



EVASÃO ESCOLAR: ANÁLISE E ESTUDO DE MODELOS PREDITIVOS SOBRE FREQUÊNCIA, RENDIMENTO E PERMANÊNCIA

SCHOOL DROPOUT: ANALYSIS AND STUDY OF PREDICTIVE MODELS ON ATTENDANCE, PERFORMANCE, AND RETENTION

DOI: 10.5281/zenodo.19447417



Moacir Roger Santos Ribeiro¹

Gleison Guardia²

Jackson Henrique da Silva Bezerra³

RESUMO

Este artigo investiga a evasão escolar, um problema multifatorial e persistente que afeta sistemas educacionais globalmente, trazendo prejuízos acadêmicos, sociais e econômicos. Partindo de uma revisão da literatura, destaca-se que fatores como desempenho acadêmico, condições socioeconômicas e dificuldades no acesso à educação digital são preditores recorrentes de evasão, especialmente em contextos de desigualdade social. Utilizando mineração de dados educacionais (EDM) como abordagem metodológica, o estudo empregou técnicas avançadas de análise de dados para explorar padrões em conjuntos de dados educacionais. A metodologia incluiu o uso de ferramentas como R e bibliotecas como tidyverse, ggplot2 e kableExtra, combinadas a modelos estatísticos e de aprendizado de máquina, como regressões lineares e Random Forest. Os dados foram submetidos a um processo rigoroso de limpeza, transformação e análise, garantindo robustez aos resultados. As análises realizadas indicaram que variáveis como rendimento acadêmico inicial, frequência escolar e renda familiar possuem alta relevância na predição de evasão, enquanto fatores psicossociais e contextuais também desempenham papéis significativos. Resultados adicionais apontam para a eficácia de modelos baseados em ensemble no aprimoramento da precisão preditiva. Este estudo contribui para o campo da educação ao oferecer uma abordagem integrada e baseada em dados, possibilitando

1. Graduando em Análise e Desenvolvimento de Sistemas no Instituto Federal de Rondônia. E-mail: moacirroger@hotmail.com

2. Mestre em Educação. Docente no Instituto Federal de Rondônia. E-mail: gleison.guardia@ifro.edu.br

3. Doutor em Desenvolvimento Regional. Docente no Instituto Federal de Rondônia. E-mail: Jackson.henrique@ifro.edu.br





intervenções mais eficazes para mitigar a evasão escolar. Além disso, destaca a importância de integrar modelos preditivos aos sistemas institucionais para maximizar seu impacto, especialmente em países com maior necessidade de soluções educacionais inovadoras. O artigo reforça a urgência de políticas públicas e práticas pedagógicas que combinem tecnologia e análise de dados para promover a permanência e o sucesso escolar em diferentes contextos.

Palavras-chave: Evasão escolar; Mineração de dados educacionais; Desempenho acadêmico; Aprendizado de máquina; Políticas públicas educacionais.

ABSTRACT

This article investigates school dropout, a multifactorial and persistent issue affecting educational systems globally, with significant academic, social, and economic repercussions. Based on a literature review, the study highlights that factors such as academic performance, socioeconomic conditions, and challenges in accessing digital education are recurring predictors of dropout, particularly in contexts of social inequality. Employing educational data mining (EDM) as a methodological approach, the study utilized advanced data analysis techniques to explore patterns in educational datasets. The methodology included the use of tools such as R and libraries like tidyverse, ggplot2, and kableExtra, combined with statistical and machine learning models, such as linear regressions and Random Forest. The data underwent a rigorous process of cleaning, transformation, and analysis, ensuring the robustness of the results. The analyses revealed that variables such as initial academic performance, school attendance, and household income are highly relevant in predicting dropout, while psychosocial and contextual factors also play significant roles. Additional results point to the effectiveness of ensemble-based models in enhancing predictive accuracy. This study contributes to the field of education by offering an integrated and data-driven approach, enabling more effective interventions to mitigate school dropout. Moreover, it underscores the importance of integrating predictive models into institutional systems to maximize their impact, especially in countries with a greater need for innovative educational solutions. The article reinforces the urgency of public policies and pedagogical practices that combine technology and data analysis to promote school retention and success across diverse contexts.

Keywords: School dropout; Educational data mining; Academic performance; Machine learning; Educational public policies.

1. INTRODUÇÃO

A evasão escolar é um dos desafios mais complexos e persistentes enfrentados por sistemas educacionais ao redor do mundo. Esse fenômeno possui consequências de longo alcance, afetando não apenas os indivíduos diretamente impactados, mas também as comunidades e economias como um todo. A perda de alunos antes da conclusão de seus





estudos compromete o desenvolvimento humano e limita as oportunidades de ascensão social, ao mesmo tempo em que agrava desigualdades estruturais existentes. Diante desse contexto, compreender os fatores que contribuem para a evasão escolar emerge como um passo essencial para o desenvolvimento de estratégias eficazes que promovam a permanência e o sucesso acadêmico.

Embora a literatura acadêmica já tenha identificado diversos fatores associados à evasão escolar, como condições socioeconômicas, desempenho acadêmico, contexto familiar e características institucionais, a complexidade dessas interações exige análises cada vez mais sofisticadas. No Brasil, onde o sistema educacional enfrenta desafios significativos relacionados à desigualdade de acesso, qualidade do ensino e infraestrutura escolar, a identificação de indicadores de evasão torna-se ainda mais importante para o direcionamento de políticas públicas e intervenções institucionais.

Com o avanço das tecnologias de análise de dados e da ciência estatística, tornou-se possível explorar com maior precisão padrões e tendências ocultos em grandes conjuntos de dados educacionais. Nesse contexto, este estudo utiliza ferramentas estatísticas avançadas e aprendizado de máquina para identificar os principais indicadores de evasão escolar em um conjunto de dados acadêmicos. A escolha dessa abordagem justifica-se pela capacidade dessas metodologias de lidar com a complexidade e a variabilidade dos dados, permitindo análises mais robustas e inferências mais confiáveis.

O presente trabalho segue um rigoroso processo metodológico para garantir a qualidade das análises e a validação dos resultados. Através do uso do *software* R, amplamente reconhecido pela sua eficiência em análises estatísticas e manipulação de dados, de acordo com a R Core Team (2024), foi possível aplicar técnicas avançadas como regressões lineares múltiplas, testes de hipóteses, análise de correlações e modelos de aprendizado supervisionado, como *Random Forest*. Estas técnicas permitiram não apenas a identificação dos fatores mais influentes na evasão escolar, mas também a visualização de padrões complexos de relacionamento entre variáveis, como renda per capita, frequência, índice de rendimento acadêmico (IRA) e situação acadêmica dos alunos.





A originalidade deste estudo reside na integração de técnicas estatísticas tradicionais com ferramentas de aprendizado de máquina, oferecendo uma abordagem multidimensional para a análise de dados educacionais. Além disso, o uso de gráficos interativos e tabelas detalhadas, gerados por bibliotecas como *tidyverse* e *kableExtra*, conferiu clareza e acessibilidade aos resultados, permitindo interpretações mais assertivas e facilitando a tomada de decisão baseada em dados.

Dessa forma, este artigo busca contribuir para o campo da educação ao fornecer evidências sólidas sobre os fatores que influenciam a evasão escolar, propondo uma abordagem replicável e adaptável a diferentes contextos institucionais, respondendo a 3 perguntas principais: Quais dados interferem diretamente no estado de permanência do(a) estudante? De que maneira esses dados correlacionam-se? Como as evidências anteriores podem contribuir para mitigar a taxa de evasão?

Os resultados obtidos não apenas reforçam a importância de políticas educacionais focadas na retenção de alunos, mas também destacam a necessidade de um olhar integrado sobre os múltiplos fatores que impactam o percurso educacional. Espera-se que este estudo sirva como um ponto de partida para futuras pesquisas e iniciativas que visem transformar a realidade educacional brasileira, promovendo a equidade, a inclusão e a excelência acadêmica.

2. TRABALHOS CORRELATOS

A evasão escolar refere-se ao afastamento do aluno das atividades escolares, seja por sua vontade ou por motivos variados, o que impede que ele conclua o processo de formação e acesse as oportunidades que teria. Araújo e Lima (2021) ressaltam que a evasão escolar existe desde períodos muito antigos, tanto no Brasil quanto em diversos outros locais do mundo, sendo debatida ao longo dos anos como uma questão problemática, porém, ainda sem solução até os dias atuais. De acordo com os mesmos autores, não se pode definir um único motivo para a evasão escolar nos institutos federais do país. Diferentes regiões e perfis populacionais





terão maior ou menor incidência de alguns fatores. No entanto, é muito comum o relato da dificuldade de conciliar estudos e trabalho, dificuldades de compreensão e uso dos recursos e materiais disponibilizados, necessidade de dedicar-se a outras atividades relacionadas à família, entre outros aspectos. Essas razões podem ser tanto pessoais quanto decorrentes da forma como as escolas organizam e oferecem os cursos.

O estudo de Silva Júnior et al. (2016) enfatiza que o aluno que abandona a escola pode não sentir os impactos imediatamente, mas, no futuro, perceberá menores índices de oportunidades no mercado de trabalho, justamente pela falta de respaldo educacional que teria recebido no curso. Isso leva a dificuldades econômicas e financeiras, comprometimento da qualidade de vida, exclusão social e aumento dos índices de desigualdade, que já são expressivos no país e em muitos lugares do mundo.

Ramos Neto (2019) aponta que, diante da percepção de que a evasão chegou a patamares expressivos e extremamente prejudiciais, muitos institutos federais do Brasil têm desenvolvido ferramentas para a redução desses números. Essas iniciativas incluem cartilhas que instruem escolas sobre como avaliar o risco de evasão, identificar sinais de que os alunos estão próximos de desistir e implementar ações para ouvir esses indivíduos, entender suas razões e oferecer suporte para evitar sua saída antes da conclusão do curso.

