Análise do Desempenho no Enade dos Concluintes de Computação usando Técnica de Agrupamento

Alisson da Silva Vieira[1], **Diego Bertolini**[1], **André Luis Schwerz**[1]

[1]Universidade Tecnológica Federal do Paraná (UTFPR)
Campo Mourão – PR – Brasil

alisson.v3@hotmail.com, {diegobertolini, andreluis}@utfpr.edu.br

Abstract. *Data mining is a well-known method for evaluating many aspects of society and its evolution over the years. Specifically, data mining has the potential to discover patterns in education that can contribute to the teaching-learning process. Thus, the main objective of this paper is to use the data clustering technique known as K-means in the Enade microdata that regularly evaluates Brazilian higher education. We have filtered students' microdata in computing-related courses across all exam editions and performed extensive pre-processing to avoid statistical bias in clusters. The results show an evolutionary analysis of the clusters throughout the editions, observing aspects relevant to students' performance. The data and results can be helpful for decision-making and further studies about higher education.*

Resumo. *A mineração de dados é um proeminente método para avaliar vários aspectos da sociedade e sua evolução ao longo dos anos. Especificamente, a mineração de dados tem potencial para descoberta de padrões na educação que podem contribuir para o processo de ensino-aprendizagem. Desta forma, o principal objetivo deste artigo é utilizar a técnica de agrupamento dos dados conhecida como K-means nos microdados do Enade que avalia regularmente o ensino superior brasileiro. Os microdados dos estudantes dos cursos de computação em todas as edições do exame foram selecionados e um extenso pré-processamento a fim de evitar viés estatístico nos agrupamentos foi realizado. Os resultados mostram uma análise evolutiva dos grupos ao longo das edições, observando aspectos relevantes para o desempenho dos estudantes. Os dados e os resultados podem ser úteis para tomada de decisões e outros estudos em relação ao ensino superior.*

1. Introdução

As avaliações educacionais proveem índices de aprendizado dos discentes e têm como um dos principais objetivos garantir a qualidade do ensino. Elas permitem observar o desempenho dos estudantes, as deficiências e potencialidades das instituições, além de possibilitar o cálculo de indicadores de qualidade que podem auxiliar a proporcionar melhorias no processo de ensino-aprendizagem. O Exame Nacional de Desempenho dos Estudantes (Enade) é um dos procedimentos de avaliação do Sistema Nacional de Avaliação da Educação Superior (Sinaes) [BRASIL 2004], organizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), órgão do Ministério da Educação [BRASIL 2022]. O Enade avalia o desempenho dos alunos concluintes em

DOI: 10.5753/sbie.2022.224710

relação aos conteúdos curriculares do curso. Cada área de conhecimento é periodicamente avaliada e os dados referentes aos participantes são disponibilizados em formato aberto que podem ser analisados e explorados pela comunidade [INEP 2022].

A mineração de dados educacionais é um método que busca padrões e regras a partir de dados educacionais para gerar conhecimentos úteis e compreensíveis que subsidiem a tomada de decisão aos diversos níveis de gestores educacionais [Baker et al. 2011, Costa et al. 2012]. Entre as técnicas, o agrupamento é usado na mineração de dados para encontrar amostras similares [Wu 2012]. Técnicas de agrupamento têm sido usadas em diferentes trabalhos [Webber et al. 2013, Vista et al. 2018, Francelino e Machado 2020, Lima et al. 2020]. Entretanto, nenhum destes trabalhos realizam um amplo estudo com os estudantes de computação ao longo das várias edições do exame.

Este artigo propõe a adoção de uma técnica de agrupamento sobre os microdados do Enade dos concluintes dos cursos de área da computação a fim de identificar padrões que possam caracterizar o desempenho dos estudantes no exame. Mais especificamente, o algoritmo de aprendizado não-supervisionado K-means é utilizado para agrupar os concluintes de acordo com a similaridade de todas suas características. Um minucioso pré-processamento que inclui seleção, limpeza, transformação dos dados foi realizado a fim de evitar viés no agrupamento. Como resultado, observou-se que há uma significativa diferença entre a média do desempenho dos participantes dos agrupamentos, o que permitiu classificar os grupos como alto, médio e baixo desempenho. Em seguida, uma análise evolutiva dos grupos foi realizada, observando questões relevantes como o turno, o tipo do ensino médio, a categoria administrativa da IES, a escolaridade dos pais, a participação do sistema de cotas para ingresso e a participação dos concluintes nos diferentes cursos.

