UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO
PRÓ-REITORIA DE GRADUAÇÃO
DEPARTAMENTO DE ENGENHARIAS E TECNOLOGIA
CURSO INTERDISCIPLINAR EM TECNOLOGIA DA INFORMAÇÃO

Davi da Silva Soares

Análise da Influência do Perfil Socioeconômico no Desempenho de Estudantes em Programação Introdutória: um estudo de caso

PAU DOS FERROS
2025

Davi da Silva Soares

Análise da Influência do Perfil Socioeconômico no Desempenho de Estudantes em Programação Introdutória: um estudo de caso

Monografia apresentada a Universidade Federal Rural do Semi-Árido como requisito para obtenção do título de Bacharel em Tecnologia da Informação.

Orientadora: Laysa Mabel de Oliveira Fontes, Profa. Dra.

PAU DOS FERROS

2025

S676a Soares, Davi da Silva . Análise da Influência do Perfil Socioeconômico no Desempenho de Estudantes em Programação Introdutória: um estudo de caso / Davi da Silva Soares. - 2025.
48 f. : il.

Orientadora: Laysa Mabel de Oliveira Fontes.
Monografia (graduação) - Universidade Federal Rural do Semi-árido, Curso de Tecnologia da Informação, 2025.

1. Algoritmos. 2. Ciência de dados. 3. Desempenho acadêmico. 4. Fatores socioeconômicos. 5. Programação introdutória. I. Fontes, Laysa Mabel de Oliveira , orient. II. Título.

Ficha catalográfica elaborada por sistema gerador automáto em conformidade com AACR2 e os dados fornecidos pelo) autor(a). Biblioteca Campus Pau dos Ferros
Bibliotecária: Erlanda Maria Lopes da Silva
CRB: 15/966

Davi da Silva Soares

Análise da Influência do Perfil Socioeconômico no Desempenho de Estudantes em Programação Introdutória: um estudo de caso

Monografia apresentada a Universidade Federal Rural do Semi-Árido como requisito para obtenção do título de Bacharel em Tecnologia da Informação.

Defendida em: 22/07/2025.

BANCA EXAMINADORA

Laysa Mabel de Oliveira Fontes, Profa. Dra. (UFERSA)
Presidente

João Batista de Souza Neto, Prof. Dr. (UFERSA)
Membro Examinador

Reudismam Rolim de Sousa, Prof. Dr. (UFERSA)
Membro Examinador

AGRADECIMENTOS

Aos meus pais, pelo apoio incondicional, incentivo constante e por sempre acreditarem em mim. Sou imensamente grato por todo o carinho, compreensão e força que me deram ao longo dessa jornada.

Aos meus amigos, pela parceria, escuta e companheirismo. Ter vocês por perto tornou esse percurso mais leve e significativo.

À minha orientadora, Profa. Mabel, pela dedicação, paciência e pelas orientações valiosas ao longo do desenvolvimento deste trabalho. Sua contribuição foi essencial para que este TCC se concretizasse com qualidade e seriedade.

“Sabia que as palavras estavam ali, por trás de tudo, uma espécie de ponte entre mim e o mundo”.

Marília Arnoud

RESUMO

Este trabalho investiga a influência do perfil socioeconômico no desempenho dos estudantes no componente curricular de Algoritmos, do curso de Bacharelado Interdisciplinar em Tecnologia da Informação da Universidade Federal Rural do Semi-Árido, Campus Pau dos Ferros. Trata-se de um estudo de caso com abordagem quantitativa, no qual foram aplicadas técnicas de ciência de dados, como o Coeficiente de Correlação de Pearson e cinco algoritmos de classificação supervisionada (J48, REPTree, LMT, Random Forest e Naive Bayes) para identificar os fatores mais associados à aprovação ou reprovação discente. Os resultados revelaram que fatores como idade do aluno, posse de computador e trabalho concomitante ao estudo foram decisivos em diversos modelos, reforçando a relação entre aspectos socioeconômicos e o desempenho acadêmico. Além disso, o município de residência emergiu como um achado relevante, sugerindo que desigualdades regionais e logísticas também podem interferir no sucesso do estudante. A partir desses resultados, conclui-se que o sucesso no componente curricular de Algoritmos está condicionado a fatores socioeconômicos que transcendem a sala de aula, reforçando a necessidade de políticas institucionais de apoio, como o acesso facilitado a equipamentos tecnológicos e suporte a estudantes que conciliam trabalho e estudo, a fim de promover a equidade e a permanência estudantil. A pesquisa contribui não apenas para validar a aplicação de técnicas de ciência de dados na análise do desempenho acadêmico, mas também para evidenciar os desafios socioeconômicos enfrentados pelos estudantes.

Palavras-chave: algoritmos; ciência de dados; desempenho acadêmico; fatores socioeconômicos; programação introdutória.

ABSTRACT

This study examines the influence of socioeconomic background on student performance in the Algorithms course of the Interdisciplinary Bachelor's Degree in Information Technology at the Federal Rural University of the Semi-Arid Region, Pau dos Ferros Campus. It is a case study with a quantitative approach, employing data science techniques such as Pearson's correlation coefficient and five supervised classification algorithms (J48, REPTree, LMT, Random Forest, and Naive Bayes) to identify the factors most associated with student success or failure. The findings reveal that variables such as student age, access to a personal computer, and employment while studying were decisive across multiple models, reinforcing the relationship between socioeconomic conditions and academic performance. Additionally, the municipality of residence emerged as a significant factor, suggesting that regional and logistical disparities may also impact student success. These results indicate that success in the Algorithms course is influenced by socioeconomic factors that go beyond the classroom, underscoring the need for institutional support policies, such as improved access to technological resources and assistance for students who balance work and study, in order to promote equity and student retention. This research contributes not only by validating the use of data science techniques in the analysis of academic performance but also by highlighting the socioeconomic challenges faced by higher education students.

Keywords: algorithms; data science; academic performance; socioeconomic factors; introductory programming.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

ACM	*Association for Computing Machinery*
ARFF	*Attribute-Relation File Format*
BITI	Bacharelado Interdisciplinar em Tecnologia da Informação
CSV	*Comma Separated Values*
ENADE	Exame Nacional de Desempenho de Estudantes
FN	Falsos Negativos
FP	Falsos Positivos
LMT	*Logistic Model Tree*
MDE	Mineração de Dados Educacionais
MEC	Ministério da Educação
SBC	Sociedade Brasileira de Computação
SHAP	*SHapley Additive exPlanations*
SMOTE	*Synthetic Minority Over-sampling Technique*
UFERSA	Universidade Federal Rural do Semi-Árido
VN	Verdadeiros Negativos
VP	Verdadeiros Positivos
WEKA	*Waikato Environment for Knowledge Analysis*

SUMÁRIO

1 INTRODUÇÃO

Este capítulo apresenta uma visão geral do trabalho e está organizado da seguinte forma: a Seção 1.1 trata da problematização; a Seção 1.2 apresenta os argumentos que justificam a realização deste estudo; a Seção 1.3 define os objetivos; e, por fim, a Seção 1.4 expõe a estrutura geral do trabalho.

1.1 Problematização

O ensino de programação na educação superior é um processo complexo, influenciado por diversos fatores interligados que podem tanto favorecer quanto dificultar a aprendizagem, a depender de suas características (Moraes, 2022). Entre os principais desafios enfrentados pelos estudantes está o componente curricular de programação introdutória, frequentemente associada a elevados índices de reprovação e evasão (Silva et al., 2020).

Segundo Pereira, Carvalho e Souto (2019), o desempenho acadêmico dos estudantes em componentes curriculares introdutórios de programação pode ser influenciado por fatores que excedem o ambiente acadêmico, como a origem escolar, experiências prévias com programação, gênero e condições socioeconômicas.

Em nível nacional, dados do Mapa do Ensino Superior no Brasil 2023 indicam que, na última década, a taxa de evasão dos cursos presenciais da área de tecnologia da informação tem se mantido acima da média dos demais cursos (Instituto Semesp, 2023). A edição de 2024 do mesmo estudo aponta um aumento de 28,1% nas matrículas nos cursos das áreas de computação e tecnologias da informação e comunicação, sinalizando tanto o crescimento do interesse pela área quanto os desafios relacionados à permanência estudantil (Instituto Semesp, 2024).

Em âmbito local, Sousa e Leite (2020) observaram que, no curso de Bacharelado Interdisciplinar em Tecnologia da Informação (BITI) da Universidade Federal Rural do Semi-Árido (UFERSA), Campus Pau dos Ferros, há um crescimento progressivo nos números de trancamentos, cancelamentos e reprovações ao longo dos semestres. Ainda segundo os autores, nos períodos letivos de 2016.2 e 2017.1, o componente curricular de programação introdutória, denominado de Algoritmos, apresentou taxas de insucesso de 85% e 70%, respectivamente.

Além disso, Andrade (2023) aponta que a evasão nos cursos da UFERSA, Campus Pau dos Ferros, entre os semestres de 2019.1 e 2022.2, está relacionada a fatores socioeconômicos, demandas do mercado de trabalho, estrutura curricular, disparidades de gênero e questões ligadas à saúde mental.

Diante desse cenário, emerge a questão central de pesquisa deste trabalho, que busca responder: o perfil socioeconômico dos estudantes exerce influência sobre o desempenho no componente curricular de Algoritmos do curso de BITI da UFERSA, Campus Pau dos Ferros?

1.2 Justificativa

Os componentes curriculares de Algoritmos e Programação estão entre os componentes curriculares fundamentais nos cursos das áreas de computação, independentemente de sua modalidade (tecnólogos, bacharelados ou licenciaturas). Esses conteúdos geralmente são introduzidos logo nos primeiros períodos dos cursos, estabelecendo as bases para o desenvolvimento acadêmico e técnico dos discentes (Silva e Moreira, 2021).

Para Pereira (2021), compreender a forma como o estudante se insere no contexto familiar e social, considerando aspectos como responsabilidade profissional, estrutura financeira e dinâmica familiar, é o ponto de partida para substituir abordagens superficiais comumente adotadas nas ações voltadas à permanência em componentes curriculares introdutórios de programação nos cursos das engenharias e ciências exatas.

Diversos estudos apontam a existência de relação significativa entre aspectos socioeconômicos, como renda familiar, escolaridade dos responsáveis, tipo de escola cursada no ensino médio e gastos com transporte, e o desempenho acadêmico (Júnio et al., 2018; Oliveira, 2021; Rezende et al., 2022).

Essa relação destaca a importância de análises aprofundadas sobre essas variáveis, sobretudo no contexto dos cursos da área de Tecnologia da Informação. Compreender essas influências pode subsidiar a formulação de políticas públicas e estratégias institucionais mais justas e eficazes, com foco na promoção da equidade e no fortalecimento da permanência estudantil.

Diante das opiniões expostas, corrobora-se que o processo de ensino e aprendizagem de programação introdutória é complexo e constituído por muitas e diferentes variáveis, dentre elas, fatores socioeconômicos, e, portanto, faz-se necessário investigar se o perfil socioeconômico impacta no desempenho de estudantes no componente curricular de Algoritmos do curso de BITI da UFERSA, Campus Pau dos Ferros.

1.3 Objetivos

Esta seção apresenta o objetivo geral e os objetivos específicos que norteiam este trabalho.

1.3.1 Objetivo Geral

O objetivo geral deste trabalho é investigar a influência do perfil socioeconômico no desempenho de estudantes no componente curricular de Algoritmos do curso de BITI da UFERSA, Campus Pau dos Ferros, considerando fatores como faixa etária, tipo de escola cursada no ensino médio, município de residência, entre outros.

