerta precoce, onde a detecção de estudantes em risco é prioritária.

Em matemática, embora tenha ocorrido melhoria no modelo, a resposta foi menos consistente, com alguns algoritmos apresentando deterioração do desempenho. Esta diferença sugere que o desbalanceamento em português é mais tratável via ponderação de classes, enquanto em matemática podem existir fatores estruturais adicionais que limitam a efetividade desta estratégia.

Considerando os melhores modelos por disciplina, o *trade-off* (balanço entre a precisão e o recall) mostrou-se mais acentuado em matemática: a precisão (reprovados) reduziu-se de 0,667 para 0,595 (variação de $-0{,}072$), queda superior à observada em português, de 0,588 para 0,531 (variação de $-0{,}057$). O ganho de *recall* em matemática foi menor, passando de 0,487 para 0,641 (variação de $+0{,}154$), enquanto, em português, passou de 0,333 para 0,833 (variação de $+0{,}500$) (Tabela 4.18). Esse padrão dificulta o equilíbrio entre a detecção de risco e a acurácia geral em matemática.

4.3.3 Diagnóstico de estabilidade

O diagnóstico de estabilidade foi conduzido mediante comparação entre o desempenho no conjunto de teste e na validação cruzada. Foram considerados problemáticos os modelos com diferença superior a 10% entre o desempenho no conjunto de teste e o desempenho médio na validação cruzada em qualquer métrica principal (F_1Score Macro, AUC ROC, Precisão ou *Recall*).

Distribuição dos diagnósticos

A Tabela 4.19 apresenta a distribuição dos diagnósticos de estabilidade por disciplina.

Tabela 4.19: Distribuição dos diagnósticos de estabilidade dos modelos por disciplina

Diagnóstico	português		matemática		Total
	N	%	N	%	
Ajuste adequado	28	70,0	31	77,5	59
Underfitting potencial	10	25,0	9	22,5	19
Overfitting potencial	2	5,0	0	0,0	2

Fonte: Elaborado pela autora, com base nos dados analisados (2025).

Padrões de instabilidade identificados

Matemática apresentou maior estabilidade geral (77,5% de modelos com ajuste adequado, comparado a 70,0% em português), possivelmente devido ao menor desempenho absoluto, o que reduz a probabilidade de sobreajuste. Os casos de *overfitting* potencial concentraram-se exclusivamente em português, especificamente em modelos *AdaBoost* e *Random Forest* sem balanceamento.

Os casos de *underfitting* potencial distribuíram-se similarmente entre as disciplinas, associando-se principalmente a modelos com seleção de atributos muito restritiva ou hiperparâmetros excessivamente conservadores. Este padrão sugere que a complexidade mínima necessária para capturar os padrões nos dados é similar entre as disciplinas, embora o potencial de sobreajuste seja maior em português.

A maior estabilidade observada em matemática, paradoxalmente, reflete suas limitações preditivas: modelos que não conseguem capturar adequadamente os padrões complexos tendem a ser mais estáveis, porém menos úteis para aplicações práticas.

4.4 Comparação de desempenho entre modelos

A comparação sistemática dos modelos considerou múltiplas dimensões de avaliação: métricas específicas por disciplina, capacidade de generalização e identificação dos modelos mais efetivos para diferentes contextos de aplicação. A análise contemplou desempenho absoluto, estabilidade e interpretabilidade, visando oferecer recomendações práticas para implementação.

4.4.1 Métricas por disciplina

A análise das métricas de desempenho revelou padrões distintos de efetividade dos algoritmos entre as disciplinas, conforme sintetizado na Tabela 4.20.

Tabela 4.20: Comparação dos melhores modelos por disciplina e estratégia

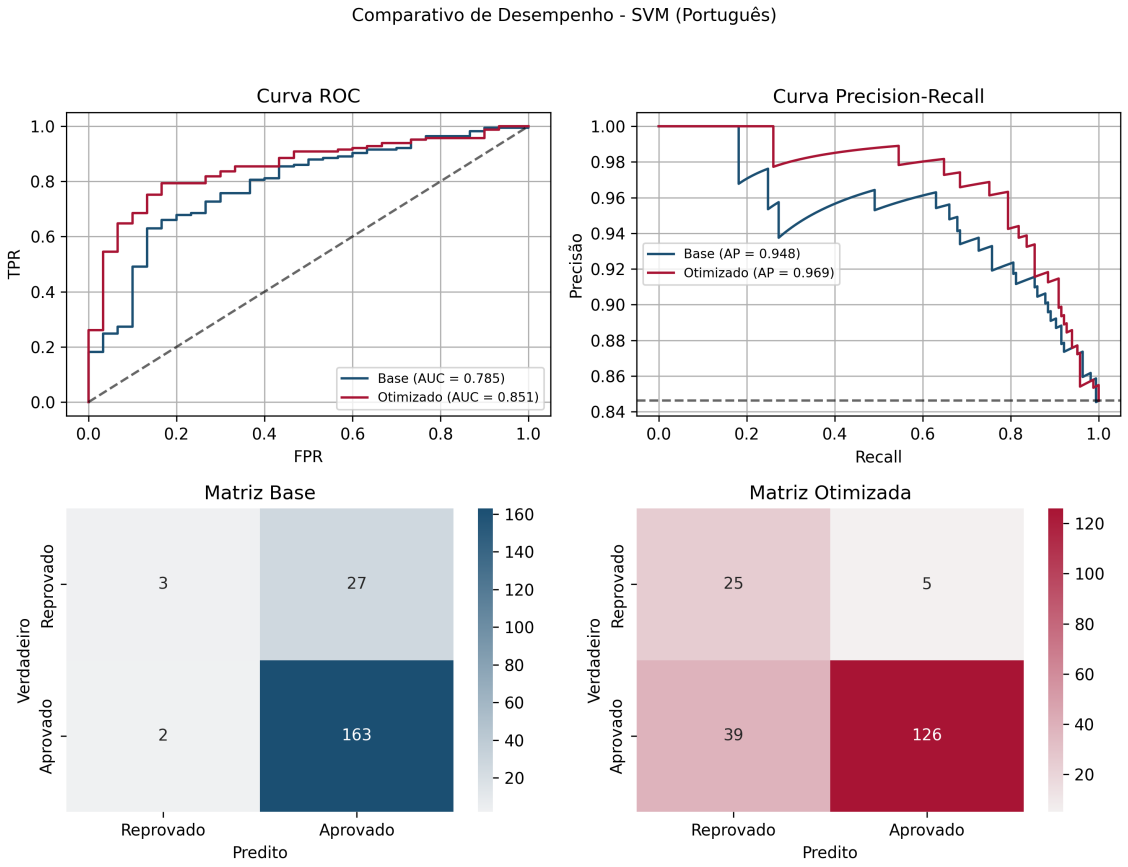
Disciplina	Estratégia	Modelo	F_1-Score Macro	AUC ROC	Precisão(Reprovados)	*Recall*(Reprovados)	Diagnóstico
português	Sem seleção	*Random Forest*	0,673	0,816	0,588	0,333	Ajuste adequado
	AED	SVM	0,692	0,851	0,391	0,833	Ajuste adequado
	Regressão	SVM	0,732	0,822	0,531	0,567	Ajuste adequado
	Inferência	*Logistic Reg.*	0,706	0,835	0,435	0,667	Ajuste adequado
matemática	Sem seleção	*Decision Tree*	0,654	0,643	0,559	0,487	Ajuste adequado
	AED	*Logistic Reg.*	0,644	0,692	0,593	0,410	Ajuste adequado
	Regressão	SVM	0,710	0,694	0,595	0,641	Ajuste adequado
	Inferência	*Random Forest*	0,665	0,673	0,667	0,410	Ajuste adequado

Fonte: Elaborado pela autora, com base nos dados analisados (2025).

Análise visual dos modelos representativos

Para ilustrar os padrões de desempenho identificados, foram selecionados modelos representativos de cada disciplina para análise detalhada das curvas ROC, *Precision-Recall* e matrizes de confusão.

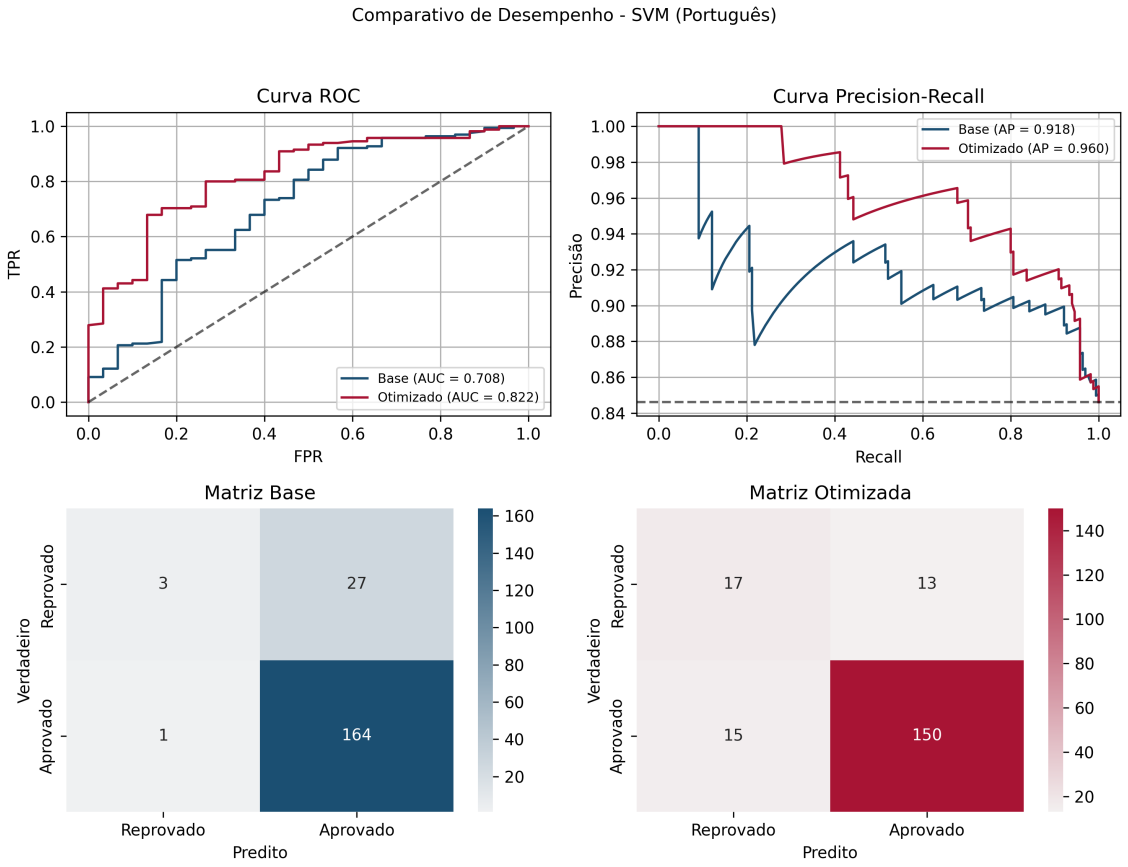
Figura 4.8: Modelo com melhor AUC ROC em português: SVM com seleção baseada na AED e balanceamento



Fonte: Elaborado pela autora, com base nos dados analisados (2025).

A Figura 4.8 apresenta o modelo com superior capacidade discriminativa em português (AUC ROC = 0,851). A otimização resultou em transformação significativa na detecção de reprovações, aumentando os verdadeiros positivos de 3 para 25 casos. Este modelo demonstra excelente sensibilidade (83,3%) para identificação precoce de estudantes em risco, sendo ideal para sistemas de alerta.