Este artigo está organizado da seguinte forma. Na Seção 2, são descritos os trabalhos relacionados. Na Seção 3, discute-se o método de pesquisa. Os resultados são apresentados e discutidos na Seção 4. Por fim, na Seção 5, apresenta-se a conclusão desse trabalho.

2. Trabalhos Relacionados

A mineração de dados educacionais tem como objetivo estudar dados de contexto educacional que podem incluir alunos, docentes, instituições, entre outros participantes para descoberta de informações úteis que visem melhorar a qualidade do ensino [Costa et al. 2012, Baker et al. 2011]. Nesta seção, são apresentados trabalhos que aplicam técnicas de agrupamento para mineração de dados educacionais.

Em [Lima et al. 2020], o algoritmo K-means foi aplicado para encontrar agrupamentos nos dados dos estudantes do Exame Nacional do Ensino Médio (ENEM) do período entre 2012 e 2017. Em seus resultados, o desempenho dos estudantes no exame foi observado nos agrupamentos produzidos pelo K-means. Os grupos foram classificados em alto, médio e baixo desempenho. Além de observar as características de cada grupo, análises socioeconômicas foram realizadas, considerando os agrupamentos nas cinco regiões do Brasil.

Uma análise restrita aos alunos do Bacharelado em Ciência da Computação que realizaram o Enade na edição de 2014 foi proposta em [Francelino e Machado 2020]. Seu objetivo é utilizar o algoritmo K-means para encontrar indicadores ou relações entre

eles, e, então, entender os motivos que estão relacionados à nota do aluno. A ferramenta Weka foi usada para a aplicação do K-means. Em destaque, os autores mostram que o rendimento do aluno não está relacionado com a região em que ele estuda. Por ser restrita a uma edição do exame, outras características evolutivas em relação aos grupos formados não foram observadas.

O agrupamento hierárquico foi utilizado para agrupar microdados da edição de 2014 dos alunos das instituições do Consórcio das Universidades Comunitárias Gaúchas (COMUNG), que oferecem o curso de Ciência da Computação [Vista et al. 2018]. Em destaque, os autores mostraram que as instituições foram distribuídas em quatro grupos, sendo que três deles apresentaram diferenças estatísticas dos conceitos Enade, IGC e CPC.

Técnicas de aprendizado de máquina supervisionado também foram usadas para análise dos dados do Enade [de Vasconcelos e de Castro 2020, Rodrigues e Gouveia 2021, Rezende et al. 2022]. Há também outros estudos que não envolvem aprendizado de máquina. Em [Cunha et al. 2021], um software foi proposto para comparar o desempenho dos estudantes entre os cursos de Bacharelado em Ciência da Computação considerando o conteúdo abordado em cada questão. Em [Capelari e Schwerz 2021], estatística descritiva é usada para identificar o perfil socioeconômico do discente de computação da região sul do Brasil.

Pode-se concluir que, no melhor do nosso conhecimento, nenhum trabalho propõe o uso de técnicas de agrupamento nos microdados de várias edições do Enade dos cursos de computação para entender o perfil dos estudantes de acordo com o desempenho alcançado no exame.

3. Método de Pesquisa

Nesta seção, são apresentados o conjunto de dados coletados; a etapa de pré-processamento incluindo seleção, limpeza, transformação dos dados; e o algoritmo de agrupamento utilizado junto com seus parâmetros.

3.1. Microdados do Enade

O Inep disponibiliza anualmente em formato de microdados a base de dados dos participantes que realizaram o Enade. O exame possui um ciclo avaliativo de três anos, avaliando um conjunto de cursos diferentes em cada um desses anos [BRASIL 2022]. Como o foco desta pesquisa são estudantes brasileiros de computação, foram coletados os anos do ciclo avaliativo do Ano II, sendo eles: 2005, 2008, 2011, 2014 e 2017. A edição de 2020 foi realizada em 2021 por causa da pandemia COVID-19, mas o resultado ainda não está disponível no momento da escrita deste trabalho. Os cursos da área de computação que realizaram o exame nas edições coletadas são: Gestão da Tecnologia da Informação (GTI); Análise e Desenvolvimento de Sistemas (ADS); Bacharelado em Ciência da Computação (BCC); Engenharia da Computação (EC); Redes de Computadores (RC); Sistemas da Informação (SI); e Licenciatura em Ciência da Computação (LCC).