1.3.2 Objetivos Específicos

Para atingir o objetivo geral, foram definidos os seguintes objetivos específicos:

- Analisar se há relação entre fatores socioeconômicos e a aprovação ou reprovação no componente curricular de Algoritmos;
- Investigar quais fatores socioeconômicos exercem maior influência sobre o desempenho dos estudantes no componente curricular de Algoritmos;
- Propor subsídios para a formulação de estratégias institucionais que considerem os principais fatores socioeconômicos identificados como determinantes do desempenho discente no componente curricular de Algoritmos.

1.4 Organização do Trabalho

Este trabalho está estruturado da seguinte forma: o Capítulo 2 apresenta a fundamentação teórica que embasa a pesquisa; o Capítulo 3 descreve os procedimentos metodológicos adotados; o Capítulo 4 expõe os resultados obtidos e a respectiva discussão; e, por fim, o Capítulo 5 reúne as considerações finais, limitações do estudo e sugestões para trabalhos futuros.

2 REFERENCIAL TEÓRICO

Este capítulo descreve a base teórica que fundamenta a elaboração deste trabalho, estando organizado da seguinte forma: a Seção 2.1 apresenta uma contextualização sobre programação introdutória; a Seção 2.2 aborda os principais conceitos relacionados à ciência de dados; e a Seção 2.3 apresenta trabalhos correlatos ao tema desta pesquisa.

2.1 Programação Introdutória

Os componentes curriculares introdutórios de programação fazem parte da grade curricular de diversos cursos de graduação, especialmente nas áreas de ciências, tecnologias, engenharias e matemática (Medeiros et al., 2020). Embora apresentem diferentes denominações, como algoritmos, introdução à programação ou introdução à ciência da computação, esses componentes compartilham o objetivo comum de ensinar os fundamentos da programação (Bosse, 2020).

Geralmente ofertadas nos primeiros períodos dos cursos, esses componentes curriculares marcam o primeiro contato dos discentes com o pensamento sistemático e lógico. Nessa etapa, são abordados conceitos essenciais para a construção de algoritmos e programas simples, como variáveis, constantes, tipos de dados, expressões, dentre outros (Silva et al., 2022). Além disso, são desenvolvidas habilidades relacionadas à resolução de problemas, ao domínio da sintaxe e semântica das linguagens de programação e à formulação de soluções computacionais (Medeiros, 2019).

Nesse sentido, esses componentes curriculares constituem a base para o aprendizado posterior em áreas mais avançadas da computação, desempenhando papel fundamental tanto na formação acadêmica quanto na atuação profissional futura (Silva et al. *apud* Campus, 2010; Sobral, 2019; Medeiros, 2019).

A relevância dos componentes curriculares introdutórios de programação também é reconhecida por instituições como a Sociedade Brasileira de Computação (SBC), a *Association for Computing Machinery* (ACM) e o Ministério da Educação (MEC). A SBC recomenda a inserção de conteúdos como algoritmos, estruturas de dados e linguagens de programação já nos primeiros períodos dos cursos. A ACM destaca a necessidade de oferecer uma introdução ampla aos conceitos da computação, enquanto o MEC, desde 1999, considera a programação uma competência essencial na formação dos profissionais da área (Medeiros, 2019).

Dessa forma, a programação introdutória representa um componente estruturante na trajetória dos estudantes, sendo crucial para o desenvolvimento de uma base sólida de conhecimento técnico e conceitual.

2.2 Ciência de Dados

A ciência de dados é uma área interdisciplinar dedicada ao estudo dos dados em todas as suas etapas, desde a coleta e representação até o armazenamento, segurança, análise e disseminação (Matos et al., 2022). Essa área emergiu impulsionada pelos avanços das tecnologias da informação e comunicação, que transformaram os hábitos de interação com a tecnologia e levaram à produção de grandes volumes de dados em alta velocidade.

Essa área integra três domínios fundamentais: (i) programação, (ii) estatística e matemática e (iii) conhecimento específico do domínio, posicionando-se na interseção entre esses saberes (Rautenberg; Carmo, 2019).

A ciência de dados busca desenvolver modelos capazes de identificar padrões em conjuntos de dados complexos e aplicar esses modelos a problemas concretos. Seu objetivo central é transformar dados em informação útil, oferecendo suporte à tomada de decisão, à explicação de fenômenos subjacentes e à verificação de hipóteses (Moreira et al., 2019; Rautenberg; Carmo, 2019).

Com isso, destaca-se como uma área estratégica para lidar com os desafios relacionados ao volume, à variedade e à velocidade dos dados gerados atualmente. Sua natureza interdisciplinar possibilita a análise aprofundada de dados em diversos contextos e contribui significativamente para a produção de conhecimento aplicado.

2.3 Trabalhos Relacionados

A relação entre fatores socioeconômicos e o desempenho acadêmico de estudantes tem sido amplamente investigada na literatura, sobretudo no contexto dos componentes curriculares introdutórios de programação. A seguir são apresentados estudos relevantes que exploram diferentes abordagens sobre o tema.

Pereira, Carvalho e Souto (2019) analisaram a influência de atributos demográficos e educacionais no desempenho de estudantes matriculados em um componente curricular de introdução à programação na Universidade Federal do Amazonas. Para tanto, utilizaram duas bases de dados: registros acadêmicos e histórico de atividades em um sistema de juiz online,

abrangendo três semestres letivos de 14 cursos. Os autores aplicaram o teste de Shapiro-Wilk para verificar a normalidade dos dados, além do teste exato de Fisher e coeficientes de contingência de Pearson para avaliar a intensidade de associação entre variáveis. Os resultados indicaram que fatores educacionais, como a origem do ensino médio e a experiência prévia com linguagem de programação, exercem maior influência sobre o desempenho dos discentes do que atributos ligados à dimensão social, como sexo, estado civil ou ter filhos.

Caldas (2020) investigou os fatores que contribuem para o bom desempenho e avaliação de estudantes em componentes curriculares de introdução a algoritmos e programação, considerando diferentes cursos de engenharia. A pesquisa combinou dados do Exame Nacional de Desempenho de Estudantes (ENADE), registros acadêmicos de alunos entre 2007 e 2018, e informações obtidas por meio de questionários e entrevistas com estudantes e professores. Os resultados mostraram que alunos oriundos de escolas técnicas e com bom desempenho no vestibular apresentaram melhores resultados no componente curricular. O estudo também destacou a relevância do preparo prévio em áreas técnicas, além da importância de metodologias de ensino claras e alinhadas ao perfil dos estudantes, o que reforça a necessidade de práticas pedagógicas adaptadas às realidades educacionais dos discentes.

Oliveira (2021) aplicou técnicas de Mineração de Dados Educacionais (MDE) para identificar os principais fatores que impactam a trajetória de estudantes dos cursos de computação do Instituto de Informática da Universidade Federal de Goiás. Utilizando métodos como SHAP (*SHapley Additive exPlanations*) e *heatmaps*, a autora analisou a influência de diferentes atributos no status final de “Graduado”. Os fatores que se destacaram positivamente foram a percentagem de aprovação, o total de componentes curriculares cursados, o número de semestres concluídos, a frequência e a média global dos alunos. Em contrapartida, um número elevado de trancamentos e altos índices de reprovação estiveram associados negativamente ao progresso acadêmico, revelando o impacto direto desses elementos no sucesso do estudante.

Pereira (2021) desenvolveu um modelo preditivo para identificar o risco de evasão em componentes curriculares introdutórios de programação, com base em dados sociodemográficos de estudantes da Universidade Federal do Amazonas. Foram aplicados algoritmos de aprendizado de máquina, como AdaBoost, Random Forest e redes neurais, alcançando-se uma acurácia de 70%. Os principais fatores associados à evasão foram faixa etária, tipo de escola de origem e estado civil. Além disso, o estudo evidenciou que alunos com menor experiência prévia em programação e oriundos de escolas públicas apresentaram maior probabilidade de evasão. A pesquisa demonstrou a viabilidade de identificar precocemente o

risco de evasão, favorecendo a implementação de intervenções pedagógicas e institucionais mais eficazes.

Rezende et al. (2022) analisaram os microdados do ENADE de 2017 com o objetivo de identificar quais aspectos socioeconômicos mais influenciam o desempenho acadêmico de estudantes de um curso de sistemas de informação. Para isso, aplicaram um algoritmo de árvore de decisão e o coeficiente de correlação estatística de Pearson. Os resultados apontaram como variáveis mais relevantes: renda familiar, participação em programas de intercâmbio, recebimento de bolsas acadêmicas, tipo de financiamento da mensalidade e tipo de escola cursada no ensino médio. Por outro lado, fatores como a modalidade do curso e o fato de o estudante trabalhar simultaneamente aos estudos não apresentaram influência significativa no desempenho.

Diante dos estudos apresentados, este trabalho se diferencia ao realizar uma análise mais específica, centrada no componente curricular de Algoritmos do curso de BITI da UFERSA, Campus Pau dos Ferros. Além disso, incorpora novas variáveis à investigação, como a quantidade de horas trabalhadas, o local de residência e os motivos que levaram os discentes a trancar o componente. Essas adições ampliam o escopo da análise, possibilitando uma compreensão mais profunda e contextualizada dos fatores que influenciam o desempenho acadêmico nesse componente curricular. O Quadro 1 apresenta um comparativo entre os trabalhos relacionados e a proposta aqui desenvolvida, com o intuito de evidenciar suas singularidades e avanços.

Quadro 1 – Resumo dos trabalhos relacionados

REFERÊNCIA	**OBJETIVO**	**METODOLOGIA**	**VARIÁVEIS**	**RESULTADO**
Pereira, Carvalho e Souto (2019)	Analisar a influência de atributos demográficos sobre o desempenho de estudantes em introdução à programação.	Análise de registros acadêmicos e dados de um juiz online, utilizando teste de Shapiro-Wilk, teste exato de Fisher e coeficiente de correlação de Pearson.	Demográficas/Sociais: sexo, estado civil e presença de filhos. Educacionais: origem da escola de ensino médio e experiência prévia em programação.	Fatores educacionais, como escola de origem e experiência prévia, exercem maior influência no desempenho do que os fatores sociais.

(Continua)

(Continuação)

Caldas (2020)	Investigar os fatores que contribuem para o bom desempenho de estudantes de engenharia em componentes curriculares de algoritmos e programação.	Análise de dados do ENADE, registros acadêmicos, questionários e entrevistas com alunos e professores.	Desempenho no vestibular, tipo de escola de origem, e percepções de alunos e professores.	Alunos de escolas técnicas e com bom desempenho no vestibular apresentaram melhores resultados.
Oliveira (2021)	Identificar os fatores que mais impactam a trajetória e a graduação de estudantes de computação.	Aplicação de técnicas de Mineração de Dados Educacionais (MDE), como SHapley Additive exPlanations (SHAP) e heatmaps, em dados acadêmicos.	Percentual de aprovação, total de componentes curriculares cursados, semestres concluídos, frequência, média global, trancamentos e reprovações.	A aprovação, frequência e média global influenciam positivamente a graduação, enquanto trancamentos e reprovações impactam negativamente.
Pereira (2021)	Desenvolver um modelo preditivo para identificar o risco de evasão em componentes curriculares introdutórios de programação.	Aplicação de algoritmos de aprendizado de máquina, como AdaBoost, Random Forest e redes neurais, em dados sociodemográficos.	Faixa etária, tipo de escola de origem, estado civil e experiência prévia em programação.	O modelo alcançou 70% de acurácia na previsão de evasão. Menor experiência e origem de escola pública foram associados a maior risco de evasão.
Rezende et al. (2022)	Identificar os aspectos socioeconômicos que mais influenciam o desempenho de estudantes de sistemas de informação.	Análise dos microdados do ENADE 2017 com algoritmo de árvore de decisão e coeficiente de correlação de Pearson.	Renda familiar, intercâmbio, bolsa acadêmicas, tipo de financiamento, tipo de escola, modalidade do curso, e trabalhar e estudar concomitantemente.	Renda familiar, intercâmbio, bolsas acadêmicas e tipo de escola foram as variáveis mais relevantes. Trabalhar durante o curso não mostrou influência significativa.

