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

Mayk F. Choji[1], Seiji Isotani[2], Carlos D. N. Damasceno[3]

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.

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.

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

[1]Pós-Graduando em Computação Aplicada à Educação, USP, mayk@alumni.usp.br
[2]Professor Doutor, USP, sisotani@icmc.usp.br
[3]Pesquisador Doutor, USP, damascenodiego@usp.br

Cite as: Choji, M. & Isotani, S. & Damasceno, C. (2020). Mineração de dados do Enade de 2016 a 2018: uma análise sobre o município de Araçatuba/SP. Anais dos Trabalhos de Conclusão de Curso. Pós-Graduação em Computação Aplicada à Educação Instituto de Ciências Matemáticas e de Computação. Universidade de São Paulo.

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 gradua-ção 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, Cruz Junior e Fagundes 2018; Silva e 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, o objetivo deste trabalho é oferecer para pesquisadores e demais interessados, mecanismos que facilitem futuras tarefas de mineração de dados utilizando microdados do Enade. Para isso, grande parte dos procedimentos descritos na Seção 4, implementados na linguagem *Python*, são disponibilizados como parte do pacote *enade-py* [Choji 2020], fornecido com licença de *software* livre.

Paralelamente, 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: a Seção 2 apresenta um resumo do cenário da área de mineração de dados educacionais no Brasil. Na Seção 3 são discutidos alguns trabalhos relacionados à análise de microdados do Enade, enquanto que a Seção 4 descreve a metodologia aplicada no presente estudo. Os principais resultados são discutidos na Seção 5 e a Seção 6 conclui este trabalho e aponta trabalhos futuros.

2. Mineração de Dados Educacionais

Mineração de dados educacionais (MDE) é definida como a área de pesquisa científica interessada no desenvolvimento de métodos para se descobrir conhecimento dentro das especificidades de dados provenientes de cenários educacionais, e utilizar esses métodos para entender melhor os estudantes e os cenários em que eles aprendem [Baker, Isotani e Carvalho 2011].

Neste contexto, Baker, Isotani e Carvalho (2011) ainda classificam as sub-áreas de pesquisa em cinco categorias: *(i)* predição; *(ii)* agrupamento; *(iii)* mineração de relações; *(iv)* descoberta com modelos; e *(v)* destilação de dados para facilitar decisões humanas. Dessas, as três primeiras estão mais relacionadas à mineração de dados em outras áreas, ao passo que as duas últimas apresentam aplicações especiais para dados educacionais.

Este trabalho, particularmente, pode ser classificado no terceiro grupo, visto que o principal objetivo é buscar perfis dos concluintes de graduação por meio da mineração de regras de associação.

Uma regra de associação é uma expressão condicional da forma $A \Rightarrow C$, onde A e C são conjuntos de itens disjuntos chamados, respectivamente, de antecedente e consequente [Tan, Steinbach e Kumar 2013]. Aqui, os conjuntos de itens representam as respostas dos/as estudantes para questões selecionadas do questionário socioeconômico do Enade.

Várias métricas têm sido propostas para medir o quão interessante uma regra é. Por exemplo, o *suporte* para uma regra é definido como a fração de transações (*i.e.* registros do conjunto de dados) que satisfazem a união de itens no consequente e no antecedente da regra [Agrawal, Imieliski e Swami 1993]. A *confiança*, também apresentada em [Agrawal, Imieliski e Swami 1993], é a probabilidade de se encontrar o consequente em uma regra dado que ela também contenha o antecedente. A métrica *lift*, por sua vez, é usada para medir o quanto mais frequente o antecedente e consequente aparecem juntos do que apareceriam caso fossem independentes. Por último, para verificar se o consequente é altamente dependende do antecedente, pode-se utilizar a métrica *convicção* apresentada por Brin et al. (1997).

Em relação à pesquisa em MDE no Brasil, ainda em 2011, Baker, Isotani e Carvalho (2011) apontavam as oportunidades e os desafios para que o país ganhasse destaque no cenário educacional mundial. As oportunidades estavam principalmente relacionadas ao recente aumento no número de cursos na modalidade de ensino à distância (EaD), visto que as plataformas de ensino e aprendizado utilizadas nessa modalidade têm o potencial de gerar uma enorme quantidade de dados que podem ser minerados.