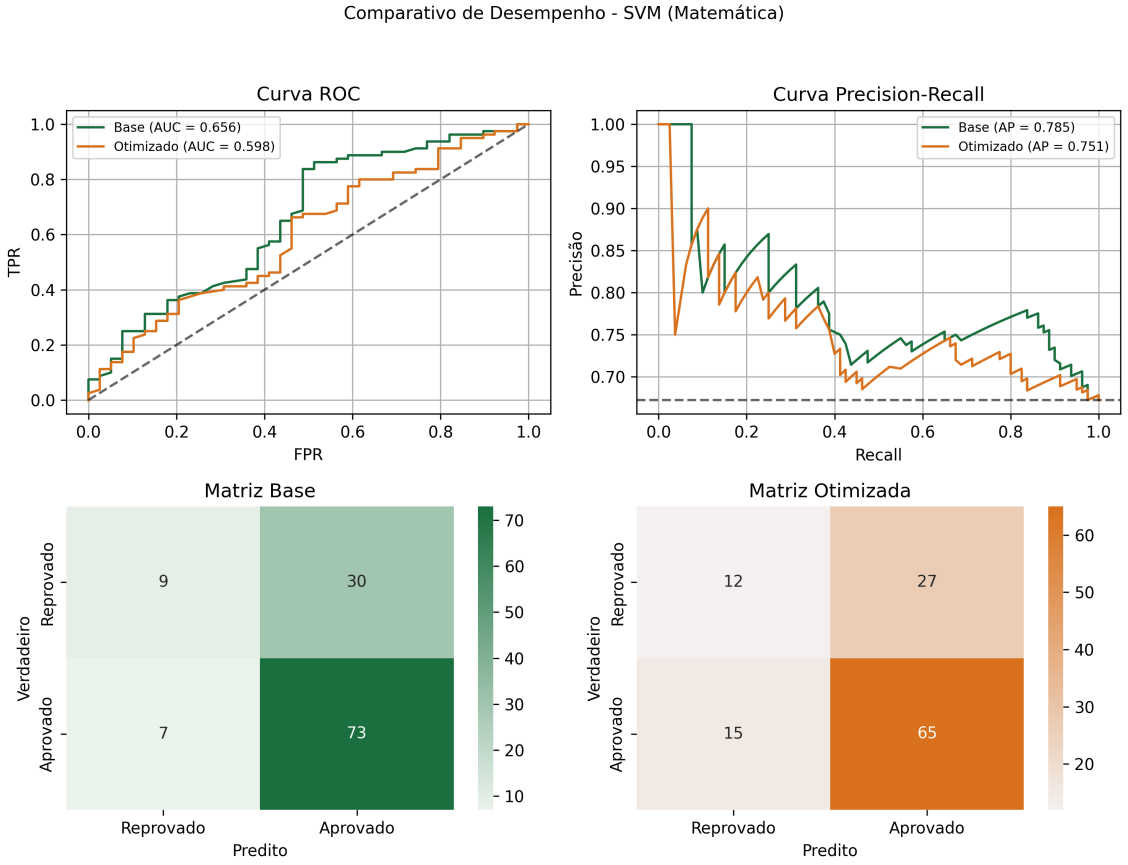
Figura 4.9: Modelo recomendado para português: SVM com seleção por regressão linear múltipla e balanceamento



Fonte: Elaborado pela autora, com base nos dados analisados (2025).

A Figura 4.9 ilustra o modelo com melhor F_1-Score Macro em português (0,732). Apresenta equilíbrio superior entre precisão e *recall*, com 17 verdadeiros positivos e apenas 15 falsos positivos, oferecendo parcimônia (9 variáveis) e interpretabilidade adequada para aplicação operacional. Este modelo representa o melhor compromisso entre desempenho e praticidade.

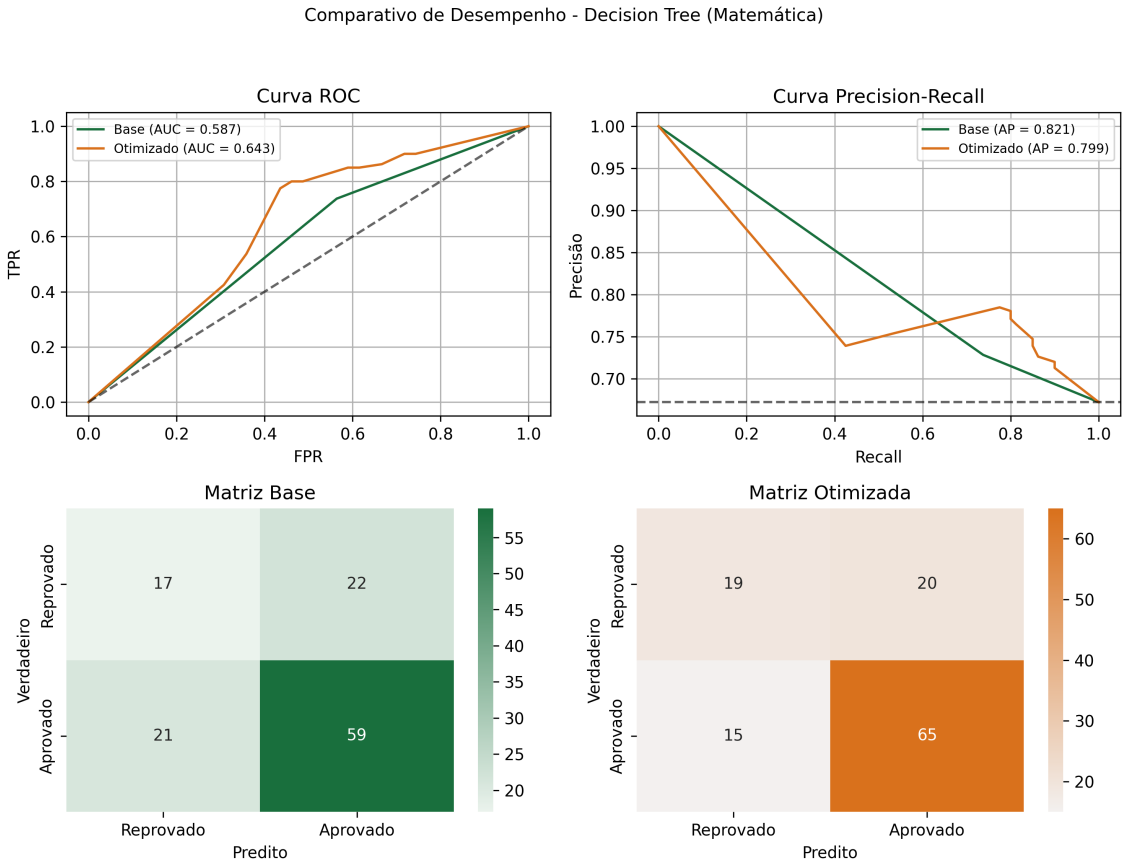
Figura 4.10: Modelo recomendado para matemática: SVM com seleção por regressão linear múltipla e balanceamento



Fonte: Elaborado pela autora, com base nos dados analisados (2025).

A Figura 4.10 apresenta o melhor modelo para matemática (F_1-Score Macro = 0,710). A otimização resultou em melhoria substancial na detecção de reprovações (9 → 25 verdadeiros positivos), embora com aumento nos falsos positivos. O modelo utiliza apenas 6 variáveis, oferecendo maior parcimônia, característica importante dado o menor poder preditivo desta disciplina.

Figura 4.11: Algoritmo simples em matemática: *Decision Tree* sem seleção de atributos com balanceamento



Fonte: Elaborado pela autora, com base nos dados analisados (2025).

A Figura 4.11 demonstra que algoritmos mais simples podem ser competitivos em matemática. O *Decision Tree* otimizado apresentou desempenho respeitável (F_1-Score Macro = 0,654) com maior interpretabilidade, adequado para contextos que priorizam transparência nas decisões, como discussões pedagógicas ou prestação de contas.

Padrões de desempenho identificados

A análise comparativa dos modelos revelou padrões distintos entre as disciplinas. Em português, observou-se maior versatilidade algorítmica: o SVM destacou-se em duas das quatro estratégias testadas, e o balanceamento das classes trouxe ganhos consistentes de desempenho. Os melhores modelos apresentaram AUC ROC superior

a 0,80, sugerindo alta capacidade discriminativa. A performance preditiva mostrou-se mais sensível a variáveis de natureza psicossocial e comportamental.

Em matemática, os resultados indicaram preferência mais restrita a determinados algoritmos, com destaque para a Árvore de Decisão e o SVM linear. O efeito do balanceamento foi menos uniforme e, mesmo nos melhores cenários, a AUC ROC não ultrapassou 0,694, indicando menor poder preditivo. A predição em matemática parece depender mais de fatores estruturais básicos, com menor influência de aspectos comportamentais ou subjetivos.

4.4.2 Análise da capacidade de generalização

A capacidade de generalização foi avaliada mediante análise da estabilidade entre validação cruzada e desempenho no conjunto de teste, complementada por métricas de dispersão das predições entre diferentes *folds*.

Estabilidade por categoria de modelo

A Tabela 4.21 apresenta indicadores de estabilidade agregados por categoria de modelo.

Tabela 4.21: Indicadores de capacidade de generalização por categoria de modelo

Categoria	Disciplina	Δ Médio AUC	Δ Médio F1	Modelos ajustados	Estabilidade
Lineares	português	-2,1%	-5,8%	85%	Excelente
	matemática	-1,3%	-3,2%	90%	Excelente
Árvores	português	-1,8%	-4,1%	75%	Boa
	matemática	+2,1%	+1,7%	80%	Boa
Ensemble	português	+3,4%	+2,8%	60%	Moderada
	matemática	+0,8%	-1,4%	75%	Boa
SVM	português	-0,9%	-2,3%	80%	Boa
	matemática	-3,1%	-4,7%	70%	Moderada

Fonte: Elaborado pela autora, com base nos dados analisados (2025).

Fatores que influenciam a generalização

Relação entre complexidade do modelo e tamanho da amostra: Modelos lineares apresentaram maior estabilidade em ambas as disciplinas, beneficiando-se da relação favorável entre complexidade e tamanho amostral. Em matemática, que possui amostra menor (395 casos, comparado a 649 em português), essa vantagem foi ainda mais pronunciada.

Estratégias de seleção de atributos: A seleção por regressão linear múltipla produziu modelos mais estáveis, seguida pela seleção baseada na AED. A seleção por testes inferenciais, embora estatisticamente rigorosa, resultou em menor estabilidade prática, possivelmente devido ao conjunto muito restrito de variáveis selecionadas.

Balanceamento e generalização: O balanceamento via *class_weight* não comprometeu a capacidade de generalização, mantendo estabilidade adequada enquanto melhorava a detecção da classe minoritária.

4.4.3 Modelos mais efetivos

Com base na análise multidimensional contemplando desempenho, estabilidade e interpretabilidade, foram identificados os modelos mais efetivos para cada contexto de aplicação.

Recomendações por contexto de uso

Para sistemas de alerta precoce , que priorizam a detecção de estudantes em risco, os melhores resultados foram obtidos com SVM em português (seleção por AED com balanceamento), alcançando AUC ROC de 0,851 e *recall* de 83,3%. Essa configuração favorece detecção ampla, mesmo com maior tolerância a falsos positivos. Em matemática, a melhor opção foi SVM com seleção por regressão e balanceamento, com AUC ROC de 0,694 e *recall* de 64,1%. O desempenho mais modesto indica a necessidade de ações complementares de apoio.