(Continua)

(Conclusão)

Este trabalho	Investigar a influência do perfil socioeconômico no desempenho de estudantes no componente curricular de Algoritmos do curso de BITI da UFERSA, Campus Pau dos Ferros.	Aplicação de técnicas de ciência de dados, incluindo cinco algoritmos de classificação supervisionada (J48, REPTree, LMT, Random Forest e Naive Bayes), além da análise estatística por meio do coeficiente de correlação de Pearson, com o objetivo de identificar os fatores mais influentes no desempenho acadêmico dos discentes.	Foram analisadas 21 variáveis, conforme especificado no Quadro 2.	A idade, a posse de computador e a conciliação entre trabalho e estudo foram fatores determinantes nas classificações geradas pelas técnicas, indicando que o sucesso no componente de Algoritmos é influenciado por condições estruturais e pessoais.

Fonte: Autoria própria (2025)

3 METODOLOGIA

Esta pesquisa caracteriza-se metodologicamente como um estudo de caso, que, segundo Yin (2001), consiste em uma investigação empírica que examina um fenômeno contemporâneo dentro de seu contexto da vida real. O caso em questão está delimitado ao universo dos discentes do componente curricular de Algoritmos, do curso de BITI da UFERSA, Campus Pau dos Ferros, nos semestres 2021.1, 2021.2 e 2022.1.

A escolha desse método possibilita uma análise focada e aprofundada dos possíveis impactos de variáveis socioeconômicas no desempenho acadêmico, considerando as especificidades do contexto estudado. Para investigar esse caso, adotou-se uma abordagem quantitativa, com a aplicação de técnicas estatísticas e algoritmos de classificação, visando identificar e analisar relações entre as variáveis socioeconômicas e o desempenho discente.

As seções a seguir descrevem os procedimentos de coleta e preparação dos dados, bem como as variáveis e técnicas de análise selecionadas.

3.1 Coleta dos Dados

Os dados utilizados neste estudo foram obtidos por meio de um questionário estruturado, elaborado com o auxílio da ferramenta Google Forms e aplicado no componente curricular de Algoritmos, do curso de BITI da UFERSA, Campus Pau dos Ferros.

Essa coleta integrou uma intervenção metodológica desenvolvida no referido componente, conforme descrito por Freitas, Fontes e Silva (2022). O questionário teve como objetivo levantar informações de natureza acadêmica e socioeconômica dos estudantes, sendo aplicado nos semestres 2021.1, 2021.2 e 2022.1.

A base de dados resultante contém dados referentes ao desempenho acadêmico, ao perfil socioeconômico e ao histórico educacional dos discentes. Ao todo, foram coletadas 137 respostas, todas provenientes de participação voluntária dos estudantes ao longo dos três semestres analisados.

3.2 Preparação dos Dados

A base de dados passou por um processo de preparação com o objetivo de torná-la adequada para as análises quantitativas. Esse processo foi realizado com o suporte das ferramentas Google Sheets, Google Colab e da linguagem de programação Python, utilizando

principalmente a biblioteca Pandas. Essa biblioteca oferece funcionalidades convenientes de indexação, reformatação, análise detalhada, agregações e seleção de subconjuntos, contribuindo significativamente para as etapas de manipulação, preparação e limpeza dos dados.

Inicialmente, as respostas coletadas foram consolidadas em uma única planilha no Google Sheets, posteriormente exportada no formato CSV (*Comma Separated Values*) para processamento no ambiente Google Colab.

Durante essa etapa, foram realizados os seguintes procedimentos:

- Renomeação das colunas para melhorar a legibilidade;
- Conversão das notas para o formato numérico padrão (ponto decimal);
- Padronização de dados textuais, com transformação para letras minúsculas e remoção de pontuações inconsistentes;
- Eliminação de dados sensíveis, como nomes e números de matrícula, para garantir o anonimato dos participantes.

Adicionalmente, foi identificado um desbalanceamento entre classes, com predominância significativa de estudantes aprovados em comparação aos reprovados. Esse tipo de problema pode comprometer a performance de algoritmos de classificação, que tendem a favorecer as classes majoritárias, conforme discutido por Soares (2020, *apud* Gu et al., 2008). Tal desbalanceamento é especialmente problemático quando a classe minoritária é de maior interesse para a pesquisa, como no caso dos estudantes reprovados (Soares, 2020 *apud* Marques-Vera et al., 2013).

Para mitigar esse problema, foi aplicado o algoritmo de balanceamento SMOTE (*Synthetic Minority Over-sampling Technique*), que cria instâncias sintéticas da classe minoritária com base no método dos k-vizinhos mais próximos (Witten et al., 2016). O uso do SMOTE é recomendado em situações com poucas amostras da classe minoritária e tem ampla aceitação na literatura por sua eficácia na melhoria da acurácia de modelos de classificação (Maione, 2020 *apud* Haixiang et al., 2017).

Após o balanceamento e os demais tratamentos, a base de dados foi exportada no formato ARFF (*Attribute-Relation File Format*), próprio do software WEKA (*Waikato Environment for Knowledge Analysis*), onde foram realizadas as análises subsequentes.

A escolha pelo WEKA justifica-se por ser um software gratuito, de código aberto, com interface amigável e ampla gama de funcionalidades. Além disso, é amplamente utilizado no meio acadêmico e em pesquisas na área de mineração de dados (Hall et al., 2009; Soares, 2020 *apud* Witten et al., 2011).

3.3 Variáveis

As variáveis analisadas neste estudo têm como objetivo investigar as possíveis relações entre o perfil socioeconômico dos estudantes e seu desempenho acadêmico no componente curricular de Algoritmos. Essas variáveis foram agrupadas em três categorias principais: (i) desempenho acadêmico, (ii) perfil socioeconômico e (iii) histórico acadêmico. A definição completa das variáveis está apresentada nos Quadros 2, 3 e 4.

Quadro 2 – Variáveis de desempenho acadêmico

VARIÁVEL	DESCRIÇÃO
nota_unidade_1	Nota da primeira unidade
nota_unidade_2	Nota da segunda unidade
nota_unidade_3	Nota da terceira unidade
nota_recuperacao	Nota da recuperação
situacao_aluno	Situação final (aprovado ou reprovado)
media_notas	Média aritmética das notas das três unidades
faltas_aluno	Total de faltas registradas
semestre_oferta_disciplina	Semestre em que o componente foi ofertado

Fonte: Autoria própria (2025)

Quadro 3 – Variáveis do perfil socioeconômico

VARIÁVEL	DESCRIÇÃO
genero_aluno	Gênero do estudante
idade_aluno	Idade do estudante
tipo_escola_ensino_medio	Tipo de escola cursada no ensino médio
mora_em_pau_dos_ferros	Se reside em Pau dos Ferros
municipio_residencia	Município de residência
trabalha_atualmente	Se trabalha atualmente
hora_trabalho_semanal	Horas de trabalho por semana
possui_computador	Se possui computador

Fonte: Autoria própria (2025)

Quadro 4 – Variáveis de histórico acadêmico

VARIÁVEL	DESCRIÇÃO
semestre_ingresso_ufersa	Semestre de ingresso na UFERSA
primeira_vez_disciplina_algoritmos	Se é a primeira vez cursando o componente
tentativa_disiciplina	Número de tentativas no componente
trancou_dedistiu_algoritmos	Se já trancou ou desistiu do componente
motivo_trancou_desistiu_algoritmos	Motivo para o trancamento ou desistência

Fonte: Autoria própria (2025)

3.4 Técnicas de Análise de Dados

As técnicas empregadas neste trabalho foram selecionadas com base na análise dos trabalhos relacionados, discutidos na Seção 2.3, os quais abordam temáticas semelhantes à proposta desta pesquisa. O Quadro 5 apresenta um resumo das técnicas identificadas nesses estudos.

Quadro 5 – Técnicas abordadas nos trabalhos relacionados

REFERÊNCIA	TÉCNICA
Pereira, Carvalho e Souto (2019)	Teste de Shapiro-Wilk
	Teste exato de Fisher
	Teste de medidas de associação baseadas no coeficiente de contingência de Pearson
Caldas (2020)	Modelo de regressão logística
	K-means
	Teste chi-quadrado
Oliveira (2021)	Naive Bayes
	CART Decision Tree
	K-Nearest Neighbors
	J48 (C4.5)
	SHapley Additive exPlanation
	Matriz de correlação
Pereira (2021)	Árvores de Decisão
	Random Forest
	Adaboost
	Neural Network
	Naive Bayes
	Support Vector Machine
	K-Nearest Neighbors
Rezende et al. (2022)	Árvores de Decisão
	Coeficiente de correlação de Pearson

Fonte: Autoria própria (2025)

Segundo Soares (2020 *apud* Wirth; Hipp, 2000), não existe uma única técnica ou algoritmo capaz de obter melhor desempenho em todos os domínios de aplicação. Isso significa que os algoritmos devem ser escolhidos com base na aderência ao problema a ser tratado, sempre com o objetivo de alcançar os resultados desejados. A escolha, portanto, depende da adequação ao problema estudado, das características da base de dados e dos objetivos da análise.

Nesse contexto, os algoritmos e métodos estatísticos adotados neste trabalho foram selecionados considerando sua recorrência e eficácia demonstradas nas pesquisas analisadas, bem como a compatibilidade dessas técnicas com a natureza dos dados, a ferramenta utilizada e o objetivo central da pesquisa: verificar se há influência de fatores socioeconômicos no desempenho acadêmico dos discentes no componente curricular de Algoritmos. As subseções a seguir apresentam as técnicas adotadas na realização deste trabalho.

3.4.1 Coeficiente de Correlação de Pearson

O Coeficiente de Correlação de Pearson (*r*) expressa uma medida de associação linear entre duas variáveis, refletindo o grau e a direção do relacionamento entre elas. Essa técnica permite estudar a relação entre dois conjuntos de valores, quantificando o quanto uma variável está relacionada com a outra, determinando, assim, a intensidade e a direção dessa relação.

O coeficiente de correlação assume valores entre -1 e +1. O sinal indica a direção (positiva ou negativa) da correlação, enquanto o valor numérico representa sua intensidade (Rezende et al., 2022). Segundo Rezende et al. (2022, *apud* Cohen, 2013), valores entre 0,10 e 0,29 indicam correlação fraca; entre 0,30 e 0,49, correlação moderada; e valores iguais ou acima de 0,50, correlação forte. Neste trabalho, as análises de correlação foram realizadas utilizando as ferramentas WEKA.

3.4.2 Árvore de Decisão

As árvores de decisão utilizam uma estrutura em forma de árvore para representar diversos caminhos de decisão e os respectivos resultados (Grus, 2016). São técnicas aplicadas para auxiliar na tomada de decisões e na inferência de valores categóricos. Essas estruturas recebem como entrada um conjunto de atributos e produzem como saída uma decisão, ou seja, a classificação da entrada em uma das categorias possíveis.

A construção de uma árvore de decisão inicia-se com a escolha do atributo mais relevante para a predição, geralmente aquele com maior relação com a classe-alvo, que se torna o nó raiz da árvore. O conhecimento é representado de forma hierárquica e visualmente estruturada, o que torna essa abordagem intuitiva e de fácil compreensão (Pereira, 2021). Neste trabalho foram aplicados os seguintes modelos de árvores de decisão no WEKA: J48 (baseado no C4.5), REPTree e LMT (*Logistic Model Tree*).

3.4.3 Random Forest

O Random Forest consiste em um conjunto de árvores de decisão combinadas de modo a gerar um classificador final robusto. Os dados são divididos aleatoriamente em subconjuntos, e cada árvore é construída a partir de um subconjunto de atributos selecionados aleatoriamente (PEREIRA, 2021). Neste trabalho, foi utilizado o modelo Random Forest, disponível na ferramenta WEKA.