De fato, desde então o número de matrículas nessa modadalidade quase que triplicou. De acordo com as sinopses estatísticas da educação superior (graduação) disponibilizadas pelo INEP [INEP 2018], o número de matrículas aumentou de 727.961 em 2008 (ano citado pelos autores) para 1.153.572 em 2013. Em 2018, último ano que se tem registro, o número de matrículas foi de 2.056.511.

Em um Mapeamento Sistemático da Literatura em MDE no Brasil realizada por Maschio et al. (2018), foram analisados 49 trabalhos publicados desde 2001. Os

resultados obtidos revelam que os trabalhos nesta área têm explorado principalmente dados oriundos de interações de estudantes do ensino superior com sistemas de gestão de aprendizado empregados na modalidade EaD. Além disso, quase 30 trabalhos empregavam algoritmos de aprendizado de máquina em problemas de predição de desempenho, evasão, aprovação *etc.*, seguidos de 10 trabalhos sobre agrupamento de dados e 5 utilizando regras de associação.

Em relação a dados públicos para MDE, muitos pesquisadores têm utilizado micro-dados disponibilizados pelo INEP, como informações referentes ao Enem (Exame Nacional do Ensino Médio) e Enade. Lima et al. (2019) apresentam uma revisão sistemática da literatura justamente sobre trabalhos que abordam análise de dados desses dois exames. No total, são considerados 54 trabalhos publicados entre 2005 e 2016, dentre artigos de periódicos ou de conferências, dissertações e teses. Desses, 37 tratam do Enade, 15 do Enem e 2 tratam de ambos. O uso de técnicas de mineração de dados, porém, é encontrado em apenas 2 dos trabalhos listados.

Embora muitos trabalhos recentes ainda tenham adotado abordagens estatísticas para análise dos microdados do Enem e do Enade, já é possível encontrar mais pesquisas envolvendo mineração de dados [vide Cretton e Gomes 2016; Nascimento, Cruz Junior e Fagundes 2018; Vista, Figueiró e Chicon 2017] do que apenas aquelas listadas por Lima et al. (2019). Sobre os temas abordados, observa-se, por exemplo, a análise para cursos ou instituições de ensino específicas, estudo sobre uma região (cidade, estado *etc.*) de interesse, identificação de perfis socioeconômicos, predição de desempenho, problema de evasão e muitos outros. Alguns relacionados com o presente estudo são discutidos a seguir.

3. 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 n^{o} 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 [De Medeiros Filho, Roseira e Pontes 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. Este tipo de análise pode ser facilitada pela ferramenta apresentada neste trabalho, com a vantagemde se obter métricas que indicam a frequência dos padrões encontrados e também a relação entre as variáveis (*e.g.*, o quão uma variável é dependente de outra).

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

Também seguindo uma abordagem estatística, Freitas, Cosme e Nascimento (2019) utilizam os microdados do Enade de 2017 para analisar o perfil das mulheres nos cursos da área de computação. Por meio das respostas do questionário socioeconômico, as autoras verificam a predominância de homens e mulheres brancos nesses cursos, estando as mulheres representando menos de 20% dos concluintes em praticamente todas as regiões do Brasil. Embora o trabalho apresente um perfil geral das mulheres envolvendo aspectos como cor e renda familiar, essas dimensões são analisadas individualmente, ou seja, não é clara a relação entre os indivíduos de cada grupo de estudo.

Araújo (2019) utiliza os microdados do Enade de 2017 para desenvolver um modelo de classificação utilizando árvores de decisão, voltado principalmente para o desempenho dos estudantes no exame. O modelo utiliza informações do exame, principalmente do questionário socioeconômico, para prever se o desempenho será alto ou baixo, dada as características de entrada. Como resultado de seu trabalho, o autor propõe uma ferramenta implementada na linguagem R com o *framework Shiny* para apresentar o modelo obtido de acordo com parâmetros de entrada fornecidos pelo usuário. A ferramenta também permite ao usuário comparar cursos e instituições estatisticamente. Tais funcionalidades, de aspectos mais administrativos, foram avaliadas por 32 entrevistados, dentre alunos, professores e coordenadores de instituições de ensino superior. Apesar de os resultados apontarem boa aceitação da ferramenta, não há referência no trabalho a respeito de sua disponibilidade e licença para uso.

