Mineração de dados do Enade de 2016 a 2018: uma análise sobre o município de Araçatuba/SP

Mayk Fernando Choji, Universidade de São Paulo, mayk@alumni.usp.br
https://orcid.org/0000-0002-3358-8109
Carlos Diego N. Damasceno, Radboud University Nijmegen, d.damasceno@cs.ru.nl
https://orcid.org/0000-0001-8492-7484
Ig Ibert Bittencourt, Universidade Federal de Alagoas, ig.ibert@ic.ufal.br
https://orcid.org/0000-0001-5676-2280
Seiji Isotani, Universidade de São Paulo, sisotani@icmc.usp.br
https://orcid.org/0000-0003-1574-0784

Resumo: Para o efetivo desenvolvimento de políticas educacionais, de inclusão e permanência é necessário ter ferramentas e métodos adequados para analisar os dados coletados. Assim, este artigo apresenta uma nova ferramenta para apoiar análises dos microdados do Enade utilizando técnicas de mineração de dados. Esta ferramenta foi desenvolvida durante um estudo de caso sobre o perfil socioeconômico dos concluintes de graduação do município de Araçatuba/SP, baseado nos microdados de 2016 a 2018. Como resultado, foram extraídas algumas regras de associação como, por exemplo, alunos brancos de IES privadas que escolheram o curso visando inserção no mercado de trabalho, tendem a ter baixas notas no Enade; enquanto alunos autodeclarados pretos, pardos ou indígenas que escolheram o curso pelo mesmo motivo apresentaram notas melhores.

Palavras-chave: mineração de dados educacionais, enade, mineração de dados, aspectos sociais.

Enade data mining from 2016 to 2018: an analysis for the city of Araçatuba/SP

Abstract: To develop effective educational, inclusion and permanence policies, it is necessary to have tools and methods to analyze the data collected. Thus, this paper presents a new tool to support the analysis of the Enade microdata based on data mining techniques. This tool was developed together with a case study about the socioeconomics profile of undergraduate students from Araçatuba/SP, according to microdata from 2016 to 2018. As a result, some association rules were extracted, such as, white students from private HEIs who chose the course aiming primarily at the job market, tend to have low grades at Enade; and self-declared black, brown or indigenous students from public HEIs who chose the course by the same motivation tend to have high grades.

Keywords: educational data mining, enade, data mining, social aspects.

1. Introdução

O ensino superior é objeto de estudo de muitos trabalhos na literatura voltados aos problemas de desigualdade social, equidade de acesso à universidade, representatividade de gêneros, entre outros. Andrade (2012) observa que indivíduos autodeclarados não brancos têm menos acesso à educação do que indivíduos autodeclarados brancos e que o acesso aos níveis mais altos de escolaridade é mais influenciado pela renda familiar do que pela cor autodeclarada. A autora destaca que "o fato de a variável renda ter maior influência no acesso aos níveis mais altos de escolaridade do que a variável cor autodeclarada é bastante importante para o formulação de programas e políticas que visam ampliar a equidade do acesso aos níveis mais altos de escolaridade".

O Exame Nacional de Desempenho dos Estudantes (Enade) é uma das principais ferramentas utilizadas hoje para avaliar o rendimento dos concluintes de cursos de graduação no Brasil, sendo aplicado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). Juntamente com o Enade, a Avaliação de cursos de graduação e a Avaliação institucional compõem o Sistema Nacional de Avaliação da Educação Superior (Sinaes) (INEP, 2019a).

Além de questões de formação específica e geral, o Enade contém questionários de aspectos socioeconômicos e de percepção do estudante sobre o exame realizado. Os resultados são utilizados em conjunto com outras avaliações como insumos para se avaliar a qualidade da educação superior brasileira, por meio dos Indicadores de Qualidade da Educação Superior (INEP, 2019b).

Os resultados do Enade impactam direta ou indiretamente todos os indicadores, não sendo raro, portanto, o interesse de gestores de instituições de ensino superior no desempenho de seus estudantes no exame. Além disso, pesquisadores têm mostrado interesse no estudo de microdados do Enade e nas informações que podem ser extraídas a partir deles (NASCIMENTO; JUNIOR; FAGUNDES, 2018; SILVA; SILVA, 2015). Conforme descrito em (INEP, 2020), "os microdados do Inep se constituem no menor nível de desagregação de dados recolhidos por pesquisas, avaliações e exames realizados".

