Análises classificatórias de aprendizado de máquina para identificação de fatores de abandono escolar

Daniella Martins Vasconcellos[1], **Rafael Ferreira Mello**[2], **Thales Vieira**[3], **Elaine H. T. Oliveira**[4,5] **e Isabela Gasparini**[1]

[1]Universidade do Estado de Santa Catarina, [2] Universidade Federal Rural de Pernambuco
[3] Núcleo de Excelência em Tecnologias Sociais (NEES) - Universidade Federal de Alagoas
[4] CESAR School e [5] Universidade Federal do Amazonas

Resumo. *A coleta de dados educacionais é essencial para a gestão eficiente de recursos e atendimento das necessidades da população, sendo amplamente utilizada por pesquisadores para compreender e melhorar a educação. Nesse contexto, este estudo visa identificar quais os fatores de risco de evasão escolar de maior impacto com base no Instrumento de Avaliação de Risco de Evasão Escolar (IAFREE). Foram realizadas análises estatísticas e classificatórias com aprendizado de máquina para compreender as variáveis com maior significância correlacional, ao final destacando-se as relações do estudante com o ambiente familiar e a distância da casa até a escola.*

Abstract. *The collection of educational data is essential for the efficient management of resources and meeting the needs of the population, and is widely used by researchers to understand and improve education. In this context, this study aims to identify which risk factors for school dropout have the greatest impact based on the School Dropout Risk Assessment Instrument (IAFREE). Statistical and classificatory analyzes were carried out with machine learning to understand the variables with greater correlational significance, in the end highlighting the student's relationships with the family environment and the distance from home to school.*

1. Introdução

Nas últimas décadas, o sistema educacional brasileiro foi alvo de diversas políticas públicas que visam garantir o acesso, permanência e qualidade do ensino em todas as suas etapas e modalidades. Entre as políticas mais relevantes estão o Plano Nacional de Educação (PNE) [BRASIL 2001], o Fundo de Manutenção e Desenvolvimento da Educação Básica (FUNDEB) [BRASIL 2020], e o Programa Nacional de Alimentação Escolar (PNAE) [BRASIL 1955], sendo o Ministério da Educação (MEC) o órgão responsável por coordenar e executar essas políticas em âmbito nacional.

Como os programas governamentais do MEC têm como objetivo o fornecimento de serviços públicos essenciais para atender a variadas necessidades educacionais da população, é fundamental que se colete informações e dados sobre os cidadãos que usufruem dos serviços para melhor utilização de recursos e retorno à população. Por exemplo, o Portal Único de Acesso ao Ensino Superior (ProUni) acessa dados socioeconômicos dos estudantes que estão em bases governamentais para poder analisar possíveis dispensas de apresentação de documentos (como comprovação de renda familiar mensal bruta ou a situação da pessoa com deficiência) [BRASIL 2005].

DOI: 10.5753/sbie.2024.242559

No âmbito acadêmico, pesquisadores utilizam os dados do MEC para diversos fins, como por exemplo, para avaliar a eficácia de programas educacionais [Vargas and Zuccareli 2021], para identificar as necessidades de formação de professores, ou para analisar as desigualdades educacionais no país e desenvolver políticas públicas que buscam melhorar a qualidade da educação [Santos et al. 2019].

Tendo conhecimento da importância das pesquisas educacionais, peça chave para a elaboração de programas educacionais eficazes e para o monitoramento contínuo da qualidade da educação brasileira, o presente artigo analisa os dados de um questionário aplicado para mais de 17 mil estudantes brasileiros afim de verificar qual categoria de prováveis motivos de abandono escolar possui maior relevância dentro da base de dados.

O presente artigo está organizado do seguinte modo: na primeira seção, a introdução, é apresentada a motivação da pesquisa. A segunda seção apresenta a fundamentação teórica. Na terceira seção descreve o contexto em que os dados foram coletados e realizada uma análise descritiva dos dados. Na quarta seção são analisadas as variáveis com maior relevância estatística. A quinta seção consiste na aplicação dos algoritmos de aprendizado de máquina para classificação preditiva das variáveis. Por fim, após todas as avaliações serem feitas, a sexta seção abrange a conclusão e trabalhos futuros.

2. Fundamentação Teórica

Educational Data Mining, ou mineração de dados educacionais, é uma disciplina dedicada ao desenvolvimento de métodos para explorar dados provenientes de ambientes educacionais, tendo como objetivo usar esses métodos para compreender melhor os alunos e os contextos em que eles aprendem [Koç 2017]. Para realizar as análises desses dados, aborda-se a área do *Learning Analytics*, ou analíticas de aprendizagem, que envolve a aplicação de métodos quantitativos e qualitativos para obter informações úteis sobre o desempenho dos estudantes, padrões de comportamento, eficácia do ensino, entre outros aspectos relacionados ao aprendizado [Aldowah et al. 2019]. Ambas as áreas têm como objetivo a melhoria de processos de ensino e aprendizagem por meio da análise de dados em larga escala, de maneira sistematizada, que possam auxiliar na ampliação de processos de avaliação, compreensão de problemas e planejamento de intervenções [Moissa et al. 2015]. Sendo assim, foram estudadas técnicas aplicáveis a ambas as áreas para decisão de quais técnicas e métricas seriam aplicadas a este estudo.

Após uma pesquisa na literatura pelos algoritmos mais utilizados para classificação de dados, foram escolhidos para comparação os seguintes modelos, pensando em representação de uma variedade de abordagens em aprendizado de máquina: **Decision Tree (Árvores de Decisão):** modelos de aprendizado de máquina que utilizam uma estrutura de árvore para tomar decisões; **Random Forest:** técnica que utiliza um conjunto de árvores de decisão para realizar a classificação ou regressão; **Naive Bayes Categórico:** classificador probabilístico baseado no Teorema de Bayes. A versão categórica assume que os atributos são discretos e seguem uma distribuição de probabilidade específica; **Adaptive Boosting (AdaBoost):** algoritmo de boosting que combina vários classificadores fracos para formar um classificador forte. Ele atribui pesos diferentes às instâncias no conjunto de treinamento, dando mais foco às instâncias classificadas incorretamente; e **K-Nearest-Neighboors (K-Vizinhos-Próximos):** algoritmo de aprendizado supervisionado que classifica uma instância com base nas clas-

ses de seus k vizinhos mais próximos no espaço de atributos. Para avaliação da comparação entre os algoritmos, foram escolhidas as seguintes métricas de análise: Acurácia, Precisão, Recall Score e F1 Score, que são métricas já amplamente utilizadas para avaliação da qualidade e o desempenho de modelos de predição da evasão de estudantes [Cechinel and da Silva Camargo 2020].