3.2. Pre-processamento

Em seguida, iniciou-se o pré-processamento para seleção, limpeza e transformações dos dados. Primeiramente, foram identificados os atributos mantidos em todas as edições do exame. Foram removidos os atributos que armazenam as respostas dos estudantes

em cada questão e notas parciais. A Tabela 1 apresenta quais são esses atributos e suas categorias.

Atributo	Categoria
Categoria do ensino médio cursado; tipo do ensino médio; sexo; raça; se obteve cota; se trabalha; quantidade de livros lidos; horas de estudo dedicadas; estado civil; tipo de moradia; quantidade de moradores em sua moradia; escolaridade do pai; escolaridade da mãe.	Socioeconômico
Nome do curso; categoria administrativa da instituição; unidade federativa; região; turno do curso; condições das salas para aulas práticas; plano de ensino das disciplinas; avaliação do curso pelo aluno.	Curso
Idade do aluno no ingresso da graduação; tempo entre o início da graduação e término do ensino médio; tempo de graduação.	Calculados
Nota geral.	Prova

Tabela 1. Atributos selecionados dos microdados do Enade.

Como a maioria dos atributos (ou seus valores) do exame de 2005 são incompatíveis com aqueles das demais edições, optou-se pela sua eliminação da análise. Por exemplo, em 2005, os atributos da escolaridade do pai e da mãe não se encontram presentes. Isto, entre outras incompatibilidades, inviabiliza a representação desses participantes e enviesa o método de agrupamento.

Ainda nesta fase, removeu-se os alunos com nota zero, que não haviam respondido alguma das questões observadas pelos atributos selecionados, e alunos com notas inválidas (informação passada por um atributo nos dados). A Tabela 2 sintetiza os dados antes e após o pré-processamento. Cabe ressaltar que em 2008, após o pré-processamento, houve uma redução acentuada dos dados. Isso ocorreu porque 22% dos alunos concluintes foram considerados com notas inválidas pelo próprio Inep. Outro ponto que vale ser observado na Tabela 2 é que o Enade tornou-se universal a partir de 2009, fazendo que houvesse um aumento significativo dos participantes nas edições de 2011, 2014 e 2017 em relação à edição de 2008.

Ano	Antes do pré processamento			Após o pré-processamento	
	# estudantes	# concluintes da computação	# atributos	# concluintes da computação	# atributos
2008	200.776	27.384	198	11.524	26
2011	376.180	43.035	115	32.157	26
2014	481.720	51.776	154	40.737	26
2017	537.436	46.446	150	38.803	26

Tabela 2. Conjunto de dados antes e após o pré-processamento.

O método *z-score* foi usado para normalizar o atributo da nota geral para os alunos do mesmo curso e do mesmo ano. Para atributos categóricos, adotou-se a técnica conhecida como *one-hot-encoding* [Daly et al. 2016]. Nesta técnica, cada valor do atributo original é transformado em uma nova coluna no conjunto de dados. Então, os valores desses novos atributos tornam-se binários. Após essa transformação, o conjunto de dados contém oitenta e duas colunas.

3.3. K-Means

O K-means [MacQueen 1967] é um dos mais famosos algoritmos de agrupamento baseado em partições por causa da sua simplicidade e da qualidade dos resultados. O K-means divide os m elementos do conjunto de dados em k grupos de tal forma que os elementos que estejam no mesmo grupo sejam similares entre si e dissimilares com os elementos dos outros grupos [Xu e Tian 2015]. Inicialmente, define-se o valor de k (número de agrupamentos). Após isso, o algoritmo aleatoriamente seleciona k elementos para serem centroides do grupo e, para cada elemento, é calculado a distância (similaridade) para cada centroide. Cada elemento é realocado ao agrupamento do centroide mais próximo (o da menor distância). Após a realocação, o algoritmo escolhe novamente novos centroides, porém, desta vez não de forma aleatória, mas pela média da posição de todos os elementos do grupo. Por fim, os passos são repetidos até que não haja mais a necessidade de realocação dos elementos, pois os centroides não foram alterados.