3.4.4 Naive Bayes

O Naive Bayes consiste é um classificador estatístico baseado no Teorema de Bayes, proposto por Thomas Bayes. É considerado "ingênuo" (*naive*) porque parte da suposição de independência entre os atributos, ou seja, assume que o efeito de um atributo sobre a variável de classe é independente dos efeitos dos demais atributos.

O algoritmo calcula a probabilidade de um exemplo pertencer a cada classe com base nas probabilidades observadas nos dados de treinamento. A classificação ocorre com base na maior probabilidade condicional observada (Silva et al., 2016). Neste trabalho, foi utilizado o modelo Naive Bayes, disponível no WEKA.

3.4.5 Síntese das Técnicas Selecionadas

Considerando as características da base de dados, os objetivos da pesquisa e os recursos oferecidos pelo software WEKA, foram selecionadas técnicas estatísticas e algoritmos de classificação supervisionada comumente utilizados em estudos educacionais. O Quadro 6 apresenta uma síntese dos métodos empregados neste trabalho.

Quadro 6 – Técnicas estatísticas e algoritmos de classificação utilizados na pesquisa

TIPO DE TÉCNICA	**NOME DO MÉTODO OU ALGORITMO**
Estatística Descritiva	Coeficiente de Correlação de Pearson
Classificação Supervisionada	J48
	REPTree
	LMT
	Random Forest
	Naive Bayes

Fonte: Autoria própria (2025)

3.4.6 Métricas de Avaliação dos Modelos de Classificação

Após a escolha dos algoritmos de classificação, torna-se fundamental definir as métricas utilizadas para avaliar o desempenho dos modelos gerados. Diversas métricas são aplicáveis, sendo selecionadas conforme os objetivos da análise e as características da base de dados (Soares, 2020). Neste estudo, foram adotadas métricas clássicas amplamente empregadas em problemas de classificação supervisionada, além de um método de validação para garantir a robustez das previsões.

A matriz de confusão foi o ponto de partida para a avaliação dos modelos. Ela organiza os resultados das predições em quatro categorias: Verdadeiros Positivos (VP), quando instâncias positivas são corretamente classificadas; Falsos Positivos (FP), quando instâncias negativas são incorretamente classificadas como positivas; Verdadeiros Negativos (VN), que representam instâncias negativas corretamente classificadas; e Falsos Negativos (FN), quando instâncias positivas são classificadas como negativas (Soares, 2020; Maione, 2020).

A partir desses valores, derivam-se outras métricas importantes. A acurácia, por exemplo, representa a proporção de classificações corretas sobre o total de classificações realizadas, sendo calculada como: Acurácia = (VP + VN) / (VP + VN + FP + FN) (Soares, 2020).

A *precision* (ou precisão) mede a confiabilidade das predições positivas do modelo, ou seja, quantas das instâncias classificadas como positivas realmente pertencem a essa classe. Sua fórmula é: *Precision* = VP / (VP + FP) (Bruce, 2019; Soares, 2020).

Já o *recall*, também conhecido como sensibilidade, indica a capacidade do modelo de identificar corretamente as instâncias positivas, sendo dado por: *Recall* = VP / (VP + FN) (Bruce, 2019; Silva et al., 2016).

Para equilibrar precisão e recall, utiliza-se a *F-Measure*, que corresponde à média harmônica entre essas duas métricas. Quanto mais próximo de 1, melhor o desempenho do modelo; valores próximos de 0 indicam desempenho insatisfatório. Sua fórmula é: *F-Measure* = (2 × *Recall* × *Precision*) / (*Recall* + *Precision*) (Witten et al., 2017; Soares, 2020).

Além dessas métricas, foi utilizada a estatística Kappa, que avalia o grau de concordância entre as classificações previstas pelo modelo e as observações reais, considerando também o que seria esperado por acaso. O valor de Kappa varia de -1 a 1: valores próximos de 1 indicam concordância quase perfeita; valores próximos de 0 indicam concordância nula; e valores negativos indicam discordância (Soares, 2020). O Quadro 7 apresenta a interpretação dos valores de Kappa, conforme a escala proposta por Koch e Landis (1997).

Quadro 7 – Rótulos da força de concordância do índice Kappa

ESTATÍSTICA KAPPA	FORÇA DA CONCORDÂNCIA
< 0.00	Pobre
0.00 a 0.20	Leve
0.21 a 0.40	Justa
0.41 a 0.60	Moderada
0.61 a 0.80	Substancial
0.81 a 1.00	Quase Perfeito

Fonte: Adaptado de Soares (2020 *apud* Koch; Landis, 1997)

Além da definição das métricas, foi necessário estabelecer um método de validação dos modelos. Independentemente da métrica escolhida, é inadequado avaliar um modelo com base apenas em seu desempenho sobre os dados de treinamento, pois isso pode levar ao fenômeno de sobreajuste (*overfitting*), no qual o modelo se ajusta excessivamente aos dados que já conhece e perde capacidade de generalização (Silva et al., 2016).

Para mitigar esse risco, foi adotada a técnica de validação cruzada. Nesse método, o conjunto de dados é dividido em K subconjuntos disjuntos. Em cada iteração, um dos subconjuntos é reservado para teste, enquanto os demais compõem o conjunto de treinamento. Esse processo é repetido K vezes, alternando o subconjunto de teste em cada rodada, o que assegura que todos os exemplos sejam utilizados tanto para treinamento quanto para validação (Silva et al., 2016).

4 RESULTADOS E DISCUSSÃO

Este capítulo está organizado da seguinte forma: a Seção 4.1 apresenta os resultados da análise de correlação; a Seção 4.2 descreve os resultados dos modelos de classificação; e a Seção 4.3 apresenta uma discussão integrada dos achados.

4.1 Resultados da Análise de Correlação de Pearson

A aplicação do Coeficiente de Correlação de Pearson (*r*) permitiu identificar as variáveis que apresentaram associação linear com a variável situacao_final, a qual representa a aprovação ou reprovação dos estudantes no componente curricular de Algoritmos. A Figura 1 exibe os valores dos coeficientes gerados no WEKA. Ressalta-se que foram incluídas na Figura 1 apenas as variáveis que apresentaram correlação minimamente relevante com a variável situação_final.

Figura 1 – Coeficientes *r* entre as variáveis preditoras e a variável situação_final

```
Ranked      attributes:
 0.4942     possui_computador
 0.3424     genero_aluno
 0.336      trabalha_atualmente
 0.3223     tipo_escola_ensino_medio
 0.2816     horas_trabalho_semanal
 0.2698     primeira_vez_disciplina_algoritmos
 0.2483     motivo_trancou_desistiu_algoritmos
 0.2188     idade_aluno
 0.2125     trancou_desistiu_algoritmos
 0.153      semestre_ingresso_ufersa
 0.1159     mora_em_pau_dos_ferros
 0.1121     municipio_residencia
```

Fonte: Autoria própria (2025)

O Quadro 8 resume as variáveis com correlações moderadas positivas (*r* entre 0,30 e 0,49), consideradas mais relevantes. As demais variáveis, ilustradas na Figura 1, apresentaram correlações fracas positivas.

Quadro 8 – Variáveis com correlação moderada positiva com a variável situação_final

VARIÁVEL	*r*	INTERPRETAÇÃO
possui_computador	0.4942	Correlação moderada positiva
genero_aluno	0.3424	
trabalha_atualmente	0.336	
tipo_escola_ensino_medio	0.3223	

Fonte: Autoria própria (2025)

Os resultados apresentados no Quadro 8 indicam que determinados fatores socioeconômicos apresentam associação com o desempenho dos estudantes no componente

curricular de Algoritmos. Dentre esses fatores, destacam-se: a posse de computador, o gênero do aluno, a condição de estar ou não trabalhando concomitantemente aos estudos e o tipo de escola em que o estudante concluiu o ensino médio.

4.2 Resultados dos Modelos de Classificação

As subseções a seguir apresentam os resultados obtidos pelos modelos de classificação adotados neste trabalho.

4.2.1 Modelo J48

O modelo J48 apresentou uma acurácia geral de 80,8%, indicando boa capacidade de prever corretamente a situação dos alunos. A estatística Kappa, com valor de 0,6173, reforça essa avaliação, evidenciando um nível substancial de concordância entre as previsões do modelo e os resultados reais. A Figura 2 exibe a árvore gerada pelo modelo J48 no WEKA.

Figura 2 – Resultado do modelo J48

```
possui_computador = sim
|   trabalha_atualmente = nao trabalho: apr
|   trabalha_atualmente = sim
|   |   idade_aluno = 26 - 30 anos: rep
|   |   idade_aluno = 20 - 25 anos: apr
|   |   idade_aluno = menos de 20: apr
|   |   idade_aluno = 31 - 35 anos: rep
|   |   idade_aluno = 36 - 40 anos: rep
possui_computador = nao: rep
```

Fonte: Autoria própria (2025)

A árvore de decisão gerada forneceu indícios relevantes sobre os fatores que mais influenciaram a situação final dos estudantes. As principais variáveis identificadas foram:

- possui_computador: este foi o fator mais relevante, ocupando o nó raiz da árvore. Alunos que não possuem computador foram diretamente classificados como reprovados.
- trabalha_atualmente: para os alunos que possuem computador, o fato de estarem ou não trabalhando foi determinante. Aqueles que não trabalham apresentaram maior probabilidade de aprovação.
- idade_aluno: entre os estudantes que possuem computador e trabalham, a idade teve impacto significativo. Alunos com idades entre 26 e 40 anos foram classificados como reprovados, enquanto os mais jovens (até 25 anos) foram, majoritariamente, classificados como aprovados.

A Figura 3 ilustra a matriz de confusão do modelo J48, proporcionando uma visão geral do desempenho do classificador.

Figura 3 – Matriz de confusão do modelo J48

```
 a  b   <-- classified as
59 22 |  a = apr
 9 72 |  b = rep
```

Fonte: Autoria própria (2025)

Observa-se na Figura 3 que modelo J48 classificou corretamente 59 alunos aprovados e 72 alunos reprovados. Os erros foram de 22 e 9 casos, respectivamente. O Quadro 9 resume as métricas por classe.

Quadro 9 – Métricas por classe do modelo J48

CLASSE	***PRECISION***	***RECALL***	***F-MEASURE***
Aprovado	0,868	0,728	0,792
Reprovado	0,766	0,889	0,823

Fonte: Autoria própria (2025)

Essa distribuição evidencia o comportamento do modelo em relação a cada classe, apresentando as seguintes interpretações das métricas de desempenho:

- Alunos aprovados (apr): para a esta classe, a métrica *Precision* foi de 0,868, indicando que 86,8% das previsões de aprovação estavam corretas. O *Recall* foi de 0,728, o que significa que o modelo conseguiu identificar corretamente 72,8% dos alunos que, de fato, foram aprovados. O *F-Measure* foi de 0,792, demonstrando um bom equilíbrio entre as métricas.
- Alunos Reprovados (rep): para esta classe, a *Precision* foi de 0,766, o que indica que 76,6% das previsões de reprovação estavam corretas, e que em 23,4% das vezes o modelo previu erroneamente a reprovação de alunos que haviam sido aprovados. O *Recall* foi de 0,889, revelando que o modelo identificou corretamente 88,9% dos alunos reprovados. O *F-Measure* registrado para essa classe foi de 0,823, evidenciando um desempenho consistente entre as métricas.