Para uso cotidiano , que exige equilíbrio entre desempenho e simplicidade, recomenda-se SVM com seleção por regressão e balanceamento em ambas as disciplinas. Em português, o modelo obteve F_1-Score macro de 0,732 com nove variáveis, combinando acurácia e parcimônia. Em matemática, o F_1-Score macro foi de 0,710 com apenas seis variáveis, oferecendo solução eficiente para triagens iniciais.

Em contextos que exigem transparência e interpretabilidade , como auditorias ou decisões compartilhadas com gestores, os modelos baseados em arvores de decisão são os mais indicados. Apesar de apresentarem desempenho ligeiramente inferior (F_1-Score acima de 0,65), fornecem regras claras, auditáveis e de fácil compreensão para os diferentes públicos envolvidos.

Síntese das descobertas principais

A análise comparativa evidenciou que a predição de desempenho em matemática impõe desafios significativamente maiores do que em português, exigindo abordagens metodológicas distintas. Em matemática, os fatores contextuais explicam parcela menor da variabilidade do desempenho, resultando em menor previsibilidade estrutural. Os algoritmos que melhor se adaptam a cada disciplina também são distintos: SVM com *kernel* linear foi mais eficaz em matemática, enquanto SVM com *kernel* RBF apresentou melhor desempenho em português.

O balanceamento das classes mostrou-se mais eficiente em português, promovendo ganhos consistentes entre diferentes algoritmos. Em matemática, sua efetividade foi mais instável, dependendo fortemente do modelo utilizado. Os *trade-offs* entre precisão e *recall* foram mais acentuados nesta disciplina, dificultando o equilíbrio entre detectar casos de risco e manter a acurácia geral.

A simplicidade dos modelos mostrou-se crucial para matemática: configurações mais parcimoniosas apresentaram melhor desempenho, sugerindo que modelos complexos não necessariamente agregam valor preditivo neste contexto. Uma possível explicação para esse padrão é o menor tamanho da base de dados de matemática em relação à de português, o que pode favorecer modelos com menor complexidade

para evitar sobreajuste e garantir estabilidade na generalização.

Implicações práticas: os achados orientam estratégias diferenciadas de intervenção: em português, o foco deve estar em programas de motivação e apoio psicossocial, dado que fatores comportamentais são mais preditivos; em matemática, são necessárias intervenções estruturais mais amplas, incluindo suporte familiar e redução de desigualdades socioeconômicas, uma vez que o desempenho depende mais de condições prévias e menos controláveis no curto prazo. Esta diferenciação é fundamental para o *design* de sistemas de apoio educacional efetivos e contextualizados.

Capítulo 5

Considerações finais

5.1 Retomada dos objetivos e principais resultados

Este trabalho teve como propósito central investigar a aplicação de técnicas de *Mineração de Dados Educacionais* e *Aprendizado de Máquina* na tarefa de prever o desempenho acadêmico de estudantes do ensino médio, com base em dados reais das disciplinas de português e matemática. Para isso, foram integradas diferentes estratégias de seleção de atributos e algoritmos de classificação supervisionada, buscando avaliar não apenas a acurácia preditiva, mas também a interpretabilidade dos modelos gerados e os fatores contextuais associados à aprovação ou reprovação escolar.

Ao longo das etapas do processo analítico — desde a preparação dos dados até a avaliação preditiva —, buscou-se responder às perguntas formuladas na fase de contextualização, com especial atenção às diferenças entre disciplinas, à influência das variáveis contextuais e ao impacto da seleção de atributos sobre o desempenho dos modelos.

A análise exploratória inicial contribuiu para identificar variáveis com associação consistente ao rendimento estudantil, como o histórico de reprovações, a escolaridade dos pais e aspectos comportamentais (tempo livre, consumo de álcool). A etapa de seleção de atributos — conduzida por métodos estatísticos, inferenciais e

pela heurística *PerfilScore* — possibilitou a construção de diferentes conjuntos preditivos, revelando como a escolha das variáveis afeta a qualidade e a simplicidade dos modelos.

Na modelagem supervisionada, observaram-se contrastes relevantes entre as disciplinas. Em português, os melhores desempenhos foram obtidos com modelos SVM (kernel RBF), alcançando AUC ROC acima de 0,85 e *recall* superior a 80% em cenários de alerta precoce. Em matemática, os modelos mostraram menor capacidade discriminativa (AUC ROC máxima de 0,736), maior sensibilidade à complexidade algorítmica e resposta menos uniforme às estratégias de balanceamento. Parte dessa limitação pode ser atribuída ao menor tamanho da amostra disponível (395 casos, contra 649 em português), o que afeta a estabilidade dos modelos mais complexos.

De forma geral, os objetivos propostos foram atendidos. Foram testadas e comparadas diferentes técnicas de modelagem, avaliados os efeitos de estratégias de seleção de atributos, e interpretados os fatores associados ao desempenho escolar sob uma perspectiva analítica e contextual. As evidências obtidas apontam para a necessidade de abordagens diferenciadas por disciplina e reforçam o potencial de sistemas preditivos no apoio a intervenções educativas mais precisas e baseadas em dados.

5.2 Contribuições do trabalho

As contribuições desta pesquisa concentram-se em três eixos principais. Em primeiro lugar, destaca-se a aplicação de um *framework* de ciência de dados a um problema educacional real, indicando a viabilidade técnica e metodológica do uso de algoritmos preditivos no apoio à tomada de decisão educacional. Em segundo lugar, foram exploradas estratégias de seleção de atributos categóricos, incluindo a heurística *PerfilScore*, cuja aplicação demonstrou potencial utilidade em contextos com predominância de variáveis nominais e ordinais. Em terceiro lugar, a geração de *insights* específicos sobre os fatores associados ao desempenho em língua portu-

guesa e matemática oferece uma compreensão mais contextualizada das dificuldades escolares, potencialmente útil para subsidiar futuras intervenções pedagógicas.

Embora esta pesquisa se insira no campo da Educação, sua estrutura fundamenta-se em uma perspectiva oriunda da Engenharia, combinando lógica algorítmica, modelagem quantitativa e sistematização metodológica — competências associadas à formação em Engenharia Eletrônica. A adaptação de princípios clássicos da engenharia de sistemas a um domínio social complexo, como o educacional, evidencia a versatilidade das ferramentas analíticas e o potencial da Engenharia na proposição de soluções inovadoras para desafios sociais. Nesse sentido, o trabalho reforça a relevância da interdisciplinaridade e o papel integrador da Engenharia na interface entre tecnologia e desenvolvimento humano.

5.3 Limitações do estudo e dificuldades encontradas

Este estudo apresenta limitações que devem ser consideradas na interpretação dos resultados. A principal refere-se à base de dados utilizada, proveniente de dois contextos escolares em Portugal e coletada entre 2005 e 2006, o que restringe a generalização dos achados para realidades educacionais contemporâneas ou de outros países. Além disso, a amostra da disciplina de matemática apresentou tamanho inferior (395 casos, contra 649 em português), o que pode ter impactado negativamente tanto a estabilidade quanto a capacidade preditiva dos modelos nesse domínio, favorecendo configurações mais simples e menos suscetíveis ao sobreajuste.

Outras limitações relacionam-se às variáveis disponíveis, condicionadas pelo período em que a base foi coletada. Por se tratar de dados de 2005–2006, não foram contempladas informações sobre engajamento digital, tempo de estudo supervisionado, uso de tecnologias educacionais ou indicadores mais recentes de comportamento e motivação — fatores que, no contexto educacional atual, podem ser altamente relevantes para a modelagem preditiva.

Do ponto de vista metodológico, destaca-se que a heurística *PerfilScore*, proposta

para apoiar a seleção de atributos categóricos, não passou por validação formal comparativa com métricas consagradas. Embora sua utilidade prática tenha sido parcialmente verificada por meio da coerência com outras estratégias analíticas e pelo bom desempenho dos modelos que a empregaram, sua eficácia precisa ser confirmada em estudos adicionais.

Por fim, destaca-se a reflexão crítica sobre o uso de variáveis demográficas na modelagem preditiva, tema sensível pelas implicações éticas envolvidas. Neste estudo, diferentes estratégias de seleção de atributos foram aplicadas, e observou-se que a abordagem inferencial — mais restritiva — naturalmente excluiu parte dos preditores demográficos, como o gênero. Outros, como a escolaridade dos pais e a classificação urbana/rural, foram retidos em modelos específicos, com base em sua relevância estatística e justificativa pedagógica. Ressalta-se, no entanto, que o impacto preditivo de variáveis demográficas não é universal: fatores como gênero, raça ou localização variam conforme o contexto sociocultural, histórico e econômico de cada país. Assim, sua inclusão deve considerar não apenas critérios técnicos, mas também o grau de maturidade institucional e de equidade educacional do contexto analisado. A postura adotada neste trabalho buscou, portanto, equilibrar rigor analítico com responsabilidade contextual, promovendo um uso consciente e transparente dessas variáveis.

5.4 Sugestões para trabalhos futuros

As limitações e observações deste estudo abrem diversas possibilidades de aprofundamento e ampliação. As direções sugeridas a seguir buscam tanto complementar lacunas observadas quanto antecipar demandas emergentes da educação contemporânea.

Validação e generalização dos modelos

Recomenda-se a replicação do estudo em bases de dados provenientes de diferentes contextos geográficos, socioeconômicos e culturais, a fim de avaliar a robustez e

a transferibilidade dos resultados. Também se sugere a realização de investigações longitudinais, acompanhando estudantes ao longo de vários anos letivos, para examinar a estabilidade preditiva dos modelos em diferentes fases do percurso escolar.

Inclusão de dimensões educacionais contemporâneas

Novas variáveis relacionadas à transformação digital no ensino — como engajamento online, tempo de tela, uso de plataformas, gamificação e interação com sistemas baseados em IA — devem ser incorporadas em estudos futuros. Além disso, fatores socioemocionais digitais (ansiedade tecnológica, FOMO, cyberbullying, dependência digital) podem ser relevantes para compreender o desempenho acadêmico na era pós-pandemia.

Expansão e reutilização do *framework* analítico

O *framework* desenvolvido neste trabalho demonstrou flexibilidade e potencial de generalização. Aplicações futuras incluem predição de evasão escolar, identificação de perfis de engajamento, avaliação de intervenções pedagógicas e recomendações personalizadas. Sua adaptação para diferentes níveis e modalidades de ensino (educação especial, EJA, ensino técnico, EAD) pode revelar padrões específicos e fomentar soluções mais customizadas.

Aprofundamento em interpretabilidade

Estudos futuros devem explorar técnicas de *interpretable machine learning*, como LIME, SHAP e modelos simplificados com explicabilidade global. O desenvolvimento de visualizações interativas e relatórios interpretáveis pode facilitar a adoção prática dos modelos por educadores e gestores escolares.

Ética, equidade e justiça algorítmica

A análise crítica de vieses preditivos e dos efeitos distributivos dos modelos deve ser aprofundada. Sugere-se o uso de métricas de fairness, auditoria algorítmica

contínua e técnicas de mitigação de viés, como reamostragem balanceada ou regularização. Essas abordagens são fundamentais para promover sistemas mais justos e transparentes.

