Revista ARACÊ
ISSN: 2358-2472

MINERAÇÃO DE DADOS EDUCACIONAIS APLICADA AOS MICRODADOS DO SAEB: PREDIÇÃO DE DESEMPENHO E FATORES ASSOCIADOS À PROFICIÊNCIA NO ENSINO MÉDIO EM RONDÔNIA

EDUCATIONAL DATA MINING APPLIED TO SAEB MICRODATA: PREDICTING PERFORMANCE AND FACTORS ASSOCIATED WITH PROFICIENCY IN HIGH SCHOOL IN RONDÔNIA

MINERÍA DE DATOS EDUCATIVOS APLICADA A MICRODATATOS SAEB: PREDICCIÓN DEL DESEMPEÑO Y FACTORES ASOCIADOS A LA COMPETENCIA EN LA EDUCACIÓN SECUNDARIA EN RONDÔNIA

https://doi.org/10.56238/arev7n9-157

Data de submissão: 13/08/2025 **Data de publicação:** 13/09/2025

José Claion Martins
Tecnólogo em Análise e Desenvolvimento de Sistemas
Instituição: Instituto Federal de Rondônia
Lattes: https://lattes.cnpq.br/9085093219166289
E-mail: joseclaionmartins@gmail.com

Leandro Ferrarezi Valiante
Mestre em Ciência da Computação
Instituição: Instituto Federal de Rondônia
Lattes: http://lattes.cnpq.br/3021868094199055
E-mail: leandro.valiante@ifro.edu.br

RESUMO

Este estudo investiga a aplicação de técnicas de Mineração de Dados Educacionais (MDE) aos microdados do Sistema de Avaliação da Educação Básica (SAEB), com foco nos estudantes do 3º e 4º anos do Ensino Médio do estado de Rondônia. O objetivo foi desenvolver modelos de predição de desempenho em Língua Portuguesa e Matemática e identificar variáveis socioeconômicas, escolares e individuais mais relevantes para a proficiência dos alunos. Foram utilizados algoritmos de regressão, incluindo Linear Regression, IBk (KNN), Random Forest, SMOreg e M5P, sendo este último o que apresentou melhor desempenho preditivo. A análise de relevância por meio do método ReliefF apontou fatores como tempo de estudo, sexo, tipo de escola, acesso a recursos tecnológicos e compreensão das aulas remotas como determinantes do desempenho. Os resultados evidenciam a importância das condições socioeconômicas e do contexto educacional para a aprendizagem, oferecendo subsídios para políticas públicas mais direcionadas e para estratégias pedagógicas voltadas à redução das desigualdades educacionais.

Palavras-chave: Desempenho Escolar. Educação. Ensino Médio. Mineração de Dados. SAEB.

ABSTRACT

This study investigates the application of Educational Data Mining (EDM) techniques to microdata from the Basic Education Assessment System (SAEB), focusing on third- and fourth-year high school students in the state of Rondônia. The objective was to develop performance prediction models in Portuguese and Mathematics and identify the most relevant socioeconomic, school-based, and

Revista ARACÊ
ISSN: 2358-2472

individual variables for student proficiency. Regression algorithms were used, including Linear Regression, IBk (KNN), Random Forest, SMOreg, and M5P, the latter presenting the best predictive performance. Relevance analysis using the ReliefF method identified factors such as study time, gender, school type, access to technological resources, and understanding of remote classes as determinants of performance. The results highlight the importance of socioeconomic conditions and the educational context for learning, providing support for more targeted public policies and pedagogical strategies aimed at reducing educational inequalities.

Keywords: School Performance. Education. High School. Data Mining. SAEB.

RESUMEN

Este estudio investiga la aplicación de técnicas de Minería de Datos Educativos (MDE) a microdatos del Sistema de Evaluación de la Educación Básica (SAEB), centrándose en estudiantes de tercer y cuarto año de secundaria en el estado de Rondônia. El objetivo fue desarrollar modelos de predicción del rendimiento en portugués y matemáticas e identificar las variables socioeconómicas, escolares e individuales más relevantes para el dominio del estudiante. Se utilizaron algoritmos de regresión, incluyendo Regresión Lineal, IBk (KNN), Bosque Aleatorio, SMOreg y M5P, siendo este último el que presentó el mejor rendimiento predictivo. El análisis de relevancia mediante el método ReliefF identificó factores como el tiempo de estudio, el género, el tipo de escuela, el acceso a recursos tecnológicos y la comprensión de las clases a distancia como determinantes del rendimiento. Los resultados destacan la importancia de las condiciones socioeconómicas y el contexto educativo para el aprendizaje, lo que respalda políticas públicas y estrategias pedagógicas más específicas destinadas a reducir las desigualdades educativas.

Palabras clave: Rendimiento Escolar. Educación. Secundaria. Minería de Datos. SAEB.

Revista ARACÊ
ISSN: 2358-2472

1 INTRODUÇÃO

A Mineração de Dados Educacionais (MDE) consolidou-se como um campo de investigação que aplica métodos de ciência de dados para compreender e aprimorar os processos de aprendizagem, com ênfase em tarefas como a predição de desempenho acadêmico e a descoberta de fatores associados ao aprendizado. Essas análises, frequentemente baseadas em técnicas de seleção de atributos e interpretabilidade, têm como finalidade oferecer suporte a decisões pedagógicas e de gestão (ROMERO; VENTURA, 2020; BAKER; INVENTADO, 2014).

No Brasil, esse tipo de abordagem encontra terreno fértil nos microdados do Sistema de Avaliação da Educação Básica (SAEB), que avalia periodicamente a qualidade da educação por meio de testes cognitivos e questionários contextuais aplicados a estudantes, professores e diretores. Esses instrumentos produzem indicadores comparáveis entre escolas, redes e regiões, sendo amplamente utilizados para orientar políticas públicas (INEP, 2020; INEP, 2021). Com base nesses microdados, torna-se possível estimar proficiências e mapear atributos relevantes, como condições socioeconômicas, formação docente, ambiente escolar e infraestrutura. O próprio Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), por exemplo, publica o Indicador de Nível Socioeconômico (INSE), derivado dos questionários contextuais, para contextualizar resultados e apoiar análises de equidade (INEP, 2023).

A literatura especializada mostra que modelos estatísticos e de aprendizado de máquina, como regressão, árvores de decisão e ensembles, são recorrentes em pesquisas de MDE, pois permitem prever notas e explicar variações de desempenho (ROMERO; VENTURA, 2024). No contexto brasileiro, estudos têm aplicado tais métodos aos dados do SAEB e do Índice de Desenvolvimento da Educação Básica (IDEB), revelando fatores associados às proficiências em Matemática e Língua Portuguesa (PINTO et al., 2020).

Apesar da relevância desses trabalhos, um dos principais desafios ainda enfrentados é identificar, de forma precisa, quais fatores exercem maior influência sobre o desempenho escolar. Embora o SAEB produza medidas robustas de proficiência em Língua Portuguesa e Matemática, esses resultados não podem ser analisados isoladamente. Pesquisas demonstram que o desempenho dos estudantes é fortemente condicionado por variáveis contextuais, como nível socioeconômico, escolaridade dos pais, infraestrutura escolar e formação docente (SOARES; ANDRADE, 2006; ALVES; SOARES, 2013). Nesse sentido, a simples comparação de médias entre regiões ou redes não é suficiente para explicar as desigualdades educacionais, sendo necessário adotar técnicas que permitam analisar simultaneamente múltiplas variáveis e suas interações.

Revista ARACÊ
ISSN: 2358-2472

A MDE apresenta ferramentas adequadas para lidar com a complexidade e a alta dimensionalidade dos microdados do SAEB. Por meio de algoritmos de regressão, classificação e seleção de atributos, é possível desenvolver modelos preditivos que não apenas estimam notas, mas também evidenciam quais variáveis possuem maior poder explicativo no desempenho dos estudantes (ROMERO; VENTURA, 2020; BAKER; INVENTADO, 2014). Estudos nacionais têm demonstrado a utilidade dessa abordagem, ao apontar, por exemplo, a relevância da trajetória escolar e do acesso a recursos pedagógicos para o desempenho em avaliações padronizadas (PINTO et al., 2020). Assim, identificar os fatores de maior impacto pode subsidiar políticas educacionais mais eficazes e estratégias pedagógicas direcionadas à redução das desigualdades.

A justificativa para o presente estudo reside, portanto, na necessidade de compreender de maneira mais aprofundada os fatores que explicam o desempenho dos estudantes brasileiros nas avaliações do SAEB. Embora os relatórios oficiais forneçam indicadores agregados, eles não permitem identificar com clareza a interação entre variáveis individuais, escolares e contextuais. A aplicação de técnicas de MDE possibilita explorar esses microdados em profundidade, permitindo não apenas prever proficiências, mas também identificar atributos mais relevantes, oferecendo subsídios para políticas públicas baseadas em evidências e práticas pedagógicas eficazes (ROMERO; VENTURA, 2020; ALVES; SOARES, 2013; INEP, 2021).

Diante desse cenário, o objetivo deste trabalho é aplicar técnicas de mineração de dados aos microdados do SAEB, visando desenvolver modelos de predição de desempenho e identificar atributos contextuais, escolares e individuais de maior relevância. Busca-se, assim, compreender padrões de desempenho em larga escala e oferecer suporte à formulação de políticas públicas e práticas pedagógicas orientadas por evidências, contribuindo para a melhoria da qualidade e da equidade na educação básica brasileira.

2 TRABALHOS RELACIONADOS

A Mineração de Dados Educacionais constitui um campo interdisciplinar voltado para a extração de padrões relevantes a partir de grandes bases de dados educacionais. Diferentemente da mineração de dados tradicional, a MDE considera a natureza hierárquica das informações educacionais, organizadas em múltiplos níveis como alunos, turmas e escolas, o que exige técnicas específicas de tratamento, modelagem e validação (BAKER; INVENTADO, 2014). Nesse processo, destaca-se a adoção do modelo de Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases* - KDD), que busca transformar dados brutos em conhecimento aplicável ao planejamento e à tomada de decisão educacional.

Revista ARACÊ
ISSN: 2358-2472

No contexto brasileiro, a disponibilização dos microdados pelo INEP, em especial os provenientes do SAEB, ampliou o potencial da MDE para a análise do desempenho estudantil em larga escala. De acordo com Fonseca e Namen (2016), tais microdados possibilitam a identificação de fatores associados às proficiências dos estudantes, considerando variáveis relacionadas ao perfil docente, à infraestrutura escolar e ao contexto socioeconômico das famílias.