Gaspar & Baida (2020) esclarecem que a falta de conclusão dos estudos, além de limitar as oportunidades de desenvolvimento socioeconômico e excluir os indivíduos do mercado de trabalho, eleva os riscos de condutas ilícitas, práticas de atos violentos, delinquência e uso de substâncias ilícitas. Complementando essa visão, Silva, Moraes & Costa (2018) argumentam que a decisão de evadir ou persistir em um curso é um fator psicossocial, influenciado pela aprovação da família, pela qualidade oferecida pela instituição e pelo encorajamento dos amigos. Esses fatores motivam a tomada de decisão, respondendo aos anseios individuais. Os mesmos autores destacam que, embora os cursos à distância tenham sido desenvolvidos para facilitar o acesso à educação e oferecer mais oportunidades de qualificação, eles envolvem desafios como o uso de tecnologias pouco conhecidas pelos





alunos e a falta de proximidade com professores, o que pode gerar uma sensação de abandono ou incapacidade de acompanhar os conteúdos.

Alvarez (2020) observa que a evasão escolar também significa custos significativos, já que os alunos evadidos ocupam vagas que poderiam ter sido preenchidas por outros. Além disso, o investimento em educação, especialmente em cenários de alta evasão, torna-se insustentável devido à necessidade de realocação de recursos limitados.

A falta de conhecimentos exatos sobre as razões que levam à evasão escolar gera dificuldades no estabelecimento de medidas corretivas verdadeiramente efetivas, que poderiam ser usadas como respaldo para o desenvolvimento de políticas públicas na área. Percebe-se, portanto, que a evasão escolar é uma questão multifatorial, decorrente de uma associação de fatores que interferem na decisão dos alunos de continuar seus estudos ou priorizar outras questões que, naquele momento, parecem mais urgentes.

A mineração de dados educacionais tem se consolidado como uma abordagem promissora para compreender e mitigar a evasão escolar. Colpo et al. (2024) realizaram uma revisão sistemática da literatura internacional sobre o uso de técnicas de mineração de dados educacionais (EDM) na predição de evasão. O estudo analisou trabalhos que utilizam métodos avançados, como aprendizado de máquina e algoritmos de classificação, para identificar padrões e alunos em risco de abandonar os estudos. Os autores destacaram a predominância de dados acadêmicos e socioeconômicos em pesquisas voltadas para o ensino superior, assim como a escassez de estudos em países subdesenvolvidos, onde a evasão representa um desafio crítico. Além disso, identificaram lacunas na aplicação prática dos modelos preditivos, ressaltando a importância de integrar essas soluções aos processos institucionais para maximizar seu impacto.

No artigo de Jesus e Gusmão (2024), foi realizado um mapeamento sistemático de literatura focado na aplicação de técnicas de aprendizado de máquina e mineração de dados para a análise da evasão escolar em cursos técnicos e de graduação. A pesquisa explorou diferentes metodologias, incluindo modelos de árvore de decisão, regressão logística e redes neurais artificiais, avaliando métricas como precisão, recall e AUC-ROC. Entre os resultados,





destacou-se o desempenho superior de algoritmos baseados em ensemble, como Random Forest, para a identificação de padrões em grandes volumes de dados educacionais. Os autores também enfatizaram a importância de pré-processar os dados, utilizando técnicas como *oversampling* e análise de importância de atributos, para melhorar a eficácia dos modelos preditivos.

Dekker et al. (2009) investigaram os desafios associados à predição de evasão escolar em cenários educacionais diversificados, com foco em contextos de alta desigualdade social. Utilizando um modelo de revisão narrativa, o estudo destacou os impactos da qualidade dos dados e da representatividade das variáveis no desempenho dos modelos preditivos. Os autores implementaram modelos de aprendizado supervisionado, como *Support Vector Machines* e *Gradient Boosting Trees*, para prever comportamentos de evasão em diferentes níveis educacionais. Os resultados indicaram que variáveis relacionadas ao desempenho acadêmico inicial, como notas e frequência, têm alta relevância na previsão de abandono. Além disso, os autores propuseram o uso de sistemas de alerta integrados às plataformas institucionais, permitindo intervenções mais ágeis e direcionadas.

Por fim, destaca-se que estudos aplicados também evidenciam o potencial da mineração de dados na identificação de padrões de evasão no Instituto Federal de Rondônia (IFRO). Bezerra (2019) utilizou algoritmos de classificação e regras de associação aplicados aos dados do sistema acadêmico institucional com o objetivo de identificar perfis de estudantes com maior propensão à evasão nos cursos técnicos integrados ao ensino médio. Os resultados permitiram identificar padrões acadêmicos e socioeconômicos relevantes, como maior tendência de evasão entre alunos do primeiro ano, estudantes de determinados turnos e discentes com baixa renda familiar ou dependentes de transporte coletivo. A partir dessas análises, foram propostas estratégias institucionais de intervenção voltadas à permanência e ao êxito dos estudantes, evidenciando o potencial da mineração de dados como ferramenta de apoio à gestão educacional e à formulação de políticas de enfrentamento da evasão.





3. MÉTODO

O estudo aqui relatado foi conduzido com base em dados acadêmicos do Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas (ADS), contemplando 403 registros de estudantes matriculados no período de 2019 a 2023. Os dados utilizados correspondem a uma parcela pública de questionários institucionais, disponibilizados no painel de indicadores da instituição em seu site oficial, o que dispensa a necessidade de submissão ao Comitê de Ética em Pesquisa. A base de dados, previamente tratada, encontra-se disponível para download no repositório Zenodo, por meio do DOI: <https://doi.org/10.5281/zenodo.19024776>.

A metodologia adotada neste trabalho foi cuidadosamente estruturada para garantir a qualidade da análise e a reprodutibilidade dos resultados fazendo-se o uso do *software* R. Primeiramente, as bibliotecas necessárias foram carregadas no ambiente R, incluindo *readxl*, para manipulação de arquivos do tipo Excel, *tidyverse*, que engloba um conjunto de ferramentas para manipulação e visualização de dados, e *esquisse*, utilizada para facilitar a criação de gráficos interativos.

O tratamento inicial dos dados envolveu a importação de dois conjuntos de informações, um armazenado no arquivo [dados.xls](#) e outro em [dados_completos.xlsx](#). Após a leitura, os dados foram submetidos a uma série de etapas de limpeza e transformação. Para padronizar os valores, substituições foram realizadas para corrigir formatações inadequadas, como a troca de vírgulas por pontos em números decimais. Além disso, campos com informações ausentes ou incompletas foram preenchidos com valores padrão ou ajustados, por exemplo, substituindo caracteres como hífen por alternativas mais informativas, como "N" ou "Brasil".

O processo de limpeza também incluiu a conversão de variáveis textuais para formatos numéricos, o que permitiu sua utilização em análises posteriores. A variável "Progresso", por exemplo, foi padronizada, preenchendo valores ausentes com zeros e convertendo-a para o tipo numérico. Outros campos, como "Frequência", "I.R.A." e "Renda_Per_Capita", também





passaram por ajustes similares para garantir consistência nos dados. Após a etapa de limpeza, o conjunto de dados transformado ficou apto para ser utilizado nas análises subsequentes.

Após o tratamento inicial dos dados, foram conduzidas as principais análises para explorar padrões e relações entre as variáveis do estudo. A biblioteca *tidyverse* desempenhou um papel central nessa etapa, fornecendo ferramentas poderosas para a manipulação de dados e geração de gráficos. As funções do pacote permitiram realizar filtragens, agrupamentos e sumarizações que facilitaram a compreensão das tendências presentes no conjunto de dados.

Para a construção de gráficos exploratórios, a biblioteca *ggplot2*, integrada ao *tidyverse*, foi amplamente utilizada. Essa ferramenta foi fundamental para criar visualizações claras e informativas, possibilitando a análise visual das distribuições das variáveis, correlações e tendências temporais. Além disso, o pacote *esquisse* complementou essa abordagem, oferecendo uma interface interativa que permitiu ajustar as visualizações de forma mais intuitiva e eficiente.

No que diz respeito às análises estatísticas, a metodologia incluiu testes de hipóteses para verificar associações significativas entre variáveis. Para esses testes, bibliotecas adicionais foram utilizadas conforme necessário, com foco em funções estatísticas amplamente aceitas na literatura acadêmica. Modelos de regressão foram implementados para explorar relações preditivas entre as variáveis. A escolha desses modelos foi orientada pelo tipo de dados e pela natureza das variáveis de resposta e explicativas. O ajuste dos modelos e a interpretação dos resultados permitiram identificar os fatores mais relevantes no contexto analisado.

As variáveis tratadas, como "Progresso", "I.R.A." e "Frequência", foram usadas em combinações estratégicas nos modelos, garantindo que as conclusões fossem respaldadas por análises estatísticas sólidas. As bibliotecas específicas para regressão incluíram pacotes como *stats*, para análises mais gerais, e outras bibliotecas de modelagem avançada que auxiliaram na parametrização e validação dos modelos.

Por fim, a construção de tabelas e gráficos finais para a apresentação dos resultados foi realizada com recursos avançados do *tidyverse* e com o auxílio do pacote *kableExtra*, que





permitiu formatar tabelas de forma elegante e adequada aos padrões de apresentação acadêmica. Essa abordagem integrativa e metódica garantiu que as análises fossem não apenas estatisticamente corretas, mas também visualmente claras, contribuindo para uma interpretação mais acessível e objetiva dos dados. Abaixo temos uma tabela que resume todo este processo.