Integração com políticas públicas

Estudos futuros podem investigar como os insights gerados pelos modelos preditivos podem subsidiar políticas educacionais baseadas em evidências. A avaliação longitudinal do impacto de intervenções informadas por modelos, incluindo sua viabilidade e sustentabilidade, pode fortalecer a articulação entre ciência de dados e gestão educacional.

5.5 Síntese e Perspectivas Finais

Este trabalho demonstrou o potencial das abordagens de ciência de dados para compreender e predizer o desempenho acadêmico de estudantes do ensino médio. A identificação de fatores de risco e sucesso — especialmente aqueles de natureza comportamental e contextual — mostrou-se fundamental para subsidiar intervenções educativas mais eficazes e direcionadas. Os resultados obtidos também ressaltam a importância de adaptar estratégias analíticas às particularidades de cada disciplina, respeitando suas dinâmicas e limitações estruturais.

Espera-se que os métodos, evidências e reflexões aqui apresentados contribuam para o aprimoramento das práticas de monitoramento educacional, bem como para o desenvolvimento de sistemas de apoio mais justos, responsivos e orientados por dados. A utilização responsável e contextualizada da modelagem preditiva revela-se, assim, um instrumento promissor na mitigação de desigualdades educacionais e na promoção do sucesso escolar em diferentes realidades.

Referências

AHMED, E. (2024). Student performance prediction using machine learning algorithms. *Applied Computational Intelligence and Soft Computing*, 2024:1–15.

ALALAWI, K.; ATHAUDA, R.; CHIONG, R. (2023). Contextualizing the current state of research on the use of machine learning for student performance prediction: A systematic literature review. *Engineering Reports*, 5(12):e12699.

ALALAWI, K.; ATHAUDA, R.; CHIONG, R. (2025). An extended learning analytics framework integrating machine learning and pedagogical approaches for student performance prediction and intervention. *Int J Artif Intell Educ*, 35:1239–1287.

ALBREIKI, B.; ZAKI, N.; ALASHWAL, H. (2021). A systematic literature review of students' performance prediction using machine learning techniques. *Education Sciences*, 11(9):552.

ALRESHIDI, N. A. K. (2023). Enhancing topic-specific prior knowledge of students impacts their outcomes in mathematics. *Frontiers in Education*, 8:e1050468.

ALSARIERA, Y. A. et al. (2022). Assessment and evaluation of different machine learning algorithms for predicting student performance. *Computational Intelligence and Neuroscience*, 2022:4151487.

AULAKH, K.; ROUL, R. K.; KAUSHAL, M. (2023). E-learning enhancement through educational data mining with covid-19 outbreak period in backdrop: A review. *International journal of educational development*, 101:102814.

BAKER, R. S. et al. (2023). Using demographic data as predictor variables: a questionable choice. *Journal of Educational Data Mining*, 15(2):22–52.

BAKER, R. S. J. D.; CARVALHO, A. M. J. B. D. (2010). Mineração de dados educacionais: oportunidades para o brasil. *Revista Brasileira de Informática na Educação*, 18(1):1–18.

BAKER, R. S. J. D.; YACEF, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1):3–17.

BATISTA, M. R.; FAGUNDES, R. A. D. A. (2023). Mineração de dados educacionais aplicada à performance de estudantes: uma revisão sistemática da literatura. *Revista Novas Tecnologias na Educação*, 21(1):271–280.

CAMBRUZZI, W. L. (2014). Mineração de dados educacionais e learning analytics: Aplicações para o monitoramento da evasão escolar. Dissertação de mestrado, Universidade do Vale do Rio dos Sinos, São Leopoldo, RS.

CHAPMAN, P. et al. (2000). *CRISP-DM 1.0*. CRISP-DM Consortium.

CHEN, Y. et al. (2025). Machine learning-driven student performance prediction for enhancing tiered instruction. *arXiv preprint arXiv:2502.03143*.

CORTES, C.; VAPNIK, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

CORTEZ, P.; SILVA, A. M. (2008a). Student Performance Data Set. `https://archive.ics.uci.edu/dataset/320/student+performance`. Acesso em: 30 mar. 2025.

CORTEZ, P.; SILVA, A. M. (2008b). Using data mining to predict secondary school student performance. In *Proceedings of 5th Annual Future Business Technology Conference*, páginas 5–12, Porto. EUROSIS.

CUNHA, J. P. Z. (2019). Um estudo comparativo das técnicas de validação cruzada aplicadas a modelos mistos. Dissertação de mestrado, Universidade de São Paulo, São Paulo.

DE WINTER, J. C. F.; GOSLING, S. D.; POTTER, J. (2016). Comparing the pearson and spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological Methods*, 21(3):273–290.

Escola Politécnica de Saúde Joaquim Venâncio (EPSJV) (2009). Capital cultural. `https://www.sites.epsjv.fiocruz.br/dicionario/verbetes/capcul.html`. Acesso em: 06 jun. 2025.

GARDNER, J.; BROOKS, C.; BAKER, R. S. (2019). Evaluating the fairness of predictive student models through simulation. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, páginas 225–234, New York, NY, USA. Association for Computing Machinery.

GRAMS, D. B. O. (2024). Uso de estratégias de reamostragem para correção de desbalanceamento entre classes em modelos de classificação. Trabalho de Conclusão de Curso (Graduação em Estatística).

GUNASEKARA, S.; SAARELA, M. (2024). Explainability in educational data mining and learning analytics: An umbrella review. In PAASSEN, B.; EPP, C. D., editors, *Proceedings of the 17th International Conference on Educational Data Mining (EDM 2024)*, páginas 887–892, Atlanta, Georgia, USA. International Educational Data Mining Society.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd. edição.

IGUAL, L.; SEGUÍ, S. (2017). *Introduction to Data Science*. Springer International Publishing.

IMRAN, M. et al. (2019). Student academic performance prediction using supervised learning techniques. *International Journal of Emerging Technologies in Learning (iJET)*, 14(14):92–104.

Instituto de Avaliação Educativa (IAVE) (2016). PISA 2015 - Portugal: Volume I: Literacia Científica, Literacia de Leitura & Literacia Matemática. Technical report, IAVE - Instituto de Avaliação Educativa, Lisboa, Portugal. Relatório Nacional.

JIN, Y. et al. (2024). Fate in mmla: A student-centred exploration of fairness, accountability, transparency, and ethics in multimodal learning analytics. *arXiv preprint arXiv:2402.19071*.

JOLLIFFE, I. T. (2002). *Principal Component Analysis*. Springer, 2. edição.

KIZILCEC, R. F.; LEE, H.; SALTARELLI, A. J. (2020). Algorithmic fairness in education. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, páginas 30–41.

KOLLMANNSBERGER, S. et al. (2021). Machine learning in physics and engineering. In *Deep Learning in Computational Mechanics: An Introductory Course*, páginas 47–54. Springer International Publishing, Cham.

LI, H.; LIU, W.; ZHANG, Y. (2024). Predicting students' learning performance using machine learning with feature selection. *Electronics*, 13(3):659.

LIAO, S. (2022). The application of penalized logistic regression for fraud detection: studying measures of prediction performance for class imbalanced and

high-dimensional data. Dissertação de mestrado, University of Oslo, Department of Mathematics, Oslo.

MALIK, S. et al. (2025). Advancing educational data mining for enhanced student performance prediction: a fusion of feature selection algorithms and classification techniques with dynamic feature ensemble evolution. *Sci Rep.*, 15(1):8738.

MCKINNEY, W. (2018). *Python para análise de dados: tratamento de dados com Pandas, NumPy e IPython*. Novatec Editora.

MEHRABI, N. et al. (2022). A survey on bias and fairness in machine learning.

MISHRA, B. B.; SAHOO, M. (2016). Application of feature selection methods in educational data mining. *Procedia Computer Science*, 85:74–80.

NAFURI, M. et al. (2022). Clustering analysis for classifying student academic performance in higher education. *Applied Sciences*, 12(19):9467.

NAWANG, H.; MAKHTAR, M.; WAN HAMZAH, W. M. A. F. (2022). Comparative analysis of classification algorithm evaluations to predict secondary school students' achievement in core and elective subjects. *International Journal of Advanced Technology and Engineering Exploration*, 9(89):430–441.

OCHOA, X.; WISE, A. F.; KNIGHT, S. (2017). Towards a concern for fairness in learning analytics. *Journal of Learning Analytics*, 4(3):1–6.

OLIVEIRA, C. F. D. et al. (2021). How does learning analytics contribute to prevent students' dropout in higher education: A systematic literature review. *Big Data and Cognitive Computing*, 5(4):64.

ORDONEZ-AVILA, R. et al. (2023). Data mining techniques for predicting teacher evaluation in higher education: A systematic literature review. *Heliyon*, 9(3).

Organisation for Economic Co-operation and Development (OECD) (2007). *PISA 2006: Science Competencies for Tomorrow's World. Volume 1: Analysis*. OECD Publishing, Paris.

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

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

ROMERO, C.; VENTURA, S. (2024). Educational data mining and learning analytics: An updated survey. https://arxiv.org/abs/2402.07956.

ROSLAN, M. H. B.; CHEN, C. J. (2022). Educational data mining for student performance prediction: A systematic literature review (2015–2021). *International Journal of Emerging Technologies in Learning (iJET)*, 17(05):148–174.

ROY, K.; FARID, D. M. (2024). An adaptive feature selection algorithm for student performance prediction. *IEEE Access*.

ROZGONJUK, D. et al. (2020). Mathematics anxiety among STEM and social sciences students: the roles of mathematics self-efficacy, and deep and surface approach to learning. *International Journal of STEM Education*, 7(1):46.

SCHRÖER, C.; KRUSE, F.; GÓMEZ, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, 181:526–534.

SHAFIQ, D. A. et al. (2021). Student retention using educational data mining and predictive analytics: a systematic literature review. *IEEE Access*, 9:106904–106924.

SHAQIRI, M. S. et al. (2023). Differences between the correlation coefficients pearson, kendall and spearman. *Journal of Natural Sciences and Mathematics of UT*, 8(15-16):392–397.

SILVA FILHO, R. L. C.; BRITO, K.; ADEODATO, P. J. L. (2023). Leveraging causal reasoning in educational data mining: An analysis of brazilian secondary education. *Applied Sciences*, 13(8).

SINGH, S.; RATHI, A. (2016). Study and analysis of filter feature selection algorithms in educational data mining. *International Journal of Computer Applications Technology and Research (IJCATR)*, 5(4).

SLADE, S.; PRINSLOO, P. (2013). Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist*, 57(10):1510–1529.

TANG, Z.; ZHANG, J.; ZHANG, K. (2023). What-is and how-to for fairness in machine learning: A survey, reflection, and perspective. *ACM Comput. Surv.*, 55(13s).

TONG, T.; LI, Z. (2025). Predicting learning achievement using ensemble learning with result explanation. *PLoS ONE*, 20(1):e0312124.

WANG, Y. et al. (2021). A framework for explainable AI in education. In *Companion Proceedings of the 11th International Conference on Learning Analytics & Knowledge (LAK21)*, páginas 195–203.

XU, W. et al. (2013). A comparative analysis of spearman's rho and kendall's tau in normal and contaminated normal models. *Signal Processing*, 93(1):261–276.

YANG, Z. et al. (2024). Inherently interpretable tree ensemble learning. *IEEE Transactions on Neural Networks and Learning Systems*. arXiv preprint arXiv:2410.19098.