A literatura nacional tem explorado, de maneira crescente, a relação entre atributos socioeconômicos e resultados acadêmicos. Brito Júnior et al. (2022), ao analisarem os microdados do SAEB em Pernambuco, identificaram correlações significativas entre o Indicador de Nível Socioeconômico (INSE) e o desempenho em Matemática de alunos do 9º (nono) ano do Ensino Fundamental. De modo semelhante, pesquisas realizadas em Sergipe e na Paraíba (FARIAS et al., 2020; FARIAS et al., 2023) também ressaltam que fatores externos à escola, como escolaridade dos pais, renda familiar e acesso a recursos educacionais, exercem impacto expressivo sobre os resultados no IDEB.

Do ponto de vista metodológico, diferentes técnicas têm sido aplicadas nos estudos em MDE, incluindo árvores de decisão, regras de associação e algoritmos de agrupamento. Essas abordagens permitem tanto a classificação de estudantes em grupos de desempenho quanto a identificação de padrões ocultos entre características socioeconômicas e resultados de aprendizagem (SOARES, 2016; ALMEIDA et al., 2019). A aplicação desses métodos aos microdados educacionais evidencia o potencial da MDE em revelar relações complexas que dificilmente seriam percebidas por análises estatísticas convencionais.

Apesar dos avanços observados na literatura, ainda existem lacunas significativas. Embora diversas investigações tenham sido realizadas em estados das regiões Nordeste e Sudeste, ainda são escassos os estudos que utilizam MDE para compreender o impacto dos atributos socioeconômicos sobre o desempenho educacional na região Norte. Em particular, o estado de Rondônia carece de análises sistemáticas fundamentadas nos microdados do SAEB, o que limita a compreensão dos fatores que influenciam o rendimento escolar dos estudantes locais.

Diante desse cenário, o presente trabalho busca contribuir para o preenchimento dessa lacuna, aplicando técnicas de mineração de dados aos microdados do SAEB com o objetivo de identificar quais variáveis socioeconômicas exercem maior influência sobre o desempenho em Matemática e Língua Portuguesa no estado de Rondônia.

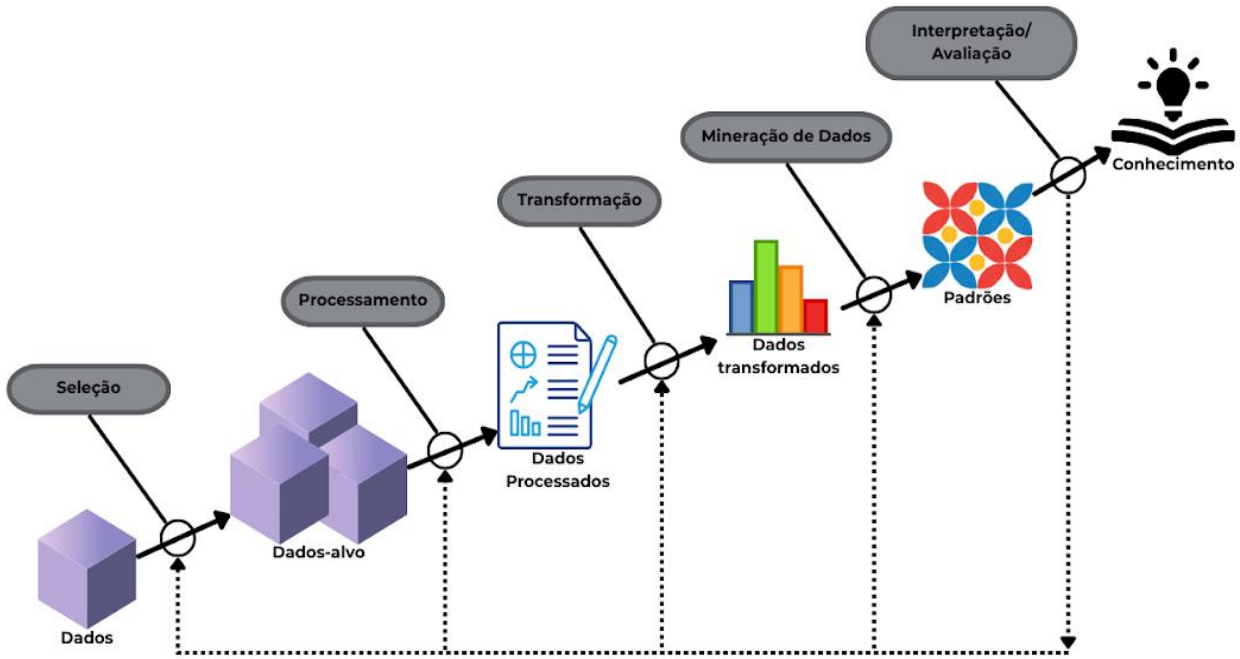
Revista ARACÊ
ISSN: 2358-2472

3 METODOLOGIA

O presente estudo caracteriza-se como uma pesquisa exploratória e quantitativa, fundamentada na aplicação de técnicas de MDE. Para a condução das análises, adotou-se como referência metodológica o processo de KDD, proposto por Fayyad et al. (1996), amplamente utilizado para a descoberta de padrões ocultos em grandes volumes de dados.

Conforme ilustrado na Figura 1, o modelo KDD organiza-se em cinco etapas principais: (i) seleção dos dados relevantes; (ii) pré-processamento para tratamento de inconsistências; (iii) transformação dos dados em formatos adequados à análise; (iv) mineração, com a aplicação de algoritmos e técnicas de exploração; e (v) interpretação e avaliação dos resultados obtidos.

Figura 1. Etapas do ciclo de KDD



Fonte: FAYYAD; PIATETSKY-SHAPIRO; SMYTH (1996, adaptado de FIGUEIREDO, 2010).

Ao adotar o modelo de KDD como guia metodológico, este trabalho assegura um fluxo estruturado de análise, desde a seleção e preparação dos dados até a interpretação dos resultados. Essa abordagem contribui para garantir maior confiabilidade às etapas investigativas e favorece a identificação de padrões significativos no contexto educacional estudado, alinhando-se aos objetivos propostos pela pesquisa.

3.1 CONJUNTO DE DADOS

A pesquisa utilizou os microdados do SAEB, disponibilizados pelo INEP, referentes ao ano de 2021. A base de dados utilizada tem como título “Microdados SAEB”, contendo 2.288.747 (dois

Revista ARACÊ
ISSN: 2358-2472

milhões, duzentos e oitenta e oito mil, setecentos e quarenta e sete registros), com 106 (cento e seis) atributos. O arquivo com os dados está disponível publicamente no site do Governo Federal (INEP, 2021).

Para este estudo, foram selecionados os registros dos estudantes do 3º e 4º anos do Ensino Médio, especificamente nas disciplinas de Matemática e Língua Portuguesa. Além do índice de proficiência, consideraram-se variáveis contextuais de diferentes naturezas (socioeconômicas, escolares e individuais), conforme disponíveis no conjunto de dados original.

Com o objetivo de facilitar a organização e a análise, os atributos foram agrupados em sete categorias principais, apresentadas no Quadro 1.

Quadro 1. Agrupamento dos atributos contidos no conjunto de dados original

GRUPO	DESCRIÇÃO
Identificação e Contexto Escolar	Reúne informações administrativas e geográficas sobre os estudantes e suas escolas, incluindo códigos de região, município, série, turma, localização (urbana/rural) e dependência administrativa.
Provas	Inclui variáveis referentes à aplicação das provas de Língua Portuguesa e Matemática.
Proficiência e Pesos	Abrange medidas de desempenho, como proficiência estimada em Língua Portuguesa e Matemática, além de pesos amostrais utilizados na análise estatística do SAEB.
Dados Pessoais e Familiares	Contém informações fornecidas pelos estudantes sobre características individuais (sexo, idade, cor/raça) e aspectos familiares (escolaridade dos pais, ocupação, renda).
Moradia e Bens Materiais	Reúne dados sobre condições habitacionais e posse de bens, como número de cômodos, acesso a saneamento, internet, computador, celular e outros recursos domésticos.
Transporte e Histórico Escolar	Inclui informações sobre o deslocamento até a escola, trajetória educacional do aluno, repetência, abandono, frequência e apoio escolar recebido.
Ensino Remoto e Pandemia	Abrange variáveis específicas sobre o período da pandemia da Covid-19, como acesso a atividades remotas, dificuldades de aprendizagem e estratégias adotadas pela escola.

Fonte: Elaborado pelos autores

No Quadro 2 são apresentados e descritos os atributos referentes à identificação e ao contexto escolar dos estudantes.

Quadro 2. Atributos do grupo Identificação e Contexto Escolar

ATRIBUTO	DESCRIÇÃO	VALORES
ID_SAEB	Ano de aplicação do Saeb	2021
ID_REGIAO	Código da Região	1 a 5, conforme região
ID_UF	Código da Unidade da Federação (UF)	Número indicando UF
ID_MUNICIPIO	Máscaras dos Códigos de Municípios	Códigos fictícios
ID_AREA	Área	1 - Capital; 2 - Interior
ID_ESCOLA	Máscaras dos Códigos de Escola	Códigos fictícios
IN_PUBLICA	Indica se a escola é pública ou não	0 - Privada; 1 - Pública
ID_LOCALIZACAO	Localização	1 - Urbana; 2 - Rural
ID_TURMA	Código da turma no SAEB	Valor numérico
ID_SERIE	Ano Escolar	3ª/4ª série
ID_ALUNO	Código do aluno no Saeb	Valor numérico
IN_SITUACAO_CENSO	Indicador de consistência entre os dados da aplicação do SAEB 2021 com o Censo da Educação Básica 2021 finalizado	0 - Não consistente 1 - Consistente

Revista ARACÊ
ISSN: 2358-2472

IN_AMOSTRA	Indicador de participação da amostra	0 – Não ; 1 – Sim
ESTRATO	Descrição dos estratos	Características da escola
IN_PREENCHIMENTO_QUESTIONARIO	Indicador de preenchimento do questionário	0 - Não preenchido 1 - Preenchido
IN_INSE	Indicador para cálculo do INSE	0 – Não ; 1 – Sim

Fonte: elaborado pelos autores

O Quadro 3 reúne os atributos relacionados às provas de Língua Portuguesa e Matemática.