Tabela 1: Esquema metodológico da análise

Etapa	Descrição
1. Configuração Inicial	Carregar bibliotecas necessárias: readxl, tidyverse, esquisse.
2. Importação dos Dados	Importar os arquivos Excel utilizando read_excel().
3. Tratamento de Dados	Aplicar transformações e correções, incluindo:
	Substituir vírgulas por pontos para padronizar números decimais.
	Preencher valores ausentes com padrões adequados ("N", "0", "Brasil").
4. Exploração Inicial	Converter variáveis textuais em numéricas (e.g., as.numeric()).
	Resumir dados e identificar padrões usando funções como summary() e glimpse().
	Realizar testes de hipóteses e ajustes de modelos estatísticos:
5. Análises Estatísticas	Aplicar regressões lineares e outros modelos preditivos.
	Avaliar significância estatística com testes apropriados.
	Criar gráficos exploratórios com ggplot2, como histogramas, scatter plots e boxplots.
6. Visualização de Dados	
7. Construção de Tabelas	Gerar tabelas formatadas para apresentação com kableExtra e tidyverse.
8. Validação Final	Verificar a consistência das análises e ajustar parâmetros ou gráficos conforme necessários.

Assim, passaremos a discutir na sequência os resultados e propostas encontradas por meio do trabalho desenvolvido.

4. RESULTADOS E DISCUSSÕES

Inicialmente foram analisados os valores da renda per capita de cada estudante, utilizando como parâmetro os sexos existentes na base de dados disponível no Zenodo.



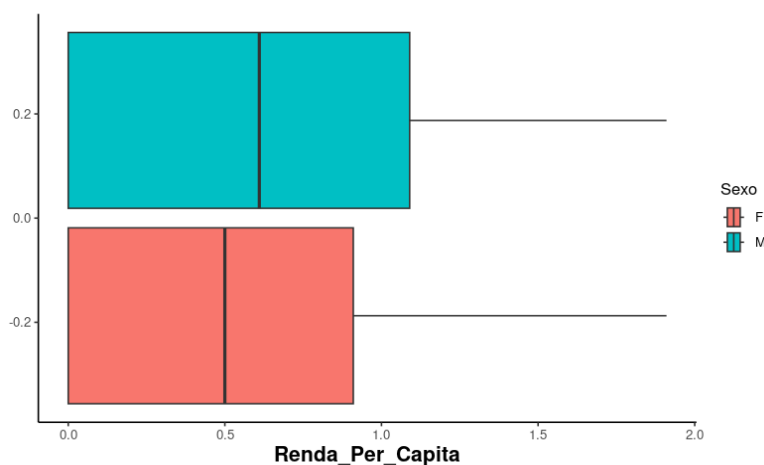


Figura 1: Correlação entre Renda e Sexo

Nesta análise, é possível inferir que a grande maioria dos estudantes do sexo masculino possui uma renda per capita em torno de 0,7 salários-mínimos, enquanto as estudantes do sexo feminino apresentam uma renda aproximada de 0,5 salários-mínimos. A seguir, observa-se um gráfico que ilustra, lado a lado, a quantidade de alunos matriculados no curso, utilizando como parâmetros o sexo e o ano de ingresso de cada estudante.

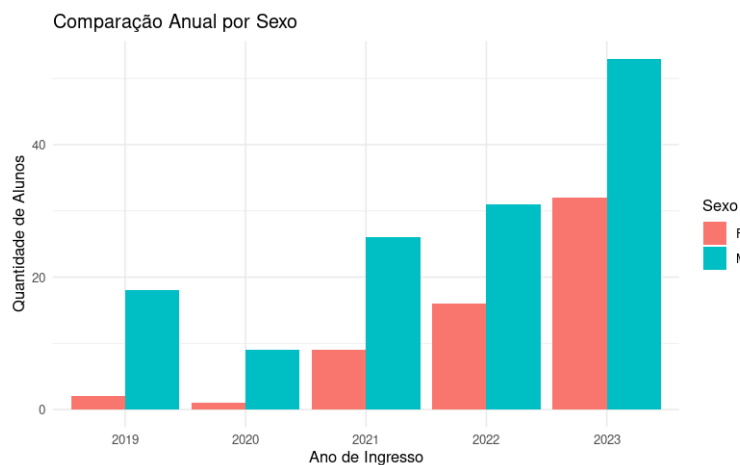


Figura 2: Alunos egressos no curso, classificados pelo sexo



De acordo com as análises realizadas, nos anos iniciais do curso, a quantidade de estudantes matriculadas do sexo feminino é significativamente menor em comparação aos do sexo masculino. Entretanto, ao longo dos anos, até o final do período analisado, embora o número de estudantes do sexo masculino continue predominando, observa-se um crescimento constante no número de mulheres matriculadas, indicando uma tendência de equilíbrio entre os gêneros. O gráfico abaixo apresenta a distribuição dos estudantes no curso até o fim do período analisado.

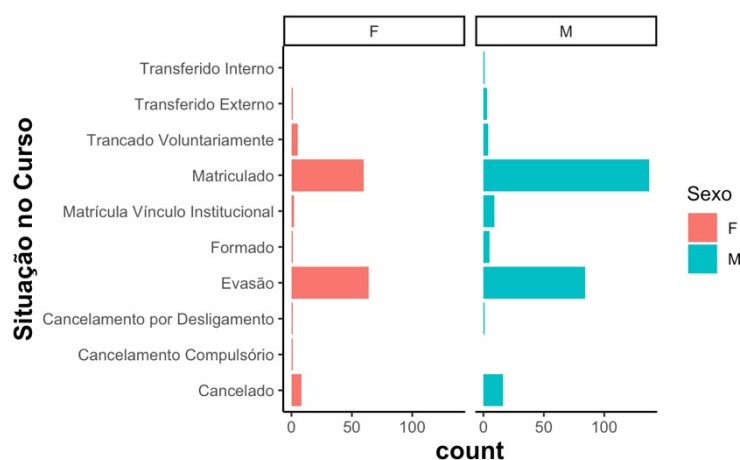


Figura 3: Situação dos alunos no curso classificada pelo sexo

A partir de uma análise visual, observa-se um alto índice de evasão entre ambos os sexos, com uma predominância proporcionalmente maior entre as alunas em relação ao total de matriculados no curso. O gráfico também evidencia que o número de alunos que concluíram o curso é significativamente inferior ao de alunos que evadiram ou permaneceram matriculados, além de mostrar uma quantidade expressiva de cancelamentos de matrículas após o ingresso no curso. A relação entre essas categorias é detalhada no gráfico a seguir.

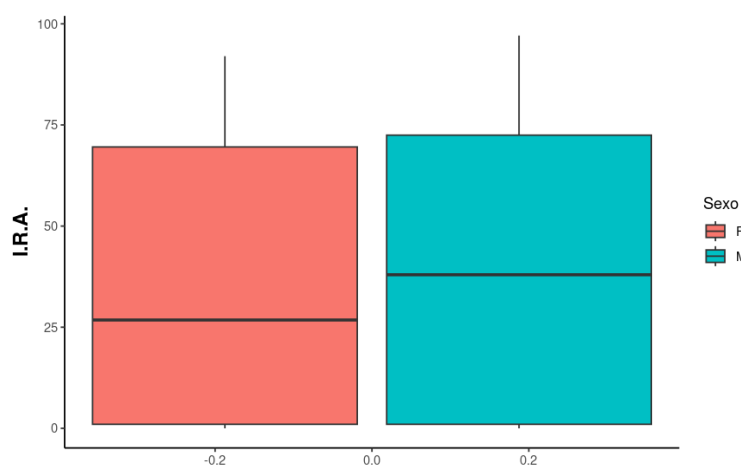


Figura 4: Índice de Rendimento Acadêmico classificado pelo sexo

O Índice de Rendimento Acadêmico (I.R.A.) utilizado como parâmetro no gráfico mede o desempenho acadêmico de cada estudante. À esquerda, encontram-se os dados das estudantes do sexo feminino, cuja mediana do I.R.A. é de aproximadamente 27, representada pela linha que divide o retângulo correspondente. À direita, estão os dados dos estudantes do sexo masculino, cuja mediana do I.R.A. é cerca de 40, indicada de forma similar. A média do I.R.A. para as alunas é de 35,91, enquanto para os alunos é de 39,20. Apesar de as medianas estarem dentro do intervalo esperado, as médias estão abaixo do ideal, considerando que o curso tem um número maior de estudantes do sexo masculino matriculados e que a evasão feminina, conforme mostrado em análises anteriores, é proporcionalmente superior ao número de mulheres matriculadas.

Ressalta-se, entretanto, que as notas máximas alcançadas por ambos os sexos são teoricamente equivalentes, apesar dos altos índices de evasão em ambos os grupos. O gráfico também evidencia a dispersão dos estudantes, utilizando o sexo como parâmetro de separação, onde cada ponto representa um estudante matriculado. Essa visualização permite identificar padrões de distribuição no desempenho acadêmico entre os gêneros.

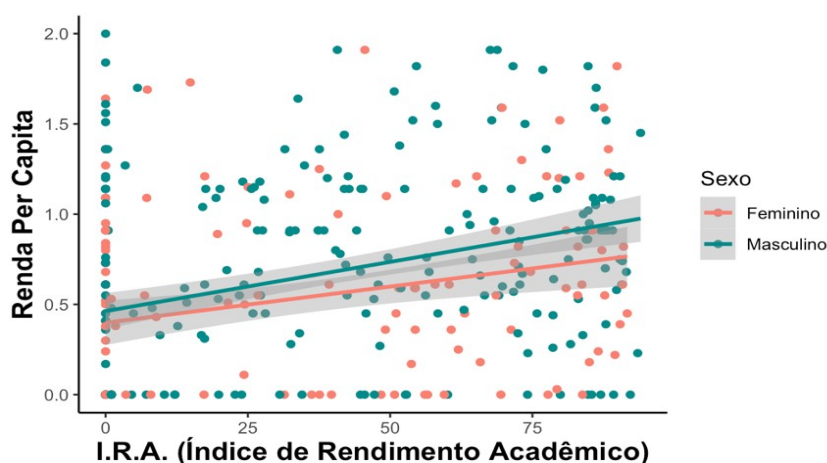


Figura 5: Gráfico de Dispersão entre Renda Per Capita e I.R.A., com linhas de Regressão, classificada pelo Sexo

No eixo X do gráfico está representado o Índice de Rendimento Acadêmico (I.R.A.) dos alunos, enquanto o eixo Y indica a Renda Per Capita. As linhas internas que cruzam os dados representam a média de crescimento ou o ponto médio da concentração de alunos. Essas linhas possuem uma faixa cinza ao redor, que reflete a dispersão dos dados utilizados para calcular a linha de tendência.