Dentre os trabalhos voltados para o estudo de cursos específicos, Vista, Figueiró e Chicon (2017) e 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 e 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 dificuldate do componenete específico, muito embora seus rendimentos tenham sido negativo, de acordo com a classificação dos autores.

O que observa-se, em geral, dos trabalhos que aplicam mineração de dados e/ou aprendizagem de máquina sobre os microdados do Enade, é que as análises são feitas sobre um contexto restrito (*e.g.*, um curso, um *campus*) e algumas etapas são comuns entre eles. Por exemplo, é preciso carregar corretamente os microdados em uma estrutura de dados adequada na linguagem de programação ou ferramenta utilizada, selecionar um subconjunto dos dados baseado em curso, região de interesse *etc.*, selecionar os atributos de interesse, dentre outras. No melhor do conhecimento do autor, este é o primeiro trabalho a disponibilizar publicamente uma biblioteca de funções para auxiliar autores em futuros trabalhos utilizando os microdados do Enade.

4. 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. O código fonte e a documentação da ferramenta podem ser acessados, respectivamente, em https://github.com/mchoji/enade-py e https://enade-py.rtfd. io.

4.1. 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 tem informações disponíveis na página oficial do INEP [INEP 2020]. 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.

O questionário do estudante, por sua vez, pode ser dividido em três partes. A primeira, correspondente às variáveis de `QE_I01` a `QE_I26`, contém questões de múltipla escolha de caráter socioeconômico. A segunda parte, correspondente às variáveis de `QE_-I27` a `QE_I68`, contém questões que medem o nível de concordância do estudante sobre assertivas referentes ao curso e à instituição de ensino. A terceira e última parte é exclusiva para cursos de licenciatura e compreende questões de múltipla escolha identificadas pelas variáveis de `QE_I69` a `QE_I81`.

4.2. 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.

Em relação às variáveis dos microdados, o conjunto referente ao ano de 2016 apresenta inconsistências comparadas aos dois anos mais recentes. Por exemplo, a variável que representa o ano em que o estudante terminou o ensino médio é chamada de `ANO_FIM_2G` no mais antigo e de `ANO_FIM_EM` nos mais novos. Outro exemplo diz respeito à indicação do período do curso. Nos dados de 2016, existe uma variável binária para cada turno (*i.e.*, matutino, diurno e noturno), ao passo que nos dados de 2017 e 2018 essa informação é representada por uma única variável categórica (`CO_TURNO_GRADUACAO`). A função `align_microdata_2016` implementada no pacote *enade-py* [Choji 2020] realiza as transformações necessárias para que os dados de 2016 possam ser alinhados com os demais.

Estando os três conjuntos definidos pelas mesmas variáveis, 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 possuiam 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 4.1 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 4.2 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[4]. As categorias das

[4] https://emec.mec.gov.br/

Tabela 4.1: 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?"

instituições foram utilizadas para dividir o conjunto de dados em dois grupos: instituições públicas e instituições privadas.

Um estudo preliminar, dividindo-se os participantes apenas por categoria da IES, apresentou pouca variação nas informações obtidas das regras de associação, principalmente pelo fato de estudantes de cor branca serem muito mais frequentes nos dados do que os demais.

Por conta disso, uma segunda divisão dos dados fora realizada, desta vez de acordo com a cor ou raça autodeclarada. Para efeitos de estudo e após analisar a distribuição exibida na Figura 5.2, 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 quatro subconjuntos resultantes dessas divisões.

4.3. 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

Tabela 4.2: 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

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.

4.4. Mineração

Na primeira parte da etapa de mineração de dados, o algoritmo *FP-Growth* [Han, Pei e 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, Imieliski e Swami 1993] e o ECLAT [Zaki 2000]. Diversos trabalhos recomendam o *FP-Growth* por 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 frequentese 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.