Quadro 3. Atributos do grupo "Provas"

ATRIBUTO	DESCRIÇÃO	VALORES
IN_PREENCHIMENTO_LP IN_PREENCHIMENTO_MT	Indicador de preenchimento da prova	0 - Prova não preenchida 1 - Prova preenchida
IN_PRESENCA_LP IN_PRESENCA_MT	Indicador de presença na prova	0 – Ausente 1 – Presente
ID_CADERNO_LP ID_CADERNO_MT	Número do caderno de prova	Prova Regular (1 a 21) Macrotipo 18 (22)
ID_BLOCO_1_LP	Identificador do Bloco 1 de Língua Portuguesa	Valores de 1 a 7
ID_BLOCO_2_LP	Identificador do Bloco 2 de Língua Portuguesa	Valores de 1 a 7
ID_BLOCO_1_MT	Identificador do Bloco 1 de Matemática	Valores de 1 a 7
ID_BLOCO_2_MT	Identificador do Bloco 2 de Matemática	Valores de 1 a 7
TX_RESP_BLOCO1_LP TX_RESP_BLOCO2_LP TX_RESP_BLOCO1_MT TX_RESP_BLOCO2_MT	Resposta do aluno ao Bloco e Disciplina correspondente	A, B, C, D, (branco), * (nulo)
IN_PROFICIENCIA_LP IN_PROFICIENCIA_MT	Indicador para cálculo da proficiência (no mínimo três itens respondidos no caderno de prova)	0 - Não 1 - Sim

Fonte: elaborado pelos autores

No Quadro 4 estão dispostos os atributos referentes à proficiência e aos pesos amostrais.

Quadro 4. Atributos do grupo Proficiência e Pesos

ATRIBUTO	DESCRIÇÃO	VALORES
PESO_ALUNO_LP PESO_ALUNO_MT	Pesos amostrais dos alunos, derivados do Censo Escolar 2021, utilizados para expandir os resultados das provas para a população-alvo.	Valor numérico
PROFICIENCIA_LP PROFICIENCIA_MT	Proficiência do aluno calculada na escala única do SAEB, com média = 0 e desvio = 1 na população de referência	Valor numérico
ERRO_PADRAO_LP ERRO_PADRAO_MT	Erro padrão da proficiência	Valor numérico
PROFICIENCIA_LP_SAEB PROFICIENCIA_MT_SAEB	Proficiência transformada na escala única do SAEB, com média = 250, desvio = 50 (do SAEB/97)	Valor numérico
ERRO_PADRAO_LP_SAEB ERRO_PADRAO_MT_SAEB	Erro padrão da proficiência transformada	Valor numérico
INSE_ALUNO	Indicador para cálculo do INSE	0 - Não ; 1 - Sim
NU_TIPO_NIVEL_INSE	Classificação do Indicador de Nível Socioeconômico em 8 Grupos	Entre 1 e 8
PESO_ALUNO_INSE	Peso do Aluno para cálculo do INSE 2021	Valor numérico

Fonte: elaborado pelos autores

Revista ARACÊ
ISSN: 2358-2472

O Quadro 5 apresenta os atributos relativos às características pessoais e familiares dos estudantes.

Quadro 5. Atributos do grupo Dados Pessoais e Familiares

ATRIBUTO	DESCRIÇÃO	VALORES
TX_RESP_Q01	Qual é o seu sexo?	* Nulo, Branco, A Masc. B Fem.
TX_RESP_Q02	Qual é a sua idade?	* Nulo, Branco, ou valores de A à F representando intervalo de idade
TX_RESP_Q03	Qual língua que seus pais falam com mais frequência em casa?	* Nulo, Branco, A Português. B Espanhol. C Outra língua
TX_RESP_Q04	Qual é a sua cor ou raça?	* Nulo, Branco, A Branca. B Preta. C Parda. D Amarela. E Indígena. F Não quero declarar
TX_RESP_Q05	Você possui algum tipo de necessidade especial?	* Nulo, Branco A Sim. B Não.
TX_RESP_Q06a	Normalmente, quem mora na sua casa? - Mãe ou madrasta.	* Nulo, Branco, A Não. B Sim
TX_RESP_Q06b	Normalmente, quem mora na sua casa? - Pai ou padrasto.	* Nulo, Branco, A Não. B Sim
TX_RESP_Q06c	Normalmente, quem mora na sua casa? - Irmão(s) ou irmã(s).	* Nulo, Branco, A Não. B Sim.
TX_RESP_Q06d	Normalmente, quem mora na sua casa? - Avô ou avó.	* Nulo, Branco, A Não. B Sim.
TX_RESP_Q06e	Normalmente, quem mora na sua casa? - Outros (tios, primos etc.).	* Nulo, Branco, A Não. B Sim.
TX_RESP_Q07	Qual é a maior escolaridade da sua mãe (ou mulher responsável por você)?	* Nulo, Branco, ou valores de A à F representando a escolaridade
TX_RESP_Q08	Qual é a maior escolaridade de seu pai (ou homem responsável por você)?	* Nulo, Branco, ou valores de A à F representando a escolaridade
TX_RESP_Q09a	Com que frequência seus pais ou responsáveis costumam: - Ler em casa	* Nulo, Branco, ou valores de A à C representando a frequência.
TX_RESP_Q09b	Com que frequência seus pais ou responsáveis costumam: - Conversar com você sobre o que acontece na escola.	* Nulo, Branco, ou valores de A à C representando a frequência.
TX_RESP_Q09c	Com que frequência seus pais ou responsáveis costumam: - Incentivar você a estudar.	* Nulo, Branco, ou valores de A à C representando a frequência.
TX_RESP_Q09d	Com que frequência seus pais ou responsáveis costumam: - Incentivar você a fazer a tarefa de casa	* Nulo, Branco, ou valores de A à C representando a frequência.
TX_RESP_Q09e	Com que frequência seus pais ou responsáveis costumam: - Incentivar você a comparecer às aulas.	* Nulo, Branco, ou valores de A à C representando a frequência.
TX_RESP_Q09f	Com que frequência seus pais ou responsáveis costumam: - Ir às reuniões de pais na escola.	* Nulo, Branco, ou valores de A à C representando a frequência.

Fonte: elaborado pelos autores

No Quadro 6, a seguir, são listados os atributos associados às condições de moradia e aos bens materiais.

Quadro 6. Atributos do grupo Moradia e Bens

ATRIBUTO	DESCRIÇÃO	VALORES
TX_RESP_Q10a	Na rua em que você mora tem: - Asfalto ou calçamento.	* Nulo, Branco, A Não. B Sim.
TX_RESP_Q10b	* Nulo . Branco A Não. B Sim.	* Nulo, Branco, A Não. B Sim.
TX_RESP_Q10c	Na rua em que você mora tem: - Iluminação.	* Nulo, Branco, A Não. B Sim.
TX_RESP_Q11a	Dos itens relacionados abaixo, quantos existem na sua casa? - Geladeira.	* Nulo, Branco, A Nenhum. B 1. C 2. D 3 ou mais.
TX_RESP_Q11b	Dos itens relacionados abaixo, quantos existem na sua casa? - Tablet.	* Nulo, Branco, A Nenhum. B 1. C 2. D 3 ou mais.
TX_RESP_Q11c	Dos itens relacionados abaixo, quantos existem na sua casa? - Computador (ou notebook).	* Nulo, Branco, A Nenhum. B 1. C 2. D 3 ou mais.

Revista ARACÊ
ISSN: 2358-2472

TX_RESP_Q11d	Dos itens relacionados abaixo, quantos existem na sua casa? - Quartos para dormir.	* Nulo, Branco, A Nenhum. B 1. C 2. D 3 ou mais.
TX_RESP_Q11e	Dos itens relacionados abaixo, quantos existem na sua casa? - Televisão.	* Nulo, Branco,A Nenhum. B 1. C 2. D 3 ou mais.
TX_RESP_Q11f	Dos itens relacionados abaixo, quantos existem na sua casa? - Banheiro.	* Nulo, Branco, A Nenhum. B 1. C 2. D 3 ou mais.
TX_RESP_Q11g	Dos itens relacionados abaixo, quantos existem na sua casa? - Carro.	* Nulo, Branco, A Nenhum. B 1. C 2. D 3 ou mais.
TX_RESP_Q11h	Dos itens relacionados abaixo, quantos existem na sua casa? - Celular com internet (smartphone).	* Nulo, Branco, A Nenhum. B 1. C 2. D 3 ou mais.
TX_RESP_Q12a	Na sua casa tem: - Tv por internet (Netflix, GloboPlay, etc.).	* Nulo, Branco, A Não. B Sim.
TX_RESP_Q12b	Na sua casa tem: - Rede Wi-Fi.	* Nulo, Branco, A Não. B Sim.
TX_RESP_Q12c	Na sua casa tem: - Um quarto só seu.	* Nulo, Branco, A Não. B Sim.
TX_RESP_Q12d	Na sua casa tem: - Mesa para estudar.	* Nulo, Branco, A Não. B Sim.
TX_RESP_Q12e	Na sua casa tem: - Forno de microondas.	* Nulo, Branco, A Não. B Sim.
TX_RESP_Q12f	Na sua casa tem: - Aspirador de pó	* Nulo, Branco, A Não. B Sim.
TX_RESP_Q12g	Na sua casa tem: - Máquina de lavar roupa.	* Nulo, Branco, A Não. B Sim.
TX_RESP_Q12h	Na sua casa tem: - Freezer (independente ou segunda porta da geladeira).	* Nulo, Branco, A Não. B Sim.
TX_RESP_Q12i	Na sua casa tem: - Garagem.	* Nulo, Branco, A Não. B Sim.

Fonte: elaborado pelos autores

O Quadro 7 reúne os atributos ligados ao transporte e ao histórico escolar.