3. Estudo da base

Os dados analisados foram providos pelo Plano de Trabalho do Termo de Execução Descentralizada (TED) nº 10974/2022, projeto celebrado entre o Núcleo de Excelência em Tecnologias Sociais (NEES), coordenado pela Universidade Federal de Alagoas (UFAL), e o Ministério da Educação (MEC). O TED possui como meta geral o desenvolvimento de um sistema de alerta preventivo dos riscos de abandono e evasão escolar [SAP 2022]. Para tal utilizou-se um questionário validado pela literatura nomeado "Instrumento de Avaliação dos Fatores de Risco de Evasão Escolar" (IAFREE) [de Vasconcelos et al. 2023]. O IAFREE é formado por 36 questões distintas com respostas baseadas na escala de Likert, com valores inteiros entre 1 (Discordo Totalmente) e 7 (Concordo Totalmente), ou 0 caso o estudante optasse por não responder. Para maior detalhamento e compreensão do escopo multifacetado de influências do abandono escolar, foram elaboradas cinco dimensões de risco, cada uma associada a um tipo diferente de envolvimento do estudante com seu ambiente escolar:

- **Estudante-Escola (E-ESC):** trata de como o estudante se sente impactado pelas condições materiais da sua instituição de ensino, não apenas relacionadas à infraestrutura física mas também aos recursos didáticos do estudante. Os fatores relacionados à esta dimensão são: Condições Materiais da Escola (E-ESC1), envolvendo estrutura física, acesso a internet, alimentação, limpeza, e Condições Materiais do(a) Estudante (E-ESC2), embarcando condições materiais de acesso, uniforme, espaço próprio para estudo, etc.
- **Estudante-Profissionais da Escola (E-PROF):** aborda como o tratamento dos professores e outros agentes pedagógicos dentro das instituições de ensino impactam o engajamento do estudante. Os fatores relacionados são Inflexibilidade Pedagógica (E-PROF1) e Qualidade Pedagógica (E-PROF2).
- **Estudante-Família (E-FAM):** trata do impacto do núcleo familiar na vida acadêmica do estudante, e da influência do apoio para continuação dos estudos. Os fatores relacionados são: Suporte Familiar (E-FAM1) e Gravidez/parentalidade/atividades de cuidado (E-FAM2).
- **Estudante-Comunidade (E-COM):** explora como medidas socioeducativas e a prevalência de violência no entorno podem criar barreiras ao acesso e à frequência escolar, além de como a desconexão entre a escola e a comunidade pode acentuar a sensação de isolamento dos estudantes. Os fatores relacionados são: Medidas socioeducativas e contextos de violência (E-COM1), Acessibilidade e frequência escolar (E-COM2) e Distanciamento escola – comunidade (E-COM3).
- **Estudante-Estudante (E-EST):** observa-se o impacto do engajamento escolar e os significados atribuídos à escolarização, bem como as interações emocionais e afetivas entre os estudantes. Os fatores relacionados são: Significados da Escolarização/ Engajamento (E-EST1), Aspectos Emocionais e Afetivo (E-EST2) e Reprovações e distorção idade – série (E-EST3).

Cada pergunta foi associada a um fator de risco (cada fator possui três perguntas associadas), e dois ou três fatores foram agrupados em uma única dimensão analisada. Utilizou-se este questionário para a realização da coletânea de entrevistas. Os dados disponibilizados são referentes a esta coletânea de entrevistas realizadas em dezembro de 2022, realizadas de forma presencial, com 17.110 estudantes do primeiro ao nono ano do ensino fundamental, sendo 51,7% do sexo masculino e 48,3% do sexo feminino. Destes, a maior participação foi dos alunos do 6º ano, com 4912 respostas, seguida por uma ligeira diminuição para 4418 no 7º ano e uma redução mais acentuada para 4103 no 8º ano. O 9º ano apresenta um número ainda menor de respondentes, com 3506 alunos. Uma queda drástica é notada nos anos iniciais do ensino fundamental: apenas 61 estudantes do 5º ano, 59 do 1º ano, 23 do 3º ano, 18 do 2º ano e 9 do 4º ano participaram do questionário. Os alunos vieram de 308 escolas diferentes, que variavam em tamanho entre 50 matrículas até mais de 1000 matrículas. Todas eram públicas, sendo 62,3% municipais e 37,7% estaduais, de todas as regiões do Brasil, notando-se uma forte representatividade das escolas municipais na maioria dos estados, com exceção de Goías, Minas Gerais, Roraima, São Paulo e Tocantins, como é possível observar na Tabela 1.

UF	AL	AM	BA	CE	DF	ES	GO	MA	MG	MS	PA	PB	PE	PI	RJ	RN	RR	SC	SE	SP	TO
Estadual	0	1	0	0	1	0	29	0	13	1	16	0	0	0	0	0	4	0	2	5	27
Municipal	7	1	26	19	0	1	0	21	1	1	85	4	2	16	2	8	0	1	3	0	11

Tabela 1. Dependência administrativa das escolas dividida por estado

Uma característica interessante a se observar nos estudantes entrevistados é sua etnia declarada, já que estudantes identificados como pardos, pretos, indígenas e amarelos podem enfrentar desafios adicionais que afetam sua jornada educacional como disparidades socioeconômicas, preconceito e racismo estrutural, que podem influenciar negativamente a experiência desses estudantes nas instituições de ensino [Heringer 2018]. Sendo assim, foi construído o gráfico apresentado na Figura 1 para mostrar a distribuição étnica dos estudantes que responderam o questionário. A categoria com maior número de estudantes é a "Pardo/a", com 12.450 respostas, seguida consideravelmente pela categoria "Branco/a" com 2.443 estudantes. A terceira categoria mais numerosa é a "Preto/a", representando 1.758 estudantes. Em menor quantidade, aparecem os estudantes que se identificam como "Indígena", somando 290, e a categoria "Amarelo/a (Origem Asiática)" com 168 estudantes. A representação visual destes dados permite uma compreensão imediata das proporções relativas de cada grupo étnico dentro do conjunto dos respondentes, destacando a predominância do grupo "Pardo/a" neste contexto específico.

Observando os microdados das etnias do Censo da Educação Básica do ano de 2023, apresentados na Figura 2, nota-se que proporcionalmente, os números de representatividade das etnias são muito semelhantes, com a exceção dos declarados brancos e dos estudantes não declarados, que no Censo são o terceiro maior grupo porém não é uma categoria existente na base estudada, indicando que todos os 17110 respondentes declararam suas etnias.