Neste contexto, este trabalho busca realizar um estudo de caso sobre o perfil socioeconômico dos concluintes de graduação das instituições de ensino superior (IES) do município de Araçatuba, São Paulo, de acordo com informações do Enade dos anos de 2016 a 2018. A escolha do estudo de caso dá-se principalmente por interesse do autor em entender as características das IES do município onde atua como docente em uma das instituições. Além de estatística descritiva, técnicas de mineração de dados são utilizadas para se identificar regras de associação que ajudem a explicar as principais características desse conjunto de dados.

O restante deste artigo está organizado da seguinte forma: na Seção 2 são discutidos alguns trabalhos relacionados à análise de microdados do Enade, enquanto que a Seção 3 descreve a metodologia aplicada no presente estudo. Os principais resultados são discutidos na Seção 4 e a Seção 5 conclui este trabalho.

2. Trabalhos Relacionados

Ristoff (2014) faz uma análise estatística dos três primeiros ciclos completos do Enade, ou seja, de 2004 a 2012, para verificar o perfil socioeconômico do estudante de graduação. O trabalho utiliza majoritariamente o questionário socioeconômico do próprio exame e busca entender, principalmente, como as políticas de inclusão impactaram a representação no ensino superior das classes historicamente excluídas deste nível de ensino.

Devido à própria natureza do estudo realizado, Ristoff (2014) utiliza apenas quatro dimensões do questionário socioeconômico: *(i)* a cor do estudante; *(ii)* a renda mensal da família do estudante; *(iii)* a origem escolar do estudante; e *(iv)* a escolaridade dos pais do estudante. O trabalho, baseado em estatística descritiva, traz informações sobre o perfil do *campus* brasileiro, como proporção de brancos, pretos e pardos em cursos selecionados pelo autor, a renda familiar dos estudantes e alguns efeitos do sistema de cotas instaurado pela Lei nº 12.711/2012. Embora o autor ainda apresente alguns perfis de estudantes relacionados aos indicadores socioeconômicos, a correlação entre as variáveis estudadas não é apresentada formalmente.

Em (FILHO; ROSEIRA; JR, 2020) os autores utilizam-se de estatística descritiva para verificar o perfil socioeconômico e o desempenho de estudantes de licenciatura em

educação física, baseado nos microdados do Enade de 2017. Simultaneamente, realizam uma revisão integrativa da literatura sobre as condições socioeconômicas dos estudantes e suas relações com o desempenho no exame. Segundo os autores, é possível observar maior desempenho entre os estudantes que possuem renda e recebem ajuda da família ou de outras pessoas para financiar os gastos durante a formação.

Nos casos analisados por Ristoff (2014), Filho, Roseira e Jr (2020), estudantes cujos pais tiveram maior nível de escolaridade apresentaram melhores resultados no exame dos que os cujos pais tiveram menor nível.

Dentre os trabalhos voltados para o estudo de cursos específicos, Vista, Figueiró e Chicon (2017), Cretton e Gomes (2016) utilizam técnicas de mineração de dados sobre microdados do Enade para avaliar, respectivamente, cursos de Ciência da Computação do Rio Grande do Sul e cursos de medicina do país. Os dois trabalhos têm como objetivo principal o estudo do desempenho dos concluintes.

Vista, Figueiró e Chicon (2017) utilizam a técnica de Agrupamento Hierárquico com Aglomeração para agrupar as diferentes IES do Rio Grande do Sul de acordo com as notas obtidas pelos estudantes no Enade. Como resultado, identificaram-se quatro grupos principais de IES, estando PUCRS, UFRGS e UFPEL no grupo de melhor desempenho e UCPEL isolada em um grupo com pior desempenho. Os autores acreditam que este tipo de análise possa ser utilizado em tomadas de decisões que resultem em melhorias na qualidade de ensino das IES brasileiras.

Já no trabalho desenvolvido por (CRETTON; GOMES, 2016), árvores de decisão foram utilizadas para avaliar relações entre a percepção dos estudantes sobre o componente específico do Enade e seus desempenhos no exame. Embora árvores de decisão sejam usadas mais comumente em tarefas de predição e classificação, os autores as utilizaram como uma forma de se analisar agrupamentos. Baseados nos resultados, os autores observaram influência das variáveis referentes à categoria e tipo das IES na criação dos perfis dos estudantes e que, no Estado de São Paulo, estudantes de IES municipais responderam como "fácil" o nível de dificuldade do componente específico, muito embora seus rendimentos tenham sido negativo, de acordo com a classificação dos autores.

De forma geral, os resultados encontrados neste trabalho são similares a trabalhos anteriores, no sentido de que aspectos sociais têm impacto sobre o desempenho no Enade. Porém, ao contrário de métodos estatísticos tradicionais, nos quais as variáveis de interesse são escolhidas e analisadas manualmente, técnicas de mineração de dados como a utilizada aqui são capazes de extrair relações e métricas de maneira semi-automática, podendo revelar informações outrora não-triviais.