Quadro 7. Atributos do grupo Transporte e Histórico Escolar

ATRIBUTO	DESCRIÇÃO	VALORES
TX_RESP_Q13	Quanto tempo você demora para chegar à sua escola?	* Nulo, Branco, A Menos de 30 minutos. B Entre 30 minutos e uma hora. C Mais de uma hora.
TX_RESP_Q14	Considerando a maior distância percorrida, normalmente de que forma você chega à sua escola?	* Nulo, Branco, valores de A à G representando o tipo de transporte
TX_RESP_Q15	Você se utiliza de transporte escolar, ou passe escolar, para ir à escola?	* Nulo, Branco, A Não. B Sim.
TX_RESP_Q16	Com que idade você entrou na escola?	* Nulo, Branco, ou valores de A à D.
TX_RESP_Q17	A partir do primeiro ano do ensino fundamental, em que tipo de escola você estudou?	* Nulo, Branco, A: pública. B: particular. C: misto.
TX_RESP_Q18	Você já foi reprovado(a)?	* Nulo, Branco, A Não. B Sim, uma vez. C Sim, duas vezes ou mais.
TX_RESP_Q19	Alguma vez você abandonou a escola deixando de frequentá-la até o final do ano escolar?	* Nulo, Branco, A Nunca. B Sim, uma vez. C Sim, duas vezes ou mais
TX_RESP_Q20a	Fora da escola em dias de aula, quanto tempo você usa para: - Estudar (lição de casa, trabalhos escolares, etc.).	* Nulo, Branco, ou valores de A à D representando o tempo.
TX_RESP_Q20b	Fora da escola em dias de aula, quanto tempo você usa para: - Fazer cursos.	* Nulo, Branco, ou valores de A à D representando o tempo.
TX_RESP_Q20c	Fora da escola em dias de aula, quanto tempo você usa para: - Trabalhar em casa (lavar louça, limpar quintal, ...).	* Nulo, Branco, ou valores de A à D representando o tempo.
TX_RESP_Q20d	Fora da escola em dias de aula, quanto tempo você usa para: - Trabalhar fora de casa (recebendo ou não um salário).	* Nulo, Branco, ou valores de A à D representando o tempo.
TX_RESP_Q20e	Fora da escola em dias de aula, quanto tempo você usa para: - Lazer (TV, internet, brincar, música etc.).	* Nulo, Branco, ou valores de A à D representando o tempo.
TX_RESP_Q21	Quando terminar o Ensino Médio você pretende:	* Nulo, Branco, ou valores de A à D.
TX_RESP_Q22	Você concluiu o Ensino Fundamental na Educação de Jovens e Adultos (EJA), antigo supletivo?	* Nulo, Branco, A Não. B Sim.

Fonte: elaborado pelos autores

Revista ARACÊ
ISSN: 2358-2472

O Quadro 8 a seguir, apresenta os atributos referentes ao ensino remoto e ao período da pandemia da Covid-19.

Quadro 8. Atributos do grupo Ensino Remoto e Pandemia

ATRIBUTO	DESCRIÇÃO	VALORES
TX_RESP_Q23a	Durante a pandemia, indique a frequência com que os seguintes fatos ocorreram: - Eu possuía equipamento adequado para acompanhar o ensino remoto.	* Nulo, Branco, ou valor de A à D indicando a frequência.
TX_RESP_Q23b	Durante a pandemia, indique a frequência com que os seguintes fatos ocorreram: - Tive conexão de internet para acesso às aulas remotas.	* Nulo, Branco, ou valor de A à D indicando a frequência.
TX_RESP_Q23c	Durante a pandemia, indique a frequência com que os seguintes fatos ocorreram: - Tive facilidade em usar os programas de comunicação.	* Nulo, Branco, ou valor de A à D indicando a frequência.
TX_RESP_Q23d	Durante a pandemia, indique a frequência com que os seguintes fatos ocorreram: - Recebi material impresso da escola.	* Nulo, Branco, ou valor de A à D indicando a frequência.
TX_RESP_Q23e	Durante a pandemia, indique a frequência com que os seguintes fatos ocorreram: - Os professores me auxiliaram a entender o conteúdo.	* Nulo, Branco, ou valor de A à D indicando a frequência.
TX_RESP_Q23f	Durante a pandemia, indique a frequência com que os seguintes fatos ocorreram: - Eu compreendi o conteúdo das aulas remotas.	* Nulo, Branco, ou valor de A à D indicando a frequência.
TX_RESP_Q23g	Durante a pandemia, indique a frequência com que os seguintes fatos ocorreram: - Em casa havia um lugar tranquilo para eu assistir às aulas.	* Nulo, Branco, ou valor de A à D indicando a frequência.
TX_RESP_Q23h	Durante a pandemia, indique a frequência com que os seguintes fatos ocorreram: - Meus familiares apoiaram o meu estudo.	* Nulo, Branco, ou valor de A à D indicando a frequência.
TX_RESP_Q23i	Durante a pandemia, indique a frequência com que os seguintes fatos ocorreram: - Meus colegas me apoiaram durante o ensino remoto.	* Nulo, Branco, ou valor de A à D indicando a frequência.

Fonte: elaborado pelos autores

Os quadros apresentados anteriormente fornecem uma visão detalhada dos atributos utilizados neste estudo, organizados por grupos temáticos que permitem compreender a abrangência e diversidade das informações contidas nos microdados do SAEB. Essa estruturação facilita a análise estatística e a aplicação das técnicas de MDE, garantindo maior clareza na exploração dos padrões observados.

3.2 TRATAMENTO DOS DADOS

Antes da aplicação dos métodos de análise, os dados passaram por um rigoroso processo de tratamento. Inicialmente, a base original foi dividida em duas versões distintas, correspondentes às disciplinas de Matemática (MT) e Língua Portuguesa (LP), permitindo análises específicas para cada área do conhecimento.

Durante o tratamento, foram identificadas e removidas inconsistências, registros com questionário socioeconômico em branco e casos com informações ausentes ou incompletas em variáveis essenciais ao estudo. Além disso, atributos irrelevantes ou redundantes foram excluídos, mantendo-se apenas aqueles que contribuem efetivamente para a análise da proficiência e para a compreensão das variáveis contextuais dos estudantes.

Revista ARACÊ
ISSN: 2358-2472

A seguir, apresenta-se a descrição do tratamento aplicado aos dados em duas etapas: alterações realizadas antes da divisão das bases e alterações realizadas após a divisão, quando foram criadas as versões específicas para Matemática e Língua Portuguesa.

Na primeira etapa de tratamento dos dados, detalham-se os procedimentos gerais de padronização, correção de inconsistências e exclusão de registros incompletos. Na segunda etapa, para cada versão da base, é apresentado o bloco de tratamento específico e, em seguida, um quadro com os atributos removidos e os respectivos motivos, garantindo a qualidade e confiabilidade das análises subsequentes.

Modificações gerais aplicadas à base de dados:

- Substituição de todos os pontos e vírgulas (‘;’) por vírgulas (‘,’), para facilitar o manuseio da base em formato CSV;
- Seleção dos registros em que ID_UF = 11, focando exclusivamente nos estudantes de Rondônia;
- Remoção dos registros sem questionário socioeconômico preenchido: garantiu consistência na análise de fatores contextuais; e
- Remoção dos registros com inconsistências em relação ao Censo da Educação Básica: assegurou coerência com os dados oficiais.

Atributos removidos:

- ID_SAEB, ID_TURMA, ID_ALUNO, ID_REGIAO, ID_UF, ID_MUNICIPIO, ID_ESCOLA: atributos irrelevantes ou fictícios, sem utilidade analítica para o experimento.
- ESTRATO, IN_SITUACAO_CENSO, IN_PREENCHIMENTO_QUESTIONARIO: atributos irrelevantes após a aplicação dos filtros; não contribuem para a proposta de análise.
- IN_INSE, INSE_ALUNO, NU_TIPO_NIVEL_ALUNO, PESO_ALUNO_INSE: valores adquiridos pós-preenchimento, sem relevância analítica para o experimento.
- ID_CADERNO_LP, ID_BLOCO_1_LP, ID_BLOCO_2_LP, ID_CADERNO_MT, ID_BLOCO_1_MT, ID_BLOCO_2_MT: representam apenas os códigos técnicos da prova.
- TX_RESP_BLOCO_1_LP, TX_RESP_BLOCO_2_LP, TX_RESP_BLOCO_1_MT, TX_RESP_BLOCO_2_MT: respostas individuais, sem relevância analítica para o experimento.

Revista ARACÊ
ISSN: 2358-2472

Após a etapa de modificações e remoções gerais, procedeu-se à adaptação específica das duas versões da base de dados, correspondentes às disciplinas de Matemática (MT) e Língua Portuguesa (LP). Para cada versão, foram aplicados tratamentos adicionais e removidos atributos considerados irrelevantes ou redundantes para a respectiva análise.

Para os dados relacionados à disciplina LP, foram selecionados apenas os registros com **IN_PRESENCA_LP = 1** e **IN_PREENCHIMENTO_LP = 1**, posteriormente sendo removidos os registros com **IN_PROFICIENCIA_LP = 0**. Adicionalmente, foram removidos os seguintes atributos:

- Indicadores e variáveis de MT: IN_PREENCHIMENTO_MT, IN_PRESENCA_MT, IN_PROFICIENCIA_MT, PESO_ALUNO_MT, PROFICIENCIA_MT, PROFICIENCIA_MT_SAEB, ERRO_PADRAO_MT, ERRO_PADRAO_MT_SAEB.
- PROFICIENCIA_LP, ERRO_PADRAO_LP, ERRO_PADRAO_LP_SAEB, PESO_ALUNO_LP. Estes valores são "pós-resultados" ou óbvios em relação ao alvo, o que atrapalharia o desempenho dos algoritmos.
- IN_PREENCHIMENTO_LP, IN_PROFICIENCIA_LP, IN_PRESENCA_LP: indicadores usados para filtrar os registros, inúteis após a aplicação do filtro.

Para os dados relacionados à disciplina MT, foram selecionados apenas os registros com **IN_PRESENCA_MT = 1** e **IN_PREENCHIMENTO_MT = 1**, posteriormente sendo removidos os registros com **IN_PROFICIENCIA_MT = 0**. Foram removidos os seguintes atributos:

- Indicadores e variáveis de Língua Portuguesa: IN_PREENCHIMENTO_LP, IN_PRESENCA_LP, IN_PROFICIENCIA_LP, PESO_ALUNO_LP, PROFICIENCIA_LP, PROFICIENCIA_LP_SAEB, ERRO_PADRAO_LP, ERRO_PADRAO_LP_SAEB.
- Colunas pós-cálculo e alvo de MT: PROFICIENCIA_MT, ERRO_PADRAO_MT, ERRO_PADRAO_MT_SAEB. Valores pós-resultados ou óbvios em relação ao alvo, sem utilidade para análise.
- Indicadores de MT redundantes após filtros: IN_PREENCHIMENTO_MT, IN_PROFICIENCIA_MT, IN_PRESENCA_MT, PESO_ALUNO_MT. Irrelevantes após filtragem.