Apêndice A

Funções de pré-processamento e preparação de dados

Este apêndice apresenta as funções implementadas no módulo 'pre_modelagem.py', que concentram as rotinas de importação, padronização, transformação e preparação dos dados brutos para as etapas subsequentes de análise exploratória e modelagem preditiva.

A divisão estratificada dos dados em conjuntos de treino e teste, por sua vez, foi implementada separadamente, na raiz do projeto, com o objetivo de garantir a integridade e a reprodutibilidade dos subconjuntos utilizados ao longo do pipeline analítico.

A.1 Função 'importar_base'

Descrição

Função responsável pela importação padronizada dos conjuntos de dados das disciplinas de Matemática ou Português. Realiza a leitura do arquivo 'CSV', a renomeação das colunas para o português, a tradução dos valores categóricos e a criação da variável-alvo binária 'aprovacao', com base na nota final.

Parâmetros

- 'materia': 'str' — código da disciplina. Aceita 'mat', 'por', 'matematica' ou 'portugues'.
- 'caminho_completo': 'str', opcional — caminho completo do arquivo 'CSV'. Se não informado, utiliza um caminho padrão.

Retorno

'pd.DataFrame' — DataFrame com colunas traduzidas e variável-alvo 'aprovacao'.

A.1.1 Padronização de colunas e valores categóricos

Para assegurar a consistência e a clareza analítica, foi implementada uma padronização sistemática dos nomes das colunas e dos valores categóricos do conjunto de dados original. A seguir, apresenta-se o código correspondente aos dicionários utilizados na tradução:

```
colunas_renomeadas = {
    'school': 'escola', 'sex': 'genero', 'age': 'idade', 'address': 'endereco',
    'famsize': 'tamanho_familia', 'Pstatus': 'status_parental',
    'Medu': 'escolaridade_mae', 'Fedu': 'escolaridade_pai',
    'Mjob': 'profissao_mae', 'Fjob': 'profissao_pai',
    'reason': 'motivo_escolha_escola', 'guardian': 'responsavel_legal',
    'traveltime': 'tempo_transporte', 'studytime': 'tempo_estudo',
    'failures': 'reprovacoes', 'schoolsup': 'apoio_escolar',
    'famsup': 'apoio_familiar', 'paid': 'aulas_particulares',
    'activities': 'atividades_extracurriculares', 'nursery': 'frequentou_creche',
    'higher': 'interesse_ensino_superior', 'internet': 'acesso_internet',
    'romantic': 'relacionamento_romantico', 'famrel': 'relacao_familiar',
    'freetime': 'tempo_livre', 'goout': 'frequencia_saidas',
    'Dalc': 'alcool_dias_uteis', 'Walc': 'alcool_fim_semana',
    'health': 'saude', 'absences': 'faltas',
    'G1': 'nota1', 'G2': 'nota2', 'G3': 'nota_final'
# Dicionário para renomear colunas
df.rename(columns=colunas_renomeadas, inplace=True)
```

```
substituicoes = {
    'escola': {'GP': 'Gabriel Pereira', 'MS': 'Mousinho da Silveira'},
    'genero': {'F': 'Mulher', 'M': 'Homem'},
    'endereco': {'U': 'Urbano', 'R': 'Rural'},
    'tamanho_familia': {'GT3': 'Mais de 3 membros', 'LE3': '3 membros ou menos'},
    'status_parental': {'A': 'Separados', 'T': 'Juntos'},
    'profissao_mae': {'at_home': 'Dona de casa', 'health': 'Área da saúde',
                      'other': 'Outra profissão', 'services': 'Serviços',
                      'teacher': 'Professor(a)'},
    'profissao_pai': {'at_home': 'Dono de casa', 'health': 'Área da saúde',
                      'other': 'Outra profissão', 'services': 'Serviços',
                      'teacher': 'Professor(a)'},
    'motivo_escolha_escola': {'course': 'Curso específico', 'other': 'Outro motivo',
                              'home': 'Próximo de casa', 'reputation': 'Reputação da escola'},
    'responsavel_legal': {'mother': 'Mãe', 'father': 'Pai', 'other': 'Outro responsável'},
    'apoio_escolar': {'yes': 'Sim', 'no': 'Não'},
    'apoio_familiar': {'yes': 'Sim', 'no': 'Não'},
    'aulas_particulares': {'yes': 'Sim', 'no': 'Não'},
    'atividades_extracurriculares': {'yes': 'Sim', 'no': 'Não'},
    'frequentou_creche': {'yes': 'Sim', 'no': 'Não'},
    'interesse_ensino_superior': {'yes': 'Sim', 'no': 'Não'},
    'acesso_internet': {'yes': 'Sim', 'no': 'Não'},
    'relacionamento_romantico': {'yes': 'Sim', 'no': 'Não'}
}# Dicionário para traduzir valores categóricos

# Aplica as substituições
for coluna, mapa_valores in substituicoes.items():
    if coluna in df.columns:
        df[coluna].replace(mapa_valores, inplace=True)
```

A.1.2 Divisão estratificada e salvamento dos conjuntos de dados

A divisão dos dados em conjuntos de treino e teste foi realizada previamente, de forma estratificada, com o objetivo de manter a proporção das classes da variável-alvo e assegurar a integridade metodológica. Para evitar modificações acidentais dos dados originais e garantir reprodutibilidade, este procedimento foi implementado em um módulo separado, posicionado na raiz do projeto.

O código a seguir apresenta a função responsável por essa divisão e pelo salva-

mento dos conjuntos em arquivos 'CSV':

```
from modulos.pre_modelagem import importar_base
from sklearn.model_selection import train_test_split
import pandas as pd

def dividir_e_salvar(df, nome_base, target='aprovacao', test_size=0.3, random_state=42):
    """
    Realiza a divisão estratificada da base em conjuntos de treino e teste,
    e salva os resultados em arquivos CSV.
    """
    X = df.drop(columns=[target])
    y = df[target]

    X_train, X_test, _, _ = train_test_split(
        X, y, test_size=test_size, stratify=y, random_state=random_state
    )
    df_treino = df.loc[X_train.index].copy()
    df_teste = df.loc[X_test.index].copy()

    df_treino.to_csv(f'data/dados_treino_{nome_base}_rs{random_state}.csv', index=False)
    df_teste.to_csv(f'data/dados_teste_{nome_base}_rs{random_state}.csv', index=False)
    print(f"{nome_base.title()}: treino e teste salvos com random_state={random_state}")

if __name__ == "__main__":
    df_por = importar_base('portugues')
    df_mat = importar_base('matematica')

    dividir_e_salvar(df_por, nome_base='portugues')
    dividir_e_salvar(df_mat, nome_base='matematica')
```

Descrição do procedimento

- **Divisão estratificada**: Mantém a proporção das classes da variável-alvo ('aprovacao') nos conjuntos de treino e teste.
- **Persistência**: Os arquivos resultantes foram salvos na pasta 'data/', nomeados conforme o identificador da base e a semente de aleatoriedade ('random_state'), garantindo reprodutibilidade.
- **Isolamento do processo**: A divisão foi realizada fora do fluxo de análise,

prevenindo alterações indevidas nos conjuntos e assegurando a integridade das avaliações subsequentes.

Arquivos gerados

- 'data/dados_treino_portugues_rs42.csv'
- 'data/dados_teste_portugues_rs42.csv'
- 'data/dados_treino_matematica_rs42.csv'
- 'data/dados_teste_matematica_rs42.csv'

Esses conjuntos serviram de base para todas as análises exploratórias, modelagens e avaliações descritas neste trabalho, em conformidade com as boas práticas recomendadas para experimentos de Ciência de Dados.

A.1.3 Tabela de atributos do conjunto original

A Tabela A.1 apresenta os 33 atributos disponíveis nos conjuntos de dados originais, com nomes e descrições traduzidos para o português, conforme a documentação da base "*Student Performance*" CORTEZ; SILVA (2008a).

Quadro A.1: Descrição dos atributos da base

Atributos	**Descrição (Domínio)**
sexo	Sexo do estudante (binário: feminino ou masculino)
idade	Idade do estudante (numérico)
escola	Escola do estudante (binário: 'Gabriel Pereira' ou 'Mousinho da Silveira')
endereco	Tipo de meio habitacional (binário: urbano ou rural)
status_parental	Status de coabitação dos pais (binário: morando juntos ou separados)
escolaridade_mae	Nível de escolaridade da mãe (numérico: 0 a 4)
profissao_mae	Profissão da mãe (nominal)
escolaridade_pai	Nível de escolaridade do pai (numérico: 0 a 4)
profissao_pai	Profissão do pai (nominal)
responsavel_legal	Tutor do estudante (nominal: pai, mãe ou outro)
tamanho_familia	Tamanho da família (binário) ($\leq$3 ou >3)
relacao_familiar	Qualidade da relação familiar (numérico: 1– muito ruim até 5– excelente)
motivo_escolha_escola	Motivo da escolha da escola (nominal)
tempo_transporte	Tempo de viagem de casa até a escola (numérico)(1: < 15min; 2: 15–30min; 3: 30min–1h; 4: >1h)
tempo_estudo	Tempo de estudo semanal (numérico)(1: < 2h; 2: 2–5h; 3: 5–10h; 4: >10h)
reprovacoes	Número de reprovações anteriores (numérico)
apoio_escolar	Apoio escolar extra (binário: sim ou não)
apoio_familiar	Suporte escolar familiar (binário: sim ou não)
atividades_extracurriculares	Atividades extra-curriculares (binário: sim ou não)
aulas_particulares	Reforço escolar pago (binário: sim ou não)
acesso_internet	Acesso residencial à Internet (binário: sim ou não)
frequentou_creche	Frequentou creche (binário: sim ou não)
interesse_ensino_superior	Pretende ingressar no ensino superior (binário: sim ou não)
relacionamento_romantico	Está em um relacionamento amoroso (binário: sim ou não)
tempo_livre	Tempo livre depois da escola (numérico: 1 a 5)
frequencia_saidas	Sair com amigos (numérico: 1 a 5)
alcool_fim_semana	Consumo de álcool em finais de semana (numérico: 1 a 5)
alcool_dias_uteis	Consumo de álcool em dias de semana (numérico: 1 a 5)
saude	Status de saúde (numérico: 1– muito ruim até 5– muito bom)
faltas	Número de faltas escolares (numérico: de 0 a 93)
nota1, nota2, nota_final	Notas do primeiro e segundo período e nota final (numérico: de 0 a 20)

Fonte: Adaptado pela autora. Dados obtidos do conjunto "Student Performance Data Set", disponível no UCI Machine Learning Repository CORTEZ; SILVA (2008a).

Vale destacar que os atributos 'escolaridade_mae' e 'escolaridade_pai' seguem a codificação ordinal do sistema educacional português, em que: 0 representa ausência de escolaridade formal; 1 corresponde ao 1º ciclo do ensino básico (até o 4º ano); 2 abrange o 2º e 3º ciclos (do 5º ao 9º ano); 3 refere-se ao ensino secundário (10º ao 12º ano, equivalente ao ensino médio no Brasil); e 4 indica formação em nível superior. As profissões dos responsáveis foram agrupadas em cinco categorias: professor(a), área da saúde, serviços públicos (como administração ou polícia), do lar e outras ocupações. Essa codificação foi mantida com base no dicionário de dados original, de forma a preservar a comparabilidade com estudos correlatos.