3. Metodologia

Esta Seção descreve o conjunto de dados e as técnicas de processamento e mineração de dados utilizados neste trabalho. A parte computacional foi realizada utilizando-se a linguagem de programação *Python* e várias funções para tratamento do conjunto de dados utilizados aqui estão disponíveis por meio do pacote *enade-py* (CHOJI, 2020), desenvolvido como parte deste estudo.

3.1. Descrição dos Dados

Neste trabalho, são utilizados os microdados do Enade referentes aos anos de 2016 a 2018, último ciclo de avaliação para o qual se tinha informações disponíveis na página oficial do INEP (INEP, 2020) quando o estudo foi conduzido. Os microdados constituem um conjunto de informações detalhadas dos estudantes participantes e também dos cursos e IES avaliadas.

Além do arquivo contendo as informações dos participantes, o INEP fornece arquivos auxiliares no formato de dicionários, isto é, explicação das variáveis definidas nos microdados. As variáveis podem ser divididas nos seguintes grupos:

(i) informações da instituição de ensino superior e do curso; *(ii)* informações do estudante; *(iii)* informações sobre número de itens da parte objetiva; *(iv)* vetores que representam gabaritos, escolhas e acertos da parte objetiva; *(v)* informações sobre tipos de presença; *(vi)* tipos de situação das questões da parte discursiva; *(vii)* notas na formação geral e componente específico; *(viii)* questionário de percepção da prova; e *(ix)* questionário do estudante.

3.2. Seleção

Dos grupos de variáveis descritos anteriormente, apenas os grupos *(i)*, *(ii)*, *(vii)* e *(ix)* são utilizados neste estudo. Mais ainda, apenas as questões de aspectos socioeconômicos do questionário do estudante são consideradas na análise. A seleção das variáveis dá-se pelo objetivo deste trabalho e de forma alguma diminui a relevância das que não foram incluídas. De fato, em trabalhos futuros espera-se incluir outras variáveis para estudo.

Do conjunto total de dados, foram selecionadas as amostras referentes ao município de Araçatuba, estado de São Paulo. Esse subconjunto refere-se às entradas cujo valor da variável `CO_MUNIC_CURSO` é igual a 3502804 e abrange um total de 2075 registros.

A partir dessa primeira seleção, verificou-se que 273 registros apresentavam o valor 222 para a variável `TP_PRES`, que corresponde a estudante ausente no exame. Visto que ausência implica em falta de informação para a variável `NT_GER` (nota geral), eles foram removidos do conjunto de dados. Na sequência, foi removido um único registro que não continha informações para as variáveis do questionário socioeconômico, resultando em 1.801 registros válidos. A variável `QE_I26` foi removida após verificar-se que 600 registros não possuíam informação válida neste atributo. Esta variável diz respeito à principal razão para a estudante ter escolhido sua instituição de ensino superior.

Por fim, algumas variáveis do questionário socioeconômico foram desconsideradas neste estudo principalmente por dois motivos: umas por estarem mais relacionadas às instituições do que aos estudantes, e outras por apresentarem pouca variação nas respostas, poluindo as regras de associação obtidas e impedindo que fossem encontrados padrões mais complexos (*i.e.*, aqueles que não seriam descobertos facilmente por estatística descritiva). A Tabela 1 descreve as variáveis selecionadas para estudo. Conforme já mencionado anteriormente, esta escolha não descarta a importância das variáveis desconsideradas, que devem ser incluídas em trabalhos futuros que utilizem outras técnicas de mineração de dados e/ou aprendizado de máquina.

Baseado no conjunto de dados selecionado, a Tabela 2 descreve o número de cursos, de acordo com a área de enquadramento no Enade, e de estudantes participantes do exame entre 2016 a 2018, para cada instituição de ensino superior do município de Araçatuba/SP. A sigla e a categoria administrativa de cada IES foram obtidas a partir da variável `CO_IES` dos microdados, por meio de consulta ao portal e-MEC.[*] As categorias das instituições foram utilizadas para dividir o conjunto de dados em dois grupos: instituições públicas e instituições privadas.

Após analisar a distribuição dos estudantes por cor autodeclarada, exibida na Figura 1, os grupos foram divididos entre estudantes autodeclarados brancos e estudantes autodeclarados pardos, pretos ou indígenas.