No presente estudo, diferentes algoritmos de regressão foram empregados com o objetivo de avaliar sua capacidade preditiva sobre os microdados do SAEB, entre eles Linear Regression (LR),

Revista ARACÊ
ISSN: 2358-2472

IBk (KNN), Random Forest (RF), SMOreg e M5P. A escolha desses algoritmos justifica-se por sua ampla utilização em pesquisas de MDE, dada a diversidade de abordagens que representam — desde modelos lineares de fácil interpretação até métodos baseados em árvores de decisão e aprendizado por vizinhança, capazes de capturar relações mais complexas entre variáveis.

Para mensurar o desempenho de cada modelo, adotou-se como métrica a "Raiz do Erro Quadrático Médio" (*Root Mean Squared Error* - RMSE), que permite avaliar a precisão das predições ao considerar a magnitude média dos erros em relação aos valores observados. Essa métrica foi escolhida por ser amplamente reconhecida em estudos de modelagem preditiva, fornecendo uma medida robusta e comparável entre algoritmos distintos, o que possibilita identificar não apenas o modelo mais eficaz, mas também aquele que apresenta maior potencial de aplicação prática no contexto educacional.

A avaliação preditiva foi conduzida por validação cruzada *k-fold* ($k = 10$), calculando-se o RMSE médio e o desvio padrão. Para a execução dos algoritmos (LR, IBk, RF, SMOreg, M5P) utilizou-se o software Weka, com parâmetros de configuração descritos nas legendas das figuras correspondentes, no próximo capítulo. Essa estratégia garante comparabilidade entre modelos e mitigação de sobreajuste.

Concluídas as fases de seleção, processamento e transformação, partiu-se então para o experimento de mineração de dados, que é descrito no capítulo seguinte.

4 RESULTADOS E DISCUSSÕES

Para compreender a efetividade dos métodos aplicados e a relevância dos atributos analisados, este capítulo apresenta os resultados obtidos a partir da aplicação dos algoritmos de regressão sobre os microdados do SAEB, organizados nas bases de Língua Portuguesa e Matemática. Inicialmente, são expostos os desempenhos comparativos dos modelos, destacando métricas de precisão e erros associados. Em seguida, procede-se à análise da importância relativa das variáveis, buscando identificar os fatores que mais influenciam a proficiência dos estudantes. Por fim, os achados são discutidos à luz da literatura, evidenciando convergências e divergências em relação a estudos anteriores e apontando implicações práticas para o contexto educacional de Rondônia e para a formulação de políticas públicas baseadas em evidências.

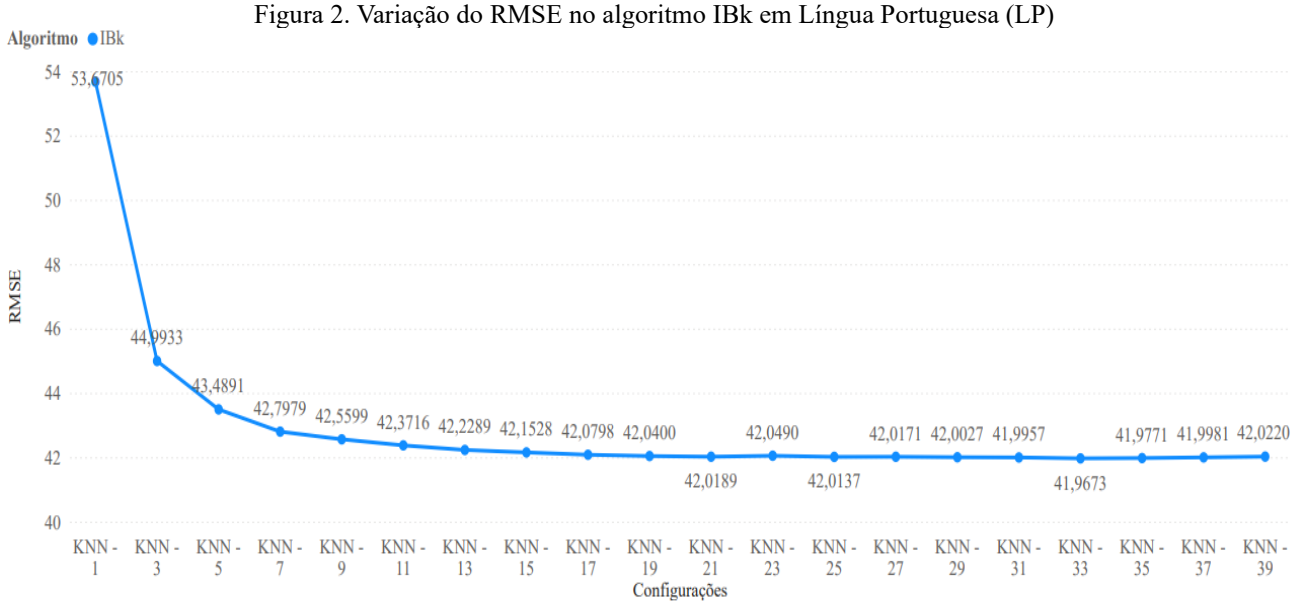
4.1 DESEMPENHO DOS ALGORITMOS

Para avaliar a capacidade preditiva dos modelos, foram aplicados diferentes algoritmos de regressão às bases ajustadas de Matemática e Língua Portuguesa. Inicialmente, o Linear Regression

Revista ARACÊ
ISSN: 2358-2472

foi utilizado como modelo de referência, fornecendo um ponto de comparação para os demais métodos. Os valores de RMSE obtidos – 39,09 em Língua Portuguesa e 40,40 em Matemática – fornecem um ponto de partida para avaliar o impacto de modelos mais complexos e da seleção de atributos. Apesar de sua simplicidade, o LR apresentou o melhor desempenho em comparação com IBk, Random Forest e SMOreg, evidenciando que relações lineares entre os atributos já explicam boa parte da variação nas proficiências dos alunos.

Figura 2. Variação do RMSE no algoritmo IBk em Língua Portuguesa (LP)

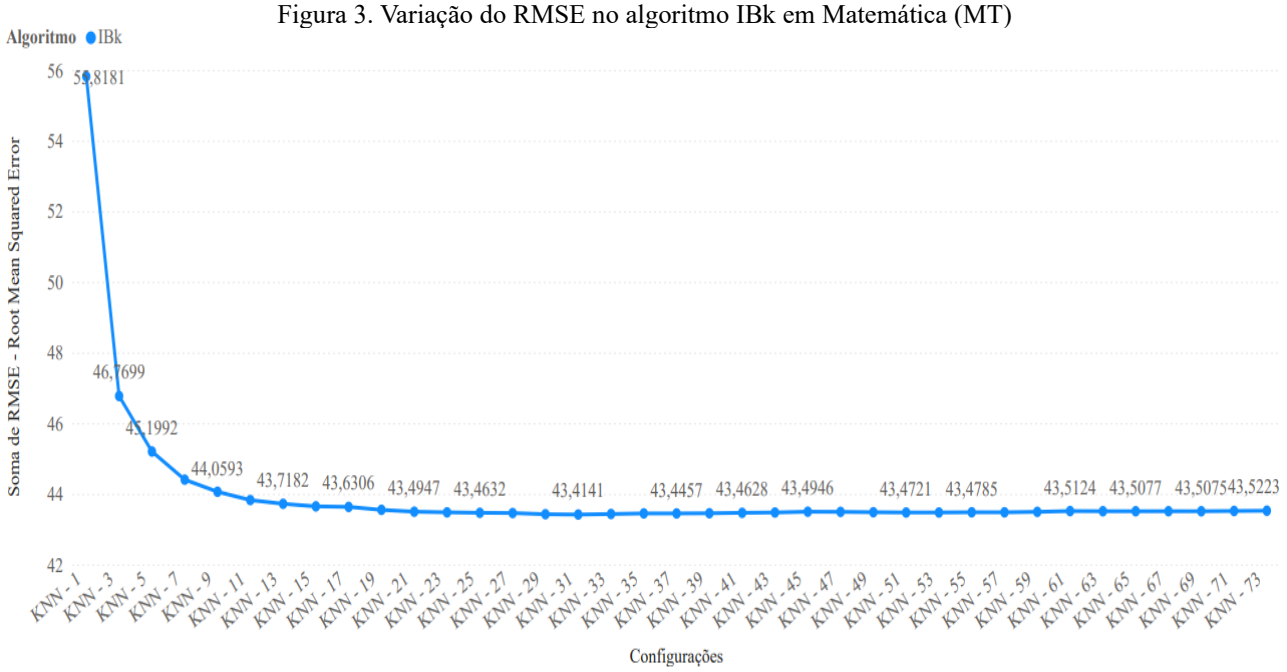


Fonte: Elaborado pelos autores

O algoritmo IBk foi avaliado com valores de K variando de 1 a 39 na versão de Língua Portuguesa e de 1 a 73 na versão de Matemática, em incrementos de 2, com o objetivo de analisar a influência do número de vizinhos no desempenho preditivo do modelo. Os resultados para o conjunto de dados da disciplina de LP estão demonstrados no gráfico da Figura 2, enquanto os resultados para MT estão na Figura 3.