Aqui, foram implementadas as funções `filter_rules`, `find_itemsets_-any` e `find_itemsets_without` para filtrar regras redundantes e buscar regras que contenham itens de interesse do pesquisador, com suporte aos operadores lógicos de *e*, *ou* e *negação*. Por exemplo, é possível buscar regras onde o antecedente contenha o item `TP_SEXO_F` OU `TP_SEXO_M`, regras que não contenham o item `QE_I06_A`, e assim por diante.

A função `filter_rules`, citada anteriormente, considera duas regras como redundantes caso a união dos itens do antecedente e consequente seja igual para as duas. Neste caso, é mantida apenas aquela que tiver os maiores valores para as métricas definidas como argumentos da função. A documentação completa do pacote *enade-py* [Choji 2020] pode ser encontrada em https://enade-py.readthedocs.io.

Utilizando esses procedimentos, foi possível analisar com mais facilidade as inúmeras regras de associação que foram geradas, algumas das quais são discutidas a seguir.

5. 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 5.1, essa relação se mostra ainda mais evidente no grupo das instituições públicas, no qual representam mais de 70%.

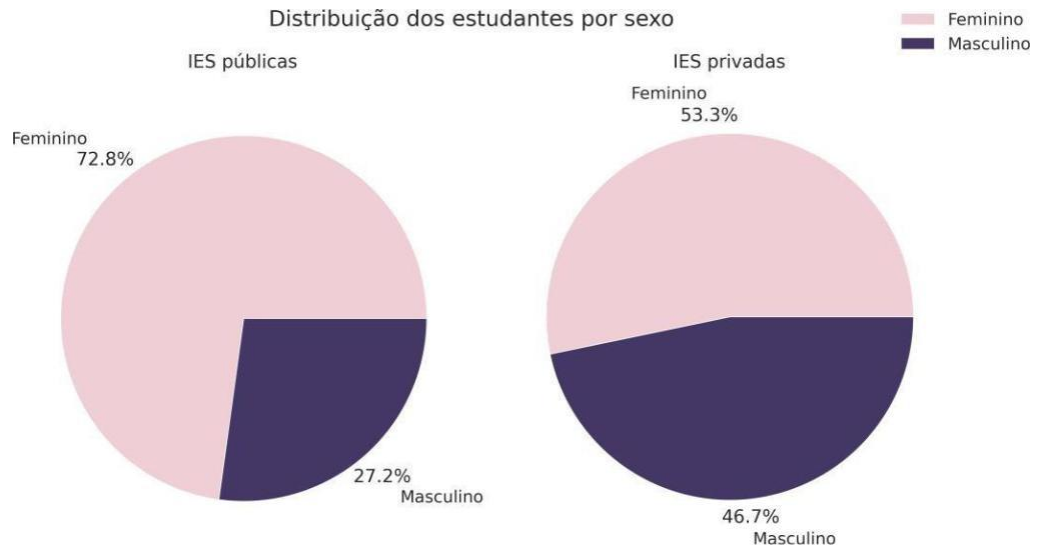

Figura 5.1: Distribuição de alunos por sexo, nas instituições públicas (à esquerda) e privadas (á direita).

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. Apesar dos estudos do IBGE classificarem a população brasileira como 42,7% branca e 46,8% parda [IBGE 2020], a proporção que se apresenta no município de Araçatuba é de mais de 70% de brancos e menos de 20% de pardos, tanto nas instituições públicas quanto privadas. A distribuição dos estudantes por cor ou raça, segundo as declarações de 2016 a 2018, é mostrada na Figura 5.2.