Os métodos descritos a seguir, portanto, foram aplicados separadamente para os

[*] ⟨https://emec.mec.gov.br/⟩

Tabela 1: Variáveis dos microdados do Enade selecionadas para o processo de mineração de dados. As variáveis do questionário do estudante estão descritas conforme consta nos dicionários de variáveis, disponíveis em (INEP, 2020).

Variável	Descrição
TP_SEXO	"Sexo."
NT_GER	"Nota bruta da prova."
QE_I02	"Qual é a sua cor ou raça?"
QE_I04	"Até que etapa de escolarização seu pai concluiu?"
QE_I05	"Até que etapa de escolarização sua mãe concluiu?"
QE_I06	"Onde e com quem você mora atualmente?"
QE_I07	"Quantas pessoas da sua família moram com você? Considere seus pais, irmãos, cônjuge, filhos e outros parentes que moram na mesma casa com você."
QE_I08	"Qual a renda total de sua família, incluindo seus rendimentos?"
QE_I09	"Qual alternativa a seguir melhor descreve sua situação financeira (incluindo bolsas)?"
QE_I17	"Em que tipo de escola você cursou o ensino médio?"
QE_I22	"Excetuando-se os livros indicados na bibliografia do seu curso, quantos livros você leu neste ano?"
QE_I23	"Quantas horas por semana, aproximadamente, você dedicou aos estudos, excetuando as horas de aula?"
QE_I25	"Qual o principal motivo para você ter escolhido este curso?"

Tabela 2: Descrição do número de cursos e estudantes participantes do Enade, entre 2016 e 2018, para cada instituição de ensino superior.

IES	Categoria Administrativa	Cursos	Participantes
FATEC	Pública estadual	1	8
UCESP	Privada sem fins lucrativos	2	43
FAC-FEA	Pública municipal	3	97
UNESP	Pública estadual	2	149
UNIP	Privada com fins lucrativos	15	301
UNISALESIANO	Privada sem fins lucrativos	22	550
UNITOLEDO	Privada com fins lucrativos	19	653

quatro subconjuntos resultantes dessas divisões.

3.3. Pré-processamento

Os algoritmos para encontrar os chamados *conjuntos de itens frequentes* utilizados neste estudo requerem que os dados estejam representados por variáveis binárias. Cada variável deve indicar a presença ou ausência daquele item (atributo) nas amostras dos dados. Por exemplo, uma variável que representa uma questão com três escolhas possíveis (*e.g.*, A, B, C), deve ser representa como três variáveis binárias, onde aquela que representa a escolha original é marcada com valor 1 ou Verdadeiro, e as demais com 0 ou Falso.

Assim, as variáveis de estudo foram transformadas em variáveis binárias de acordo com os valores encontrados nos dados para cada uma. A variável `NT_GER`, antes de passar por este processo, foi transformada em variável categórica de acordo com a mediana dos valores presentes no conjunto, resultando em duas categorias que representam as notas mais baixas e as notas mais altas.

3.4. Mineração

Na primeira parte da etapa de mineração de dados, o algoritmo *FP-Growth* (HAN; PEI; YIN, 2000) foi utilizado para gerar os conjuntos de itens frequentes, configurando-se o parâmetro *suporte mínimo* para $0,05$. Esta configuração permite que se obtenha todos os conjuntos de itens que aparecem juntos em ao menos 5% dos registros.

A escolha por este algoritmo dá-se principalmente por sua comprovada eficiência, em termos de recursos computacionais e tempo de processamento, quando comparado a algoritmos de mesmo propósito como o *Apriori* (AGRAWAL; IMIELIŃSKI; SWAMI, 1993) e o ECLAT (ZAKI, 2000). Diversos trabalhos recomendam o *FP-Growth* por

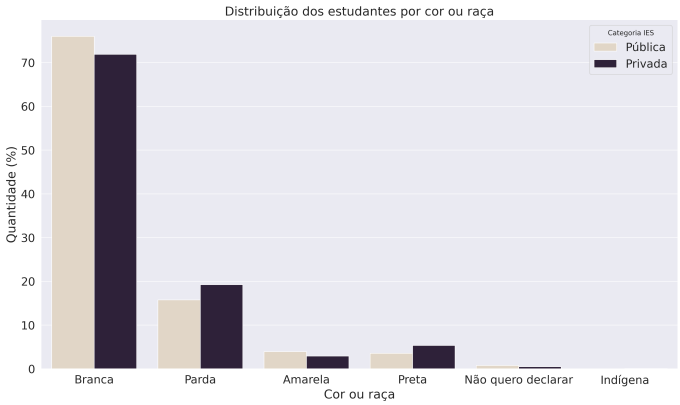

Figura 1: Distribuição de alunos por cor ou raça, nas IES públicas e privadas.

conta de seu desempenho otimizado para diferentes tipos de conjuntos de dados e sua escalabilidade para tratar grandes conjuntos (HEATON, 2016; BORGELT, 2012).