Revista ARACÊ
ISSN: 2358-2472

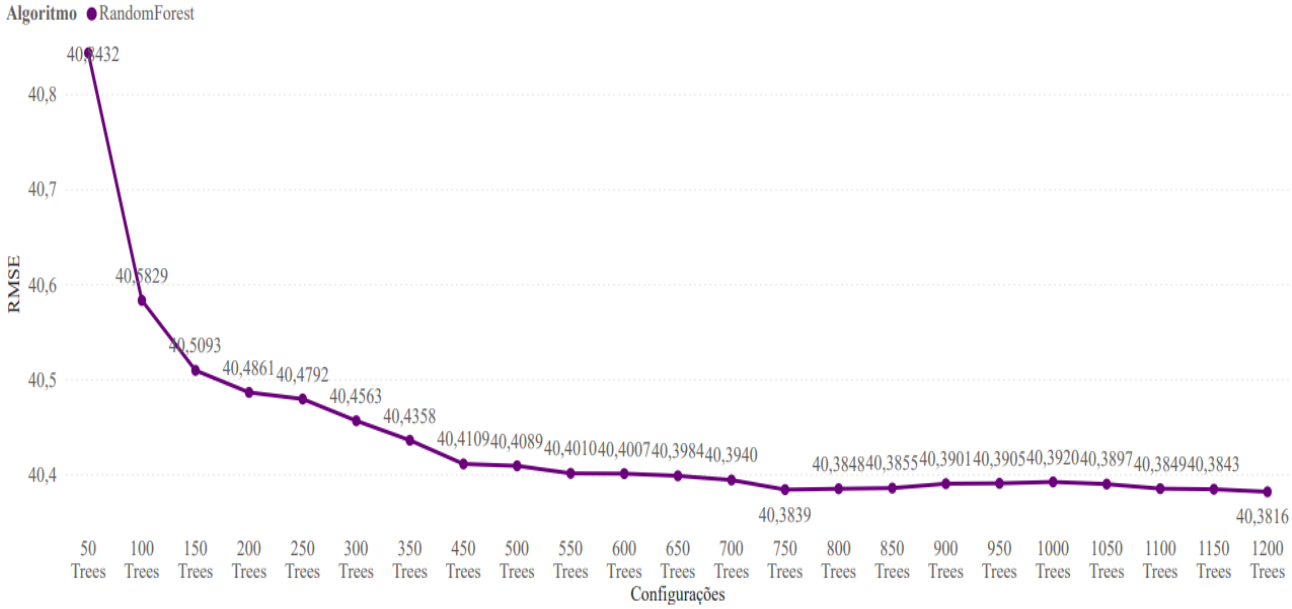
Figura 3. Variação do RMSE no algoritmo IBk em Matemática (MT)



Fonte: Elaborado pelos autores

O algoritmo IBk demonstrou sensibilidade ao número de vizinhos K, com diferenças de desempenho entre as versões. A melhor configuração foi K=33 para Língua Portuguesa (RMSE = 41,9673) e K=31 para Matemática (RMSE = 43,4141). Apesar de ajustar o número de vizinhos, o IBk apresentou resultados piores que o Linear Regression em ambas as disciplinas, indicando que a abordagem baseada em vizinhos não conseguiu capturar de forma eficiente as relações entre os atributos e as proficiências dos alunos neste estudo.

Figura 4. Variação do RMSE no algoritmo Random Forest com os dados de LP

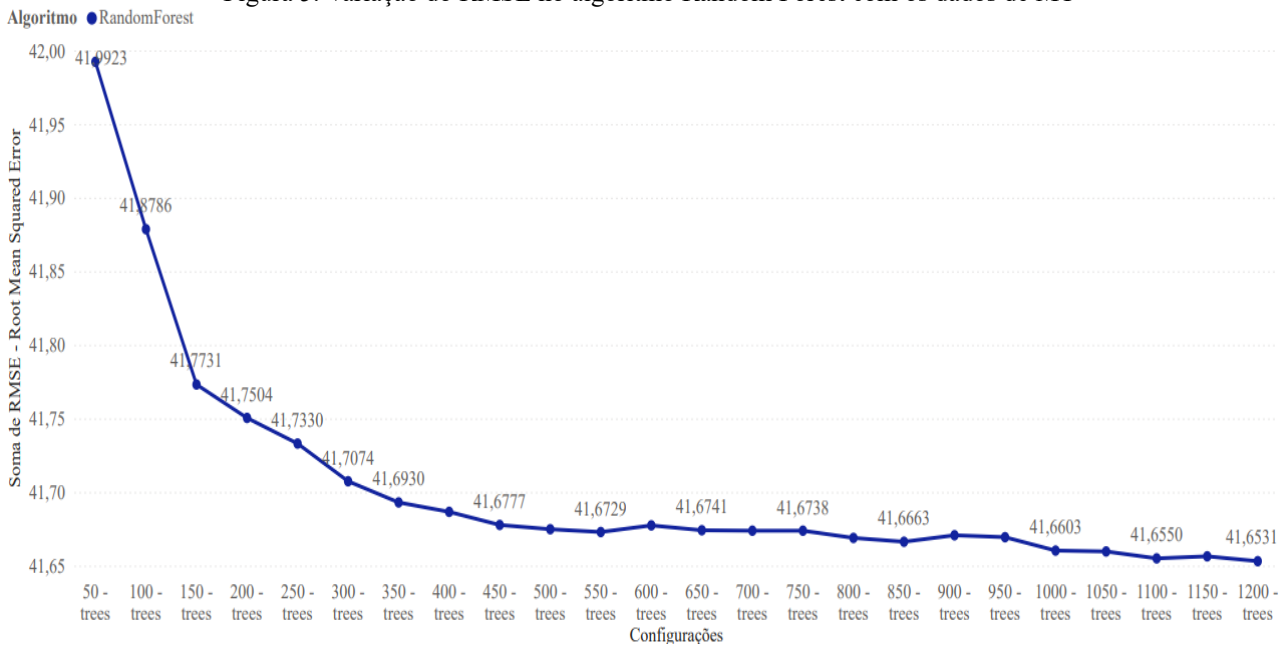


Fonte: Elaborado pelos autores

Revista ARACÊ
ISSN: 2358-2472

O algoritmo Random Forest foi executado com diferentes números de árvores, variando de 50 a 1200, em incrementos de 50, permitindo identificar a configuração mais eficiente com precisão.

Figura 5. Variação do RMSE no algoritmo Random Forest com os dados de MT



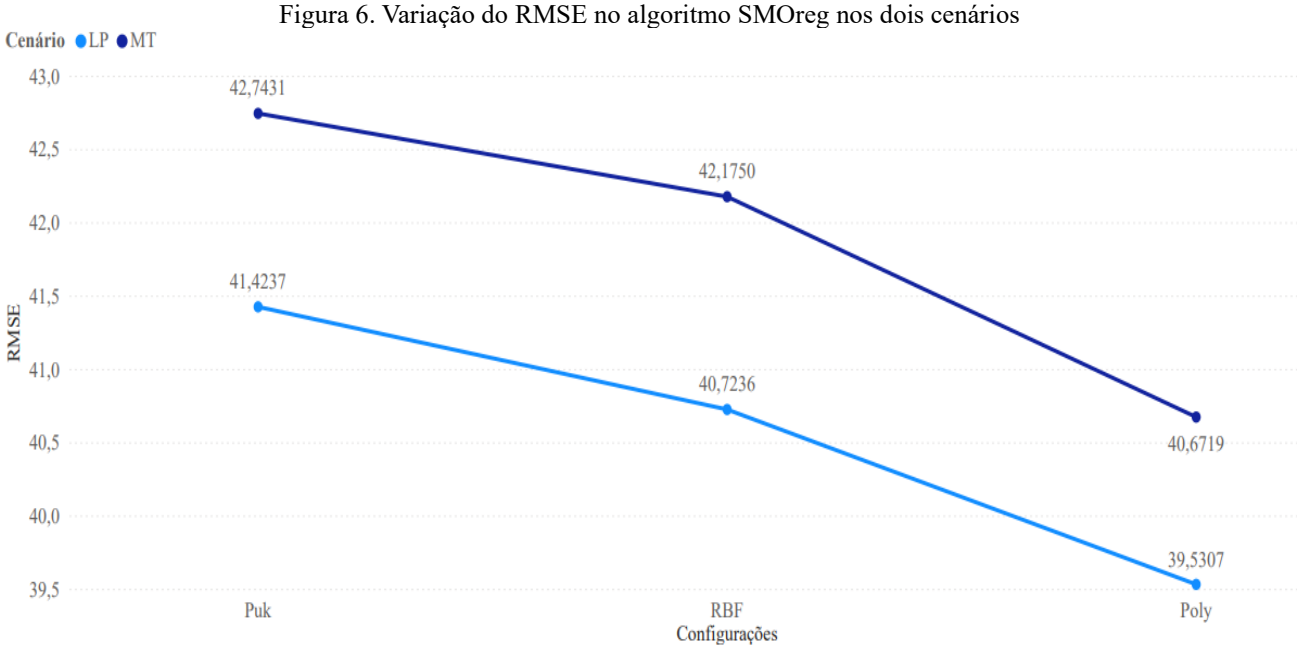
Fonte: Elaborado pelos autores

O Random Forest foi avaliado com diferentes números de árvores, variando de 50 a 1200. A configuração que apresentou o menor RMSE foi com 1200 árvores, resultando em RMSE = 40,3816 para Língua Portuguesa e RMSE = 41,6531 para Matemática. Apesar de a agregação de múltiplas árvores reduzir a variância do modelo, os resultados ainda ficaram ligeiramente piores que os obtidos pelo Linear Regression, indicando que, para este conjunto de dados, o Random Forest não conseguiu superar a simplicidade e a eficácia do modelo linear.

O SMOreg foi avaliado utilizando diferentes funções *Kernel* (Poly, Puk e RBF), possibilitando comparar o impacto da escolha do *kernel* na capacidade preditiva.