4. Análise Estatística

Para entender a relevância estatística das dimensões e fatores nas influências das variáveis socioeconômicas observadas na Seção 3, foi calculado o valor p para cada uma delas. O valor p é uma medida que ajuda a avaliar a significância estatística em um teste de

Figura 1. Etnias dos estudantes que responderam o IAFREE

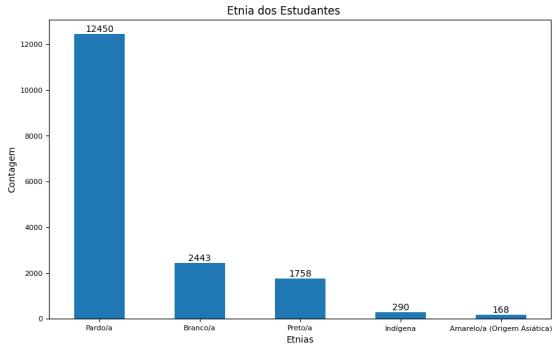
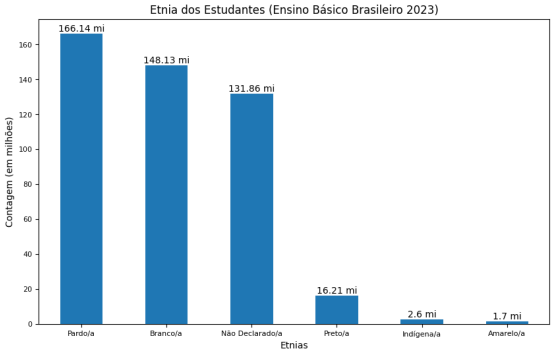


Figura 2. Etnias dos estudantes - Censo 2023



hipótese, mostrando a probabilidade de um resultado igual ou mais extremo ocorrer sob a hipótese nula. Para alcançar esse resultado, define-se a Hipótese Nula (H0), que assume que não há efeito significativo, e a Hipótese Alternativa (H1), que sugere um efeito significativo. Um valor p baixo ($< 0{,}05$) sugere a rejeição da hipótese nula, enquanto um valor p alto indica que não há evidências suficientes para rejeitá-la [Wasserstein and Lazar 2016].

Para analisar essas relações das variáveis socioeconômicas com as dimensões e fatores, a Tabela 2 foi construída. Ela apresenta os p-valores em destaque para todos os resultados menores que 0.05, ou seja, os resultados com significância estatística. Alguns destaques são observados na dimensão E_ESCV (Estudante-Escola), apontando "porte da escola" e "renda familiar" têm forte associação, da mesma forma que a dimensão E_PROFV (Profissionais da Escola) apresenta "sexo", "ano_turma" e "outras_ofertas_educacionais" com p-valores baixos, indicando forte associação entre essas variáveis. Então, observa-se que as variáveis socioeconômicas possuem um impacto distinto nos fatores e dimensões, com algumas se relacionando com mais áreas que outras, para as quais foram geradas alguns gráfios para análise mais aprofundada:

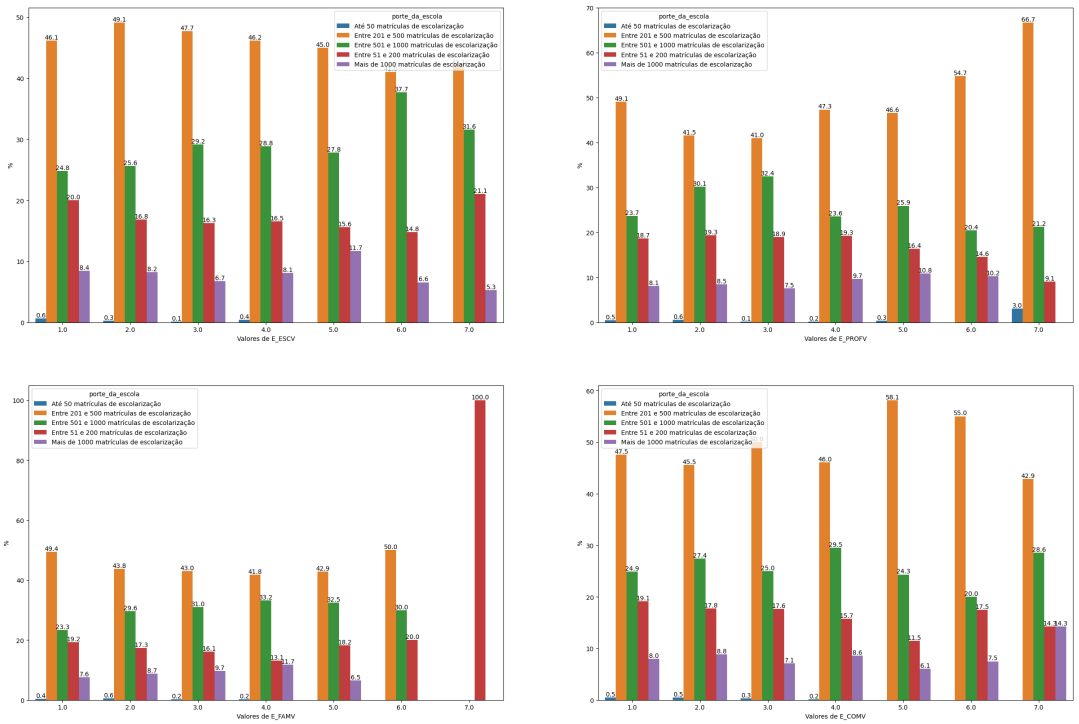
- Dimensões: porte_da_escola, com 4 relações; e sexo, uf, localizacao, ano_turma e id_renda_familiar com 3 relações cada;
- Fatores: dependencia_administrativa com 11 relações; e sexo, com 9 relações.