A segunda etapa consistiu em extrair-se regras de associação a partir dos conjuntos de itens frequentes. As funções básicas para geração dos conjuntos de itens frequentes e regras de associação foram providas pelo pacote *mlxtend* (RASCHKA, 2018). O pacote *enade-py* (CHOJI, 2020) estende tais funções para fornecer informações adicionais sobre os conjuntos de itens frequentes, antecedentes e consequentes das regras, indicando o número de itens em cada conjunto e se este é um *conjunto de itens frequente fechado*.

Das regras obtidas, foram selecionadas para avaliação manual aquelas com maiores valores para as métricas *suporte* e *convicção*. A primeira por permitir identificar os padrões que aparecem com mais frequência, e a segunda por identificar as relações mais fortes entre antecedentes e consequentes.

4. Discussão e Resultados

De acordo com a análise descritiva dos dados realizadas como passo inicial do estudo, verifica-se que as mulheres são maioria ao se considerar os exames do Enade de 2016 a 2018. Conforme exibido na Figura 2, essa relação se mostra ainda mais evidente no grupo das instituições públicas, no qual representam mais de 70%.

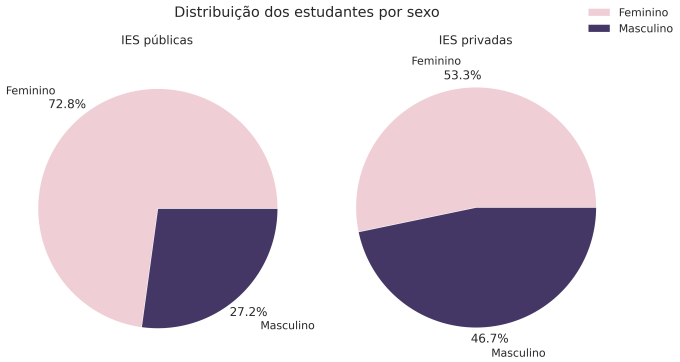

Figura 2: Distribuição de alunos por sexo, nas instituições públicas (à esquerda) e privadas (á direita).

Em relação à cor ou raça declarada pelos estudantes, observa-se que o *campus* araçatubense é predominantemente branco, em consonância com o estudo realizado por Ristoff (2014) no nível nacional. A distribuição dos estudantes por cor ou raça, segundo as declarações de 2016 a 2018, é mostrada na Figura 1.

Além das duas variáveis descritas anteriormente (sexo e cor), uma breve análise das idades dos estudantes mostra uma distribuição semelhante para os dois grupos de instituições. Conforme os histogramas da Figura 3, a maioria dos valores encontra-se na faixa entre 20 e 30 anos, notando-se uma população levemente mais jovem nas instituições privadas.

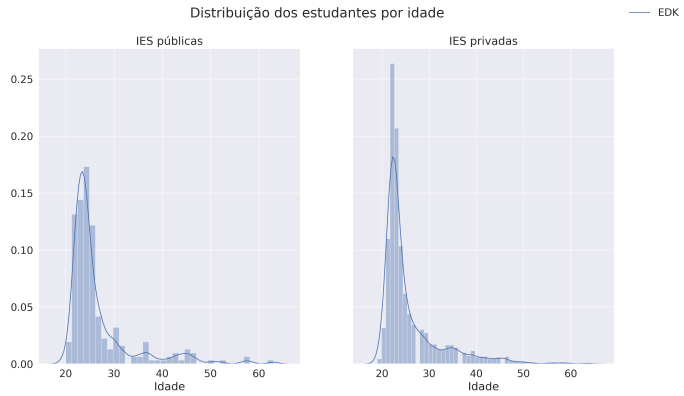

Figura 3: Histograma da distribuição dos estudantes por idade, nas IES públicas (à esquerda) e privadas (á direita). A curva indicada em cada gráfico representa a estimativa de densidade por Kernel (EDK) da variável `NU_IDADE`.

Seja s, l, c reais não-negativos onde s é o valor de suporte, l o *lift* e c a convicção de uma dada regra de associação, com $s \in [0, 1]$ e $l, c \in [0, \infty]$. Nos parágrafos seguintes são discutidas algumas regras de associação obtidas para cada grupo de estudo, resultado do processo discutido na Subseção 3.4, acompanhadas das respectivas métricas.