A.2 Função 'preparar_treino_e_teste'

Descrição

Esta função é responsável pela preparação sistemática dos dados para a modelagem supervisionada, assegurando que estejam adequadamente estruturados e codificados para serem utilizados pelos algoritmos de aprendizado de máquina. Inicialmente, realiza-se a remoção opcional das colunas de notas ('nota1', 'nota2' e 'nota_final'), a fim de evitar vazamento de informação e garantir que apenas variáveis contextuais e comportamentais sejam utilizadas como preditores, conforme o delineamento metodológico deste estudo.

Em seguida, aplica-se a codificação das variáveis categóricas: os atributos binários ('Sim' ou 'Não') são convertidos para valores numéricos ('1' ou '0'),respectivamente, facilitando sua interpretação pelos algoritmos. As variáveis nominais, por sua vez, são codificadas por meio de One-Hot Encoding, o que permite representar categorias qualitativas sem pressupor qualquer ordenação ou relação hierárquica entre elas, preservando a informação e evitando vieses.

Na sequência, realiza-se a imputação de valores ausentes, utilizando a média para preenchimento, o que evita a perda de dados e mantém a coerência estatística das variáveis numéricas. Por fim, opcionalmente, aplica-se o escalonamento dessas variáveis por meio do 'StandardScaler', padronizando-as com média zero e desvio

padrão unitário. Esta etapa é fundamental para garantir a comparabilidade entre os atributos, evitando que variáveis com escalas distintas influenciem desproporcionalmente o desempenho dos modelos.

Após as transformações, os dados são separados em conjuntos X (preditores) e y (variável-alvo), estruturando-os conforme os requisitos das técnicas de modelagem supervisionada empregadas neste trabalho.

Parâmetros

- 'df_train', 'df_test': 'pd.DataFrame' — conjuntos de treino e teste.
- 'target': 'str' — nome da variável-alvo. Padrão: 'aprovacao'.
- 'drop_notas': 'bool' — se 'True', remove as colunas de notas.
- 'scaling': 'bool' — se 'True', aplica imputação e escalonamento.

Retorno

- 'X_train', 'X_test': conjuntos de preditores.
- 'y_train', 'y_test': variáveis-alvo.
- 'scaler', 'imputer': objetos utilizados para transformação, ou 'None' se 'scaling=False'.

Apêndice B

Funções de estatísticas descritivas e exploratórias

Este apêndice documenta as principais funções implementadas para apoiar a análise estatística descritiva e exploratória dos dados, bem como procedimentos específicos para o perfilamento e a seleção de atributos. As funções descritas viabilizam um processo estruturado de análise preliminar, essencial para a compreensão dos dados e para fundamentar decisões metodológicas subsequentes.

B.1 Função 'add_features_describe_pd'

Esta função foi desenvolvida para gerar um conjunto completo de estatísticas descritivas, tanto para variáveis numéricas quanto categóricas, com foco na análise exploratória de dados educacionais. Ela integra métricas tradicionais (como média, desvio padrão e quartis) e medidas mais avançadas, como o Coeficiente de Variação (CV), o teste de normalidade de Shapiro-Wilk e a Entropia de Shannon.

Além disso, permite a geração de relatórios adaptados para estudos de frequência, com foco na diversidade e predominância de categorias, sendo essencial para o perfilamento de atributos categóricos e posterior aplicação no Índice de Perfil Composto.

Parâmetros de entrada:

- 'colunas': lista de colunas a serem analisadas.

- 'estudo_frequencia': define se o tratamento será categórico ('True') ou numérico ('False').
- 'shapiro_values': indica se o p-valor de Shapiro-Wilk será calculado.
- 'shannon': ativa ou não o cálculo da Entropia de Shannon.

Principais operações:

Para variáveis **categóricas**, converte os dados para string e calcula, além das estatísticas básicas, a frequência relativa da categoria dominante e a diversidade percentual:

```
resumo['freq rel. top (%)'] = (resumo['freq'] / resumo['count'] * 100)
resumo['% únicas'] = (resumo['unique'] / resumo['count'] * 100)
```

Se 'shannon=True', calcula também a entropia de Shannon:

```
entropies[col_name] = entropy(counts, base=2) if not counts.empty else np.nan
```

Para variáveis **numéricas**, além das estatísticas descritivas padrão, calcula a Moda, o p-valor do teste de Shapiro-Wilk (quando 'shapiro_values=True'), e o Coeficiente de Variação:

```
resumo['CV'] = np.where(resumo['mean'] > 0, (resumo['std'] / resumo['mean']).round(3), np.nan)
```

As colunas do relatório podem ser renomeadas conforme um dicionário fornecido pelo usuário ou com nomes padrão em português, como por exemplo:

```
'mean': 'Média', 'std': 'Desvio Padrão', 'CV': 'Coeficiente de Variação (CV)'
```

Por fim, a função remove a coluna de contagem para apresentação mais clara:

```
resumo_final = resumo_renomeado.drop(columns=[col_contagem_nome])
```

Retorno: um 'DataFrame' contendo as estatísticas descritivas detalhadas para as colunas selecionadas, com foco em facilitar a interpretação e o perfilamento dos dados.

Exemplo de chamada:

```
add_features_describe_pd(df, colunas=['idade', 'frequencia'],
                         estudo_frequencia=False, shapiro_values=True)
```

Nota: a função adota práticas de segurança para evitar erros em casos de dados ausentes, pequenas amostras ou ausência de variabilidade, assegurando robustez na análise.

B.2 Implementação do *PerfilScore* para seleção de variáveis categóricas

Este procedimento automatiza a avaliação de variáveis categóricas com base em dois critérios principais: a **diversidade informacional**, representada pela Entropia Relativa, e a **variação no desempenho médio**, medida pelo gap entre categorias. A combinação ponderada dessas métricas resulta no *PerfilScore*, que orienta a priorização de atributos para perfilamento estudantil.

A função central que operacionaliza esse procedimento é 'avaliacao_variacao_pontuacao_media_por_categoria', implementada em Python.

Resumo das etapas e operações principais:

Geração de estatísticas descritivas: utiliza a função auxiliar 'add_features_describe_pd', que calcula, entre outras métricas, a Entropia de Shannon para cada variável categórica.

```
df_describe = add_features_describe_pd(df, colunas=atributos,
                                       estudo_frequencia=True, shannon=True)
```

Cálculo da entropia relativa: a Entropia de Shannon é normalizada pelo logaritmo do número de categorias, conferindo comparabilidade entre variáveis com diferentes cardinalidades.

```
df_describe['Entropia Relativa'] = df_describe['Entropia (Shannon)'] /
                                   np.log2(df_describe['Total de Categorias'])
```

Cálculo do gap de desempenho: para cada variável, estima-se a diferença máxima entre as médias da variável dependente (e.g., 'nota_final') nas diferentes categorias.

```
medias = df.groupby(col)[coluna_avaliada].mean()
gaps[col] = medias.max() - medias.min()
```

Aplicação de filtros heurísticos: são selecionadas apenas variáveis que atendem simultaneamente a:

- Frequência mínima.
- Entropia Relativa ≥ 0.4.

Cálculo do *PerfilScore*: combinação ponderada da Entropia Relativa e do Gap de Desempenho, com pesos iguais.

```
df_filtrado['PerfilScore'] = 0.5 * escore_entropia + 0.5 * escore_gap
```

Implementação de alerta de dispersão: identifica variáveis com distribuição excessivamente dispersa, que podem comprometer a interpretabilidade.

```
dispersao = (df_filtrado['Diversidade de Categorias (%)'] > 80.0) &
            (df_filtrado['Total de Categorias'] > 2)
df_filtrado['Alerta Dispersão'] = dispersao.map({True: 'Alta dispersão (>80%)', False: ''})
```

Seleção final: mantêm-se apenas variáveis que, além de passarem nos filtros anteriores, apresentem:

- Entropia Relativa ≥ 0.4.
- Gap de Desempenho ≥ 1.0.

Observação: Considerou-se inicialmente uma abordagem mais rígida, restringindo o limiar de entropia relativa a valores superiores. No entanto, variáveis com um gap de desempenho elevado não deveriam necessariamente ser excluídas apenas pela baixa diversidade interna. Assim, optou-se por flexibilizar o critério de Entropia Relativa para ≥ 0.4 e manter um alerta explícito sobre dispersão para permitir uma análise caso a caso mais segura e fundamentada.

Fórmulas Utilizadas: As fórmulas a seguir sintetizam os principais indicadores utilizados ao longo do trabalho. Ressalta-se que algumas, como a equação da entropia relativa (Equação B.2), já foram previamente apresentadas no Capítulo 2, sendo aqui retomadas para fins de consolidação.

$$H(X) = -\sum_{i=1}^{k} p(x_i) \log_2 p(x_i) \tag{B.1}$$

$$\textit{Entropia Relativa} = \frac{H(X)}{\log_2(k)} \quad \textit{(ver também Seção 2.5)} \tag{B.2}$$

$$\textit{Gap} = \max(\mu_{\textit{categoria}}) - \min(\mu_{\textit{categoria}}) \tag{B.3}$$

$$\textit{PerfilScore} = 0.5 \times \frac{\textit{Entropia Relativa}}{\max(\textit{Entropia Relativa})} + 0.5 \times \frac{\textit{Gap}}{\max(\textit{Gap})} \tag{B.4}$$

Nota: Este procedimento é um componente central do pipeline de perfilamento e seleção de atributos categóricos, combinando fundamentos informacionais

com critérios estatísticos robustos. Sua implementação permite uma avaliação sistemática e transparente das variáveis, subsidiando tanto análises exploratórias quanto modelagens preditivas.

B.3 Funções para análise de grupos extremos

Este conjunto de funções foi desenvolvido para apoiar a análise exploratória de perfis extremos de desempenho estudantil, com foco na identificação e comparação de categorias de variáveis qualitativas entre grupos com notas significativamente distintas. A abordagem baseia-se na segmentação dos dados em dois grupos: **baixo desempenho** e **alto desempenho**, definidos a partir de quantis da variável numérica 'nota_final', e na posterior análise das diferenças de proporções das categorias entre esses grupos.

Estas funções não integram diretamente o pipeline automatizado de seleção de atributos, mas fornecem **insumos qualitativos importantes** para a compreensão das diferenças contextuais entre os estudantes, orientando a interpretação e a eventual decisão de incluir variáveis na modelagem preditiva.

B.3.1 Função 'comparar_grupos_extremos'

Esta função executa a comparação direta das proporções de categorias de variáveis qualitativas entre os grupos de **baixo** e **alto desempenho**. Para cada variável categórica, calcula a diferença absoluta de proporções das categorias nos dois grupos, destacando aquelas cuja diferença supera um limiar mínimo definido pelo usuário.

Principais parâmetros:

- 'variavel_numerica': variável utilizada para definir os grupos extremos, usualmente 'nota_final'.
- 'variaveis_categoricas': lista de variáveis categóricas a serem comparadas.
- 'Q1', 'Q3': limites inferior e superior para definição dos grupos.

- 'min_diferenca': diferença mínima de proporção para considerar uma categoria relevante.

Resultado: Um 'DataFrame' contendo, para cada variável e categoria, a proporção observada em cada grupo e a diferença absoluta percentual.

B.3.2 Função 'identificar_extremos_comparaveis'

Esta função atua como um *wrapper* da função anterior ('comparar_grupos_extremos'), adicionando mecanismos de **automação** e **otimização** na escolha dos limites para definição dos grupos de desempenho extremo.