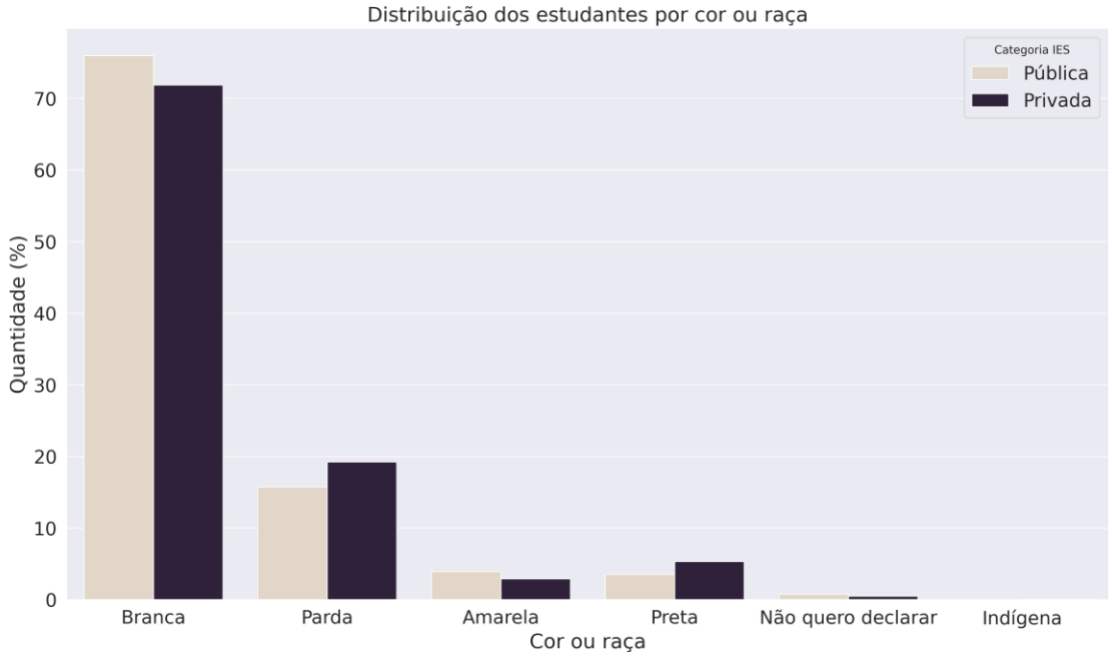

Figura 5.2: Distribuição de alunos por cor ou raça, nas IES públicas e privadas.

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 5.3, a maioria dos valores encontra-se na faixa entre 20 e 30 anos, notando-se uma população levemente mais jovem nas instituições privadas.

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* [∈0, 1] e *l, c* [0∈,∞]. Nos parágrafos seguintessão discutidas algumas regras de associação obtidas para cada grupo de estudo, resultadodo processo discutido na Subseção 4.4, acompanhadas das respectivas métricas.

Cada subgrupo analisado neste trabalho apresenta características que refletem na força das métricas selecionadas para estudo. Assim, análises descritivas dos valores das métricas são apresentadas em forma de tabelas para facilitar o entendimento das regras selecionadas para discussão e seus valores frente às demais.

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. As distribuições dos valores das métricas são apresentadas na Tabela 5.3. Nela, observa-se que as regras com maiores valores de suporte aproximam-se de 50% e que foi possível obter regras com valor máximo de convicção (∞).

Algumas regras interessantes, tanto do ponto de vista objetivo (valor alto em uma ou mais métricas) quanto subjetivo, são apresentadas na Tabela 5.5. 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