Para os experimentos realizados neste trabalho, os alunos são representados por suas características contidas no conjunto de dados do Enade. Alunos similares são aqueles que possuem características similares no espaço n dimensional (26 dimensões listadas na Tabela 1). Ao considerar o espaço n dimensional, o K-means provê partições que observam todas as características dos estudantes. Esta abordagem é mais adequada do que apenas estabelecer limiares artificiais (como quartis, mediana) baseadas em apenas um atributo, para categorizar os estudantes pela nota obtida no exame.

Os experimentos foram implementados com a linguagem de programação Python, versão 3.9.7, e a biblioteca sklearn. Para manter a reprodutibilidade da pesquisa, o código-fonte e os dados do experimento pode ser encontrado anonimizado a partir do Github[1].

3.4. Método de Elbow

Uma dificuldade comum em se utilizar o K-means é a definição adequada do número de agrupamentos k, dado como entrada para o algoritmo. Se k for demasiadamente pequeno, haverá no mesmo agrupamento elementos distintos. Por outro lado, se k for muito elevado, elementos similares estarão em agrupamentos distintos. O método de Elbow é uma técnica bem aceita para identificar a melhor quantidade de agrupamentos (o valor de k). O método propõe encontrar o valor k de agrupamentos que minimizará a soma das distâncias de cada elemento para o seu centroide. Neste trabalho, aplicou-se a soma dos erros quadráticos (SSE, do inglês *Sum Square Error*) de cada elemento para o seu centroide. Então, o melhor valor de k agrupamentos é aquele que produzirá a melhor redução no SSE. Para isso, o método de Elbow incrementa o número de k agrupamentos até que se observe uma grande redução do valor SSE, caracterizando um "cotovelo" no gráfico.

A Figura 1 ilustra o resultado da aplicação do método Elbow para cada uma das edições do Enade. Embora o "cotovelo" não tenha se destacado no gráfico, $k = 3$ foi o agrupamento que reduziu mais acentuadamente o SSE enquanto os demais valores de k produziram apenas uma suave redução. Entre os trabalhos relacionados, a pesquisa realizada por [Lima et al. 2020] também usou $k = 3$, tendo apresentando um comportamento semelhante ao ilustrado na Figura 1.

[1]Código-fonte: https://github.com/als-v/Analise-Enade

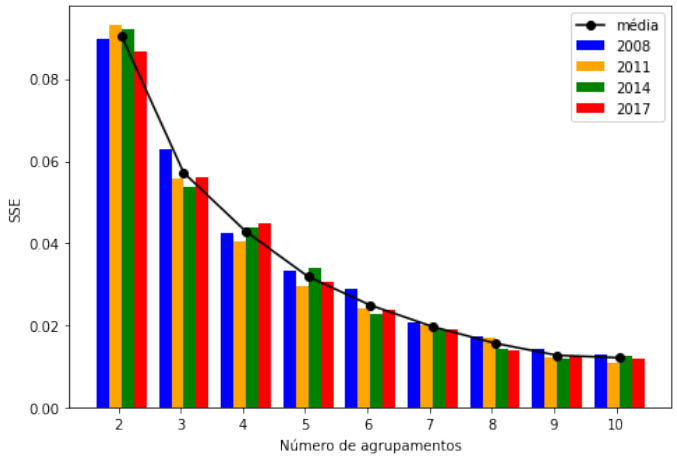

Figura 1. Melhor valor de k agrupamentos.

4. Resultados e Discussões

Nesta seção apresenta-se os resultados obtidos a partir da aplicação do K-means usando $k = 3$ conforme foi discutido anteriormente. A seguir são apresentadas análises estatísticas e evolutivas que se tornaram possíveis a partir dos agrupamentos.

A Figura 2 ilustra três gráficos que observam características dos três grupos. Primeiramente, na Figura 2(a) pode-se observar a média das notas dos alunos de cada grupo dando, assim, origem às classes alto, médio e baixo desempenho. Foi realizado o teste t-student para cada um dos agrupamentos entre si, e foi constatado $p < 0,05$ em todas as comparações. Por outro lado, a Figura 2(b) mostra que as classes não contam com o mesmo número de concluintes. A classe de alunos com baixo desempenho é predominante e tem 44% dos concluintes, enquanto há 25% e 30% dos concluintes nas classes de médio e alto desempenho, respectivamente. A Figura 2(c) mostra a distribuição da nota geral em cada grupo, complementando a Figura 2(a). Uma das vantagens do K-means é que o método leva em consideração todas as características dos alunos para avaliar a similaridade, não apenas a nota geral. Por causa disso, a distribuição ilustrada na Figura 2(c) indica uma sobreposição das notas entre os grupos. Em outras palavras, no grupo de baixo desempenho com notas predominantemente menores, há uma minoria de alunos que tem notas para estarem em outros grupos (médio ou alto), mas as demais características os levam a estarem no grupo de baixo desempenho. Essa classificação é mais abrangente do que separar os alunos dado uma nota (ou duas notas) de corte.