Pode operar em três modos distintos:

- **Manual**: utilizando limites explícitos fornecidos pelo usuário ('entrada=(Q1, Q3)').
- **Quantil fixo**: utilizando um quantil padrão, como 25% ('q_limite=0.25').
- **Otimização automática**: testando múltiplos quantis para encontrar aqueles que proporcionem grupos com tamanhos equilibrados e que respeitem critérios internos de qualidade.

Critérios de otimização considerados:

- **Tamanho mínimo** de cada grupo: $n \geq 30$.
- **Diferença relativa máxima** entre tamanhos dos grupos: até 20%.
- **Limites de notas adequados** para caracterização dos grupos extremos:
  - Nota mínima do grupo alto: ≥ 14.0.
  - Nota máxima do grupo baixo: ≤ 9.0.

Quando configurada com 'otimizar=True', a função percorre uma lista de quantis candidatos:

```
quantis_teste = [0.10, 0.125, 0.15, 0.175, 0.20, 0.225, 0.25, 0.275, 0.30]
```

Para cada quantil, calcula-se os limites:

```
Q1_teste = df[variavel_numerica].quantile(q_atual)
Q3_teste = df[variavel_numerica].quantile(1 - q_atual)
```

Em seguida, verifica-se se esses limites atendem aos critérios estabelecidos. Se forem satisfatórios, realiza-se a comparação entre os grupos via 'comparar_grupos_extremos'.

Exemplo do filtro por critérios:

```
if Q3_teste < nota_minima_grupo_alto or Q1_teste > nota_maxima_grupo_baixo:
    continue  # Descartado
if n_baixo_teste < tamanho_minimo_grupo or n_alto_teste < tamanho_minimo_grupo:
    continue
if diff_tamanho_abs / max(n_baixo_teste, n_alto_teste) > max_diff_relativa_tamanho:
    continue
```

Caso os critérios sejam satisfeitos, a função compara a diferença absoluta de tamanhos entre os grupos e, se for menor que a melhor encontrada até então, atualiza a solução ótima:

```
if diff_tamanho_abs < melhor_diff_abs_tamanho:
    melhor_diff_abs_tamanho = diff_tamanho_abs
    melhores_resultados_tupla = (df_dif_teste, n_baixo_teste,
                                 n_alto_teste, Q1_teste, Q3_teste)
```

Resultado: Ao final, retorna o melhor conjunto de limites identificado, juntamente com a tabela de categorias com diferenças significativas entre os grupos, além dos tamanhos dos grupos e os limites utilizados:

- DataFrame com categorias relevantes.

- Tamanho dos grupos baixo e alto.
- Valores de Q_1 e Q_3 selecionados.

Caso nenhum par de quantis satisfaça todos os critérios simultaneamente, a função retorna 'None' e emite uma mensagem informativa.

B.3.3 Função 'plot_top_diferencas_extremos'

Esta função permite a **visualização gráfica** das principais diferenças percentuais identificadas entre os grupos de desempenho. A partir do 'DataFrame' gerado pelas funções anteriores, cria gráficos de barras horizontais destacando as categorias com maior diferença absoluta de proporção.

Recursos do gráfico:

- Anotação direta do valor percentual sobre cada barra.
- Ajuste dinâmico de cores conforme o contexto (e.g., disciplina analisada).
- Exibição de informações adicionais sobre os tamanhos dos grupos e os limites utilizados para a segmentação.
- Opção de salvar automaticamente o gráfico em diretórios organizados.

B.3.4 Aplicação no Estudo

Estas funções foram aplicadas na etapa de AED, com o objetivo de:

- Identificar categorias discriminantes com base na variável 'nota_final', contrastando grupos de baixo e alto desempenho;
- Apoiar a interpretação qualitativa dos resultados, enriquecendo a compreensão sobre os perfis estudantis;
- Subsidiar decisões sobre a manutenção ou exclusão de variáveis no processo de modelagem preditiva, com base na magnitude das diferenças identificadas;

- Ranquear categorias com maior variação percentual entre os grupos, utilizando como critério a **diferença absoluta de proporção**.

Exemplo de uso da função de identificação de grupos extremos

```
df_dif, n_baixo, n_alto, Q1, Q3 = identificar_extremos_comparaveis(
    df, variavel_numerica='nota_final',
    variaveis_categoricas=['atividade_extra', 'apoio_familiar']
)
plot_top_diferencas_extremos(df_dif, materia='portugues',
                             q1_lim=Q1, q3_lim=Q3,
                             n_baixo=n_baixo, n_alto=n_alto,
                             top_n=10, salvar=True)
```

B.3.5 Considerações

A utilização desta abordagem de análise de grupos extremos proporcionou insights adicionais sobre as diferenças contextuais entre estudantes, contribuindo para uma análise mais rica e orientada a evidências. Trata-se de uma ferramenta **exploratória e interpretativa**, que complementa os demais procedimentos analíticos descritos neste trabalho.

B.4 Procedimentos de regressão exploratória e seleção de atributos

Embora a regressão seja tradicionalmente associada à modelagem preditiva, neste trabalho foi utilizada exclusivamente como ferramenta de **análise exploratória**, com foco na compreensão das relações entre variáveis contextuais e o desempenho escolar. As funções implementadas viabilizaram o ajuste e avaliação de modelos de regressão linear múltipla, a seleção automática de atributos e o diagnóstico da qualidade dos modelos ajustados.

B.4.1 Funções implementadas

Foram desenvolvidas as seguintes funções auxiliares que compõem o núcleo desta análise exploratória:

- 'regressao_multipla': ajusta modelos de regressão linear múltipla com base na técnica de Mínimos Quadrados Ordinários (OLS), retornando o sumário estatístico completo e os objetos de predição.
- 'stepwise_selection': realiza seleção automatizada de variáveis preditoras, utilizando o método **stepwise** (bidirecional), com base em critérios estatísticos como p-valor, AIC ou BIC. Esse procedimento adiciona e remove variáveis iterativamente, buscando um modelo parcimonioso e estatisticamente robusto.
- 'avaliar_residuos_regressao': fornece diagnóstico gráfico e estatístico dos resíduos do modelo, com geração de visualizações (resíduos vs. preditos, histograma, Q-Q plot) e execução de testes formais de normalidade (Shapiro-Wilk), homoscedasticidade (Breusch-Pagan) e autocorrelação (Durbin-Watson).
- 'comparar_modelos_regressao': sintetiza e compara múltiplos modelos de regressão, com base em métricas como R^2, AIC, BIC, log-verossimilhança e número de variáveis significativas, apoiando a escolha do modelo mais adequado.

B.4.2 Procedimento analítico

O procedimento analítico conduzido com essas funções seguiu os seguintes passos:

1. Ajuste inicial de modelos completos utilizando todas as variáveis contextuais como preditoras, com inspeção dos coeficientes e p-valores.
2. Aplicação do **stepwise selection** para automatizar a escolha de variáveis com maior significância estatística ou melhor adequação aos critérios de informação (AIC/BIC). O processo de seleção funciona da seguinte forma:

- **Etapa Forward**: são testadas as variáveis ainda não incluídas no modelo. A variável que melhora o critério (redução do AIC/BIC ou p-valor abaixo do limiar) é adicionada.
- **Etapa Backward**: são testadas as variáveis já incluídas. Caso alguma perca significância (por exemplo, p-valor acima do limiar), é removida.
- O processo se repete iterativamente até que nenhuma variável possa ser adicionada ou removida segundo os critérios definidos.

3. Comparação entre diferentes modelos ajustados, observando indicadores de ajuste e complexidade, para embasar decisões analíticas.

B.4.3 Exemplo de aplicação

O uso típico dessas funções no estudo consistiu no seguinte fluxo:

Fluxo típico de aplicação das funções de regressão exploratória

```
# Ajuste do modelo inicial
modelo, X_ctx, y_ctx = regressao_multipla(
                                    df_pp,
                                    target='nota_final',
                                    variaveis=variaveis_contexto
)
# Seleção automática de variáveis
variaveis_selecionadas = stepwise_selection(df_pp, target='nota_final',
                                            variaveis_candidatas=variaveis_contexto,
                                            criterion='aic')
# Ajuste final com variáveis selecionadas
modelo_final, X_sel, y_sel = regressao_multipla(df_pp, target='nota_final',
                                               variaveis=variaveis_selecionadas)
# Diagnóstico dos resíduos
resultados_residuos = avaliar_residuos_regressao(y_sel, modelo_final.predict(X_sel))
# Comparação de modelos
df_comparacao = comparar_modelos_regressao([modelo, modelo_final],
                                          nomes=['Completo', 'Selecionado'])
```

B.4.4 Papel no Estudo

Estas ferramentas foram fundamentais para:

- Explorar as associações entre variáveis contextuais e o desempenho final ('nota_final'), sem risco de **vazamento de informação** para modelos futuros de classificação.
- Guiar a **seleção de atributos** com base empírica, considerando significância estatística, ausência de multicolinearidade e estabilidade dos coeficientes.
- Fornecer um **suporte analítico rigoroso**, permitindo verificar a robustez dos resultados e a adequação das variáveis ao contexto educacional estudado.

Importante destacar que, embora a regressão tenha subsidiado o processo de **seleção exploratória de atributos**, os modelos de regressão em si não foram empregados diretamente nas tarefas de **modelagem preditiva** de classificação binária, mas sim como instrumento de investigação das relações contextuais e de apoio às decisões de pré-processamento.

Apêndice C

Funções para Seleção Exploratória de Atributos

Este apêndice reúne as funções desenvolvidas para apoiar a seleção exploratória de atributos, abrangendo a identificação de redundâncias, associações e relevância estatística entre variáveis numéricas e categóricas. Essas ferramentas foram essenciais na etapa de preparação e escolha de atributos com maior potencial explicativo ou preditivo, contribuindo para a robustez e interpretabilidade dos modelos.

As funções aqui descritas foram aplicadas na **Estratégia 3 – Seleção por Testes Estatísticos Inferenciais** (Seção 4.2.3), que envolveu:

- **Variáveis ordinais**: análise mediante Correlação de Spearman e Teste de Kruskal-Wallis, com seleção baseada em correlação $\geq 0{,}15$ e p-valor $< 0{,}05$.
- **Variáveis nominais**: avaliação via Teste de Qui-Quadrado e V de Cramér, com inclusão das variáveis que apresentaram p-valor $< 0{,}05$ e V $\geq 0{,}10$.

Além disso, o cálculo do Fator de Inflação da Variância (VIF) e a análise de multicolinearidade auxiliaram no controle de redundâncias entre variáveis numéricas, promovendo a estabilidade dos modelos.

A seguir, são apresentadas as funções que operacionalizam esses procedimentos.

C.1 Análise de Multicolinearidade

C.1.1 Função 'relatorio_multicolinearidade'

Esta função gera um relatório detalhado de multicolinearidade entre variáveis numéricas, combinando o cálculo do Fator de Inflação de Variância (VIF) com a análise da matriz de correlação absoluta.

Aspectos principais:

- Cálculo do VIF para cada variável numérica.
- Identificação de pares de variáveis com correlação elevada.
- Geração de avaliação textual sobre o nível de redundância.

Trecho ilustrativo:

```
vifs = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
df_vif = pd.DataFrame({'variavel': colunas, 'vif': vifs})


def avaliar(row):
    if row['vif'] >= limite_vif and row['Alta correlação com'] != '-':
        return 'VIF alto + correlação alta'
    elif row['vif'] >= limite_vif:
        return 'VIF elevado'
    elif row['Alta correlação com'] != '-':
        return 'Correlação elevada'
    else:
        return 'Sem alerta'
```

Este relatório orienta a exclusão ou transformação de variáveis redundantes, promovendo a parcimônia do modelo.