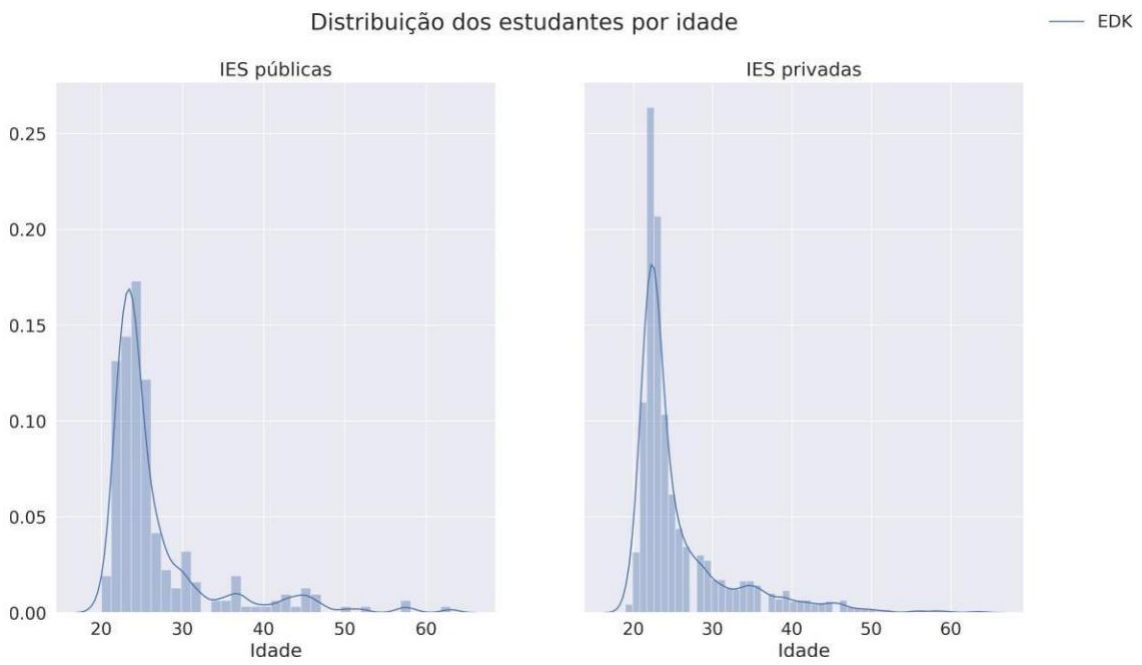

Figura 5.3: Histograma da distribuição dos estudantes por idade, nas IES públicas (à esquerda) e privadas (á direita). A curva indicada em cada gráfico representa a estimativa de densidade por Kernel (EDK) da variável `NU_IDADE`.

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

Tabela 5.3: Distribuição dos valores das métricas de suporte, *lift* e convicção, referentes às regras de associação obtidas para os subgrupos de estudantes brancos e pretos, pardos ou indígenas das IES públicas.

Cor Autodeclarada	Métrica	mean	std	min	25%	50%	75%	max
Brancos	suporte	0,07	0,03	0,05	0,05	0,06	0,07	0,46
	lift	1,63	0,49	1,10	1,27	1,49	1,83	5,85
	convicção	inf	NaN	1,01	1,07	1,17	1,44	inf
Pretos, pardos ou indígenas	suporte	0,07	0,02	0,06	0,06	0,06	0,08	0,49
		2,74	1,77	1,10	1,62	2,15	3,27	16,33
	convicção	inf	NaN	1,02	1,14	1,39	2,19	inf

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 5.5).

As regras obtidas para o grupo dos estudantes brancos das IES privadas não apresentaram métricas com valores tão altos quanto as anteriores. De fato, os valores máximos de *suporte* e *convicção* foram 0, 28 e 9, 71, respectivamente, conforme consta na Tabela 5.4. 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 5.6.

Tabela 5.4: Distribuição dos valores das métricas de suporte, *lift* e convicção, referentes às regras de associação obtidas para os subgrupos de estudantes brancos e pretos, pardos ou indígenas das IES privadas.

Cor Autodeclarada	Métrica	mean	std	min	25%	50%	75%	max
Brancos	suporte	0,07	0,02	0,05	0,05	0,06	0,08	0,28
	lift	1,29	0,40	1,10	1,15	1,20	1,33	14,99
	convicção	1,19	0,37	1,01	1,04	1,08	1,18	9,71
Pretos, pardos ou indígenas	suporte	0,07	0,02	0,05	0,05	0,06	0,08	0,31
		1,36	0,41	1,10	1,16	1,25	1,40	5,83
	convicção	inf	NaN	1,01	1,04	1,09	1,22	inf

Tabela 5.5: 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 escoal 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 5.6: 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 5.6, 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.

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.

6. Conclusão e Trabalhos Futuros

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 [De Medeiros Filho, Roseira e Pontes Jr 2020; Freitas, Cosme e 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 caractéricas 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.

Mais importante do que os resultados sobre o município escolhido para estudo de caso, este trabalho oferece o pacote *enade-py* [Choji 2020] publicamente para que outros trabalhos possam utilizar seus recursos em novas análises sobre os microdados do Enade. Diminuindo-se o tempo gasto em operações comuns como leitura dos dados e seleção de variáveis, o pesquisador pode se dedicar às especificidades de seu estudo, seja voltado

a um curso no âmbito nacional, uma instituição de ensino específica, ou outro contexto qualquer.