Revista ARACÊ
ISSN: 2358-2472

Figura 6. Variação do RMSE no algoritmo SMOreg nos dois cenários



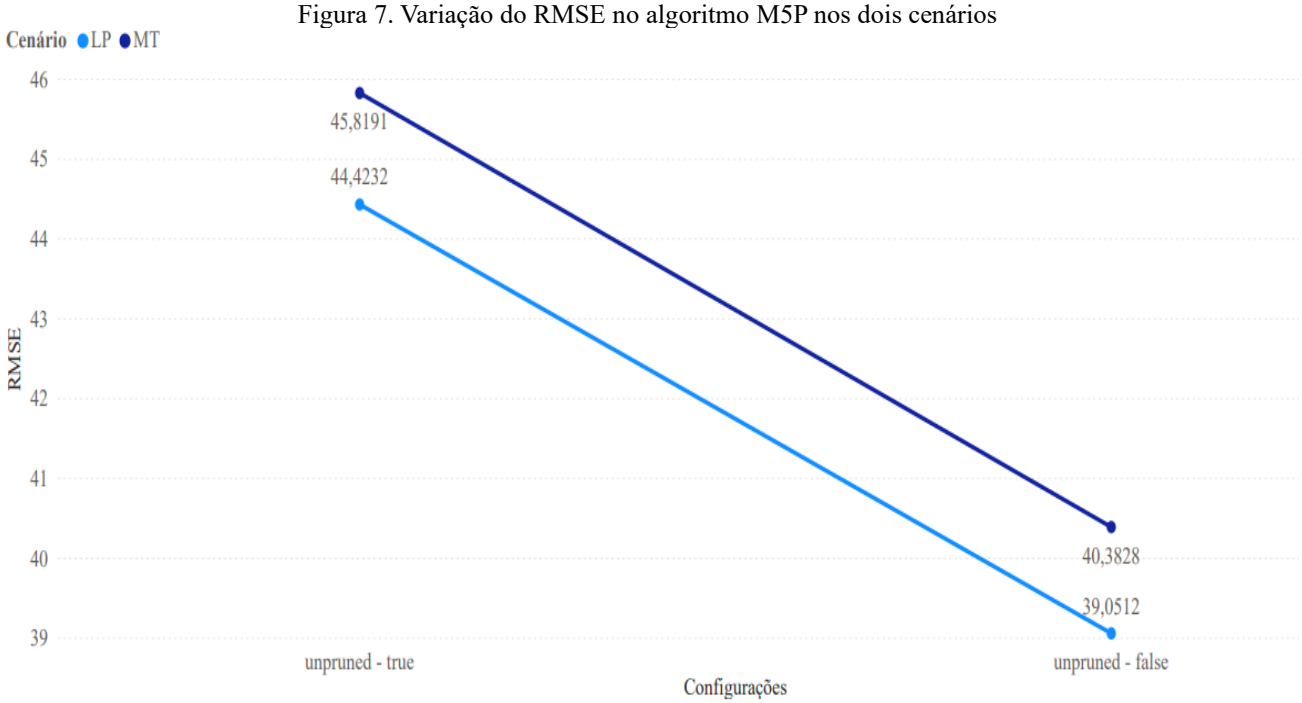
Fonte: Elaborado pelos autores

O SMOreg foi avaliado com diferentes funções *kernel*, e o *kernel* Poly apresentou os menores valores de RMSE: 39,53 em Língua Portuguesa e 40,67 em Matemática. Embora o modelo permita capturar relações não lineares, seu desempenho não superou o Linear Regression em nenhuma das versões. Isso indica que, apesar da flexibilidade do SMOreg, a complexidade adicional do *kernel* não proporcionou vantagem significativa para este conjunto de dados.

O algoritmo M5P foi executado nas versões “com poda” e “sem poda”, permitindo analisar como a poda da árvore influencia o desempenho das predições.

Revista ARACÊ
ISSN: 2358-2472

Figura 7. Variação do RMSE no algoritmo M5P nos dois cenários



Fonte: Elaborado pelos autores

Conforme se pode avaliar no gráfico da Figura 7, a melhor configuração foi a "com poda", obtendo o resultado RMSE de 39,05 em LP e resultado RMSE de 40,38 em MT. Essa configuração apresentou o menor RMSE entre todos os algoritmos avaliados, evidenciando que a poda da árvore contribui para um modelo mais robusto e preciso. Considerando o desempenho observado, o M5P foi selecionado como o modelo mais eficiente para predição dos valores de proficiência em ambos os cenários experimentados.

Após a execução e análise individual de cada algoritmo, foi elaborado um comparativo das melhores configurações obtidas para cada técnica em ambas as versões (LP e MT). O Quadro 9 apresenta os resultados consolidados, destacando o menor valor de RMSE de cada algoritmo e permitindo uma visão clara de seu desempenho relativo.

Quadro 9. Melhores configurações e valores RMSE em ambos os cenários avaliados

	Melhor Configuração LP	RMSE LP	Melhor Configuração MT	RMSE MT
LR	padrão	39,09	padrão	40,40
KNN	$k = 33$	41,96	$k = 31$	43,41
RF	1200 árvores	40,38	1200 árvores	41,65
SMOreg	*Kernel* Poly	39,53	*Kernel* Poly	40,67
M5P	Com poda	39,05	Com poda	40,38

Fonte: Elaborado pelos autores

A análise do Quadro 9 evidencia que, embora algoritmos como o KNN e o RF tenham apresentado desempenhos competitivos em determinadas configurações, suas melhores versões ainda

Revista ARACÊ
ISSN: 2358-2472

resultaram em valores de RMSE superiores aos observados no algoritmo LR, utilizada como modelo de referência. Por outro lado, os algoritmos SMOreg e, sobretudo, o M5P mostraram-se consistentemente superiores, alcançando os menores valores de RMSE em ambas as disciplinas.

Dessa forma, o M5P destacou-se como o modelo mais adequado para este estudo, apresentando RMSE de 39,0512 para Língua Portuguesa e 40,3828 para Matemática, configurando-se como a melhor técnica preditiva entre as avaliadas. Com base nessa melhor configuração, a etapa seguinte envolve a análise de relevância dos atributos, a fim de identificar quais variáveis têm maior impacto na previsão da proficiência dos alunos, servindo como base para a seleção de variáveis nos estudos subsequentes.

4.2 RANKING DOS ATRIBUTOS

Após a avaliação do desempenho dos algoritmos de regressão, tornou-se relevante verificar quais atributos são mais determinantes para a previsão da proficiência no SAEB. Para essa finalidade, foi aplicado o método ReliefF, em conjunto com validação cruzada, a fim de estimar a relevância relativa de cada variável no processo de modelagem.

O procedimento adotado consistiu na execução iterativa do ranqueamento, com a remoção progressiva do atributo de menor importância em cada cenário (LP e MT), até que restassem apenas um atributo preditor e a variável-alvo. Em seguida, os modelos foram novamente avaliados, permitindo observar o impacto da redução de atributos no desempenho preditivo.

Com base nos valores de RMSE obtidos em cada etapa, foi definida uma linha de corte de perda aceitável, de modo a equilibrar simplicidade do modelo e capacidade preditiva. Os resultados desse processo são apresentados nas subseções seguintes, destacando os atributos mais relevantes para a previsão da proficiência dos estudantes.

4.2.1 Dados de Língua Portuguesa

Com base no ranqueamento realizado, os atributos foram eliminados progressivamente, tomando como referência a variação do RMSE. Essa métrica representa o desempenho do modelo ao ser treinado sem o atributo considerado, permitindo avaliar a importância de cada variável na previsão da proficiência.

A partir da definição de uma linha de corte correspondente à perda aceitável de RMSE, selecionaram-se os atributos mais relevantes para o modelo final. Esses atributos, considerados os que mais influenciam a proficiência em Língua Portuguesa, estão apresentados no Quadro 10, juntamente com os valores de RMSE obtidos em cada etapa de eliminação.

Revista ARACÊ
ISSN: 2358-2472

Quadro 10. Atributos mais significativos no cenário de LP

Posição	Atributo	Mérito Médio (±DP)	Ranking Médio (±DP)
1	TX_RESP_Q18	0.011 ± 0.001	1 ± 0
2	TX_RESP_Q23a	0.008 ± 0.001	3.6 ± 2.06
3	TX_RESP_Q22	0.008 ± 0.001	3.7 ± 0.9
4	TX_RESP_Q21	0.008 ± 0.001	3.8 ± 1.83
5	TX_RESP_Q23f	0.008 ± 0.001	4.9 ± 1.81
6	TX_RESP_Q05	0.007 ± 0	5.3 ± 1.19
7	TX_RESP_Q11b	0.005 ± 0.001	8.7 ± 2.1
8	TX_RESP_Q11a	0.005 ± 0.001	9.4 ± 3.04
9	TX_RESP_Q20d	0.005 ± 0.001	9.4 ± 3.26
10	TX_RESP_Q11c	0.005 ± 0.001	11 ± 3.87
11	TX_RESP_Q11g	0.004 ± 0.001	12.7 ± 3.85
12	TX_RESP_Q20a	0.004 ± 0.001	14.7 ± 5.04

Fonte: Elaborado pelos autores

O Quadro 10 apresenta, além da posição no ranking, o mérito médio (±DP) e o ranking médio (±DP) de cada atributo, calculados a partir da avaliação realizada pelo algoritmo ReliefF. O mérito médio indica a relevância estatística do atributo no processo de predição, enquanto o ranking médio reflete sua posição relativa ao longo das iterações da validação cruzada, considerando a variabilidade expressa pelo desvio padrão. A definição de um ponto de corte baseou-se na análise do impacto de cada atributo sobre o RMSE: aqueles cuja remoção provocou perdas significativas de desempenho foram mantidos no modelo final, enquanto os demais foram descartados. Dessa forma, a tabela não apenas evidencia os atributos mais significativos para a proficiência em Língua Portuguesa, mas também explicita o critério metodológico adotado para a seleção das variáveis.

4.2.2 Dados de Matemática

No caso da proficiência em Matemática, foi realizado o ranqueamento dos atributos por meio do algoritmo de seleção, seguido da eliminação progressiva de variáveis, tomando como referência a variação do RMSE. É importante destacar que o RMSE não representa o erro associado a cada atributo em si, mas sim o desempenho do modelo ao ser treinado sem a presença do atributo.

A partir desse procedimento, definiu-se uma linha de corte correspondente à perda aceitável de desempenho preditivo. Assim, apenas os atributos situados acima desse ponto de corte foram considerados mais relevantes para o modelo final. Esses atributos estão presentes no Quadro 11, juntamente com os valores de RMSE obtidos em cada etapa de eliminação.

Revista ARACÊ
ISSN: 2358-2472

Quadro 11. Atributos mais significativos no cenário de MT

Posição	Atributo	Mérito Médio (±DP)	Ranking Médio (±DP)
1	TX_RESP_Q18	0.011 ± 0.001	1.4 ± 0.66
2	TX_RESP_Q17	0.011 ± 0.001	1.9 ± 0.7
3	TX_RESP_Q21	0.009 ± 0.001	3.2 ± 0.87
4	TX_RESP_Q11b	0.008 ± 0.001	4.2 ± 1.17
5	TX_RESP_Q02	0.008 ± 0.001	4.8 ± 1.08
6	TX_RESP_Q20c	0.007 ± 0.001	7.1 ± 1.58
7	TX_RESP_Q05	0.006 ± 0.001	7.9 ± 1.22
8	TX_RESP_Q11c	0.006 ± 0.001	9.7 ± 4.03
9	TX_RESP_Q11e	0.006 ± 0.001	10.2 ± 2.4
10	TX_RESP_Q23f	0.006 ± 0.001	10.3 ± 3.87
11	TX_RESP_Q20a	0.005 ± 0.001	11.2 ± 3.43
12	TX_RESP_Q01	0.005 ± 0.001	12 ± 2.14

Fonte: Elaborado pelos autores

A seguir são apresentadas as discussões com base nas análises dos dois cenários.