C.1.2 Função 'calcular_vif'

Função auxiliar que calcula o VIF para um conjunto específico de variáveis, facilitando análises pontuais de multicolinearidade.

Cálculo de VIF para variáveis especificadas

```
vif_data = pd.DataFrame()
vif_data['variavel'] = X_with_const.columns
vif_data['VIF'] = [variance_inflation_factor(X_with_const.values, i)
                   for i in range(X_with_const.shape[1])]
```

C.2 Seleção estatística de variáveis categóricas

C.2.1 Função 'selecionar_nominais_relevantes'

Seleciona variáveis nominais com associação estatisticamente significativa ao alvo, combinando o teste Qui-quadrado e o Coeficiente de Contingência V de Cramér.

Critérios de seleção:

- *P-valor* do teste Qui-quadrado < 0.05.
- V de Cramér superior a um limiar (default: 0.3).

Cálculo do V de Cramér para variáveis categóricas

```
v_cramer = (chi2 / (n * min_dim)) ** 0.5
if p < 0.05:
    results.append({'Variable': column, 'V de Cramér': v_cramer})
```

Esta função permite priorizar variáveis com maior capacidade discriminativa sobre o alvo.

C.2.2 Função 'selecionar_ordinais_relevantes'

Identifica variáveis ordinais associadas ao alvo por meio da Correlação de Spearman e do teste de Kruskal-Wallis.

Procedimentos principais:

- Seleção por significância na correlação de Spearman ($p < 0.05$).
- Cálculo complementar do teste de Kruskal-Wallis para apoio à análise.

Correlação de Spearman e teste de Kruskal-Wallis

```
coef, p_spear = spearmanr(temp_df[var], temp_df[target])
h_stat, p_kruskal = kruskal(*grupos)
```

A ordenação final das variáveis se dá pela magnitude absoluta da correlação de Spearman.

Apêndice D

Procedimentos de modelagem e avaliação de classificadores

Este apêndice documenta as funções implementadas no módulo 'modelagem.py', utilizadas na avaliação, otimização e comparação de classificadores binários no contexto educacional.

D.1 Fluxograma do processo de avaliação e comparação de classificadores

A Figura D.1 ilustra o processo sistemático adotado para a avaliação e comparação de múltiplos classificadores binários, conforme implementado no módulo 'modelagem.py'.

O fluxo operacional abrange desde o pré-processamento dos dados até a geração de relatórios de desempenho, incorporando práticas recomendadas de otimização e validação cruzada.

As principais etapas são:

- **Divisão dos dados**: separação estratificada em conjuntos de treino e teste.
- **Balanceamento (opcional)**: aplicação do SMOTE-Tomek para mitigar desequilíbrios de classes.

- **Treinamento do modelo base**: avaliação do desempenho inicial sem otimização.
- **Otimização de hiperparâmetros (opcional)**: uso do 'GridSearchCV' para ajustar parâmetros e melhorar o desempenho.
- **Avaliação**: geração de métricas de desempenho no conjunto de teste e via validação cruzada.
- **Comparação**: análise comparativa entre o modelo base e o otimizado, incluindo curvas ROC, Precision-Recall e matrizes de confusão.
- **Diagnóstico de overfitting**: comparação entre métricas de teste e validação cruzada.
- **Exportação**: geração de tabelas e gráficos para documentação e análise.

Este fluxo foi implementado de forma modular, garantindo reprodutibilidade e flexibilidade para aplicação em diferentes contextos e conjuntos de dados.

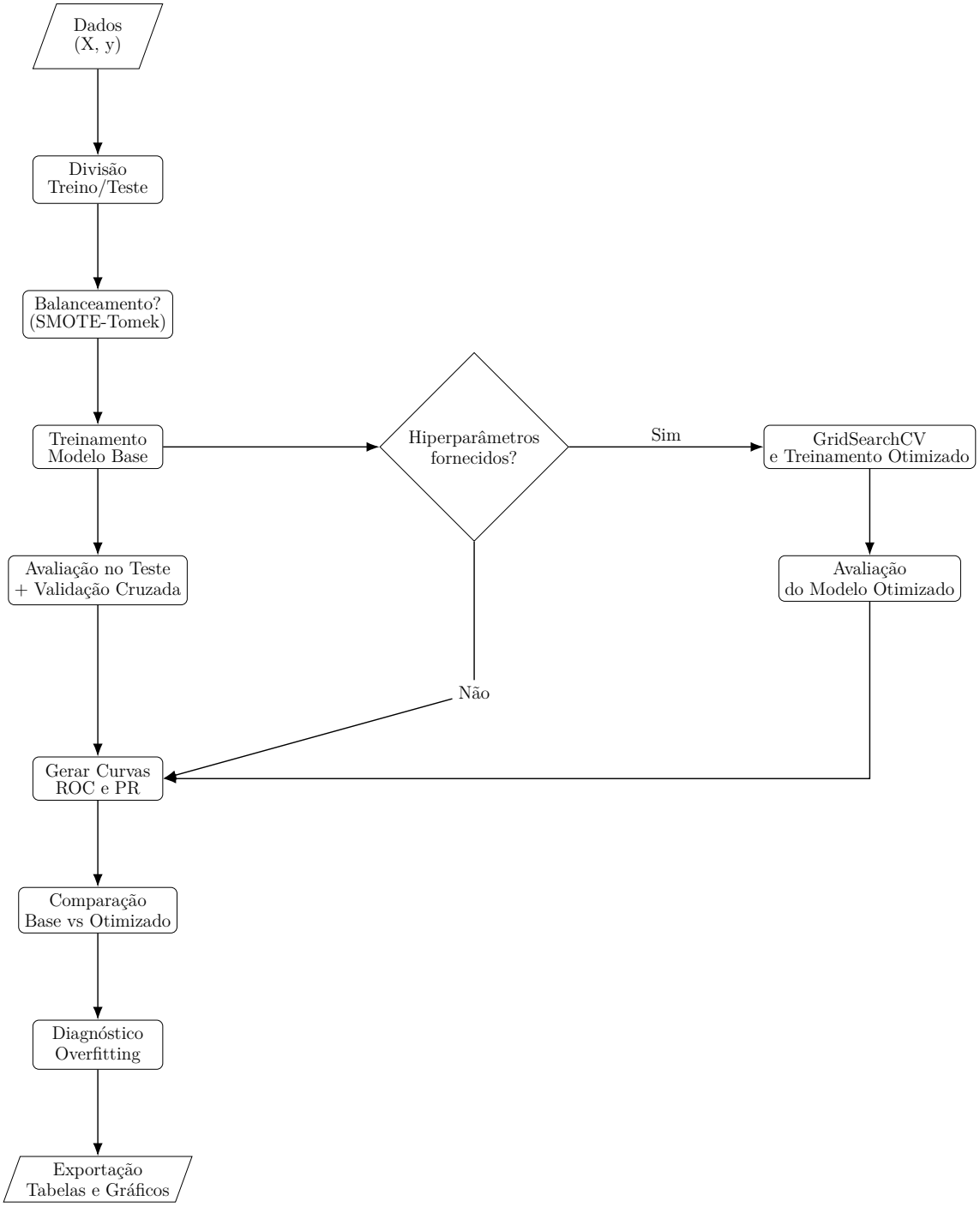


Figura D.1: Fluxograma do processo de avaliação e comparação de classificadores Binários

Fonte: Elaborado pela autora.

Este fluxograma sistematiza o pipeline descrito, assegurando clareza na replicação do processo e na interpretação dos resultados.

D.2 Avaliação e otimização de modelos

Função 'avaliar_classificadores_binarios_otimizados' Esta função executa o pipeline completo de avaliação e otimização de modelos, incluindo treinamento, validação cruzada, ajuste via 'GridSearchCV', e visualização.

Trecho chave - Loop principal de avaliação

```
for nome_modelo, base in classificadores.items():
    print(f"\nProcessando modelo: {nome_modelo}")
    modelo = base.class(**base.get_params())
    modelo.set_params(random_state=42)

    # --- Treinamento e Avaliação Base ---
    modelo.fit(X_train, y_train)
    y_pred = modelo.predict(X_test)
    if hasattr(modelo, 'predict_proba'):
        y_prob = modelo.predict_proba(X_test)[:, 1]
    else:
        y_prob = np.nan
```

Trecho chave - Otimização via GridSearchCV:

```
if param_spaces and nome_modelo in param_spaces:
grid = GridSearchCV(estimator=modelo, param_grid=param_spaces[nome_modelo],
cv=cv, scoring='f1_macro', n_jobs=-1)
grid.fit(X_train, y_train)
best = grid.best_estimator_
best_params_list.append({'Modelo': nome_modelo, 'Melhores Parâmetros': grid.best_params_})
```

Trecho chave - Plotagem comparativa (Curva ROC):

```
fpr, tpr, _ = roc_curve(y_test, y_prob)
fpr_opt, tpr_opt, _ = roc_curve(y_test, y_prob_opt)
axs[0, 0].plot(fpr, tpr, label=f"Base (AUC = {auc(fpr, tpr):.3f})", color=cor_0)
axs[0, 0].plot(fpr_opt, tpr_opt, label=f"Otimizado (AUC = {auc(fpr_opt, tpr_opt):.3f})", color=cor_1)
```

Em suma os processos realizados na função automatizam o ciclo completo de avaliação, produzindo métricas, diagnósticos e visualizações comparativas entre modelos base e otimizados.

D.3 Diagnóstico de overfitting:

'Função verificar_overfitting'

Esta função compara métricas de teste e validação cruzada, sinalizando possíveis casos de overfitting ou underfitting. Fornecendo assim um diagnóstico automatizado de estabilidade dos modelos, sendo fundamental para a tomada de decisão.

Trecho chave - Cálculo de diferença percentual relativa:

```
for met in metricas:
vt = t_row[met]
vc = c_row[f"Validação Cruzada ({met})"]
diff_rel = (vt - vc) / vc if vc != 0 else np.inf
res_modelo[f"delta_{met}"] = f"{100 * diff_rel:.1f}%"
```

Trecho chave - Diagnóstico baseado na média:

```
if media_diff > limite_diferenca:
res_modelo["Diagnóstico"] = "Overfitting Potencial"
elif media_diff < -limite_diferenca:
res_modelo["Diagnóstico"] = "Underfitting Potencial / Teste Ruim"
else:
res_modelo["Diagnóstico"] = "OK"
```

D.4 Comparação visual:

Função 'comparar_resultados_classificacao'

Esta função gera gráficos de barras comparando métricas de teste e validação cruzada, com pontos sobrepostos para facilitar a visualização.

Trecho chave - Transformação para formato longo:

```
for _, row in df_merged.iterrows():
for m in metrics_list:
teste_val = row[m]
cv_val = row[f"{m}_CV"]
diff_perc = (cv_val - teste_val) * 100
records.append({
'Modelo': row[model_col_name],
'Métrica': m,
'Teste': teste_val,
'CV': cv_val,
'Diferença (%)': diff_perc
})
```

Trecho chave - Plotagem com seaborn:

```
sns.barplot(data=df_comp, x='Métrica', y='Teste', hue='Modelo', palette=palette, ax=ax)
sns.pointplot(data=df_comp, x='Métrica', y='CV', hue='Modelo', markers='D',
linestyles='--', dodge=True, ax=ax, errorbar=None)
```