Uma das restrições, porém, é que o processo para encontrar regras com características distintas, ou seja, itens diferentes no antecedente e/ou consequente das regras, requer algum esforço manual e é, portanto, passível de erro. O motivo é que os itens mais frequentes tendem a aparecer em muitas regras de associação, gerando regras mais ou menos redundantes em alguns casos. Possíveis alternativas neste sentido seriam a implementação de métodos que automatizem a extração de regras de interesse dentro do conjunto total de regras obtidas, ou a utilização de técnicas de aprendizado de máquina em conjunto com mineração de dados. Futuramente, portanto, espera-se incluir novas funcionalidades ao pacote *enade-py* [Choji 2020], como apoio à criação de modelos de árvores de decisão, agrupamentos, e suporte a versões mais antigas dos microdados.

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.

No entanto, para que ferramentas de mineração de dados e aprendizado de máquina possam ser exploradas por docentes e gestores, é essencial que ofereçam facilidades para operação, principalmente em relação a aspectos de interface e experiência de usuário. Assim, planeja-se desenvolver uma interface gráfica *web*, também de licença livre, para que usuários finais possam interagir com esses conjuntos de dados e entender melhor os perfis dos estudantes de suas instituições e cursos específicos de maneira simples e intuitiva, na esperança de que isso resulte, de fato, em melhores políticas educacionais, de inclusão e permanência.

Embora a ferramenta desenvolvida neste trabalho, *enade-py* [Choji 2020], não tenha sido avaliada por outros profissionais da área como parte do estudo, ela encontra-se disponível publicamente para uso, avaliação e contribuição da comunidade, a partir do repositório localizado em https://github.com/mchoji/enade-py. No melhor do conhecimento do autor, esta é a primeira iniciativa do tipo para estudo sobre os dados do Enade. Uma avaliação formal deve ser conduzida a partir do momento que se tenha uma interface voltada para usuários finais.

Referências

Agrawal, Rakesh, Tomasz Imieliski e Arun Swami (1993). “Mining association rules between sets of items in large databases”. Em: *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pp. 207–216.

Andrade, Cibele Yahn de (2012). “Acesso ao ensino superior no Brasil: equidade e desigualdade social”. Em: *Revista Ensino Superior Unicamp* 6, pp. 18–27.

Araújo, Rodrigo Alexandrino (2019). “Análise dos microdados do Enade: Proposta de uma ferramenta de exploração utilizando mineração de dados”. Diss. de mestr. Universidade Federal de Goiás. 69 pp.

Baker, Ryan, Seiji Isotani e Adriana Carvalho (2011). “Mineração de Dados Educacionais: Oportunidades para o Brasil”. Em: *Revista Brasileira de Informática na Educação* 19.02, pp. 3–13. URL: https://br-ie.org/pub/index.php/rbie/article/view/1301.

Borgelt, Christian (2012). “Frequent item set mining”. Em: *Wiley interdisciplinary reviews: data mining and knowledge discovery* 2.6, pp. 437–456.

Brin, Sergey et al. (1997). “Dynamic Itemset Counting and Implication Rules for Market Basket Data”. Em: *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*. SIGMOD 97. Tucson, Arizona, USA: Association for Computing Machinery, pp. 255–264. URL: https://doi-org.ez67.periodicos. capes.gov.br/10.1145/253260.253325.

Choji, M. (out. de 2020). *mchoji/enade-py: v0.1.0*. Versão v0.1.0. URL: https://doi.org/10.5281/zenodo.4082026.

Cretton, Nicollas Nogueira e Georgia Rodrigues Gomes (2016). “Aplicação de técnicas de mineração de dados na base de dados do ENADE com enfoque nos cursos de medicina”. Em: *Acta Biomedica Brasiliensia* 7.1, pp. 74–89.

De Medeiros Filho, Antonio Evanildo Cardoso, Ítalo Breno Rocha Roseira e Jose Airton Freitas Pontes Jr (2020). “Perfil socioeconômico e desempenho de estudantes de licenciatura em educação fsica no ENADE/BRASIL”. Em: *Tendencias pedagógicas* 35, pp. 90–101.