A Figura 3 ilustra a evolução ao longo dos anos da proporção dos alunos em cada agrupamento uma vez que nossa investigação abrange concluintes das últimas quatro edições do Enade para computação. Apesar da maioria dos alunos concentrar-se no grupo de baixo desempenho, com esse gráfico observa-se que há uma tendência de redução desse grupo e, por outro lado, um aumento dos alunos de alto desempenho. Embora seja um dado positivo para o ensino de computação, os motivos que sustentam essa tendência são incertos e carecem de maiores investigações.

Outro fato interessante está na proporção de alunos em relação à tipo da escola de conclusão do Ensino Médio (EM) e a categoria administrativa da IES da graduação. Em relação ao EM, observa-se na Figura 4(a) que, dos alunos que concluíram o EM em

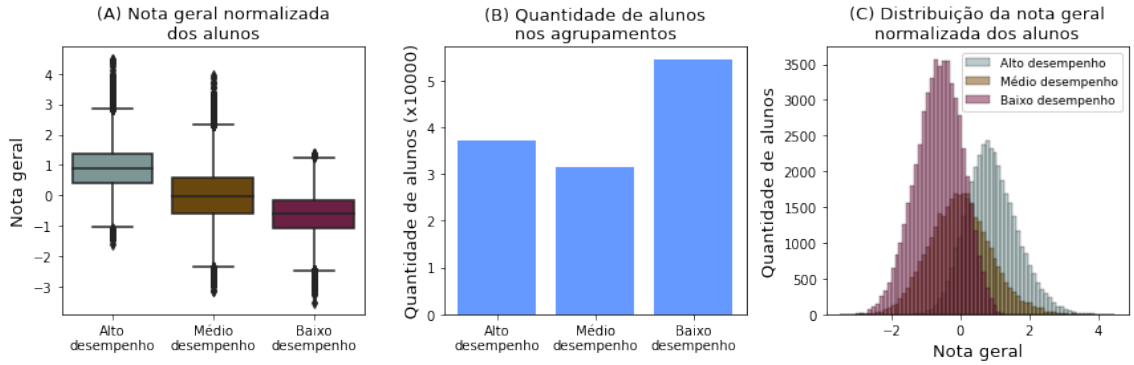

Figura 2. Características dos agrupamentos: (a) nota geral média por grupo; (b) quantidade de aluno por grupo; e (c) distribuição da nota geral.

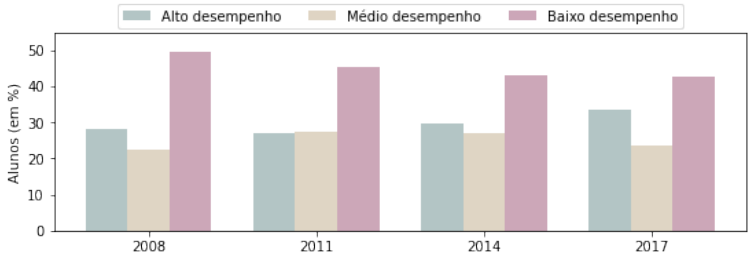

Figura 3. Evolução temporal dos agrupamentos ao longo dos anos.

escolas privadas, 43,15% estão no grupo de alto desempenho enquanto 22,03% e 34,80% estão no grupo de médio e baixo desempenho. Há uma clara diferença quando observa-se os concluintes de EM de escolas públicas, em que há uma predominância no grupo de baixo desempenho. Outros trabalhos também notaram essa diferença entre o desempenho dos alunos [Sampaio e Guimarães 2009, Esteves de Moraes e Belluzzo Júnior 2014]. Esse fato levanta a questão sobre a qualidade do ensino público quando comparado ao ensino privado.

Em relação à categoria administrativa, a partir da Figura 4(b), nota-se que é predominantemente maior os concluintes de alto desempenho em IES públicas enquanto aqueles de IES privadas estão predominantemente no grupo de baixo desempenho. A categoria administrativa da instituição também foi notada como relevante para o desempenho do aluno em [Fagundes et al. 2015]. Em [Bielschowsky 2020], observa-se também o problema da alta concentração de matrículas em IES privadas que mantém baixo desempenho no Enade em seus cursos.