A análise do gráfico revela que a renda per capita não possui uma influência significativa no desempenho acadêmico dos alunos, visto que a maioria se encontra na faixa entre 0 e 1 salário-mínimo. Observa-se, ainda, casos de alunos sem renda aparente com I.R.A. elevado, acima da média, assim como alunos com renda mais alta, mas com I.R.A. inferior à média. Embora a linha média apresente uma tendência crescente, a dispersão elevada dos dados indica uma menor precisão na definição dessa tendência.

Outro ponto relevante é a inclinação das linhas representativas para ambos os sexos ao longo do tempo, sugerindo uma melhora gradual no desempenho acadêmico (I.R.A.) e na Renda Per Capita conforme os alunos progridem no curso. Isso pode indicar um desenvolvimento positivo na atitude dos estudantes durante sua trajetória acadêmica.



```
Call:
lm(formula = df$I.R.A. ~ df$Renda_Per_Capita)
Coefficients:
(Intercept) df$Renda_Per_Capita
37.89835      0.08776
```

Código 1: Resultado da primeira Regressão linear

A fórmula utilizada nesta equação apresenta semelhanças com a anterior em alguns aspectos, mas permite observar, de forma mais detalhada, como cada indicador contribui para o aumento relativo em comparação com os outros. A primeira parte dos resultados da regressão linear simples fornece os coeficientes utilizados na análise, relacionando o Índice de Rendimento Acadêmico (I.R.A.) com a Renda Per Capita. Já a segunda parte apresenta os resultados dessa análise, indicando que, quando a Renda Per Capita é igual a zero, o valor esperado do I.R.A. é de aproximadamente 37,90. Além disso, para cada aumento de uma unidade na Renda Per Capita, estima-se um incremento médio de 0,08776 no I.R.A.

Portanto, conclui-se que um aumento na Renda Per Capita tende a resultar em uma ligeira elevação no I.R.A. No entanto, o impacto desse aumento é considerado pequeno, com um incremento de apenas 0,08776 no I.R.A. por unidade de Renda Per Capita. Esses dados reforçam que, embora exista uma relação positiva entre os indicadores, a magnitude dessa influência é limitada.

```
[1] Situação do Período e Frequência => -0.6187501
[2] Índice de Rendimento Acadêmico e Renda => 0.3051202
[3] Situação do Período e Índice de Rendimento Acadêmico => -0.2927299
```

Código 2: Correlação entre os dados

Os resultados apresentados foram obtidos a partir de análises de correlação entre diferentes variáveis. Na primeira análise, a correlação entre a Situação do Período e a Frequência revelou um valor negativo de $-0,6187501$, indicando que, com o passar do tempo, há uma tendência de redução na frequência dos alunos (como menor presença às aulas). Na segunda análise, a correlação entre o Índice de Rendimento Acadêmico (I.R.A.) e a Renda mostrou um valor positivo de $0,3051202$, sugerindo que, em média, estudantes com maior



renda tendem a apresentar um I.R.A. ligeiramente mais elevado. Contudo, essa relação não é suficientemente forte para ser considerada um parâmetro decisivo.

Na terceira análise, foi identificada uma correlação negativa entre a Situação do Período e o I.R.A., indicando que, à medida que o rendimento acadêmico diminui, a situação do período do estudante tende a se deteriorar. Isso sugere que estudantes com menor desempenho acadêmico estão mais propensos a abandonar ou trancar o curso.

Com base nesses resultados, será conduzida uma regressão múltipla para explorar mais detalhadamente as relações entre as variáveis e identificar os principais fatores que influenciam a situação dos estudantes.

```
Call:
lm(formula = situacao_periodo ~ frequencia + transporte + renda +
    indice_rendimento + escolaridade + tipo_veiculo + residencia +
    sexo + etnia + estado_civil, data = alunos)
Residuals:
    Min       1Q   Median       3Q      Max
-1.5965 -0.5356 -0.3847 -0.1033  5.9191
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.601126   0.235908   6.787 4.24e-11 ***
frequencia    -0.015178   0.001396 -10.876 < 2e-16 ***
transporte     0.973209   1.219803   0.798  0.425
renda         -0.004426   0.006315  -0.701  0.484
indice_rendimento 0.003453   0.001802   1.916  0.056 .
escolaridade  -0.191101   0.211047  -0.905  0.366
tipo_veiculo  -0.792602   1.157645  -0.685  0.494
residencia    -0.070355   0.191266  -0.368  0.713
sexo          -0.049885   0.120536  -0.414  0.679
etnia          0.043774   0.049644   0.882  0.378
estado_civil   0.013364   0.103001   0.130  0.897
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.131 on 392 degrees of freedom
Multiple R-squared:  0.2459, Adjusted R-squared:  0.2267
F-statistic: 12.78 on 10 and 392 DF, p-value: < 2.2e-16
```

Código 3: Resultados da regressão múltipla feita da Situação em que o estudante se encontra no período, em comparação com os demais indicadores



Neste exemplo, foi utilizada uma regressão linear múltipla para identificar quais fatores exercem maior influência sobre a Situação do Período, com o objetivo de determinar os principais indicadores da evasão estudantil. Este modelo foi selecionado devido à sua relevância estatística na análise dos dados.

Na área de Coeficientes, são apresentados os valores estimados para cada variável (indicados por Estimate), o erro padrão associado a essas estimativas (Std. Error), a relevância estatística em termos de desvios padrão (t value), e a probabilidade de a relação observada ocorrer ao acaso ($\text{Pr}(>|t|)$). Relações com menores probabilidades de ocorrência ao acaso são consideradas mais significativas no modelo.

Os resultados dessa regressão permitiram inferir que a Frequência e o Índice de Rendimento Acadêmico (I.R.A.) são as variáveis mais importantes e estatisticamente significativas na previsão da Situação do Período. Contudo, o impacto do I.R.A., embora relevante, é menor em comparação à Frequência. Com base nesses achados, análises subsequentes foram conduzidas para explorar mais detalhadamente os resultados e implicações desse modelo.

```
Call:
lm(formula = frequencia ~ situacao_periodes + transporte + renda +
    indice_rendimento + escolaridade + tipo_veiculo + residencia +
    sexo + etnia + estado_civil, data = alunos)
Residuals:
    Min       1Q   Median       3Q      Max
-75.053 -30.409  -1.301   21.934  141.742
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   49.93714    7.49797   6.660 9.27e-11 ***
situacao_periodes -15.27299    1.40425 -10.876 < 2e-16 ***
transporte     79.72524   38.51539   2.070  0.0391 *
renda          -0.07471    0.20041  -0.373  0.7095
indice_rendimento 0.37400    0.05423   6.896 2.15e-11 ***
escolaridade   -3.41207    6.69951  -0.509  0.6108
tipo_veiculo  -64.15283   36.60098  -1.753  0.0804 .
residencia     1.46660    6.06784   0.242  0.8091
sexo          -5.88131    3.81287  -1.542  0.1238
etnia           0.52749    1.57611   0.335  0.7380
estado_civil   -6.61615    3.25029  -2.036  0.0425 *
---
```



Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 35.87 on 392 degrees of freedom
Multiple R-squared: 0.3508, Adjusted R-squared: 0.3342
F-statistic: 21.18 on 10 and 392 DF, p-value: < 2.2e-16

Código 4: Resultados da regressão múltipla feita da Frequência que o estudante teve durante no período, em comparação com os demais indicadores

Seguindo o exemplo da análise anterior, foi realizada uma regressão linear múltipla com os dados identificados como mais relevantes para determinar quais fatores exercem maior influência sobre a Frequência dos alunos matriculados.

Os resultados indicaram que as variáveis com maior impacto sobre a Frequência são a Situação do Período, o Transporte, o Índice de Rendimento Acadêmico (I.R.A.), o Tipo de Veículo e o Estado Civil. No entanto, considerando que o Tipo de Veículo apresentou uma influência relativamente menor em comparação às demais variáveis, e que a Situação do Período já havia sido analisada em detalhes, essas foram descartadas para esta etapa. Assim, a análise se concentrou nas variáveis Transporte, I.R.A. e Estado Civil, que demonstraram maior relevância na predição da Frequência.

Residual standard error: 31.55 on 392 degrees of freedom
Multiple R-squared: 0.1547, Adjusted R-squared: 0.1332
F-statistic: 7.175 on 10 and 392 DF, p-value: 2.045e-10

Código 5: Erros e Estatísticas dos testes realizados acima

Dando continuidade à análise previamente realizada, foi aplicada uma regressão linear múltipla utilizando os dados identificados como mais significativos, com o objetivo de determinar os fatores que exercem maior influência sobre o Índice de Rendimento Acadêmico (I.R.A.) dos alunos matriculados.