Freitas, Barbara, Luciana Cosme e Mayara Nascimento (2019). “Exame Nacional de Desempenho de Estudantes (ENADE): Análise do Perfil das mulheres dos cursos da área de computação”. Em: *Anais do XIII Women in Information Technology*. Belém: SBC, pp. 179–183. URL: https://sol.sbc.org.br/index.php/wit/article/view/6733.

Han, Jiawei, Jian Pei e Yiwen Yin (2000). “Mining frequent patterns without candidate generation”. Em: *ACM sigmod record* 29.2, pp. 1–12.

Heaton, Jeff (2016). “Comparing dataset characteristics that favor the Apriori, Eclat or FP-Growth frequent itemset mining algorithms”. Em: *SoutheastCon 2016*. IEEE, pp. 1–7.

IBGE (2020). *Conheça o Brasil - População: cor ou raça*. URL: https://educa.ibge.gov.br/jovens/conheca-o-brasil/populacao/18319-cor-ou-raca.html (acesso em 09/07/2020).

INEP (20 de set. de 2018). *Sinopses Estatísticas da Educação Superior Graduação*. URL: http://portal.inep.gov.br/web/guest/sinopses-estatisticas-da-educacao-superior (acesso em 25/06/2020).

INEP (23 de ago. de 2019a). *Exame Nacional de Desempenho dos Estudantes (Enade)*. URL: http://portal.inep.gov.br/enade (acesso em 12/06/2020).

INEP (11 de jul. de 2019b). *Indicadores de Qualidade da Educação Superior*. URL: http://portal.inep.gov.br/web/guest/indicadores-de-qualidade (acesso em 13/06/2020).

INEP (23 de mar. de 2020). *Microdados*. URL: http://portal.inep.gov.br/enade (acesso em 25/06/2020).

Lima, Priscila da Silva Neves et al. (mai. de 2019). “Análise de dados do Enade e Enem: uma revisão sistemática da literatura”. pt. Em: *Avaliação: Revista da Avaliação da*

Educação Superior (Campinas) 24, pp. 89–107. URL: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1414-40772019000100089&nrm=iso.

Maschio, Pedro et al. (2018). "Um panorama acerca da mineraçao de dados educacionais no Brasil". Em: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. Vol. 29. 1, pp. 1936–1940.

Nascimento, Rafaella Leandra Souza do, Geraldo Gomes da Cruz Junior e Roberta Andrade de Araújo Fagundes (2018). "Mineração de Dados Educacionais: Um estudo sobre indicadores da educação em bases de dados do INEP". Em: *RENOTE-Revista Novas Tecnologias na Educação* 16.1.

Raschka, Sebastian (abr. de 2018). "MLxtend: Providing machine learning and data science utilities and extensions to Pythons scientific computing stack". Em: *The Journal of Open Source Software* 3.24. URL: http://joss.theoj.org/papers/10.21105/joss.00638.

Ristoff, Dilvo (nov. de 2014). "O novo perfil do campus brasileiro: uma análise do perfil socioeconômico do estudante de graduação". pt. Em: *Avaliação: Revista da Avaliação da Educação Superior (Campinas)* 19, pp. 723–747. URL: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1414-40772014000300010&nrm=iso.

Silva, Leandro A. e Luciano Silva (2015). "Fundamentos de Mineração de Dados Educacionais". Em: *Anais dos Workshops do Congresso Brasileiro de Informática na Educação* 3.1. URL: https://br-ie.org/pub/index.php/wcbie/article/view/3281.

Tan, Pang-Ning, Michael Steinbach e Vipin Kumar (2013). *Introduction to data mining*. Pearson Education Limited.

Vista, Nicolas Pastorio Boa, Michele Ferraz Figueiró e Patricia Mariotto Mozzaquatro Chi-con (2017). "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". Em: *I Seminário de Pesquisa Cientfica e Tecnológica* 1.1.

Zaki, Mohammed Javeed (2000). "Scalable algorithms for association mining". Em: *IEEE transactions on knowledge and data engineering* 12.3, pp. 372–390.