A escolaridade dos pais dos concluintes também foi observada nos agrupamentos. Embora existam pequenas variações, as mesmas tendências são observadas tanto para a escolaridade do pai quanto da mãe do concluinte. Desta forma, os resultados ilustrados na Figura 5 são resumidos considerando tanto pais quanto mães. No gráfico, verifica-se que quanto maior é a escolaridade dos pais, maior é a proporção do grupo de alto desempenho. O inverso também é constatado, isto é, aumenta a proporção do grupo de baixo desempenho quanto menor é a escolaridade dos pais. Vários são os estudos que trabalham sobre a influência da escolaridade dos pais sobre os alunos [Vautero et al. 2017, Elisa Lovison e Glasenapp 2021]. Esses resultados mostram que,

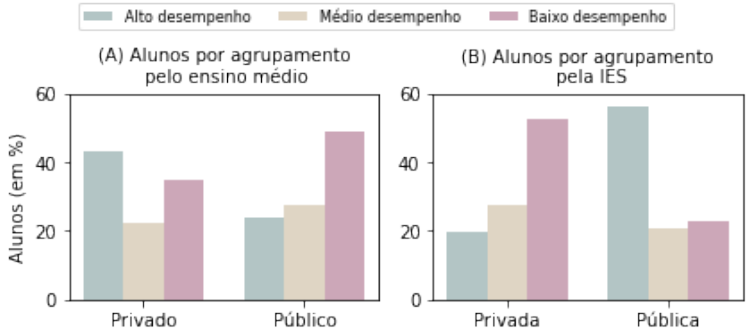

Figura 4. Distribuição dos concluintes a partir do (a) tipo de escola e (b) IES cursada.

quanto maior for a escolaridade, maior será o investimento dos pais na educação dos filhos, e isso implica diretamente no desempenho do aluno.

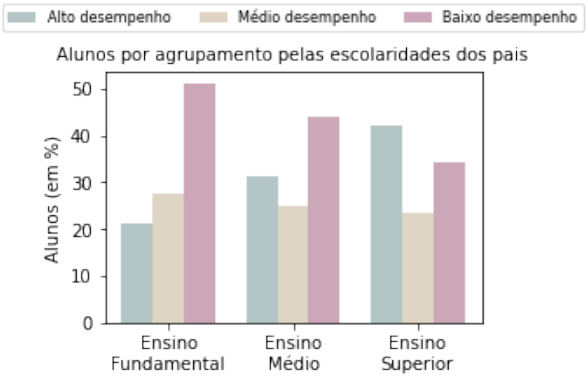

Figura 5. Comparação entre as escolaridades dos pais dos alunos.

Em relação às cotas, observa-se na Figura 6 que alunos não cotista são predominantes tanto da IES pública quanto na privada. Especificamente, na Figura 6(a), observa-se que a porcentagem de alunos que usufruíram de algum tipo de cota vem tendo um pequeno aumento no período analisado para IES públicas, mas para IES privadas, como mostra a Figura 6(b), permanece praticamente inalterada. Observa-se também positivamente que a proporção de alunos de alto desempenho (ambos cotistas na 6(a) e não cotistas na 6(c)) nas IES públicas tem crescido. Por outro lado, nas IES privadas, a distribuição dos concluintes entre os grupos permaneceu praticamente constante para os cotistas como visto na Figura 6(c). Para os não cotistas, na Figura 6(d), a proporção de alunos de baixo desempenho tem diminuído a cada ano. Estudos como o de [Cavalcanti et al. 2019] evidenciam a diferença entre o desempenho dos alunos não cotistas em relação aos alunos cotistas.