4.2.2 Modelo REPTree

A aplicação do modelo REPTree resultou em um modelo com desempenho satisfatório na predição da situação final dos estudantes no componente de Algoritmos. A acurácia geral foi de 74,6%, indicando que a maior parte das previsões realizadas pelo modelo correspondeu

à realidade observada. Além disso, a estatística Kappa de 0,4938 aponta para um nível moderado de concordância entre os valores previstos e os reais, reforçando a confiabilidade do modelo. A Figura 4 apresenta a estrutura da árvore de decisão gerada no WEKA.

Figura 4 – Resultado do modelo REPTree

```
municipio_residencia = alexandria-rn : apr
municipio_residencia = doutor-severiano-rn : apr
municipio_residencia = vicosa-rn : apr
municipio_residencia = severiano-melo : apr
municipio_residencia = pau-dos-ferros-rn
|   semestre_ingresso_ufersa = 2020_2 : rep
|   semestre_ingresso_ufersa = 2021_1
|   |   idade_aluno = 26 - 30 anos : rep
|   |   idade_aluno = 20 - 25 anos : rep
|   |   idade_aluno = menos de 20 : apr
|   |   idade_aluno = 31 - 35 anos : rep
|   |   idade_aluno = 36 - 40 anos : rep
|   semestre_ingresso_ufersa = 2020_1 : rep
|   semestre_ingresso_ufersa = 2021_2 : apr
|   semestre_ingresso_ufersa = 2022_1 : apr
|   semestre_ingresso_ufersa = 2019_2 : rep
|   semestre_ingresso_ufersa = 2016_2 : apr
municipio_residencia = parana-rn : rep
municipio_residencia = encanto-rn : apr
municipio_residencia = piloes-rn
|   idade_aluno = 26 - 30 anos : rep
|   idade_aluno = 20 - 25 anos : apr
|   idade_aluno = menos de 20 : rep
|   idade_aluno = 31 - 35 anos : rep
|   idade_aluno = 36 - 40 anos : rep
municipio_residencia = iracema-ce : apr
municipio_residencia = riacho-da-cruz-rn : rep
municipio_residencia = belem-do-brejo-do-cruz-pb : apr
municipio_residencia = portalegre-rn : apr
municipio_residencia = erere-ce : apr
municipio_residencia = martins-rn : apr
municipio_residencia = antonio-martins-rn : apr
municipio_residencia = riacho-da-cruz- rn : rep
municipio_residencia = sao-miguel-rn : apr
municipio_residencia = lucrecia-rn : apr
municipio_residencia = tenente-ananias-rn : apr
municipio_residencia = angicos-rn : apr
municipio_residencia = jose-da-penha-rn : rep
municipio_residencia = frutuoso-gomes-rn : apr
municipio_residencia = castanhal-pa : apr
municipio_residencia = sao-joao-do-sabugi-rn : apr
municipio_residencia = riacho-de-santana-rn : apr
municipio_residencia = uirauna-pb : apr
municipio_residencia = luis-gomes-rn : rep
municipio_residencia = messias-targino-rn : rep
municipio_residencia = potiretama-ce : rep
municipio_residencia = agua-nova-rn : apr
municipio_residencia = rafael-fernandes-rn : rep
municipio_residencia = apodi-rn : apr
municipio_residencia = sao-paulo-sp : apr
municipio_residencia = itau-rn : apr
municipio_residencia = francisco-dantas-rn : apr
municipio_residencia = ico-ceará : rep
municipio_residencia = major-sales-rn : rep
```

Fonte: Autoria própria (2025)

A estrutura da árvore de decisão gerada pelo modelo REPTree permitiu identificar relações relevantes entre variáveis socioeconômicas e acadêmicas associadas à situação final dos estudantes. As principais variáveis identificadas foram:

- municipio_residencia: foi a variável mais relevante, ocupando o nó raiz da árvore. Para a maioria dos municípios, o modelo realizou classificações diretas quanto à aprovação ou reprovação. No entanto, nos casos dos municípios de Pau dos Ferros e Pilões, a árvore apresentou uma estrutura mais aprofundada, revelando relações com outras variáveis. Esse aprofundamento sugere a necessidade de uma análise complementar, com o objetivo de investigar os fatores específicos que contribuíram para esse comportamento no modelo.
- semestre_ingresso_ufersa: para alunos residentes em Pau dos Ferros, o semestre de ingresso na UFERSA tornou-se o próximo critério decisório. Alunos que ingressaram em semestres como 2019.2, 2020.1 e 2020.2 foram, majoritariamente, classificados como reprovados. Por outro lado, os ingressantes de semestres como

2016.2, 2021.2 e 2022.1 apresentaram tendência à aprovação. Esse resultado aponta para a necessidade de uma investigação mais aprofundada, especialmente no que se refere ao possível impacto do contexto da pandemia da COVID-19 sobre o desempenho acadêmico dos estudantes que ingressaram nesse período.

- idade_aluno: essa variável também apareceu como relevante para alunos de Pau dos Ferros e Pilões. No primeiro caso, estudantes que ingressaram no semestre 2021.1 e tinham menos de 20 anos foram classificados como aprovados, enquanto aqueles com 20 anos ou mais tenderam à reprovação. Situação semelhante foi observada em Pilões: estudantes com idade entre 20 e 25 anos foram classificados como aprovados, ao passo que faixas etárias superiores apresentaram maior probabilidade de reprovação.

A matriz de confusão do modelo, apresentada na Figura 5, permite visualizar o desempenho do classificador em relação a cada classe.

Figura 5 – Matriz de confusão do modelo REPTree

```
 a  b   <-- classified as
61 20 |  a = apr
21 60 |  b = rep
```

Fonte: Autoria própria (2025)

Observa-se na Figura 5 que modelo REPTree classificou corretamente 61 alunos aprovados e 60 alunos reprovados. Os erros foram de 20 e 21 casos, respectivamente. O Quadro 10 resume as métricas por classe.

Quadro 10 – Métricas por classe do modelo REPTree

CLASSE	***PRECISION***	***RECALL***	***F-MEASURE***
Aprovado	0,744	0,753	0,748
Reprovado	0,750	0,741	0,745

Fonte: Autoria própria (2025)

Essa distribuição evidencia o comportamento do modelo em relação a cada classe, apresentando as seguintes interpretações das métricas de desempenho:

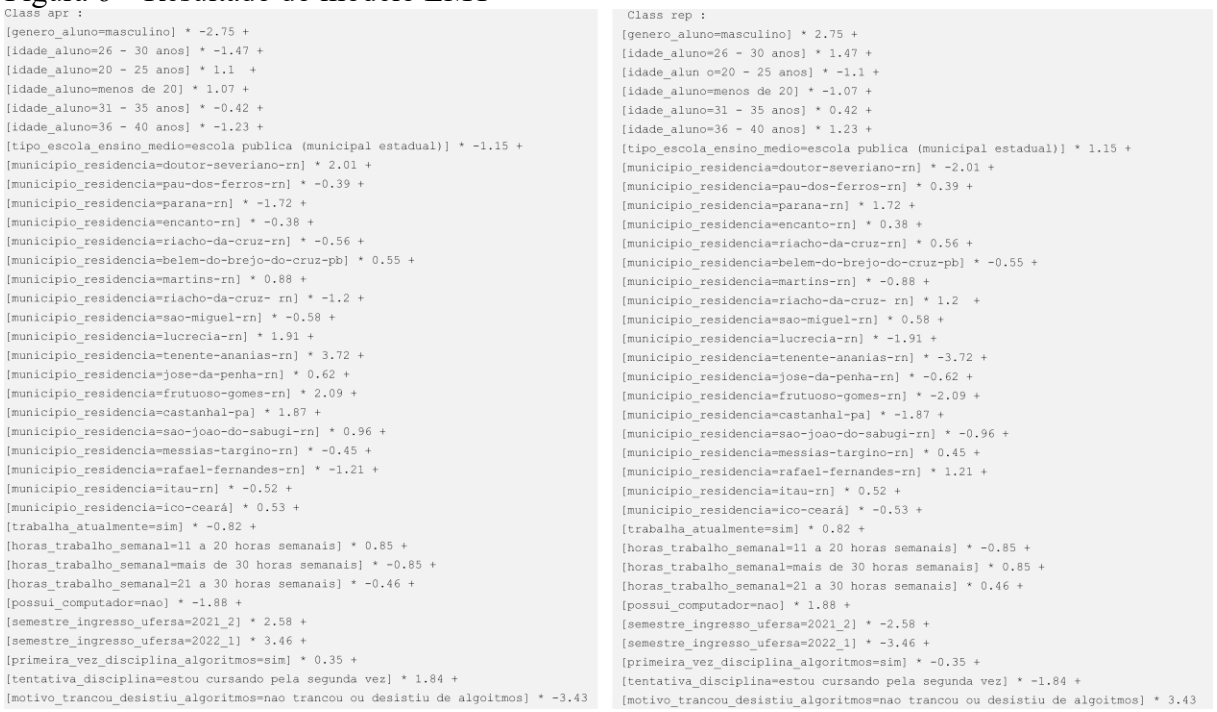
- Alunos aprovados (apr): a *Precision* foi de 0,744, indicando que 74,4% das previsões de aprovação estavam corretas. O *Recall* foi de 0,753, o que significa que o modelo conseguiu identificar corretamente 75,3% dos alunos que, de fato, foram aprovados. O *F-Measure* foi de 0,748, evidenciando um desempenho consistente nessa classe.

- Alunos reprovados (rep): a *Precision* foi de 0,750, o que indica que 75,0% das previsões de reprovação estavam corretas, e que em 25,0% das vezes o modelo previu erroneamente a reprovação de alunos que haviam sido aprovados. O *Recall* foi de 0,741, revelando que o modelo identificou corretamente 74,1% dos alunos reprovados. O *F-Measure* registrado para essa classe foi de 0,745, demonstrando também um bom equilíbrio entre as métricas.

4.2.3 Modelo LMT

O modelo LMT gerou uma estrutura simplificada, resultando em uma única regressão logística aplicada a todo o conjunto de dados. O algoritmo apresentou uma acurácia geral de 87,6%, evidenciando alta capacidade de prever corretamente a situação final dos alunos. A estatística Kappa foi de 0,7531, valor que, segundo os critérios de interpretação da métrica, indica um nível substancial de concordância entre as previsões do modelo e os dados observados. A Figura 6 apresenta o resultado do modelo LMT gerado no WEKA.