Os resultados indicaram que, excluindo as variáveis já analisadas, o Estado Civil é o fator com maior impacto sobre o I.R.A. Sua relevância se deve à frequência com que aparece como determinante em múltiplos indicadores, tornando-se um dos principais focos nas análises subsequentes. Esses achados sugerem que a abordagem adotada pela regressão linear





múltipla segue uma ordem lógica e consistente, destacando variáveis de maior influência dentro do contexto analisado.

```
Call:
lm(formula = transporte ~ situacao_periodo + indice_rendimento +
    renda + frequencia + escolaridade + tipo_veiculo + residencia +
    sexo + etnia + estado_civil, data = alunos)
Residuals:
    Min       1Q   Median       3Q      Max
-0.85807 -0.00561 -0.00130  0.00248  0.07347
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.292e-03  1.032e-02  -0.319   0.7498
situacao_periodo  1.666e-03  2.088e-03   0.798   0.4254
indice_rendimento -4.376e-05  7.487e-05  -0.585   0.5592
renda          1.415e-05  2.614e-04   0.054   0.9569
frequencia     1.356e-04  6.552e-05   2.070   0.0391 *
escolaridade    1.097e-04  8.741e-03   0.013   0.9900
tipo_veiculo    9.325e-01  8.853e-03 105.336 <2e-16 ***
residencia      3.384e-03  7.913e-03   0.428   0.6691
sexo           -2.313e-03  4.987e-03  -0.464   0.6430
etnia           -4.682e-04  2.056e-03  -0.228   0.8200
estado_civil    1.820e-03  4.261e-03   0.427   0.6694
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.04679 on 392 degrees of freedom
Multiple R-squared:  0.9671, Adjusted R-squared:  0.9662
F-statistic: 1151 on 10 and 392 DF, p-value: < 2.2e-16
```

Código 6: Resultados da regressão múltipla feita do Transporte que o estudante utilizou para chegar até a instituição, em comparação com os demais indicadores no período analisado

Foi realizada uma regressão linear múltipla para identificar os fatores que mais influenciam o Tipo de Transporte utilizado pelos alunos matriculados. Os resultados indicaram que, dentre os fatores analisados, o Tipo de Veículo do estudante é o que exerce maior influência sobre o Transporte, excluindo os dados previamente investigados. Apesar de a variável Frequência ter sido excluída nesta etapa devido à sua análise anterior, os resultados mostraram que o tipo de transporte utilizado por um aluno impacta significativamente sua frequência, evidenciando uma forte relação entre essas variáveis.



Com base nos resultados obtidos, foi conduzida uma nova regressão linear múltipla, desta vez utilizando os dados mais relevantes para identificar os fatores que exercem maior influência sobre o Estado Civil dos alunos matriculados. Esse processo permitiu uma análise mais detalhada das interações entre as variáveis, ampliando a compreensão sobre os elementos que influenciam o comportamento acadêmico dos estudantes.

```
Call:
lm(formula = estado_civil ~ situacao_periodo + indice_rendimento +
    renda + frequencia + escolaridade + tipo_veiculo + residencia +
    sexo + etnia + transporte, data = alunos)
Residuals:
    Min       1Q   Median       3Q      Max
-0.61958 -0.31268 -0.18207 -0.06284  2.85638
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.6104867  0.1183337   5.159 3.95e-07 ***
situacao_periodo 0.0032133  0.0247658   0.130  0.8968
indice_rendimento -0.0020154  0.0008818  -2.286  0.0228 *
renda         -0.0006093  0.0030984  -0.197  0.8442
frequencia    -0.0015809  0.0007766  -2.036  0.0425 *
escolaridade  -0.1934058  0.1031333  -1.875  0.0615 .
tipo_veiculo   -0.2688003  0.5678260  -0.473  0.6362
residencia     -0.0555369  0.0937613  -0.592  0.5540
sexo           0.0067045  0.0591167   0.113  0.9098
etnia          -0.0224869  0.0243404  -0.924  0.3561
transporte     0.2556858  0.5984742   0.427  0.6694
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.5545 on 392 degrees of freedom
Multiple R-squared:  0.05571,    Adjusted R-squared:  0.03162
F-statistic: 2.313 on 10 and 392 DF,  p-value: 0.01192
```

Código 7: Resultados da regressão múltipla feita do Estado Civil em que o estudante encontrava-se durante o período analisado, em comparação com os demais indicadores

Como resultado desta análise, concluiu-se que os fatores que mais influenciam o Estado Civil dos alunos são os mesmos identificados anteriormente, indicando o encerramento dessa etapa. Nesse contexto, foram levantadas três possíveis explicações para a predominância desse indicador como principal resultado. Antes de discuti-las, é importante destacar que a relação entre a Escolaridade e o Estado Civil dos alunos mostrou-se fraca,



como indicado pelo símbolo de significância estatística “.”, o que sugere que, embora não tenha grande relevância neste modelo, pode ser um ponto de interesse para análises futuras.

A primeira possibilidade é que o Estado Civil realmente seja a principal variável associada à evasão. Nesse cenário, o Estado Civil dos alunos poderia desencadear uma série de fatores, como a diminuição da frequência às aulas e a queda no índice de rendimento acadêmico (I.R.A.), resultando na deterioração da Situação do Período, que, por sua vez, leva ao abandono ou ao cancelamento da matrícula. Embora plausível, essa abordagem não exclui a possibilidade de outros fatores também contribuírem para o fenômeno.

A segunda possibilidade sugere que a predominância do Estado Civil pode não indicar uma relação causal direta com a evasão, mas sim a ausência de dados ou informações mais precisas na análise. Nesse caso, fatores externos não capturados pelo conjunto de dados atual poderiam influenciar a frequência e o desempenho dos alunos, levando-os a evadir. Essa visão abre espaço para a investigação de variáveis adicionais que possam estar fora do escopo das análises realizadas.

A terceira possibilidade relaciona a evasão à dificuldade de adaptação inicial ao curso, influenciada pela estrutura curricular. Essa hipótese considera que a combinação de uma grade curricular desafiadora com os fatores já identificados (como o Estado Civil) contribui para o abandono do curso. Apesar dessas hipóteses, a análise sugere que ainda há dados a serem explorados, o que dá continuidade ao estudo nas próximas etapas.