4.3 COMPARAÇÕES E ANÁLISES DOS RESULTADOS

A análise conduzida para as duas disciplinas – Língua Portuguesa e Matemática – evidenciou diferenças sutis, porém significativas, na relevância dos atributos preditivos. A aplicação do ranqueamento progressivo permitiu identificar os fatores socioeconômicos e educacionais mais determinantes para o desempenho dos estudantes em cada área.

Observou-se que, embora alguns atributos fossem comuns a ambas as bases, outros se mostraram mais específicos, refletindo particularidades da aprendizagem em cada disciplina. A eliminação progressiva, baseada na variação do RMSE, possibilitou estabelecer uma linha de corte, destacando os atributos essenciais para previsão da proficiência, bem como reduzindo os esforços necessários para se obter uma predição de proficiência nos resultados das provas do SAEB.

De forma geral, os resultados indicam que fatores como **tempo de estudo, sexo, tipo de escola e acesso a recursos tecnológicos** são consistentes como determinantes do desempenho, corroborando estudos anteriores sobre a influência de condições socioeconômicas na aprendizagem. Essa abordagem possibilitou não apenas comparar a importância relativa dos atributos entre as bases, permitindo definir um conjunto enxuto de variáveis-chave que pode orientar políticas educacionais e intervenções pedagógicas mais direcionadas.

A análise dos microdados evidencia que variáveis como tempo de estudo extra, tipo de escola e acesso a recursos tecnológicos exercem forte influência sobre a proficiência dos estudantes. Esses achados não apenas confirmam tendências já apontadas pela literatura, mas também oferecem subsídios concretos para práticas pedagógicas.

Revista ARACÊ
ISSN: 2358-2472

Por exemplo, a relevância do tempo dedicado ao estudo além da sala de aula sugere a necessidade de que as escolas implementem programas de reforço, monitoria ou plantões de dúvidas, capazes de ampliar o engajamento discente e apoiar aqueles com maiores dificuldades. Da mesma forma, a influência do tipo de escola e do nível socioeconômico aponta para desigualdades estruturais que podem ser mitigadas por meio de políticas de financiamento equitativo e de programas de apoio estudantil, incluindo transporte, alimentação e bolsas.

Outro ponto importante é o papel das variáveis relacionadas ao acesso à internet e a dispositivos tecnológicos, que se mostraram significativas na predição do desempenho. Isso reforça a urgência de políticas de inclusão digital que garantam condições mínimas de participação no processo de aprendizagem, sobretudo em contextos de ensino híbrido ou remoto. Para o professor, isso implica repensar estratégias didáticas que utilizem recursos digitais de forma acessível e significativa, reduzindo lacunas entre estudantes com diferentes condições de acesso.

Essas descobertas servem de base para a discussão das implicações educacionais e para a formulação de recomendações voltadas à melhoria do desempenho escolar, que serão detalhadas no capítulo seguinte.

5 CONCLUSÃO

O presente estudo teve como objetivo analisar a influência de diferentes atributos socioeconômicos, educacionais e demográficos no desempenho de estudantes do 3º e 4º anos do Ensino Médio em Língua Portuguesa e Matemática, utilizando microdados do SAEB. Para isso, foram aplicados diversos algoritmos de regressão, incluindo Linear Regression, IBk (KNN), Random Forest, SMOreg e M5P, a fim de comparar sua capacidade preditiva.

Os resultados indicaram que, embora alguns modelos apresentassem desempenho semelhante ao da regressão linear, o algoritmo M5P se destacou como o mais eficiente em termos de RMSE para ambas as disciplinas, oferecendo maior precisão na previsão da proficiência. Essa constatação reforça a utilidade de modelos baseados em árvores de decisão para análise de grandes bases educacionais com múltiplos atributos heterogêneos.

A aplicação do método ReliefF com validação cruzada permitiu identificar os atributos mais relevantes para a previsão do desempenho, a partir da eliminação progressiva dos menos importantes até uma linha de corte de RMSE aceitável. Entre os atributos mais determinantes, destacaram-se o tempo de estudo (TX_RESP_Q20a), o sexo (TX_RESP_Q01), o tipo de escola (TX_RESP_Q17), o acesso a recursos tecnológicos e a compreensão das aulas remotas, evidenciando o impacto das condições socioeconômicas e do contexto educacional no aprendizado.

Revista ARACÊ
ISSN: 2358-2472

A comparação entre as bases de Língua Portuguesa e Matemática revelou tanto atributos comuns quanto específicos, indicando que certas variáveis exercem influência transversal no desempenho escolar (como TX_RESP_Q18, TX_RESP_Q17, TX_RESP_Q21, TX_RESP_Q11b, TX_RESP_Q05, TX_RESP_Q11c, TX_RESP_Q23f, TX_RESP_Q20a e TX_RESP_Q01), enquanto outras afetam de forma distinta cada disciplina (por exemplo, TX_RESP_Q23a e TX_RESP_Q22 em Língua Portuguesa; TX_RESP_Q02 e TX_RESP_Q20c em Matemática).

Os resultados obtidos demonstram que a mineração de dados educacionais não apenas contribui para identificar variáveis preditoras de desempenho, mas também oferece elementos para intervenções pedagógicas e políticas educacionais mais eficazes. O destaque do tempo de estudo extra e do acesso a recursos tecnológicos sinaliza a importância de investir em programas de tutoria, atividades complementares e políticas de inclusão digital. Já a influência do tipo de escola e do nível socioeconômico dos estudantes evidencia a necessidade de ações voltadas à redução das desigualdades, garantindo condições mais equitativas de ensino e aprendizagem.

Tais achados podem orientar professores na personalização das estratégias de ensino, gestores escolares na implementação de projetos de apoio pedagógico, e formuladores de políticas na elaboração de iniciativas que assegurem maior equidade e qualidade na educação. Dessa forma, a análise estatística e computacional, quando interpretada sob a ótica pedagógica, transforma-se em instrumento valioso para a promoção de uma educação mais justa e eficaz.

Em síntese, o estudo contribui para a compreensão dos fatores que impactam o desempenho estudantil do Ensino Médio, oferecendo subsídios para a formulação de políticas educacionais mais direcionadas e para o desenvolvimento de modelos preditivos robustos que podem auxiliar na identificação de alunos em situação de risco. Futuros trabalhos podem ampliar a análise incluindo novas variáveis, outros anos de avaliação ou técnicas de aprendizado de máquina mais avançadas, de modo a refinar ainda mais a previsão de proficiência e apoiar estratégias de intervenção educacional. Adicionalmente, a extração de regras de associação pode explicar a relação dos atributos mais significativos com o resultado da prova do SAEB.

Como limitação, destaca-se que a análise considerou apenas microdados do SAEB de 2021 e do estado de Rondônia, o que restringe a generalização dos resultados. Futuros estudos podem incluir outras edições do SAEB, múltiplos estados e técnicas de aprendizado de máquina mais avançadas.

Revista ARACÊ
ISSN: 2358-2472

AGRADECIMENTOS

Registra-se o agradecimento ao Instituto Federal de Educação, Ciência e Tecnologia de Rondônia, por viabilizar o projeto de pesquisa "Mineração de Dados Públicos" por meio do EDITAL Nº 6/2024/REIT/PROPESP/IFRO, DE 07 DE MAIO DE 2024, bem como ao DEPESP do *Campus* Porto Velho Calama pelo auxílio na publicação do presente artigo por meio do EDITAL Nº 159/2025/PVCAL - CGAB/IFRO, DE 15 DE AGOSTO DE 2025.

Revista ARACÊ
ISSN: 2358-2472

REFERÊNCIAS

ALVES, M. T. G.; SOARES, J. F. Contexto escolar e indicadores educacionais: condições desiguais para a efetivação de uma política de avaliação educacional. Educação e Pesquisa, São Paulo, v. 39, n. 1, p. 177-194, mar. 2013.

BAKER, R. S.; INVENTADO, P. S. Educational data mining and learning analytics. In: LARUSSON, J. A.; WHITE, B. (Ed.). *Learning analytics*. Heidelberg: Springer, 2014. p. 61–75.

BRITO JÚNIOR, J. J. R. T. A relação entre nível socioeconômico e proficiência em matemática de estudantes pernambucanos do 9º ano através da Mineração de Dados Educacionais. Dialnet, 2022.

FARIAS, M. de L.; GUSMÃO, R. P. de; GUSMÃO, C. S. D. Mineração de dados educacionais: investigando a relação entre os microdados do INEP e o desempenho do IDEB. Renote, Porto Alegre, v. 20, n. 2, 2022.

FARIAS, M. L. de; GUSMÃO, R. P. de; GUSMÃO, C. S. D. Mineração de Dados para investigar o IDEB usando o Censo da Educação Básica e SAEB: um estudo de caso em Sergipe. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO (SBIE), 2020.

FONSECA, S. O.; NAMEN, A. A. Mineração em bases de dados do Inep: uma análise exploratória para nortear melhorias no sistema educacional brasileiro. Educação em Revista, Belo Horizonte, v. 32, n. 1, p. 133–157, mar. 2016.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). Microdados do Sistema de Avaliação da Educação Básica – SAEB 2021. Brasília: INEP, publicado em 16 set. 2022; atualizado em 05 maio 2023.

PINTO, G. da S. et al. Mineração de dados educacionais: um modelo de predição do perfil do aluno para melhoria do IDEB. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO (SBIE), 31., 2020, Online. Porto Alegre: Sociedade Brasileira de Computação, 2020.

ROMERO, C.; VENTURA, S. Educational Data Mining and Learning Analytics: An Updated Survey. WIREs Data Mining and Knowledge Discovery, 2020.

SOARES, J. F.; ANDRADE, R. J. Nível socioeconômico, qualidade e eqüidade das escolas de Belo Horizonte. Ensaio: Avaliação e Políticas Públicas em Educação, Rio de Janeiro, v. 14, n. 50, p. 107-126, jan./mar. 2006.