Figura 6 – Resultado do modelo LMT



```
Class apr :
[genero_aluno=masculino] * -2.75 +
[idade_aluno=26 - 30 anos] * -1.47 +
[idade_aluno=20 - 25 anos] * 1.1  +
[idade_aluno=menos de 20] * 1.07 +
[idade_aluno=31 - 35 anos] * -0.42 +
[idade_aluno=36 - 40 anos] * -1.23 +
[tipo_escola_ensino_medio=escola publica (municipal estadual)] * -1.15 +
[municipio_residencia=doutor-severiano-rn] * 2.01 +
[municipio_residencia=pau-dos-ferros-rn] * -0.39 +
[municipio_residencia=parana-rn] * -1.72 +
[municipio_residencia=encanto-rn] * -0.38 +
[municipio_residencia=riacho-da-cruz-rn] * -0.56 +
[municipio_residencia=belem-do-brejo-do-cruz-pb] * 0.55 +
[municipio_residencia=martins-rn] * 0.88 +
[municipio_residencia=riacho-da-cruz- rn] * -1.2 +
[municipio_residencia=sao-miguel-rn] * -0.58 +
[municipio_residencia=lucrecia-rn] * 1.91 +
[municipio_residencia=tenente-ananias-rn] * 3.72 +
[municipio_residencia=jose-da-penha-rn] * 0.62 +
[municipio_residencia=frutuoso-gomes-rn] * 2.09 +
[municipio_residencia=castanhal-pa] * 1.87 +
[municipio_residencia=sao-joao-do-sabugi-rn] * 0.96 +
[municipio_residencia=messias-targino-rn] * -0.45 +
[municipio_residencia=rafael-fernandes-rn] * -1.21 +
[municipio_residencia=itau-rn] * -0.52 +
[municipio_residencia=ico-ceará] * 0.53 +
[trabalha_atualmente=sim] * -0.82 +
[horas_trabalho_semanal=11 a 20 horas semanais] * 0.85 +
[horas_trabalho_semanal=mais de 30 horas semanais] * -0.85 +
[horas_trabalho_semanal=21 a 30 horas semanais] * -0.46 +
[possui_computador=nao] * -1.88 +
[semestre_ingresso_ufersa=2021_2] * 2.58 +
[semestre_ingresso_ufersa=2022_1] * 3.46 +
[primeira_vez_disciplina_algoritmos=sim] * 0.35 +
[tentativa_disciplina=estou cursando pela segunda vez] * 1.84 +
[motivo_trancou_desistiu_algoritmos=nao trancou ou desistiu de algoitmos] * -3.43
```

```
 Class rep :
[genero_aluno=masculino] * 2.75 +
[idade_aluno=26 - 30 anos] * 1.47 +
[idade_alun o=20 - 25 anos] * -1.1 +
[idade_aluno=menos de 20] * -1.07 +
[idade_aluno=31 - 35 anos] * 0.42 +
[idade_aluno=36 - 40 anos] * 1.23 +
[tipo_escola_ensino_medio=escola publica (municipal estadual)] * 1.15 +
[municipio_residencia=doutor-severiano-rn] * -2.01 +
[municipio_residencia=pau-dos-ferros-rn] * 0.39 +
[municipio_residencia=parana-rn] * 1.72 +
[municipio_residencia=encanto-rn] * 0.38 +
[municipio_residencia=riacho-da-cruz-rn] * 0.56 +
[municipio_residencia=belem-do-brejo-do-cruz-pb] * -0.55 +
[municipio_residencia=martins-rn] * -0.88 +
[municipio_residencia=riacho-da-cruz- rn] * 1.2  +
[municipio_residencia=sao-miguel-rn] * 0.58 +
[municipio_residencia=lucrecia-rn] * -1.91 +
[municipio_residencia=tenente-ananias-rn] * -3.72 +
[municipio_residencia=jose-da-penha-rn] * -0.62 +
[municipio_residencia=frutuoso-gomes-rn] * -2.09 +
[municipio_residencia=castanhal-pa] * -1.87 +
[municipio_residencia=sao-joao-do-sabugi-rn] * -0.96 +
[municipio_residencia=messias-targino-rn] * 0.45 +
[municipio_residencia=rafael-fernandes-rn] * 1.21 +
[municipio_residencia=itau-rn] * 0.52 +
[municipio_residencia=ico-ceará] * -0.53 +
[trabalha_atualmente=sim] * 0.82 +
[horas_trabalho_semanal=11 a 20 horas semanais] * -0.85 +
[horas_trabalho_semanal=mais de 30 horas semanais] * 0.85 +
[horas_trabalho_semanal=21 a 30 horas semanais] * 0.46 +
[possui_computador=nao] * 1.88 +
[semestre_ingresso_ufersa=2021_2] * -2.58 +
[semestre_ingresso_ufersa=2022_1] * -3.46 +
[primeira_vez_disciplina_algoritmos=sim] * -0.35 +
[tentativa_disciplina=estou cursando pela segunda vez] * -1.84 +
[motivo_trancou_desistiu_algoritmos=nao trancou ou desistiu de algoitmos] * 3.43
```

Fonte: Autoria própria (2025)

A regressão logística gerada pelo modelo LMT forneceu *insights* relevantes sobre os fatores que mais influenciaram a situação final dos estudantes. As principais variáveis identificadas foram:

- genero_aluno: estudantes do gênero masculino apresentaram menor probabilidade de aprovação. E resultado sugere uma análise mais aprofundada sobre as possíveis disparidades de gênero no desempenho discente no componente de Algoritmos.
- idade_aluno: alunos mais jovens, especialmente aqueles com até 25 anos, tiveram maior chance de aprovação. Em contraste, discentes nas faixas etárias de 26 a 40 anos demonstraram menor probabilidade de sucesso no componente curricular de Algoritmos.
- tipo_escola_ensino_medio: alunos que frequentaram o ensino médio em escolas públicas estaduais ou municipais apresentaram menores taxas de aprovação.
- municipio_residencia: as probabilidades de aprovação variaram conforme o município de residência. Algumas localidades apresentaram associação positiva, enquanto outras estiveram vinculadas a uma menor chance de aprovação.
- trabalha_atualmente e horas_trabalho_semanal: a variável trabalha_atualmente impactou negativamente a chance de aprovação, especialmente entre os estudantes que trabalham mais de 20 horas semanais. Já para aqueles com carga horária de trabalho entre 11 e 20 horas, o modelo associou uma probabilidade positiva de aprovação.
- possui_computador: não possuir computador foi um fator com impacto negativo na probabilidade de aprovação.
- semestre_ingresso_ufersa: esta variável indicou que os estudantes que ingressaram nos semestres 2021.1 e 2022.1 apresentaram maior probabilidade de aprovação no componente de Algoritmos.
- primeira_vez_disciplina_algoritmos e tentativa_disiciplina: estudantes que estavam cursando o componente pela primeira ou segunda vez tiveram maiores chances de aprovação.
- motivo_trancou_desistiu_algoritmos: esta variável apresentou um comportamento inesperado, exibindo coeficiente negativo em relação à aprovação. Esse resultado sugere a necessidade de uma investigação mais aprofundada quanto à sua codificação, à forma como foi interpretada no conjunto de dados e à sua possível alteração após a aplicação do algoritmo de balanceamento de classes SMOTE.

A matriz de confusão do modelo, apresentada na Figura 7, permite visualizar o desempenho do modelo em relação a cada classe.

Figura 7 – Matriz de confusão do modelo LMT

```
 a  b   <-- classified as
69 12 |  a = apr
 8 73 |  b = rep
```

Fonte: Autoria própria (2025)

Observa-se na Figura 7 que modelo LMT classificou corretamente 69 alunos aprovados e 73 alunos reprovados. Os erros foram de 12 e 8 casos, respectivamente. O Quadro 11 resume as métricas por classe.

Quadro 11 – Métricas por classe do modelo LMT

CLASSE	***PRECISION***	***RECALL***	***F-MEASURE***
Aprovado	0,896	0,852	0,873
Reprovado	0,859	0,901	0,880

Fonte: Autoria própria (2025)

Essa distribuição evidencia o comportamento do modelo em relação a cada classe, apresentando as seguintes interpretações das métricas de desempenho:

- Alunos aprovados (apr): a *Precision* foi de 0,896, indicando que 89,6% das previsões de aprovação estavam corretas. O *Recall* foi de 0,852, o que significa que o modelo conseguiu identificar corretamente 85,2% dos alunos que, de fato, foram aprovados. O *F-Measure* foi de 0,873, evidenciando um desempenho consistente nessa classe.
- Alunos reprovados (rep): a *Precision* foi de 0,859, o que indica que 85,9% das previsões de reprovação estavam corretas, e que em 14,1% das vezes o modelo previu erroneamente a reprovação de alunos que haviam sido aprovados. O *Recall* foi de 0,901, revelando que o modelo identificou corretamente 90,1% dos alunos reprovados. O *F-Measure* registrado para essa classe foi de 0,880, demonstrando também um bom equilíbrio entre as métricas.

4.2.4 Modelo Random Forest

O modelo apresentou uma acurácia de 83,9%, o que indica um bom desempenho na predição da situação final dos alunos. A estatística Kappa, com valor de 0,679, reforça essa avaliação, representando um nível substancial de concordância entre as previsões do modelo e os resultados reais.

A Figura 8 exibe o *output* com os valores de importância atribuídos a cada atributo pelo modelo.

Figura 8 – Resultado do modelo Random Forest

Attribute importance	attributes:
0.58	municipio_residencia
0.58	horas_trabalho_semanal
0.51	trancou_desistiu_algoritmos
0.5	idade_aluno
0.44	semestre_ingresso_ufersa
0.43	trabalha_atualmente
0.41	possui_computador
0.35	primeira_vez_disciplina_algoritmos
0.32	motivo_trancou_desistiu_algoritmos
0.31	tentativa_disciplina
0.3	tipo_escola_ensino_medio
0.28	genero_aluno
0.16	mora_em_pau_dos_ferros

Fonte: Autoria própria (2025)

O modelo Random Forest forneceu resultados relevantes quanto à importância relativa dos atributos na predição da situação final dos alunos no componente curricular de Algoritmos. As variáveis mais influentes foram as seguintes:

- municipio_residencia e horas_trabalho_semanal: essas duas variáveis se destacaram como as mais relevantes, indicando que tanto a localização geográfica quanto a carga horária dedicada ao trabalho exercem maior influência no desempenho dos estudantes no componente.
- trancou_desistiu_algoritmos: o modelo atribuiu importância significativa a essas variáveis, sugerindo que experiências anteriores com o componente curricular constituem indicativos relevantes de risco de reprovação.
- idade_aluno: a faixa etária também se mostrou determinante, refletindo padrões semelhantes aos observados em outros modelos.
- semestre_ingresso_ufersa, trabalha_atualmente e possui_computador: esses fatores também tiveram papel relevante nas decisões do modelo, reforçando a influência de aspectos estruturais e contextuais no processo de aprendizagem.

A matriz de confusão do modelo, apresentada na Figura 9, permite visualizar o desempenho do modelo em relação a cada classe.

Figura 9 – Matriz de confusão do modelo Random Forest

```
 a  b   <-- classified as
66 15 |  a = apr
11 70 |  b = rep
```

Fonte: Autoria própria (2025)

Observa-se na Figura 9 que modelo Random Forest classificou corretamente 66 alunos aprovados e 70 alunos reprovados. Os erros foram de 15 e 11 casos, respectivamente. O Quadro 12 resume as métricas por classe.

Quadro 12 – Métricas por classe do modelo Random Forest

CLASSE	***PRECISION***	***RECALL***	***F-MEASURE***
Aprovado	0,857	0,815	0,835
Reprovado	0,824	0,864	0,843

Fonte: Autoria própria (2025)

Essa distribuição evidencia o comportamento do modelo em relação a cada classe, apresentando as seguintes interpretações das métricas de desempenho:

- Alunos aprovados (apr): a *Precision* foi de 0,857, indicando que 85,7% das previsões de aprovação estavam corretas. O *Recall* foi de 0,815, o que significa que o modelo conseguiu identificar corretamente 81,5% dos alunos que, de fato, foram aprovados. O *F-Measure* foi de 0,835, evidenciando um desempenho consistente nessa classe.
- Alunos reprovados (rep): a *Precision* foi de 0,824, o que indica que 82,4% das previsões de reprovação estavam corretas, e que em 17,6% das vezes o modelo previu erroneamente a reprovação de alunos que haviam sido aprovados. O *Recall* foi de 0,864, revelando que o modelo identificou corretamente 86,4% dos alunos reprovados. O *F-Measure* registrado para essa classe foi de 0,843, demonstrando também um bom equilíbrio entre as métricas.

4.2.5 Modelo Naive Bayes

O modelo Naive Bayes alcançou uma acurácia geral de 85,1%, demonstrando bom desempenho na predição da situação final dos estudantes no componente curricular. A estatística Kappa foi de 0,7037, o que indica um nível substancial de concordância entre as previsões do modelo e os resultados reais.

O modelo Naive Bayes permitiu identificar a influência de algumas variáveis na aprovação ou reprovação dos estudantes no componente curricular de Algoritmos, conforme ilustrado na Figura 10.