Os concluintes também foram observados em relação ao turno cursado. Observa-se que o período noturno é o que mais possui alunos (cerca de 73,66%) em relação ao período integral (cerca de 12,64%) e ao diurno (cerca de 13,68%). A Figura 7 mostra também que os cursos noturnos são historicamente predominantes. A Figura 7(a) mostra que há um pequeno aumento dos alunos do turno integral e este aumento é mais intenso no grupo dos alunos de alto desempenho. Em relação ao turno diurno, ilustrado na Fi-

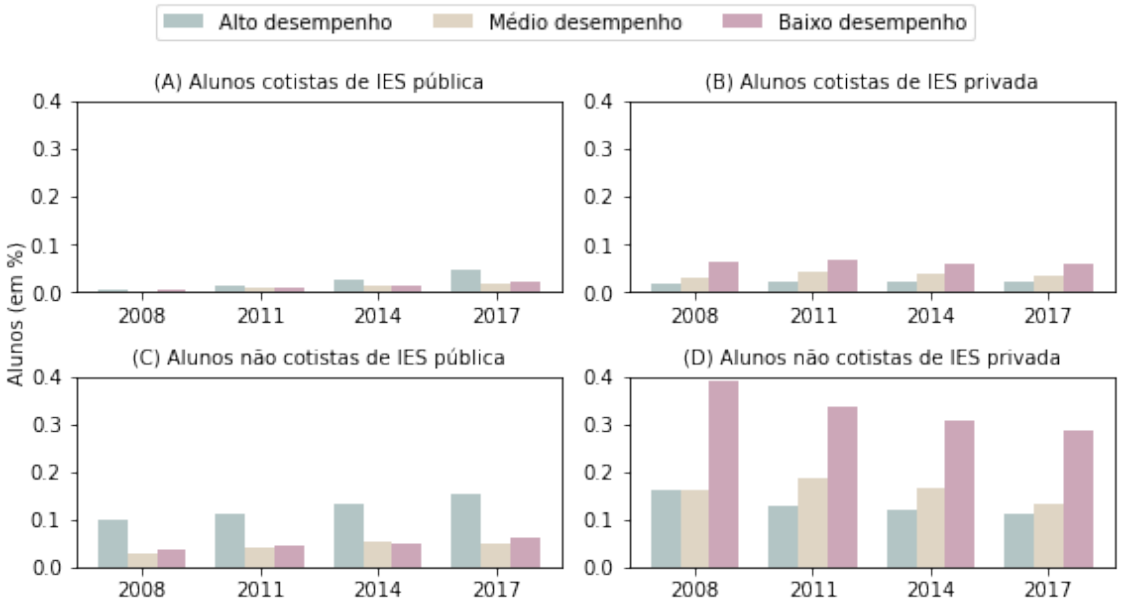

Figura 6. Evolução temporal dos (a) alunos cotistas matriculados em uma IES pública; (b) alunos cotistas matriculados em uma IES privada; (c) alunos não cotistas matriculados em uma IES pública; e (d) alunos não cotistas matriculados em uma IES privada.

gura 7(b), observa-se a predominância ao longo dos anos dos alunos de alto desempenho. Por fim, o turno noturno, apresentado pela Figura 7(c), tem tido uma diminuição gradual da quantidade de alunos com baixo desempenho, mesmo embora essa classe tenha sido predominante nos anos analisados. Outros estudos em cursos de graduação também têm constatado a defasagem do rendimento dos alunos dos cursos noturnos em relação aos demais turnos [Farias et al. 2015]. Isso se deve ao fato de que normalmente os alunos escolhem cursos noturnos para conciliar com o trabalho que ocorre nos outros períodos do dia. Existem vários estudos em outras áreas do conhecimento, em que se observa a influência do trabalho sobre o estudante [Rocha et al. 2018].

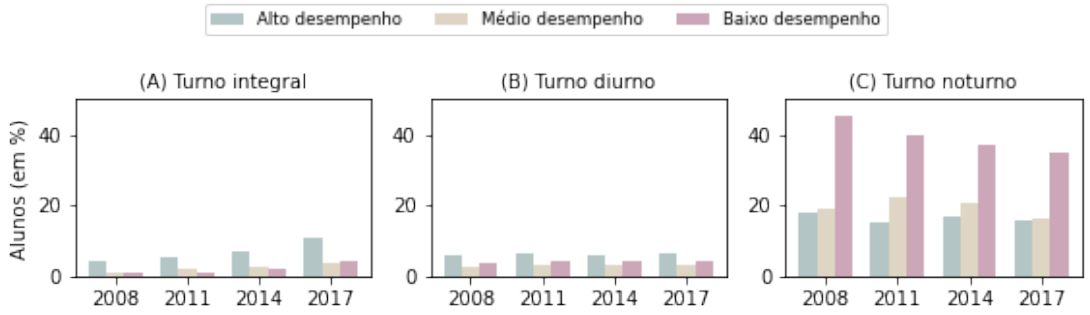

Figura 7. Evolução temporal da quantidade de alunos por agrupamento dado os turnos: (a) integral, (b) diurno e (c) noturno.