Call:

```
lm(formula = tipo_veiculo ~ situacao_periodo + indice_rendimento +  
renda + estado_civil + escolaridade + frequencia + residencia +  
sexo + etnia + transporte, data = alunos)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.04540	-0.00432	-0.00106	0.00267	0.95310

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.622e-03	1.087e-02	0.241	0.8096
situacao_periodo	-1.507e-03	2.201e-03	-0.685	0.4940
indice_rendimento	6.783e-05	7.886e-05	0.860	0.3902
renda	-5.075e-05	2.755e-04	-0.184	0.8539





estado_civil	-2.126e-03	4.490e-03	-0.473	0.6362
escolaridade	3.537e-04	9.212e-03	0.038	0.9694
frequencia	-1.212e-04	6.916e-05	-1.753	0.0804
residencia	-3.908e-03	8.339e-03	-0.469	0.6396
sexo	4.829e-03	5.251e-03	0.920	0.3584
etnia	7.226e-04	2.166e-03	0.334	0.7389
transporte	1.036e+00	9.833e-03	105.336	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04931 on 392 degrees of freedom

Multiple R-squared: 0.967, Adjusted R-squared: 0.9662

F-statistic: 1150 on 10 and 392 DF, p-value: < 2.2e-16

Código 8: Resultados da regressão múltipla feita do Tipo de Veículo utilizado pelo estudante durante o período analisado, em comparação com os demais indicadores

Com base nos resultados anteriores, foi realizada uma regressão linear múltipla para avaliar quais fatores exercem maior impacto sobre o Tipo de Veículo utilizado pelos alunos matriculados. A análise revelou que o fator com maior influência sobre o Tipo de Veículo é o Transporte. Como os dados relacionados ao Transporte já haviam sido analisados, conclui-se que o Tipo de Transporte é um dos principais indicadores que afetam a permanência dos alunos no curso.

Nesse contexto, surge a necessidade de realizar um estudo mais detalhado de cada resultado individualmente dentro do universo de análise, buscando compreender as relações e distribuições no modelo. Para isso, a tabela a seguir organiza os indicadores do banco de dados, destacando a quantidade associada a cada propriedade. Na coluna da esquerda, encontra-se a propriedade analisada, com um sumário no formato "Chave: Valor," descrevendo o significado de cada dado. Na coluna da direita, são apresentados os dados em valores numéricos, organizados pelo padrão "Código", "Tipo", "Quantidade" e "Percentual", permitindo uma análise quantitativa e proporcional dos indicadores em relação ao todo.





Tabela 2: Índice de contagem dos indicadores e relação entre cada tipo de estudante com o indicador

Propriedade	Código	Tipo	Quantidade	Percentual (%)
Estado Civil				
	0	Solteiro	317	78
	1	Casado	73	18
	2	Divorciado	7	1.7
	3	União estável	6	1.4
Tipo de Veículo				
	0	Próprio	375	93
	1	Outros	27	6
	2	Municipal	1	0.2
Tipo de Transporte				
	0	Privado	375	93
	1	Público	28	7
Situação do Curso				
	0	Matriculado	197	48
	1	Evasão	148	36
	2	Cancelamento	26	6
	3	Trancado Voluntariamente	9	2
	4	Transferência Externa	4	1
	5	Transferência Interna	1	0.2
	6	Formado	17	4
	7	Cancelamento Compulsório	1	0.2
Situação do Período				
	0	Matriculado	197	48
	1	Evasão	141	35
	2	Cancelamento	24	6
	3	Fechado	13	3
	4	Em aberto	20	5
	5	Transferência Externa	3	0.7
	6	Transferência Interna	1	0.2
	7	Formado	4	1

A análise dos dados revelou algumas distribuições importantes entre as variáveis observadas. No que diz respeito ao Estado Civil, a maioria dos alunos matriculados encontra-se na condição de “Solteiro,” representando 78% do total, enquanto os “Casados” correspondem a 18%. Em relação ao Tipo de Veículo, a maioria dos estudantes utiliza veículos “Próprios,” o que representa 93% do total, seguidos por veículos classificados como “Outros,” que compreendem 6% dos alunos.





Quanto ao Tipo de Transporte, os dados mostram que 93% dos alunos utilizam transporte “Privado,” enquanto apenas 7% utilizam transporte “Público.” Sobre a Situação no Curso, 48% dos alunos permanecem “Matriculados,” 4% estão “Formados,” e 36% já “Evadiram.” Outros 6% optaram pelo “Cancelamento,” e uma parcela menor trancou voluntariamente, foi transferida ou teve a matrícula cancelada de forma compulsória.

Ao observar a Situação do Período, 48% dos alunos ainda estão “Matriculados,” 35% estão “Evadidos,” 5% mantêm a matrícula “Em Aberto,” 6% optaram pelo “Cancelamento” e apenas 1% estão no status de “Formando.” Esses resultados refletem as tendências gerais de permanência, evasão e formatura ao longo do período analisado.

Tabela 3: Indicadores da Situação no Curso por: (a) Estado Civil e (b) Tipo de Veículo

Propriedade	Tipo	Matriculado	Evasão	Cancelamento	Trancado Voluntariamente	Transf. Externo	Transf. Interno	Formado	Cancelamento Compulsório
Estado Civil									
Tipo de Veículo	Solteiro	171	103	22	3	3	1	14	0
	Casado	21	38	3	6	1	0	3	1
	Divorciado	2	4	1	0	0	0	0	0
	União estável	3	3	0	0	0	0	0	0
	Próprio	181	141	23	9	3	1	16	1
	Público	16	6	3	0	1	0	1	0
	Outros	0	1	0	0	0	0	0	0

A tabela acima apresenta um sumário dos dados previamente analisados, destacando a quantidade que cada categoria representa em relação às demais. Na primeira parte, são exibidos os dados sobre o Estado Civil comparados à Situação do Curso dos alunos matriculados. Na segunda parte, são apresentados os dados sobre o Tipo de Veículo comparados à Situação do Curso.

Com base nesse sumário, conclui-se que, em relação à Situação do Curso, a maioria dos alunos com o Estado Civil “Casado” ou “Divorciado” está na condição de evadidos ou transferidos. Em contraste, a maioria dos alunos com o Estado Civil “Solteiro” permanece





matriculada. Entre os que utilizam veículos “Próprios,” 181 permanecem matriculados, enquanto apenas 16 alunos com veículos classificados como “Outros” estão matriculados.

O gráfico correspondente ilustra a relação entre o Rendimento Acadêmico e a Frequência dos alunos, utilizando o Tipo de Veículo como parâmetro de separação. Essa visualização destaca a interação entre esses fatores, facilitando a identificação de padrões no desempenho acadêmico em função dos meios de transporte utilizados pelos estudantes.

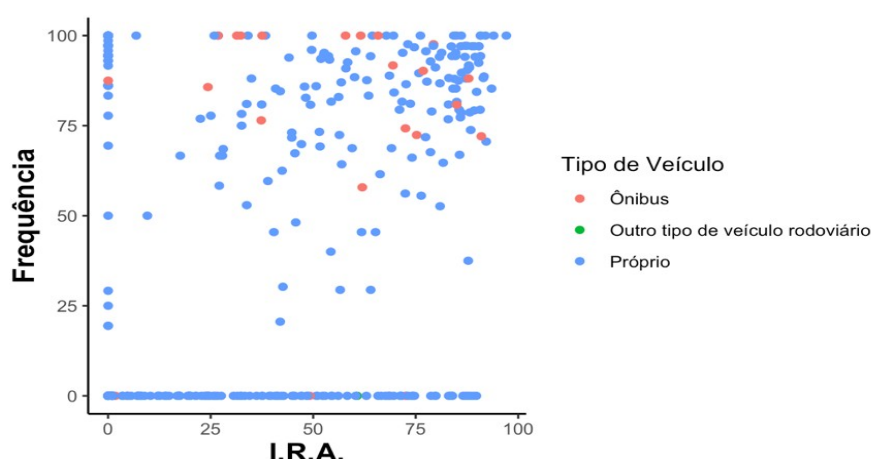


Figura 6: Dispersão da Frequência X I.R.A categorizado pelo Tipo de Veículo

A partir desta análise, é possível extrair algumas informações relevantes. Uma parcela significativa dos alunos matriculados apresenta baixa Frequência, mas mantém um Índice de Rendimento Acadêmico (I.R.A.) acima da média. Esse comportamento reflete o sistema de atribuição do I.R.A., que inicia com um valor padrão de 100 e é gradualmente reduzido à medida que o aluno não registra presença nas aulas.

Além disso, foi observado que muitos alunos com Frequência acima de 50% possuem um I.R.A. inferior a 50%, indicando que uma alta frequência nas aulas não garante, necessariamente, um bom desempenho acadêmico. Um padrão semelhante foi identificado entre alunos que utilizam ônibus como meio de transporte: aqueles com níveis de I.R.A. acima de 65% não apresentam, necessariamente, uma frequência superior a 90%. Da mesma



forma, alunos que utilizam ônibus e têm 100% de frequência nem sempre possuem um I.R.A. superior a 65%. Esse comportamento também se aplica aos alunos que utilizam veículo próprio, reforçando que outros fatores além da frequência influenciam o rendimento acadêmico. A partir desta análise, é possível extrair algumas informações relevantes. Uma parcela significativa dos alunos matriculados apresenta baixa Frequência, mas mantém um Índice de Rendimento Acadêmico (I.R.A.) acima da média. Esse comportamento reflete o sistema de atribuição do I.R.A., que inicia com um valor padrão de 100 e é gradualmente reduzido à medida que o aluno não registra presença nas aulas.

Além disso, foi observado que muitos alunos com Frequência acima de 50% possuem um I.R.A. inferior a 50%, indicando que uma alta frequência nas aulas não garante, necessariamente, um bom desempenho acadêmico. Um padrão semelhante foi identificado entre alunos que utilizam ônibus como meio de transporte: aqueles com níveis de I.R.A. acima de 65% não apresentam, necessariamente, uma frequência superior a 90%. Da mesma forma, alunos que utilizam ônibus e têm 100% de frequência nem sempre possuem um I.R.A. superior a 65%. Esse comportamento também se aplica aos alunos que utilizam veículo próprio, reforçando que outros fatores além da frequência influenciam o rendimento acadêmico.

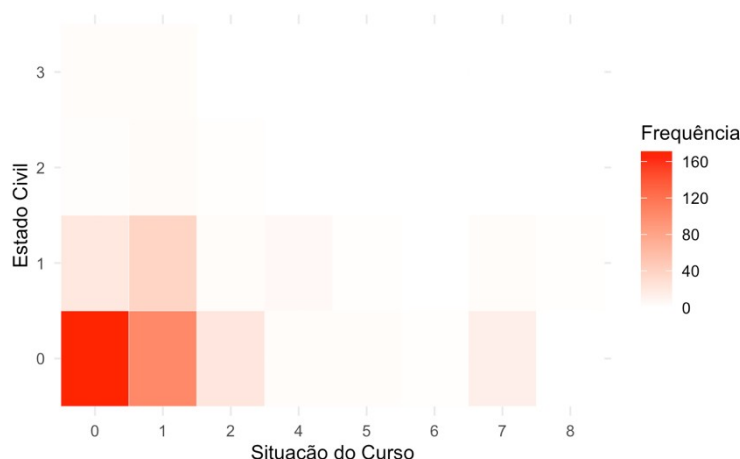


Figura 7: Mapa de Calor - Situação do Curso X Estado Civil



O gráfico apresentado ilustra o índice de Frequência dos alunos matriculados, comparando-o com seu Estado Civil e Situação no Curso. A intensidade da cor no gráfico reflete os valores dos indicadores, sendo que cores mais intensas representam maiores índices de Frequência. A análise do gráfico revela que os alunos solteiros apresentam a maior frequência entre os matriculados. No entanto, o mesmo padrão é observado para a evasão, com os solteiros também representando o maior índice de evasão. Esses resultados destacam a necessidade de investigar os fatores que contribuem para esse comportamento, considerando que, apesar de terem maior presença nas aulas, muitos desses alunos acabam evadindo do curso.