Para o grupo dos estudantes de instituições públicas, foram obtidas um total de 13.366 regras para o subgrupo dos estudantes autodeclarados brancos e 32.510 regras para o subgrupo dos estudantes autodeclarados negros, pardos ou indígenas. Algumas regras interessantes, tanto do ponto de vista objetivo (valor alto em uma ou mais métricas) quanto subjetivo, são apresentadas na Tabela 3. De forma geral, o item que aparece frequentemente nas regras obtidas sobre os estudantes autodeclarados brancos é o fato de serem oriundos de escolas privadas e não possuírem renda própria. Em contrapartida, são observados itens associados com menos privilégios socioeconômicos nas regras sobre os estudantes autodeclarados pretos, pardos ou indígenas. Identifica-se, por exemplo, que quase metade deste último grupo são mulheres que cursaram o ensino médio todo em escola pública.

Em um trabalho publicado em 2012, Andrade (2012) apontava que o efeito da renda familiar era muito mais forte do que a cor na chance de os jovens terem acesso ao ensino superior. Pelas regras obtidas, estima-se que não só isso ainda é verdade, como também a renda impacta no desempenho durante o curso. Observa-se, por exemplo, que entre os autodeclarados pretos, pardos ou indígenas, aqueles que moram com cônjuges e/ou filhos e trabalham apresentam desempenho abaixo do que os que moram com outras pessoas e são financiados pela família ou outras pessoas (regras 6 e 7 da Tabela 3).

Para o grupo dos estudantes brancos das IES privadas, os valores máximos de *suporte* e *convicção* foram $0,28$ e $9,71$, respectivamente. Ainda, a regra com valor máximo de *convicção* diz apenas que se o estudante mora sozinho, então nenhuma pessoa da família mora com ele, o que é óbvio. Esse resultado pode ser consequência da diversidade de cursos e instituições e do maior número de participantes do grupo. O grupo dos estudantes autodeclarados pretos, pardos ou indígenas apresentou valores maiores para as métricas de *suporte* e *convicção*, a última chegando inclusive ao valor máximo (∞).

No total, foram obtidas 4.576 regras de associação para o subgrupo dos estudantes autodeclarados brancos, e 7.816 para o subgrupo dos estudantes autodeclarados pretos, pardos ou indígenas. Algumas das regras obtidas são apresentadas na Tabela 4.

Tabela 3: Regras de associação obtidas para o subconjunto dos estudantes de IES públicas. São apresentados os itens que compõem o antecedente e o consequente de cada regra, acompanhados de seus valores de suporte, *lift* e convicção.

Cor Autodeclarada	Regra	Antecedente	Consequente	Sup.	*Lift*	Conv.
Brancos	1	Cursou o ensino médio todo em escola privada.	Não tem renda e os gastos são financiados pela família ou outras pessoas.	$0,46$	$1,27$	$2,89$
	2	Cursou o ensino médio todo em escola privada; Nota no exame está entre as maiores.	Não tem renda e os gastos são financiados pela família ou outras pessoas.	$0,29$	$1,37$	$8,56$
	3	Cursou o ensino médio todo em escola privada; Não tem renda e os gastos são financiados pela família ou outras pessoas; Não mora com ninguém da família; Dedicou de quatro a sete horas de estudo por semana.	Nota do exame está entre as maiores.	$0,06$	$2,03$	∞
Pretos, pardos ou indígenas	4	Mulher.	Cursou o ensino médio todo em escola pública.	$0,49$	$1,17$	$1,29$
	5	Renda total da família é de 1,5 a 3 salários mínimos; Pai concluiu até o ensino médio; Cursou o ensino médio todo em escola pública.	Dedicou de uma a três horas de estudo por semana.	$0,14$	$1,96$	∞
	6	Mora em casa ou apartamento, com cônjuge e/ou filhos; Possui renda mas recebe ajuda da família ou de outras pessoas para financiar os gastos.	Nota está entre as menores.	$0,14$	$1,96$	∞
	7	Mora em casa ou apartamento, com outras pessoas (incluindo república); Não possui renda e os gastos são financiados pela família ou por outras pessoas.	Nota está entre as maiores.	$0,12$	$2,04$	∞

Tabela 4: Regras de associação obtidas para o subconjunto dos estudantes de IES privadas. São apresentados os itens que compõem o antecedente e o consequente de cada regra, acompanhados de seus valores de suporte, *lift* e convicção.