Tabela 2. P Valor relacionado às Dimensões e Fatores

Categoria	Variável	sexo	uf	localizacao	ano_turma	localidade_ diferenciada	dependencia_ administrativa	id_raca_ etnia	porte_da _escola	outras_ofertas _educacionais	id_renda_ familiar
Dimensões	**E_ESCV**	0.039	0.938	0.486	0.372	0.066	0.568	0.141	**0.000**	0.201	**0.000**
	E_PROFV	**0.000**	0.141	**0.010**	**0.025**	0.579	**0.000**	0.951	**0.041**	**0.000**	0.150
	E_FAMV	**0.000**	**0.075**	**0.000**	**0.000**	**0.000**	0.050	**0.028**	**0.000**	**0.000**	**0.000**
	E_COMV	**0.000**	**0.000**	**0.000**	**0.000**	**0.002**	**0.000**	**0.002**	**0.000**	0.239	**0.000**
	E_ESTV	0.213	**0.030**	0.538	0.668	0.850	0.514	0.844	0.249	0.411	0.594
Fatores	**E_ESC1V**	**0.011**	0.079	0.465	0.305	0.416	**0.000**	**0.020**	0.748	0.083	**0.032**
	E_ESC2V	**0.001**	0.064	0.109	0.199	0.138	**0.009**	0.935	**0.000**	**0.002**	**0.000**
	E_PROF1V	0.533	0.177	**0.015**	0.061	0.116	**0.000**	0.385	0.713	0.443	**0.000**
	E_PROF2V	**0.000**	0.116	0.383	0.236	0.717	**0.000**	0.369	0.706	**0.000**	0.079
	E_FAM1V	**0.000**	0.159	0.093	**0.000**	**0.000**	**0.000**	**0.004**	**0.000**	**0.000**	0.065
	E_FAM2V	**0.001**	0.612	**0.000**	0.803	**0.024**	**0.000**	0.627	0.865	0.245	**0.000**
	E_COM1V	**0.000**	**0.000**	**0.000**	0.330	**0.000**	**0.000**	0.641	0.415	**0.000**	**0.000**
	E_COM2V	**0.000**	0.074	0.701	**0.000**	0.500	0.182	0.285	0.276	0.820	0.835
	E_COM3V	0.095	0.781	**0.000**	0.240	0.938	**0.000**	0.339	0.527	**0.025**	0.648
	E_EST1V	**0.000**	0.988	**0.000**	0.215	0.307	**0.000**	0.804	**0.028**	0.632	**0.002**
	E_EST2V	**0.000**	**0.000**	0.060	**0.000**	0.332	**0.001**	0.127	0.409	**0.003**	0.000
	E_EST3V	0.069	**0.000**	**0.001**	**0.000**	0.009	**0.001**	**0.033**	**0.015**	**0.006**	0.057

A Figura 3 apresenta a disparidade na quantidade de estudantes de cada tipo de escola, com as escolas entre 201 e 500 matrículas possuindo a maioria dos estudantes para todos os valores das dimensões. Para a dimensão E_ESCV observa-se que as notas mais altas, como 6 e 7 possuem uma porcentagem maior de estudantes de escolas entre 501 e 1000 matrículas do que se comparado com os valores mais baixos. Já para E_PROFV acontece o fenômeno inverso, e esse porte de escola possui uma menor porcentagem de estudantes nas notas 6 e 7 do que nas demais notas. Além disso, a porcentagem de estudantes das escolas de porte de até 50 matrículas e de 201 a 500 matrículas atingem seus máximos na nota 7 de E_PROFV. Outro ponto de destaque é o gráfico para E_FAMV, que 100% dos alunos que responderam a maior nota para essa dimensão eram de escolas de porte de 51 a 200 matrículas.

Figura 3. Correlação entre dimensões e porte da escola



Na Figura 4 é possível notar que valores mais altos das dimensões E_PROFV, E_FAMV e E_COMV apresentam uma porcentagem maior de estudantes homens do que mulheres, o que foi observado como sendo uma diferença significativa pelo p-valor. Essa diferença é especialmente grande para a dimensão E_FAMV, onde 100% dos alunos que obtiveram o valor 7.0 eram do sexo masculino.

Para a análise dos fatores, a Figura 5 apresenta a relação entre a dependência administrativa da escola do estudante com uma série de fatores distintos, cuja relação é significativa, conforme o p-valor observado. Com isso, alguns pontos de destaque desta relação são encontrados nos fatores E_COM1V (Medidas Socioeducativas e Contextos de Violência), E_EST3V (Reprovações e Distorção Idade – Série), onde a medida que a nota aumenta, a porcentagem de estudantes de escolas municipais aumenta e de escolas estaduais diminui. Outro ponto de destaque é o fator E_FAM1V (Suporte Familiar), que para a nota 7, 100% dos estudantes eram de escolas estaduais.

Figura 4. Correlação entre dimensões e sexo

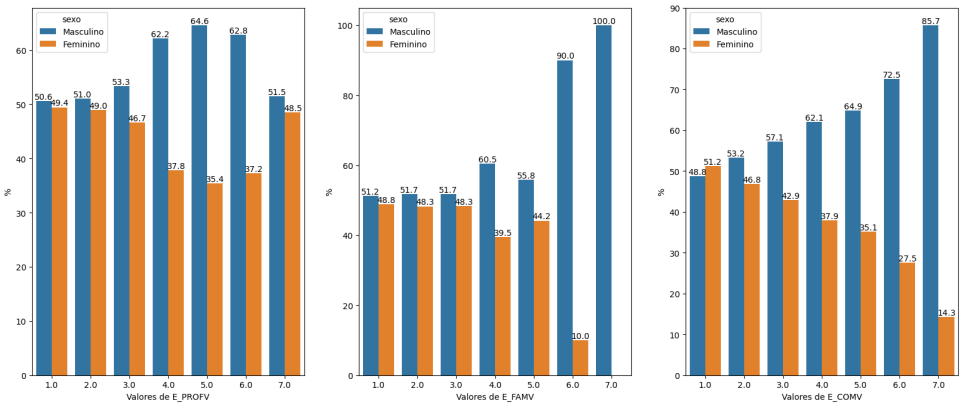
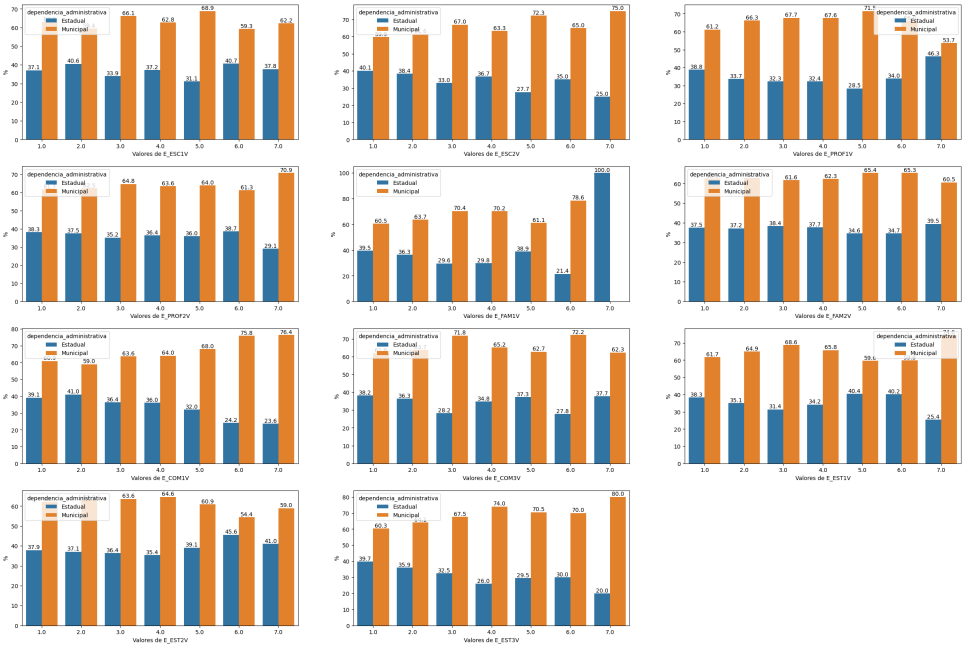