Figura 10 – Resultado do modelo Naive Bayes

	Class	
Attribute	apr	rep
possui_computador		
sim	63.0	23.0
nao	20.0	60.0
[total]	83.0	83.0
trabalha_atualmente		
nao trabalho	70.0	45.0
sim	13.0	38.0
[total]	83.0	83.0
idade_aluno		
26 - 30 anos	8.0	22.0
20 - 25 anos	31.0	24.0
menos de 20	44.0	19.0
31 - 35 anos	2.0	13.0
36 - 40 anos	1.0	8.0
[total]	86.0	86.0
genero_aluno		
feminino	18.0	1.0
masculino	65.0	82.0
[total]	83.0	83.0

Fonte: Autoria própria (2025)

Entre os resultados obtidos, destacam-se as seguintes variáveis consideradas mais relevantes:

- possui_computador: este atributo apresentou forte associação com a aprovação. Entre os alunos que possuíam computador, 63 foram aprovados e 23 reprovados. Em contraste, entre aqueles que não possuíam o equipamento, apenas 20 foram aprovados, enquanto 60 foram reprovados. Esses dados reforçam a importância do acesso a recursos tecnológicos para o bom desempenho no componente curricular.
- trabalha_atualmente: estudantes que não trabalhavam apresentaram desempenho superior (70 aprovados e 45 reprovados) em relação aos que estavam empregados, dos quais apenas 13 foram aprovados e 38 reprovados. Esses resultados sugerem que a maior disponibilidade de tempo para estudo pode impactar positivamente o rendimento acadêmico.
- idade_aluno: as faixas etárias mais jovens estiveram associadas a maiores taxas de aprovação. Alunos com menos de 20 anos apresentaram 44 aprovações e 19 reprovações, enquanto aqueles com idades entre 20 e 25 anos registraram 31 aprovações e 24 reprovações. Em contrapartida, faixas etárias mais elevadas mostraram maior tendência à reprovação: estudantes entre 26 e 30 anos tiveram 8

aprovações e 22 reprovações; já na faixa de 31 a 35 anos, foram apenas 2 aprovações contra 13 reprovações.

- genero_aluno: o desempenho dos estudantes também variou de acordo com o gênero. Alunas do gênero feminino apresentaram alto índice de aprovação (18 aprovadas e apenas 1 reprovada), enquanto entre os estudantes do gênero masculino, a reprovação foi predominante (65 aprovados e 82 reprovados).

A Figura 11 apresenta a matriz de confusão gerada pelo modelo, permitindo visualizar os acertos e erros de classificação em cada uma das categorias.

Figura 11 – Matriz de confusão Naive Bayes

```
 a  b   <-- classified as
70 11 |  a = apr
13 68 |  b = rep
```

Fonte: Autoria própria (2025)

Observa-se na Figura 11 que modelo Naive Bayes classificou corretamente 70 alunos aprovados e 68 alunos reprovados. Os erros foram de 11 e 13 casos, respectivamente. O Quadro 13 resume as métricas por classe.

Quadro 13 – Métricas por classe do modelo Naive Bayes

CLASSE	***PRECISION***	***RECALL***	***F-MEASURE***
Aprovado	0,843	0,864	0,854
Reprovado	0,861	0,840	0,850

Fonte: Autoria própria (2025)

Diferentemente dos modelos baseados em árvores de decisão, o Naive Bayes não gera uma estrutura de regras explícitas para visualização das decisões do modelo. No entanto, a análise das métricas por classe permite observar o desempenho do algoritmo na predição de cada categoria de resultado acadêmico:

- Alunos aprovados (apr): a *Precision* foi de 0,843, indicando que 84,3% das previsões de aprovação estavam corretas. O *Recall* foi de 0,864, o que significa que o modelo conseguiu identificar corretamente 86,4% dos alunos que, de fato, foram aprovados. O *F-Measure* foi de 0,854, evidenciando um desempenho consistente nessa classe.
- Alunos reprovados (rep): a *Precision* foi de 0,861, o que indica que 86,1% das previsões de reprovação estavam corretas, e que em 13,9% das vezes o modelo

previu erroneamente a reprovação de alunos que haviam sido aprovados. O *Recall* foi de 0,840, revelando que o modelo identificou corretamente 84,0% dos alunos reprovados. O *F-Measure* registrado para essa classe foi de 0,850, demonstrando também um bom equilíbrio entre as métricas.

4.2.6 Comparação entre os Modelos de Classificação

Os resultados obtidos pelos modelos estão sintetizados no Quadro 14, que resume as principais métricas de avaliação para os cinco modelos de classificação aplicados. Os valores de *Precision*, *Recall* e *F-Measure* correspondem às médias ponderadas calculadas pelo WEKA, refletindo o desempenho geral de cada modelo considerando simultaneamente as duas classes analisadas (aprovado e reprovado).

Quadro 14 – Resumo do desempenho dos modelos de classificação

MODELO	**ACURÁCIA (%)**	**KAPPA**	***PRECISION***	***RECALL***	***F-MEASURE***
LMT	87,65%	0,7531	0,877	0,877	0,876
Naive Bayes	85,19%	0,7037	0,852	0,852	0,852
Random Forest	83,95%	0,6790	0,840	0,840	0,839
J48	80,86%	0,6173	0,817	0,809	0,807
REPTree	74,69%	0,4938	0,747	0,747	0,747

Fonte: Autoria própria (2025)

A partir da análise dos resultados apresentados no Quadro 14, observa-se que o modelo LMT obteve o melhor desempenho geral, com os valores mais elevados em todas as métricas avaliadas. Esses resultados indicam que, para o conjunto de dados utilizado nesta pesquisa, o LMT foi o classificador mais eficaz.

O modelo Naive Bayes também demonstrou desempenho consistente, com métricas próximas às do LMT, evidenciando sua robustez na tarefa de predição. Os modelos Random Forest e J48 apresentaram desempenhos semelhantes entre si, ocupando uma posição intermediária em relação à acurácia e à estatística Kappa. Por outro lado, o modelo REPTree obteve os resultados mais modestos, com os menores valores entre todos os algoritmos avaliados.

4.3 Discussão Integrada dos Resultados

No que diz respeito às variáveis, observa-se uma tendência consistente na identificação de determinados atributos como fatores decisivos para a predição do desempenho dos discentes no componente curricular de Algoritmos. A recorrência dessas variáveis ao longo das diferentes técnicas indica sua relevância para o problema em análise. O Quadro 15 apresenta a frequência com que cada variável foi selecionada como significativa pelas técnicas aplicadas.

Quadro 15 – Frequência das variáveis nas técnicas aplicadas

VARIÁVEL	FREQUÊNCIA	TÉCNICAS
idade_aluno	5/6	J48, REPTree, LMT, Random Forest, Naive Bayes
possui_computador	5/6	Correlação de Pearson, J48, LMT, Random Forest, Naive Bayes
trabalha_atualmente	5/6	Correlação de Pearson, J48, LMT, Random Forest, Naive Bayes
municipio_residencia	3/6	REPTree, LMT, Random Forest
semestre_ingresso_ufersa	3/6	REPTree, LMT, Random Forest
genero_aluno	3/6	Correlação de Pearson, LMT, Naive Bayes
motivo_trancou_desistiu_algoritmos	2/6	LMT, Random Forest
tentativa_disciplina	2/6	LMT, Random Forest

Fonte: Autoria própria (2025)

A análise do Quadro 15 revela que as variáveis idade_aluno, possui_computador e trabalha_atualmente foram as mais recorrentes, sendo consideradas relevantes em cinco das seis técnicas aplicadas. Em seguida, destacam-se as variáveis municipio_residencia, semestre_ingresso_ufersa e genero_aluno, identificadas como significativas em três das seis abordagens avaliadas. Essa consistência na seleção das variáveis reforça a hipótese de que fatores socioeconômicos, como acesso a recursos tecnológicos, disponibilidade de tempo e características demográficas, exercem influência significativa sobre o desempenho discente no componente curricular de Algoritmos.

Ao confrontar os resultados desta pesquisa com os estudos da literatura, observa-se tanto convergência quanto avanços. A relevância do histórico escolar, como o tipo de escola cursada

no ensino médio e o histórico de trancamentos, está em conformidade com as conclusões de Pereira, Carvalho e Souto (2019), Caldas (2020) e Oliveira (2021), que destacam a bagagem educacional como um fator crítico. De forma semelhante, o estudo de Pereira (2021) também apontou a faixa etária como variável associada ao risco de evasão em componentes curriculares de programação, o que reforça os achados desta pesquisa.

Além dessas confirmações, o presente estudo introduz contribuições originais. Diferentemente do trabalho de Rezende et al. (2022), que não observaram correlação significativa entre trabalho e desempenho, esta pesquisa identificou que a variável trabalha_atualmente é um fator de risco para reprovação. Esse resultado evidencia que conciliar trabalho e estudo representa um desafio real para os discentes, afetando negativamente sua trajetória no componente curricular de Algoritmos. Outro achado inédito foi a variável municipio_residencia, apontada como relevante por três dos modelos, sugerindo que desigualdades regionais e logísticas também podem interferir no rendimento acadêmico.

5 CONSIDERAÇÕES FINAIS

Este trabalho teve como objetivo investigar a influência do perfil socioeconômico no desempenho dos estudantes no componente curricular de Algoritmos, componente introdutório do curso de BITI da UFERSA, Campus Pau dos Ferros. A análise foi conduzida por meio da aplicação de técnicas de ciência de dados, com base em estudos correlatos e utilizando um conjunto de variáveis socioeconômicas, acadêmicas e contextuais.

Os achados deste trabalho estão em consonância com estudos anteriores (Pereira; Carvalho; Souto, 2019; Caldas, 2020; Oliveira, 2021; Pereira, 2021), os quais também destacam o papel do histórico educacional e da faixa etária como fatores relevantes. Além disso, esta pesquisa apresenta contribuições originais, como a identificação do município de residência como um fator de influência, além da confirmação do impacto negativo da variável trabalha_atualmente, em contraste com os resultados de Rezende et al. (2022).

Dessa forma, conclui-se que o perfil socioeconômico dos estudantes exerce influência direta e significativa sobre o desempenho no componente curricular de Algoritmos. Fatores como idade, posse de computador e trabalho concomitante ao estudo foram determinantes nas classificações realizadas pelas técnicas, revelando que o sucesso no componente curricular é fortemente condicionado por fatores estruturais e pessoais.

Diante disso, reforça-se a necessidade de implementação de políticas institucionais de apoio que levem em consideração o perfil socioeconômico dos estudantes. Entre as ações recomendadas, destacam-se:

- Programas de empréstimo ou acesso facilitado a equipamentos tecnológicos, com ênfase na disponibilização de computadores para uso acadêmico;
- Apoio específico para estudantes que conciliam trabalho e estudo, como bolsas de permanência;
- Monitorias e atividades de reforço pedagógico direcionadas a perfis de estudantes com maior risco de reprovação, identificados a partir dos fatores analisados neste estudo.

Em síntese, este estudo contribui não apenas para validar a aplicação de técnicas de ciência de dados na análise do desempenho acadêmico, mas também para evidenciar os desafios socioeconômicos enfrentados pelos estudantes. Os achados reforçam que o sucesso no componente curricular de Algoritmos resulta de um processo complexo, influenciado por múltiplas variáveis. Essa complexidade demanda uma atenção das instituições de ensino no planejamento de estratégias eficazes de apoio e permanência estudantil.

5.1 Limitações do Estudo

Embora os resultados apresentados ofereçam contribuições relevantes, é importante reconhecer algumas limitações. A pesquisa foi realizada com estudantes de um único curso, o que restringe a generalização dos achados. A análise concentrou-se em um único componente curricular, limitando a visão sobre o percurso acadêmico completo. Além disso, os dados foram obtidos por meio de autodeclaração, o que pode introduzir vieses de percepção. Por fim, não foram utilizados dados longitudinais, o que impossibilita o acompanhamento do desempenho ao longo do tempo.