Cor Autodeclarada	Regra	Antecedente	Consequente	Sup.	*Lift*	Conv.
Brancos	1	Mãe concluiu até o ensino médio.	Pai concluiu até o ensino médio.	$0,24$	$1,43$	$1,45$
	2	Dedicou de 1 a 3 horas por semana aos estudos, aproximadamente, excetuando as horas de aula.	Nota está entre as menores.	$0,23$	$1,10$	$1,11$
	3	Escolheu o curso visando inserção no mercado de trabalho.	Nota do exame está entre as menores.	$0,17$	$1,19$	$1,25$
	4	Escolheu o curso por vocação.	Nota do exame está entre as maiores	$0,19$	$1,20$	$1,24.$
Pretos, pardos ou indígenas	5	Mãe concluiu até ensino médio.	Mora em casa ou apartamento, com pais e/ou parentes.	$0,31$	$1,42$	$1,51$
	6	Mulher; Possui renda mas recebe ajuda da família ou de outras pessoas para financiar os gastos; Nota no exame está entre as maiores.	Cursou o ensino médio todo em escola pública.	$0,08$	$1,37$	∞
	7	Escolheu o curso visando inserção no mercado de trabalho; Nota no exame está entre as maiores.	Cursou o ensino médio todo em escola pública.	$0,14$	$1,12$	$6,74$

Da Tabela 4, nota-se que boa parte dos concluintes entre 2016 e 2018 dedicaram apenas de 1 a 3 horas por semana aos estudos, levando a notas baixas no Enade. Itens como este, verificando-se em uma análise para uma instituição específica, pode auxiliar gestores

a entender melhor o perfil de seus estudantes e direcionar docentes a desenvolverem planos de ensino que melhor atendam à realidade de seus alunos.

Um fato interessante identificado pelas regras 3 e 7 é que estudantes autodeclarados pretos, pardos ou indígenas que escolheram o curso visando inserção no mercado de trabalho obtiveram notas melhores do que aqueles autodeclarados brancos que o escolheram pelo mesmo motivo. Um estudo complementar focado nesse aspecto poderia apresentar resultados interessantes e possivelmente úteis para todos os envolvidos. Um possível motivo poderia ser o esforço empreendido por aqueles do primeiro grupo para melhorar sua situação econômica, em geral pior que a do segundo grupo.

Nota-se, ainda, que muitas famílias cujos pais não possuem ensino superior agora têm filhas(os) com esse nível de escolaridade, graças às oportunidades oferecidas por IES privadas, cuja concorrência para ingresso costuma ser menor. A equidade de acesso às IES públicas para os grupos historicamente menos favorecidos, porém, continua sendo fundamental para uma sociedade mais justa.

5. Conclusão

O estudo realizado sobre o município de Araçatuba/SP, utilizando técnicas de mineração de dados, mostra que é possível extrair informações relevantes a partir dos microdados do Enade, tanto do ponto de vista subjetivo quanto em termos de métricas das regras de associação obtidas.

Nos estudos baseados em estatística descritiva, como em (FILHO; ROSEIRA; JR, 2020; FREITAS; COSME; NASCIMENTO, 2019; RISTOFF, 2014), geralmente são tratados números reduzidos de variáveis de estudo, e o pesquisador é responsável por direcionar a análise em busca de informações sobre um conjunto de dados, utilizando métricas como média, frequência, covariância *etc.* Assim, pode-se dizer que a quantidade de informação extraída está diretamente relacionada à quantidade de operações aplicadas sobre os dados e o número de variáveis selecionadas para estudo.

Por outro lado, técnicas de mineração de dados permitem que sejam analisadas, com igual ou menor esforço, um número maior de variáveis, e atuam de forma que os dados "revelem" suas características e relações entre as variáveis. Por exemplo, a análise feita neste trabalho revelou que aproximadamente um quarto dos estudantes brancos de IES privadas de Araçatuba possuem pais que concluíram até o ensino médio, sem que medidas estatísticas fossem calculadas explicitamente sobre as variáveis referentes ao grau de instrução dos pais. Pôde-se encontrar ainda, com um alto grau de convicção, que estudantes pretos, pardos ou indígenas de IES públicas, que moram com cônjuge e/ou filhos, possuem renda mas recebem ajuda para financiar os gastos, tendem a apresentar baixo desempenho no Enade.

O tipo de análise desenvolvido neste trabalho, se aplicado no contexto de uma instituição específica, ou até mesmo para cursos de uma instituição, tende a destacar padrões que talvez fossem difíceis de se enxergar utilizando apenas técnicas de estatística descritiva. Conhecendo os perfis socioeconômicos de seus estudantes e seus reflexos no desempenho no Enade, gestores podem desenvolver melhores políticas educacionais, de inclusão e permanência. Tal conhecimento pode ser também utilizado pelos docentes para que estes revejam suas metodologias de ensino de forma a amparar seus estudantes, principalmente aqueles em situações de vulnerabilidade social, para que estes também tirem maior proveito de suas oportunidades de cursarem o ensino superior.