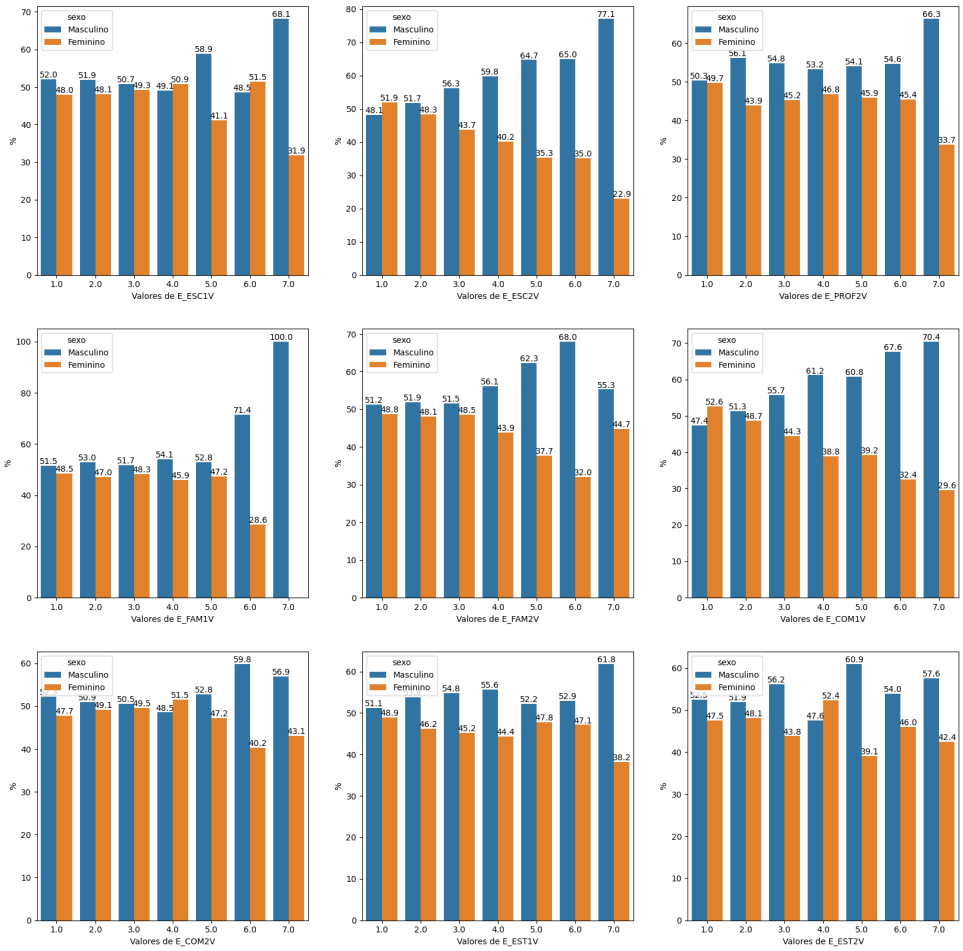
Figura 5. Correlação entre fatores e dependência administrativa



Por fim, a Figura 6 apresenta a relação de fatores diversos com a variável sexo. Nela, os fatores E_ESC2V (Condições Materiais do(a) Estudante) e E_COM1V (Medidas Socioeducativas e Contextos de Violência) apresentam uma queda clara da porcentagem de estudantes do sexo feminino em detrimento do aumento dos estudantes do sexo masculino à medida que a nota para esses fatores aumenta. Outro ponto de interesse acontece no fator E_FAM1V que para a nota 7, 100% dos estudantes são do sexo masculino.

Outra métrica analisada para identificar dependência entre variáveis foi traçar uma análise de correlação entre as variáveis das dimensões e entre as variáveis dos fatores para analisar qual delas poderia ter mais influência uma sobre as outras. Essas matrizes de correlações são representadas pela Figura 7. Nelas, é possível verificar que quando os números estão mais próximos de 1.0 significa que há uma maior correlação positiva entre as variáveis, e quando estão próximos de -1.0, há uma maior correlação negativa, com valores próximos a zero não possuindo muita significância. É importante lembrar que

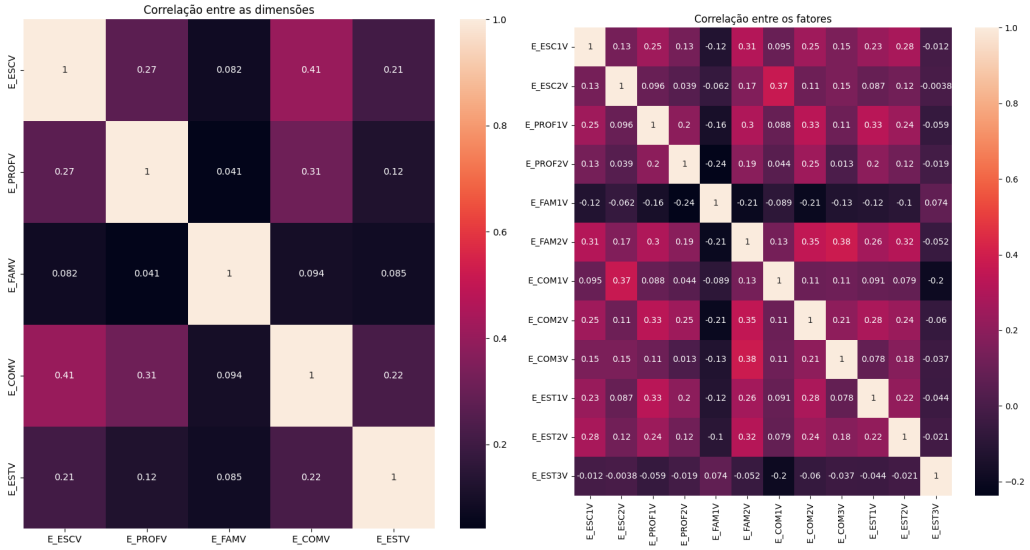
Figura 6. Correlação entre fatores e sexo



enquanto duas variáveis estão correlacionadas uma à outra, isso não significa que uma implica na causalidade da outra.