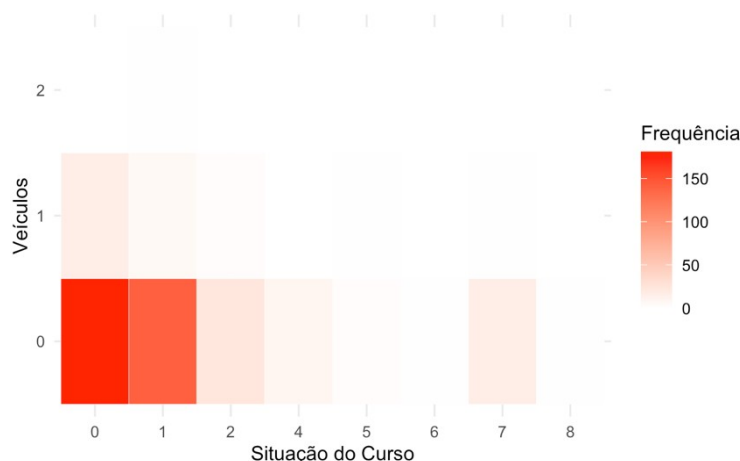


Figura 8: Mapa de Calor - Situação no Curso X Tipo de Veículos

O gráfico apresentado compara o índice de Frequência dos alunos matriculados com o tipo de Veículo utilizado e sua Situação no Curso. A intensidade da cor no gráfico representa os valores dos indicadores, com cores mais fortes indicando maior frequência.

A análise do gráfico revela que a maioria dos alunos matriculados utiliza veículos próprios, e o mesmo ocorre entre os alunos que evadiram. Entretanto, observa-se que outros tipos de veículos apresentam uma frequência maior entre os alunos matriculados do que entre os evadidos. Esses resultados sugerem que o uso de veículos próprios é predominante,



independentemente da situação acadêmica, enquanto outros meios de transporte estão mais associados à permanência no curso.

Pearson's Chi-squared test
data: tabela_contingencia_um(Estado Civil e Situação do Curso)
X-squared = 37.353, df = 21, p-value = 0.01533

Pearson's Chi-squared test
data: tabela_contingencia_dois(Transporte e Situação do Curso)
X-squared = 5.2611, df = 7, p-value = 0.6281

Código 9: Teste Qui-Quadrado

Os resultados do teste qui-quadrado estão apresentados em dois quadros: Contingência Um e Contingência Dois. Para a primeira tabela, o teste resultou em um p-value de aproximadamente 0,01533, indicando divergência significativa entre as colunas e os resultados esperados. Isso sugere que há diferenças estatisticamente significativas na relação entre o Estado Civil e a Situação no Curso, mas não aponta para uma anomalia nos dados analisados.

Na segunda tabela, o p-value foi de aproximadamente 0,6281, indicando que as colunas da tabela seguem o comportamento esperado. Assim, a relação entre o Tipo de Transporte e a Situação no Curso está alinhada com o que é descrito na literatura, sem evidências de desvios significativos.

Contudo, é importante destacar que o teste qui-quadrado fornece informações aproximadas e, como alertado pelo próprio método, os resultados podem conter imprecisões. Por isso, interpretações devem ser feitas com cautela, considerando possíveis limitações dos dados analisados.





Tabela 4: Percentual da Situação no Curso X Estado Civil e Tipo de Transporte

Propriedade	Tipo	Matricula do	Evasão	Cancelamento	Trancado Voluntariamente	Transf. Externo	Transf. Interno	Formado	Cancelamento Compulsório
Estado Civil									
Tipo de Transporte	Solteiro	0.424	0.255	0.054	0.007	0.007	0.002	0.034	0.000
	Casado	0.052	0.094	0.007	0.014	0.002	0.000	0.007	0.002
	Divorciado	0.004	0.009	0.002	0.000	0.000	0.000	0.000	0.000
	União estável	0.007	0.007	0.000	0.000	0.000	0.000	0.000	0.000
	Próprio	0.449	0.349	0.0570	0.022	0.007	0.002	0.039	0.002
	Outros	0.039	0.017	0.0074	0.000	0.002	0.000	0.002	0.000

A primeira tabela apresentada detalha o percentual de cada Estado Civil dentro do universo de Situação do Curso, permitindo compreender a distribuição proporcional dos estados civis entre diferentes situações acadêmicas. A segunda tabela, por sua vez, apresenta o percentual de diferentes Tipos de Veículos utilizados pelos alunos, também no contexto da Situação do Curso, oferecendo uma visão estatística mais abrangente dos dados analisados.

Por fim, foi realizada uma análise utilizando o método de Random Forest seguindo as métricas dispostas na tabela abaixo para calcular a importância, acurácia, sensibilidade e relevância do modelo.

Tabela 5: Métricas utilizadas para validar a precisão e importância do modelo Random Forest utilizado

Critérios de Desempenho	Cálculo
Acurácia	$(VP + VN) / (VP + VN + FP + FN)$
Precisão	$VP / (VP + FP)$
Sensibilidade (Recall)	$VP / (VP + FN)$
Especificidade	$VN / (VN + FP)$
F1-Score	$2 * Precisão * Sensibilidade / (Precisão + Sensibilidade)$

Legenda: VP (Verdadeiro Positivo): casos positivos corretamente classificados pelo modelo; VN (Verdadeiro

Negativo): casos negativos corretamente classificados; FP (Falso Positivo): casos negativos incorretamente classificados como positivos; FN (Falso Negativo): casos positivos incorretamente classificados como negativos

Procedendo assim com o objetivo de determinar a importância das variáveis em relação a uma variável dependente. No caso a seguir, a variável dependente escolhida foi a



Situação do Curso, identificada como uma das três mais relevantes nos dados analisados anteriormente. Essa abordagem possibilita uma avaliação mais detalhada da contribuição de cada variável para a explicação da Situação do Curso, destacando os fatores mais influentes e suas relações no modelo.

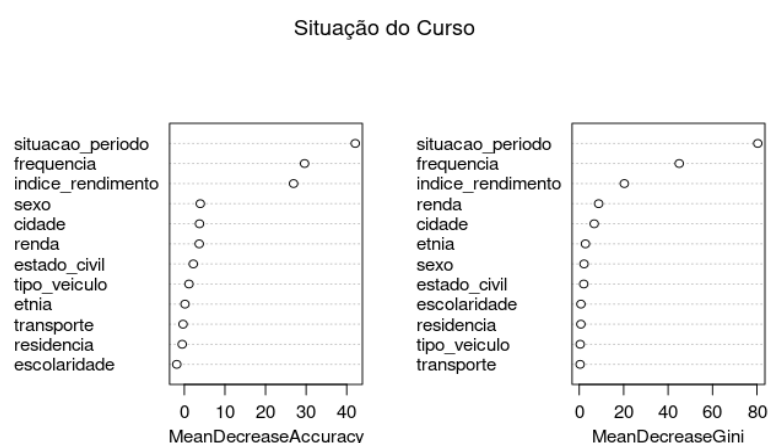


Figura 9: Floresta Randômica para Situação no Curso

No painel à esquerda, observa-se o impacto de cada variável no desempenho do modelo quando seus valores são embaralhados, enquanto o gráfico à direita mede a contribuição de cada variável para a homogeneidade, ou seja, o quão similares são as observações nos nós das árvores do modelo.

Os resultados indicam que as variáveis "Situação do Período", "Frequência" e "Índice de Rendimento" são as mais relevantes, destacando-se em ambas as métricas ($%IncMSE$ e $IncNodePurity$). Embora variáveis como "Renda", "Cidade" e "Transporte" também sejam importantes, seu impacto no modelo é relativamente menor.

A tabela 6 a seguir apresenta as métricas de desempenho alcançadas pelo modelo utilizado.



Tabela 6: Métricas alcançadas após a execução do modelo para validar a precisão e importância do modelo
Random Forest utilizado

Critérios de Desempenho	Valor
Acurácia	0.9745763
Precisão	0.9579710
Sensibilidade (Recall)	0.7261905
Especificidade	0.9940206
F1-Score	0.8953225

Dando continuidade à análise, um novo modelo Random Forest foi desenvolvido, desta vez utilizando o Índice de Rendimento Acadêmico (I.R.A.) como variável dependente. Essa abordagem permite uma análise mais focada na identificação dos fatores que influenciam diretamente o desempenho acadêmico dos alunos, ampliando as possibilidades de interpretação e intervenção.

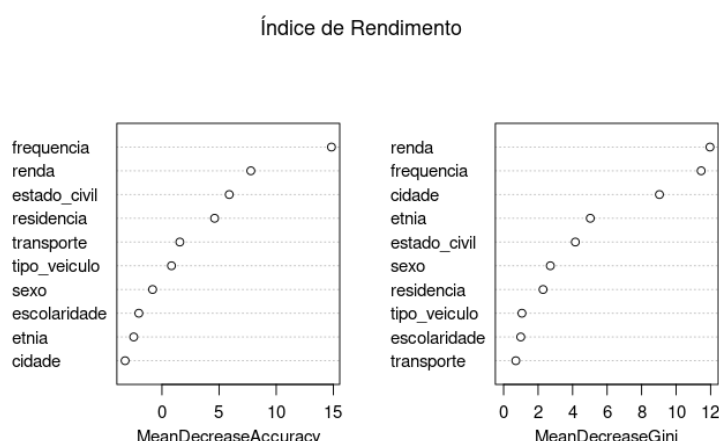


Figura 10: Floresta Randômica para Índice de Rendimento

Os resultados desta análise evidenciam que a variável "Renda" se destaca no *IncNodePurity*, contribuindo significativamente para melhorar as divisões nas árvores do modelo. Enquanto isso, variáveis como "Frequência", "Situação do Curso" e "Escolaridade"



apresentam relevância moderada em ambos os painéis, indicando que também desempenham papéis importantes, mas com menor impacto em comparação à Renda.

A tabela 7 a seguir apresenta as métricas de desempenho alcançadas pelo modelo utilizado.

Tabela 7: Métricas alcançadas após a execução do modelo para validar a precisão e importância do modelo Random Forest utilizado

Critérios de Desempenho	Valor
Acurácia	0.9189189
Precisão	0.9565217
Sensibilidade (Recall)	0.9166667
Especificidade	0.9230769
F1-Score	0.9361702

Como exemplo final, o gráfico apresentado ilustra os resultados de uma análise utilizando Random Forest para identificar as variáveis com maior importância na previsão da Renda. Essa abordagem permite compreender melhor as interações entre as variáveis e como elas influenciam o comportamento observado, fornecendo insights valiosos para futuras aplicações e intervenções baseadas nos dados.

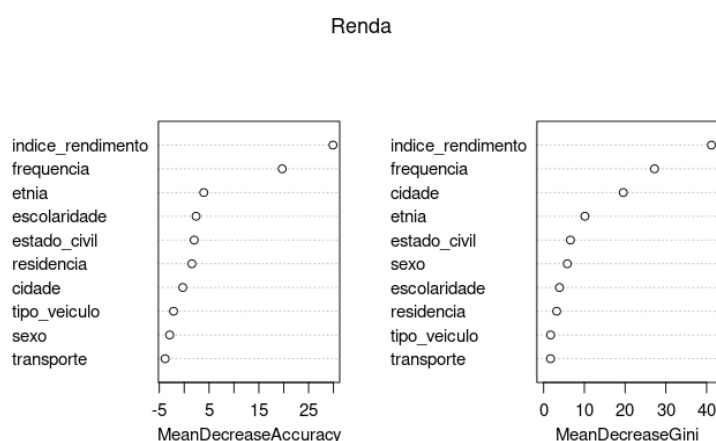


Figura 11: Floresta Randômica para Renda



Como resultado da análise, não foi possível identificar uma única variável como a mais importante para prever a Renda, pois sua relevância varia de acordo com a métrica utilizada.

No caso do *%IncMSE*, as variáveis que demonstraram maior importância quando embaralhadas foram o "Índice de Rendimento Acadêmico", "Frequência" e "Etnia". Já no *IncNodePurity*, as variáveis que mais contribuíram para melhorar a homogeneidade do modelo foram o "Índice de Rendimento Acadêmico" e a "Cidade".

A tabela 8 a seguir apresenta as métricas de desempenho alcançadas pelo modelo utilizado.

Tabela 8: Métricas alcançadas após a execução do modelo para validar a precisão e importância do modelo
Random Forest utilizado

Critérios de Desempenho	Valor
Acurácia	0.5555556
Precisão	0.5123457
Sensibilidade (Recall)	0.5082621
Especificidade	0.7652820
F1-Score	0.5047138

Os resultados alcançados e informados pelas métricas indicam que as informações evidenciadas na Figura 11 não tem um alto nível de confiabilidade ou assertividade. Essa informação torna-se perceptível principalmente no critério de Precisão e Sensibilidade, os quais informam que, neste caso, o modelo alcançou uma precisão medíocre, acertando 50% das previsões realizadas.