Referências

AGRAWAL, R.; IMIELIŃSKI, T.; SWAMI, A. Mining association rules between sets

of items in large databases. In: **Proceedings of the 1993 ACM SIGMOD international conference on Management of data**. [S.l.: s.n.], 1993. p. 207–216.
ANDRADE, C. Y. de. Acesso ao ensino superior no brasil: equidade e desigualdade social. **Revista Ensino Superior Unicamp**, v. 6, p. 18–27, 2012.
BORGELT, C. Frequent item set mining. **Wiley interdisciplinary reviews: data mining and knowledge discovery**, Wiley Online Library, v. 2, n. 6, p. 437–456, 2012.
CHOJI, M. **mchoji/enade-py: v0.1.0**. Zenodo, 2020. Disponível em: ⟨https://doi.org/10.5281/zenodo.4082026⟩.
CRETTON, N. N.; GOMES, G. R. Aplicação de técnicas de mineração de dados na base de dados do enade com enfoque nos cursos de medicina. **Acta Biomedica Brasiliensia**, v. 7, n. 1, p. 74–89, 2016.
FILHO, A. E. C. D. M.; ROSEIRA, Í. B. R.; JR, J. A. F. P. Perfil socioeconômico e desempenho de estudantes de licenciatura em educação física no enade/brasil. **Tendencias pedagógicas**, Departamento de Didáctica y Teoría de la Educación, n. 35, p. 90–101, 2020.
FREITAS, B.; COSME, L.; NASCIMENTO, M. Exame nacional de desempenho de estudantes (enade): Análise do perfil das mulheres dos cursos da área de computação. In: **Anais do XIII Women in Information Technology**. Porto Alegre, RS, Brasil: SBC, 2019. p. 179–183. ISSN 0000-0000. Disponível em: ⟨https://sol.sbc.org.br/index.php/wit/article/view/6733⟩.
HAN, J.; PEI, J.; YIN, Y. Mining frequent patterns without candidate generation. **ACM sigmod record**, ACM New York, NY, USA, v. 29, n. 2, p. 1–12, 2000.
HEATON, J. Comparing dataset characteristics that favor the Apriori, Eclat or FP-Growth frequent itemset mining algorithms. In: IEEE. **SoutheastCon 2016**. [S.l.], 2016. p. 1–7.
INEP. **Exame Nacional de Desempenho dos Estudantes (Enade)**. 2019. Disponível em: ⟨http://portal.inep.gov.br/enade⟩.
INEP. **Indicadores de Qualidade da Educação Superior**. 2019. Disponível em: ⟨http://portal.inep.gov.br/web/guest/indicadores-de-qualidade⟩.
INEP. **Microdados**. 2020. Disponível em: ⟨http://portal.inep.gov.br/enade⟩.
NASCIMENTO, R. L. S. do; JUNIOR, G. G. da C.; FAGUNDES, R. A. d. A. Mineração de dados educacionais: Um estudo sobre indicadores da educação em bases de dados do inep. **RENOTE-Revista Novas Tecnologias na Educação**, v. 16, n. 1, 2018.
RASCHKA, S. Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. **The Journal of Open Source Software**, The Open Journal, v. 3, n. 24, abr. 2018. Disponível em: ⟨http://joss.theoj.org/papers/10.21105/joss.00638⟩.
RISTOFF, D. O novo perfil do campus brasileiro: uma análise do perfil socioeconômico do estudante de graduação. **Avaliação: Revista da Avaliação da Educação Superior (Campinas)**, scielo, v. 19, p. 723–747, nov. 2014. ISSN 1414-4077. Disponível em: ⟨http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1414-40772014000300010&nrm=iso⟩.
SILVA, L. A.; SILVA, L. Fundamentos de mineração de dados educacionais. **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**, v. 3, n. 1, 2015. Disponível em: ⟨https://br-ie.org/pub/index.php/wcbie/article/view/3281⟩.
VISTA, N. P. B.; FIGUEIRÓ, M. F.; CHICON, P. M. M. Técnicas de mineração de dados aplicadas aos microdados do enade para avaliar o desempenho dos acadêmicos do curso de ciencia da computação no rio grande do sul utilizando o software r. **I Seminário de Pesquisa Científica e Tecnológica**, v. 1, n. 1, 2017.
ZAKI, M. J. Scalable algorithms for association mining. **IEEE transactions on knowledge and data engineering**, IEEE, v. 12, n. 3, p. 372–390, 2000.