É notável a correlação positiva mediana que há entre as dimensões E_PROFV e E_COMV, com 0.41, atingindo o maior número dentre as dimensões. Ou seja, isso significa que há uma tendência de quando os valores de E_COMV forem mais altos, há maior probabilidade de E_PROFV também serem, indicando que há uma interligação entre esses dois fatores dentro do questionário. Por outro lado, percebe-se que o fator E_FAM1V (Suporte Familiar) é o que possui correlação negativa com a maioria todos os outros fatores, exceto com E_EST3V (Reprovações e Distorção idade-série), o que indica que é uma variável significativa mesmo que inversamente proporcional às outras. É importante ressaltar que o fator E_EST3V também possui em sua maioria correlação negativa com os outros fatores, mas são valores mais próximos de zero que a variável anterior, indicando que possui menos significância estatística correlacional. Nota-se também que os fatores E_FAM2V (Gravidez/parentalidade/atividades de cuidado) e E_COM2V (Acessibilidade e frequência escolar) são os que possuem maior correlação positiva com a maioria dos outros fatores, com seus valores positivos ficando entre 0.13 a 0.38 e 0.11 a 0.33, respectivamente.

Figura 7. Correlação entre dimensões e fatores



5. Análise Classificatória

Antes de aplicar as técnicas ao conjunto de dados, algumas precauções tiveram que ser tomadas para que os modelos não enfrentassem problemas durante a análise e o treinamento dos dados.

Uma das necessidades foi converter os dados de números reais para inteiros. Isso ocorre porque, em uma tarefa de classificação, todos os resultados únicos representariam uma classe separada, com valores como “3.0” sendo diferentes de “3.333333”. Para identificar melhor a classe a que um aluno poderia pertencer, todos os valores de ponto flutuante foram arredondados para cima ou para baixo, dependendo se o primeiro dígito após o ponto decimal era cinco ou maior (por exemplo, 2.5 e 2.6666667 se tornaram três, enquanto 2.3333333 se tornaram 2).

Outro ponto é a observação de grandes diferenças da quantidade de respostas para determinadas classes, pois a maioria das respostas estava distribuída entre valores de 1 a 3 (cerca de 13 mil respostas), e muito poucas de 5 a 7, o que dificulta o treinamento dos modelos para esses casos. Para equilibrar o conjunto de dados e assim evitar desempenho tendencioso em relação a uma ou mais classes, durante o treinamento dos modelos para cada dimensão e fator, a classe com maior número de respostas foi selecionada como o valor máximo de linhas. Em seguida, para cada classe, elementos aleatórios foram escolhidos e duplicados para preencher as linhas até o valor máximo encontrado, resultado em todas as classes com a mesma quantidade de respostas.

Também é importante observar que durante os testes os conjuntos de dados para treinamento, teste e validação eram distintos entre si. Em cada iteração do código, 20% do conjunto de dados foi selecionado aleatoriamente para teste para evitar superestimar o desempenho. Essa precaução garantiu que os modelos não encontrassem os dados de teste durante o treinamento e que o treinamento fosse imparcial.

O desempenho do treinamento do modelo de Árvore de Decisão é mostrado na Tabela 3. Sob análise, é perceptível que os valores das métricas para todas as variáveis

são relativamente altos, variando de aproximadamente 74% a 99%, indicando, em geral, uma boa capacidade do modelo em prever corretamente as classes. Em especial, nota-se que o fator **E_ESC1V** se destaca em acurácia, e as dimensões **E_PROFV** (Estudante-Profissionais da Escola) e **E_FAMV** (Estudante-Família) também em precisão e recall. Por outro lado, a pior variável para predição, ou seja, que apresentou os piores valores nas métricas, é a dimensão **E_ESTV** (Estudante-Estudante).

Tabela 3. Desempenho do Modelo Decision Tree por Dimensões e Fatores

Categoria	Variável	Acurácia	Precisão	Recall Score	F1 Score
Dimensões	**E_ESCV**	0.751673360	0.739981653	0.751673360	0.741630082
	E_PROFV	0.783035930	0.786358820	0.783035930	0.777129550
	E_FAMV	0.830982219	0.831129806	0.830982219	0.827472004
	E_COMV	0.774183466	0.769368781	0.774183466	0.767111565
	E_ESTV	0.741793911	0.732561420	0.741793911	0.732551048
Fatores	**E_ESC1V**	0.973298932	0.974228894	0.973298932	0.972954891
	E_ESC2V	0.957975316	0.958817154	0.957975316	0.957990201
	E_PROF1V	0.985004378	0.985474842	0.985004378	0.984873283
	E_PROF2V	0.995594947	0.995617592	0.995594947	0.995594299
	E_FAM1V	0.981763607	0.982526522	0.981763607	0.981566092
	E_FAM2V	0.988539873	0.988990114	0.988539873	0.988512460
	E_COM1V	0.957073377	0.958085738	0.957073377	0.956205580
	E_COM2V	0.977282629	0.978631022	0.977282629	0.976942412
	E_COM3V	0.988506982	0.989023192	0.988506982	0.988421136
	E_EST1V	0.987947572	0.988248921	0.987947572	0.987787534
	E_EST2V	0.973289665	0.974189465	0.973289665	0.973112790
	E_EST3V	0.986756179	0.987237780	0.986756179	0.986687563

Já a Tabela 4 mostra os resultados do treinamento do modelo de Random Forest. Observando as dimensões, as variáveis **E_ESCV** (Estudante-Escola) **E_FAMV** (Estudante-Família) se destacam com os maiores valores em todas as métricas, indicando um bom desempenho global do modelo nas duas dimensões. Já em relação aos fatores, destacam-se particularmente as variáveis **E_EST1V** (Significados da Escolarização/Engajamento) e **E_COM3V** (Distanciamento escola – comunidade), que apresentam desempenho praticamente impecável em todas as métricas avaliadas.