Os resultados acima indicam que diferentes variáveis desempenham papéis importantes em contextos distintos, dependendo do aspecto do modelo analisado. Essa variabilidade ressalta a necessidade de considerar múltiplas métricas ao avaliar a influência de variáveis em modelos preditivos.





5. CONCLUSÕES E TRABALHOS FUTUROS

Este estudo analisou dados relacionados à evasão escolar, com o objetivo de identificar padrões e fatores que contribuem para o afastamento de estudantes do ambiente educacional. Com base nos trabalhos correlatos e na análise dos resultados, foi possível concluir que a evasão escolar é um fenômeno multifacetado, influenciado por fatores individuais, sociais e institucionais. Dentre os principais achados, destaca-se que a decisão de evadir está fortemente associada ao apoio familiar, à qualidade das instituições e ao encorajamento dos pares, conforme discutido por Silva, Moraes & Costa (2018).

Além disso, foi observado que a evasão limita oportunidades no mercado de trabalho e intensifica desigualdades sociais e econômicas, conforme apontado por Silva Júnior et al. (2016). Outro aspecto relevante são os desafios enfrentados nos cursos à distância, principalmente a dificuldade de acesso a tecnologias e a ausência de proximidade com professores, conforme discutido por Gaspar & Baida (2020). A comparação com trabalhos correlatos demonstra que, embora diversas iniciativas tenham sido implementadas como ferramentas preditivas e programas de suporte aos alunos, a evasão escolar permanece como um fenômeno complexo, o que reforçou a necessidade de análises específicas de variáveis socioeconômicas e de desempenho acadêmico para melhor compreensão do problema.

Os resultados da análise evidenciam que há diferenças significativas entre os sexos no que diz respeito à renda per capita, frequência e índices de rendimento acadêmico (I.R.A.). Observou-se que estudantes do sexo masculino tendem a ter maior renda e um I.R.A. mediano superior em relação ao sexo feminino, embora ambos os sexos apresentem altos índices de evasão. Em contraposto, o comportamento geral indica que fatores como renda per capita e I.R.A. têm uma relação pouco significativa, uma vez que a dispersão dos dados mostra exemplos de alunos com alta renda e baixos rendimentos acadêmicos, e vice-versa. Além disso, a frequência e o tipo de transporte utilizado também demonstraram impacto indireto no desempenho e na permanência dos estudantes no curso.





A análise estatística, incluindo regressões lineares e múltiplas, identificou que a frequência e o I.R.A. são variáveis determinantes para a situação dos alunos no curso, como matrícula ativa, evasão ou cancelamento. A correlação negativa entre frequência e situação do período sugere que, à medida que os estudantes frequentam menos as aulas, aumentam as chances de evasão. Além disso, o estado civil dos alunos aparece como um fator que pode influenciar a evasão, especialmente entre os casados e divorciados, que apresentam maior tendência a abandonar o curso. A regressão também destacou que fatores externos, como transporte e tipo de veículo, possuem menor influência direta na evasão, entretanto, contribuem diretamente para entender as dinâmicas de permanência, sendo fortemente associados a frequência e I.R.A.

Por fim, análises mais complexas, como Random Forest, identificaram variáveis críticas para explicar a situação do curso, o rendimento acadêmico e a renda. As variáveis mais significativas, como frequência, situação do período e índice de rendimento, foram consistentes em demonstrar alta relevância nos modelos preditivos. No entanto, variáveis como cidade, etnia, renda e estado civil também desempenham papéis moderados em determinados contextos. O estudo sugere que políticas institucionais voltadas para aumentar o engajamento dos alunos e melhorar o suporte social e acadêmico garantindo meios de locomoção apropriados, podem contribuir para reduzir a evasão e melhorar o desempenho geral dos estudantes.

Os resultados finais revelam que a evasão escolar pode ser influenciada por fatores que vão além das variáveis analisadas diretamente, como o Estado Civil do(a) estudante. A ausência de dados mais precisos ou a falta de inclusão de variáveis externas no modelo podem ter limitado a compreensão completa do fenômeno. Além disso, a dificuldade de adaptação inicial ao curso, possivelmente associada a uma estrutura curricular desafiadora, surge como outro fator relevante que, em conjunto com variáveis como Estado Civil e Transporte, pode aumentar as chances de abandono. Essas hipóteses reforçam a necessidade de ampliar as análises para incluir fatores contextuais e psicossociais ainda não considerados no momento de realização da pesquisa.





Para trabalhos futuros, sugere-se aprofundar os modelos preditivos para padrões de evasão escolar, incorporando variáveis psicossociais e comportamentais que ainda não foram amplamente exploradas, expandindo para os demais cursos ofertados pela instituição de ensino. É importante realizar análises específicas em diferentes regiões do estado, escalando se necessário para nível nacional, considerando as particularidades culturais, econômicas e institucionais que influenciam a evasão em cada região. Além disso, pretende-se avaliar a eficácia de programas de intervenção, como mentorias e suporte emocional, bem como examinar os desafios enfrentados por estudantes de cursos à distância e propor soluções tecnológicas para promover maior engajamento.

REFERÊNCIAS

ALVAREZ, K. R. (2020). Permanência e êxito escolar nos institutos federais. *Ensino em Foco*, 3(6), 106–115.

ARAÚJO, E. B. de, & Lima, A. M. de. (2021). O estado da arte sobre evasão escolar nos institutos federais. *Revista Labor*, 1(26), 54–75.

BEZERRA, J. H. S. (2019). Análise da evasão escolar do Instituto Federal de Rondônia – Campus Ji-Paraná utilizando técnicas de mineração de dados (Dissertação de mestrado). Instituto Politécnico do Porto. Disponível em:
<https://recipp.ipp.pt/entities/publication/25ea8a5f-5cac-497a-9b6a-6bc5abba0e74>

COLPO, M. P., THOMPSEN PRIMO, T., AGUIAR, M. S. de, & CECHINEL, C. (2024). Mineração de Dados Educacionais na Predição da Evasão Estudantil: Tendências, Oportunidades e Desafios. *Revista Brasileira De Informática Na Educação*, 32, 220–256. Disponível em: <https://journals-sol.sbc.org.br/index.php/rbie/article/view/3559>

DEKKER, Gerben W.; PECHENIZKIY, Mykola; VLEESHOUWERS, Jan M. Applying data mining techniques to predict student dropout: a case study. In: *Educational Data Mining 2009 – Proceedings of the 2nd International Conference on Educational Data Mining*. Córdoba, Espanha, 2009. Disponível em:
https://www.researchgate.net/publication/328158354_Applying_Data_Mining_Techniques_to_Predict_Student_Dropout_A_Case_Study. Acesso em: 16 mar. 2026.





GASPAR C., A.; BAIDA, C. A. (2020). Análise quantitativa das causas da evasão escolar dos cursos técnicos de nível médio integrado dos institutos federais de educação, ciência e tecnologia. Encontro Internacional de Gestão, Desenvolvimento e Inovação (EIGEDIN), 4(1). Disponível em: <https://periodicos.ufms.br/index.php/EIGEDIN/article/view/10934>

JESUS, Jeferson Andrade de; GUSMÃO, Renê Pereira de. Investigação da evasão estudantil por meio da mineração de dados e aprendizagem de máquina: um mapeamento sistemático. Revista Brasileira de Informática na Educação, 2024. Disponível em: <https://journals-sol.sbc.org.br/index.php/rbie/article/view/3466>. Acesso em: 16 mar. 2026.

RAMOS NETO, J. O. (2019). A evasão escolar nos Institutos Federais de Educação, Ciência e Tecnologia: Uma análise dos planos estratégicos de permanência e êxito. Educação em Revista, 20(2), 7–24. Disponível em: <https://revistas.marilia.unesp.br/index.php/educacaoemrevista/article/view/8999>

R CORE TEAM (2024). *_R: A Language and Environment for Statistical Computing_*. R Foundation for Statistical Computing, Vienna, Austria. Disponível em: <https://www.r-project.org>

SILVA JÚNIOR, A. S., et al. (2016). A evasão escolar no ensino à distância nas escolas públicas: Propostas de intervenções pedagógicas. Congresso ABED, Porto Velho, RO, Brasil. Disponível em: https://www.researchgate.net/publication/308316311_A_EVASAO_ESCOLAR_NO_ENSINO_A_DISTANCIA_NAS_ESCOLAS_PUBLICAS_PROPOSTAS_DE_INTERVENCOES_PEDAGOGICAS

SILVA, Rafaela Costa; MORAES, Ana Flávia de Moraes; COSTA, Geraldo Vieira da. FATORES QUE PODEM INTERFERIR NA EVASÃO ESCOLAR EM UMA INSTITUIÇÃO DE ENSINO SUPERIOR PRIVADA. Revista Cesumar – Ciências Humanas e Sociais Aplicadas, [S. l.], v. 23, n. 2, p. 205–228, 2018. Disponível em: <https://periodicos.unicesumar.edu.br/index.php/revcesumar/article/view/6486>. Acesso em: 16 mar. 2026

Recebido em: 17/03/2026

Aprovado em: 28/03/2026

Publicado em: 07/04/2026