5.2 Trabalhos Futuros

Como sugestões para trabalhos futuros, recomenda-se: (i) a análise para outros componentes curriculares e cursos, com o objetivo de verificar se os padrões identificados neste estudo se repetem em diferentes contextos acadêmicos; (ii) ampliar a base de dados por meio da inclusão de um número maior de participantes e variáveis, a fim de aumentar a robustez estatística e a capacidade de generalização dos modelos preditivos; (iii) realizar uma investigação específica sobre o comportamento inesperado da variável motivo_trancou_desistiu_algoritmos, cuja influência negativa sobre a aprovação foi evidenciada pelo modelo LMT; (iv) aprofundar a análise sobre a influência da variável municipio_residencia, dada a sua relevância nos modelos e o potencial reflexo de desigualdades regionais e logísticas; (v) conduzir uma análise exploratória sobre variáveis com coeficientes de correlação mais expressivos, como genero_aluno, a fim de compreender melhor suas possíveis associações com o desempenho acadêmico; e, por fim, (vi) empregar técnicas complementares de análise, como algoritmos de agrupamento (*clustering*) e métodos de *boosting* (como o AdaBoost), para captar padrões mais sutis e identificar perfis de estudantes com maior risco de insucesso, os quais não são imediatamente evidenciados pelos modelos utilizados neste estudo.

REFERÊNCIAS

ANDRADE, D. L. A. **Análise da evasão nos cursos da UFERSA campus Pau dos Ferros:** um estudo de tendência e causas**.** 2023, 56 f. Monografia (Graduação em Bacharelado em Tecnologia da Informação), Universidade Federal Rural do Semi-Árido, Pau dos Ferros, 2023.

BOSSE, Y. **Padrões de dificuldades relacionadas com o aprendizado de programação**. 2020. 270 f. Tese (Doutorado em Ciências) – Universidade de São Paulo, São Paulo, 2020.

BRUCE, P. BRUCE, A. Estatística Prática para Cientistas de Dados: 50 conceitos essenciais. Rio de Janeiro: Alta Books, 2019.

CALDAS, T. A. **Desempenho dos alunos numa disciplina introdutória de programação e o conhecimento longevo**. 2020. 343 f. Dissertação (Mestrado em Ciência da Computação) – Instituto de Computação, Universidade Estadual de Campinas, Campinas, 2020.

DANIEL, E. M. P.; WIVES, L. K.; LORANDI, A. Análise Bibliométrica da Revisão Sistemática da Literatura sobre *Learning Analytics* e *Educational Data Mining* focados na Predição de ocorrências no Sistema Educacional. **EDUCAÇÃO, FORMAÇÃO & TECNOLOGIAS,** [s.l.], v. 12, P. 85-101, 2024. DOI: 10.5281/zenodo.13124179. Disponível em: https://eft.educom.pt/index.php/eft/article/view/229. Acesso em: 4 mar. 2025.

FREITAS, B. C. B.; FONTES, L. M. O.; SILVA, B. G. S. Autoavaliação no Processo de Ensino e Aprendizagem de Programação. Brazilian Journal of Development, Curitiba, v. 8, n. 5, p. 39485-39506, 2022. DOI:10.34117/bjdv8n5-441. Disponível em: https://ojs.brazilianjournals.com.br/ojs/index.php/BRJD/article/view/48365/pdf. Acesso em: 17 abr. 2025.

GRUS, J. Data Science do Zero: primeiras regras com o Python. Rio de Janeiro: Alta Books, 2016.

INSTITUTO SEMESP. **Mapa do Ensino Superior: 13ª edição**. São Paulo: Instituto Semesp, 2023. Disponível em: https://www.semesp.org.br/mapa/edicao-13/. Acesso em: 2 mar. 2025.

INSTITUTO SEMESP. **Mapa do Ensino Superior: 14ª edição**. São Paulo: Instituto Semesp, 2024. Disponível em: https://www.semesp.org.br/mapa/edicao-14/. Acesso em: 2 mar. 2025.

JÚNIOR, R. S. O.; CARDINS, D. V. A.; LIMA, W. J. P.; PADILHA, P. P.; DANTAS, V. F. Análise da ralação entre perfil e desempenho acadêmico dos alunos matriculados na disciplina de introdução à programação utilizando algoritmos de classificação. *In*: WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA (WIDat), 2., 2018, Paraíba. **Anais do II Workshop de informação, Dados e Tecnologia**. Tupã: Faculdade de Ciências e Engenharia UNESP, 2018. Disponível em: https://dadosabertos.info/enhanced_publications/idt/. Acesso em: 8 mar. 2025.

MAIONE, C. **Balanceamento de dados com base em oversampling em dados transformados**. 2020, 135 f. Tese (Doutorado em Computação) – Universidade Federal de Goiás, Goiânia, 2020.

MATOS, M. T.; CONDURÚ, M. T.; BENCHIMOL, A. C. Interseções na produção científica da ciência da informação e ciência de dados. **Acervo**, [s. l.], v. 35, n. 2, p. 1-18, 2022. Disponível em: https://revista.an.gov.br/index.php/revistaacervo/article/view/1804. Acesso em: 8 abr. 2025.

MEDEIROS, R. P.; FALCÃO, T. P.; RAMALHO, G, L. Ensino e Aprendizagem de Introdução à Programação no Ensino Superior Brasileiro: Revisão Sistemática da Literatura. *In*: WORKSHOP SOBRE EDUCAÇÃO EM COMPUTAÇÃO (WEI), 28., 2020, Cuiabá. **Anais do XXVIII Workshop sobre educação em computação**. Porto Alegre: Sociedade Brasileira de Computação, 2020. Disponível em: https://sol.sbc.org.br/index.php/wei/article/view/11155. Acesso em: 28 mar. 2025.

MORAIS, C. G. B. **Ensino e aprendizagem de programação:** estudo de caso no ensino superior. 2022. 269 f. Tese (Doutorado em Ciências da Educação) – Universidade do Minho, Braga, 2022.

MEDEIROS, R. ***Hello, world***: uma análise sobre dificuldades no ensino e aprendizagem de introdução à programação nas universidades. 2019. 193 f. Tese (Doutorado em Ciência da Computação) - Universidade Federal de Pernambuco. Recife. 2019.

MOREIRA, J. M.; CARVALHO, A. C. P. L. F. **A General Introduction to Data Analytics**. Hoboken: Wiley, 2019.

OLIVEIRA, M. C. S. **Fatores de impacto no desempenho acadêmico:** um estudo de caso em cursos de computação. 2021. 81 f. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Goiás, Goiânia, 2021.

PEREIRA, A. F. S.; CARVALHO, L. S. G.; SOUTO, E. J. P. Analisando a influência de atributos demográficos no desempenho de estudantes em uma disciplina de introdução à programação. *In:* WORKSHOP SOBRE EDUCAÇÃO EM COMPUTAÇÃO (WEI), 27., 2019, Belém. **Anais do XXVII Workshop sobre Educação em Computação.** Porto Alegre: Sociedade Brasileira de Computação, 2019. Disponível em: https://sol.sbc.org.br/index.php/wei/article/view/6642. Acesso em: 30 abr. 2025

PEREIRA, A. F. S. **Previsão da evasão estudantil em disciplinas introdutórias de programação por meio de mineração de dados sociodemográficos**. 2021, 83 f. Dissertação (Mestrado em informática) – Instituto de Computação, Universidade Federal do Amazonas, Manaus, 2021.

RAUTENBERG, S.; CARMO, P. R. V. Big data e ciência de dados: complementariedade conceitual no processo de tomada de decisão. **Brazilian Journal of Information Science: research trends**, Marília, v. 13, n. 1, p. 56-67, 2019. DOI: 10.36311/1981-1640. 2019.v13n1.06.p56. Disponível em: https://revistas.marilia.unesp.br/index.php/bjis/article/view/8315. Acesso em: 8 abr. 2025.

REZENDE, C. C. S.; CANTARINO, L. A. B.; SOUZA, P. F.; ALVES, T. O. M.; CAMPOS, R. S. O impacto de aspectos socioeconômicos no desempenho de estudantes de Sistemas de Informação no Enade. **Revista Brasileira de Informática na Educação**, [s.l.], v. 30, p. 157–

181, 2022. DOI: https://doi.org/10.5753/rbie.2022.2093. Disponível em: https://journals-sol.sbc.org.br/index.php/rbie/article/view/2093/1929. Acesso em: 22 jan. 2025.

SOUSA, R. R. de; LEITE, F. T. Usando gamificação no ensino de programação introdutória. **Brazilian Journal of Development**, [s. l.], v. 6, p. 33338–33356 2020. DOI: 10.34117/bjdv6n6-043. Disponível em: https://ojs.brazilianjournals.com.br/ojs/index.php/BRJD/article/view/10995. Acesso em: 6 mar. 2025.

SILVA, D.; CLEGER, S.; PESSOA, M. PIRES, F. OLIVEIRA, D. B. F; OLIVEIRA, E. H. T.; CARVALHO, L. S. G. Minerando dados de um juiz on-line para prever a evasão de estudantes em disciplinas introdutórias de programação, *In*: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO (SIBE),31.,2020, Online**. Anais do XXXI Simpósio Brasileiro de Informática na Educação.** Porto Alegre: Sociedade Brasileira de Computação, 2020. Disponível em: https://sol.sbc.org.br/index.php/sbie/article/view/12890. Acesso em 6 mar. 2025.

SILVA, E. P.; CACEFFO, R. E.; AZEVEDO, R. J. Análise dos Tópicos Mais Abordados em Disciplinas de Introdução à Programação em Universidades Federais Brasileiras. *In*: SIMPÓSIO BRASILEIRO DE EDUCAÇÃO EM COMPUTAÇÃO (EDUCOMP), 2., 2022, Online**. Anais do II Simpósio Brasileiro de Educação em Computação.** Porto Alegre: Sociedade Brasileira de Computação, 2022. Disponível em: https://sol.sbc.org.br/index.php/educomp/article/view/19196. Acesso em: 28 mar. 2025.

SILVA, F. L.; MOREIRA, I. A. T. Análise das dificuldades na aprendizagem de programação no curso de análise e desenvolvimento de sistemas do IFRN/Pau dos Ferros. *In*: ENCONTRO UNIFICADO DE COMPUTAÇÃO DO PIAUÍ (ENUCOMPI), 14., 2021, Picos**. Anais do XIV Encontro Unificado de Computação do Piauí e XI Simpósio de Sistemas de Informação**. Porto Alegre: Sociedade Brasileira de Computação, 2021. Disponível em: https://sol.sbc.org.br/index.php/enucompi/article/view/17752. Acesso em: 5 mar. 2025.
SILVA, L. A.; PERES. S. M.; BOSCARIOLI, C. Introdução à mineração de dados: com aplicações em R. 1 ed. Rio de Janeiro: Elsevier, 2016.

SOARES, G. C. **SAM – Uma abordagem específica de mineração de dados socioeconômicos de alunos do IF amazonas para apoio ao processo de concessão de assistência estudantil**. 2020. 150 f. Dissertação (Mestrado em Ciência da Computação), Universidade Federal de Pernambuco, Recife, 2020.

WITTEN, I. H.; FRANK, E.; HALL, M. A.; PAL, C. J. **The WEKA workbench. Online apêndix for Data Mining: Pratical Machine Learning tools and Techniques**. 4ª Ed. Morgan Kaufmann, 2016.

WITTEN, I. H.; FRANK, E.; HALL, M. A.; PAL, C. J. **Data Mining: Pratical Machine Learning tools and Techniques**. 4ª Ed. Morgan Kaufmann, 2017.

YIN, Robert K. **Estudo de Caso: Planejamento e Métodos**. 2. ed. Porto Alegre: Bookman, 2001.