A Tabela 5 evidencia que as métricas de desempenho do modelo Adaboost variam significativamente entre as diferentes variáveis testadas. Com relação aos fatores, a maioria das variáveis atingiu altos níveis de desempenho, com valores de acurácia, precisão, recall e F1 score consistentemente elevados. Destacam-se particularmente as variáveis **E_EST1V** (Significados da Escolarização/Engajamento) e **E_COM3V** (Distanciamento escola – comunidade), que apresentam desempenho praticamente impecável em todas as métricas avaliadas, acima de 99%. Contudo, com relação às dimensões, as variáveis **E_ESCV** (Estudante-Escola), **E_PROFV** (Estudante-Profissionais da Escola), **E_FAMV** (Estudante-Família), **E_COMV** (Estudante-Escola), E_ESTV (Estudante-Estudante) têm desempenhos baixos em todas as métricas. A variável **E_FAMV** em particular destaca-se pela sua baixa precisão e F1 Score, indicando que o modelo tem dificuldade em fazer previsões precisas e equilibradas para essa dimensão. Por outro lado, alguns fatores, como **E_FAM1V** (Suporte Familiar), **E_COM3V** (Distanciamento Escola-Comunidade) e **E_EST1V** (Significados da Escolarização/Engajamento), mostram desempenhos mais promissores em comparação com outras, como é o caso do fator **E_FAM1V** que possui uma acurácia relativamente alta.

Tabela 4. Random Forest - Dimensões e Fatores

Categoria	Variável	Acurácia	Precisão	Recall Score	F1 Score
Dimensões	E_ESCV	0.848055295	0.842921845	0.848055295	0.843924479
	E_PROFV	0.786592718	0.790278008	0.786592718	0.781311994
	E_FAMV	0.833147431	0.831530566	0.833147431	0.829228533
	E_COMV	0.778324738	0.770605436	0.778324738	0.771282145
	E_ESTV	0.744445460	0.734167526	0.744445460	0.735079615
Fatores	E_ESC1V	0.979659186	0.980308426	0.979659186	0.979535759
	E_ESC2V	0.964458678	0.964734405	0.964458678	0.964430198
	E_PROF1V	0.988178634	0.988461239	0.988178634	0.988120343
	E_PROF2V	0.996337968	0.996351606	0.996337968	0.996338277
	E_FAM1V	0.982499858	0.983288480	0.982499858	0.982314134
	E_FAM2V	0.990855650	0.991095707	0.990855650	0.990836153
	E_COM1V	0.968210333	0.968547724	0.968210333	0.967829172
	E_COM2V	0.982635215	0.983218686	0.982635215	0.982435642
	E_COM3V	0.992910849	0.993159647	0.992910849	0.992893850
	E_EST1V	0.996836238	0.996854884	0.996836238	0.996829868
	E_EST2V	0.981295488	0.981635502	0.981295488	0.981176100
	E_EST3V	0.987244431	0.987795643	0.987244431	0.987156619

Tabela 5. Adaboost - Dimensões e Fatores

Categoria	Variável	Acurácia	Precisão	Recall Score	F1 Score
Dimensões	E_ESCV	0.310408300	0.283796729	0.310408300	0.290538426
	E_PROFV	0.271883289	0.264811380	0.271883289	0.263795898
	E_FAMV	0.300177154	0.516699197	0.300177154	0.218699253
	E_COMV	0.302390999	0.289724284	0.302390999	0.284777157
	E_ESTV	0.295144921	0.281181766	0.295144921	0.272988844
Fatores	E_ESC1V	0.358454338	0.342833509	0.358454338	0.320662206
	E_ESC2V	0.362599594	0.408169540	0.362599594	0.330145255
	E_PROF1V	0.439306042	0.444431554	0.439306042	0.430396321
	E_PROF2V	0.371563528	0.429579049	0.371563528	0.339875124
	E_FAM1V	0.517981537	0.805497664	0.517981537	0.415171182
	E_FAM2V	0.281396592	0.636934146	0.281396592	0.155494301
	E_COM1V	0.416367097	0.409562625	0.416367097	0.397870672
	E_COM2V	0.333478559	0.368666728	0.333478559	0.319560681
	E_COM3V	0.449838883	0.520024864	0.449838883	0.454163910
	E_EST1V	0.465022849	0.458395635	0.465022849	0.454569975
	E_EST2V	0.381586608	0.425480643	0.381586608	0.382408663
	E_EST3V	0.318156851	0.338854607	0.318156851	0.273703286

Por fim, a Tabela 6 mostra os resultados do treinamento do algoritmo Naive Bayes Categórigo, e nela chama a atenção a grande variabilidade nos valores de acurácia, precisão, recall e F1 Score entre as diferentes variáveis testadas. Destaca-se a variável **E_FAMV** (Estudante-Família) por apresentar os maiores valores em todas as métricas, indicando um desempenho geral mais consistente. Isso sugere que esta dimensão possui uma capacidade significativa de predição para o modelo Naive Bayes Categórico. Por outro lado, a variável **E_PROFV** (Estudante-Profissionais da Escola) mostra valores mais baixos em todas as métricas, demonstrando sua baixa eficácia. Por outro lado, as variáveis relacionadas à interação do estudante dentro da escola (**E_ESC1V**, **E_ESC2V**, **E_PROF1V**, **E_PROF2V**, **E_EST1V**, **E_EST2V**, **E_EST3V**) e com sua família (**E_FAM1V**, **E_FAM2V**) apresentam resultados gerais mais elevados em comparação com as variáveis relacionadas à relação do estudante com a comunidade (**E_COM1V**, **E_COM2V**, **E_COM3V**).

Tabela 6. Naive Bayes Categórico - Dimensões e Fatores

Categoria	Variável	Acurácia	Precisão	Recall Score	F1 Score
Dimensões	**E_ESCV**	0.457914993	0.434943418	0.457914993	0.428372704
	E_PROFV	0.411020014	0.407000983	0.411020014	0.400572047
	E_FAMV	0.618463355	0.598444859	0.618463355	0.601411496
	E_COMV	0.494374121	0.467487650	0.494374121	0.472578934
	E_ESTV	0.433299808	0.414196732	0.433299808	0.406825414
Fatores	**E_ESC1V**	0.561202448	0.555537424	0.561202448	0.556714967
	E_ESC2V	0.563427589	0.555158840	0.563427589	0.558275101
	E_PROF1V	0.637259194	0.638644712	0.637259194	0.635457652
	E_PROF2V	0.616813502	0.610137577	0.616813502	0.611076953
	E_FAM1V	0.750127428	0.744078395	0.750127428	0.745922676
	E_FAM2V	0.640163886	0.636322205	0.640163886	0.636805646
	E_COM1V	0.529216889	0.528650634	0.529216889	0.522478593
	E_COM2V	0.519138607	0.513238162	0.519138607	0.513408647
	E_COM3V	0.629699248	0.636854980	0.629699248	0.630850677
	E_EST1V	0.567317833	0.568723238	0.567317833	0.563402589
	E_EST2V	0.482168850	0.482410913	0.482168850	0.474973736
	E_EST3V	0.758010375	0.761468350	0.758010375	0.758019485

6. Conclusão

Para fazer as análises estatísticas foi utilizado p-valores para avaliar a significância estatística das relações entre variáveis socioeconômicas e dimensões educacionais. Entre as dimensões analisadas, E_ESCV (Estudante-Escola), E_PROFV (Estudante-Profissionais da Escola), E_FAMV (Estudante-Família) e E_COMV (Estudante-Comunidade) mostraram variações significativas em suas associações com as variáveis socioeconômicas. Por exemplo, o porte da escola impacta as dimensões relacionadas à escola e à comunidade, enquanto o sexo dos estudantes influencia dimensões como E_PROFV, E_FAMV e E_COMV.

A análise classificatória realizada revelou nuances importantes nos diferentes modelos avaliados, com Decision Tree e Random Forest demonstrando desempenho superior, enquanto AdaBoost e Naive Bayes Categórico mostraram resultados abaixo do esperado, sendo que K Vizinhos Próximos teve um desempenho mediano. Contudo, independente de performance, a dimensão E_FAMV (Estudante-Família) é consistentemente a mais influente, apresentando altos valores de acurácia, precisão, recall e F1 Score em todos os algoritmos, destacando-se especialmente no Naive Bayes Categórico e no Random Forest. Por outro lado, a dimensão E_ESCV (Estudante-Escola) também mostrou influência significativa, particularmente nos modelos Random Forest e Naive Bayes Categórico. Já entre os fatores, E_EST1V (Significados da Escolarização/Engajamento) demonstrou desempenho acima da média nos modelos Decision Tree e Random Forest, enquanto E_COM3V (Distanciamento escola – comunidade) destacou-se nos modelos Random Forest e Adaboost. Foi possível entender que os fatores de risco na base do IAFREE que possuem maior influência estatística ou melhor desempenho entre os algoritmos de classificação, são **E_FAM1V (Suporte Familiar)** e **E_COM3V (Distanciamento escola – comunidade)** para os fatores, e **E_FAMV (Estudante-Família)** para as dimensões. Como trabalhos futuros espera-se buscar medidas de retenção escolar focadas a dimensões e fatores específicas, e utilizar outras técnicas descritas na literatura para lidar com o problema de desbalanceamento de classe.

Agradecimentos

Agradecemos o apoio do NEES/UFAL (TED11476), do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) através dos processos 302959/2023-8 (DT2) e 303443/2023-5 (DT2) e da Fundação de Amparo à Pesquisa e Inovação do Estado de Santa Catarina (FAPESC) Nº 48/2022 - Apoio à Infraestrutura para Grupos de Pesquisa da UDESC TO n°2023TR000245.

Referências

Aldowah, H., Al-Samarraie, H., and Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics and Informatics*, 37:13–49.

BRASIL (1955). Decreto nº 37.106, de 31 de março de 1955. Diário Oficial da União. https://www2.camara.leg.br/legin/fed/decret/1950-1959/decreto-37106-31-marco-1955-332702-publicacaooriginal-1-pe.html.

BRASIL (2001). Lei n° 010172, de 9 de janeiro de 2001. Diário Oficial da União. https://www.planalto.gov.br/ccivil_03/leis/leis_2001/l10172.htm.

BRASIL (2005). Lei n° 11.096, de 13 de janeiro de 2005. Diário Oficial da União. https://www.planalto.gov.br/ccivil_03/_ato2004-2006/2005/lei/l11096.htm.

BRASIL (2020). Lei nº 14.113, de 25 de dezembro de 2020. Diário Oficial da União. https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2020/lei/l14113.htm.

Cechinel, C. and da Silva Camargo, S. (2020). Mineração de dados educacionais: avaliação e interpretação de modelos de classificação. In Jaques, P. A., Siqueira, S., Bittencourt, I., and Pimentel, M., editors, *Metodologia de Pesquisa Científica em Informática na Educação: Abordagem Quantitativa*, volume 2 of *Série Metodologia de Pesquisa em Informática na Educação*, chapter 12. Sociedade Brasileira de Computação, Porto Alegre: SBC. Disponível em: https://metodologia.ceie-br.org/livro-2.

de Vasconcelos, A. N., Freires, L. A., Loureto, G. D. L., Fortes, G., da Costa, J. C. A., Torres, L. F. F., Bittencourt, I. I., Cordeiro, T. D., and Isotani, S. (2023). Advancing school dropout early warning systems: the IAFREE relational model for identifying at-risk students. *Front. Psychol.*, 14:1189283.

Heringer, R. (2018). Democratização da educação superior no brasil: das metas de inclusão ao sucesso acadêmico. *Revista Brasileira de Orientação Profissional*, 19(1):7–17.

Koç, M. (2017). Learning analytics of student participation and achievement in online distance education: A structural equation modeling. *Educational Sciences: Theory & Practice*.

Moissa, B., Gasparini, I., and Kemczinski, A. (2015). Educational data mining versus learning analytics: estamos reinventando a roda? um mapeamento sistemático. In *Anais*

do XXVI Simpósio Brasileiro de Informática na Educação (SBIE 2015). Sociedade Brasileira de Computação - SBC.

Santos, M. C. S., Delatorre, L. R., Ceccato, M. d. G. B., and Bonolo, P. d. F. (2019). Programa bolsa família e indicadores educacionais em crianças, adolescentes e escolas no brasil: revisão sistemática. *Ciência & Saúde Coletiva*, 24(Ciênc. saúde coletiva, 2019 24(6)):2233–2247.

SAP (2022). Sistema de alerta preventivo (sap) de evasão e abandono escolar. Acesso em Mai. 2024.

Vargas, H. M. and Zuccareli, C. (2021). A nova face da docência: uma proposta de revisão do censo da educação superior. *Estudos em Avaliação Educacional*, 32.

Wasserstein, R. L. and Lazar, N. A. (2016). The asa statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2):129–133.