Na Figura 8, exibe-se a distribuição dos alunos ao longo dos anos em cada um dos agrupamentos considerando seis cursos de computação. Omitiu-se o curso GTI, pois ele possui concluintes apenas em 2017. Por outro lado, o curso LCC foi mantido mesmo não

tendo participantes em 2008.

Em termos gerais, observa-se que os cursos de tecnologia (ADS e RC) e o bacharelado (SI) têm uma predominância dos seus concluintes no agrupamento de baixo desempenho. Por outro lado, há uma menor discrepância entre o tamanho dos agrupamentos (alto e baixo) nos cursos EC, BCC e LCC. Diferente dos demais, destaca-se que o agrupamento de alto desempenho no curso EC sempre foi predominante ao longo dos anos. Para os cursos ADS, BCC, EC, SI, observa-se positivamente uma tendência de queda da quantidade de alunos de baixo desempenho. Já o agrupamento de baixo desempenho mantém-se praticamente constante para os cursos RC e LCC. Para todos os cursos, não observa-se nenhuma tendência nos agrupamentos de alto e médio desempenho.

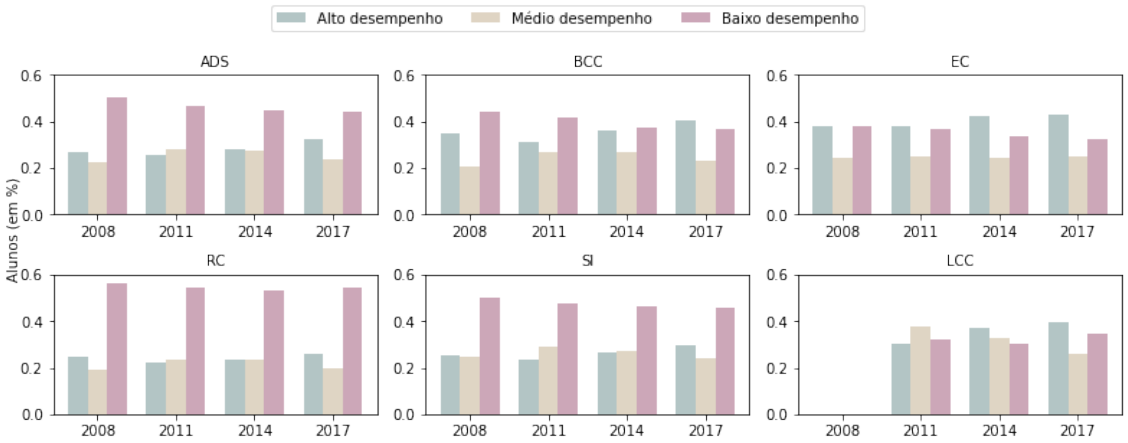

Figura 8. Evolução temporal da quantidade de alunos por agrupamento nos cursos de graduação analisados.

5. Conclusões

Este trabalho propõe a aplicação de um método de mineração sobre os microdados dos concluintes dos cursos da área de computação das edições de 2008, 2011, 2014 e 2017 do Enade. Mais especificamente, o algoritmo de agrupamento K-means foi utilizado para classificar os estudantes em três grupos nos quais a nota geral no exame foi observada para caracterizá-los em alto, médio e baixo desempenho. A partir dos grupos, observou-se o desempenho dos concluintes no exame sob diferentes perspectivas, como o tipo da escola e da IES cursada, a escolaridade dos pais, o ingresso na IES por cotas ou não, o turno e o curso graduação realizado.

Como trabalho futuro, espera-se ampliar a abrangência desse estudo para contemplar outras áreas do conhecimento e, também, avaliar a aplicação de outras técnicas de agrupamento. Além disso, um estudo envolvendo algoritmos de classificação está em andamento a fim de predizer o desempenho dos estudantes a partir do seu perfil socioeconômico.

Agradecimentos

Os autores agradecem a Universidade Tecnológica Federal do Paraná – Campo Mourão pelo apoio financeiro (Edital DIRPPG-CM nº 01/2022) e a Fundação Araucária (Edital PROPPG nº02/2021).

Referências

Baker, R., Isotani, S., e Carvalho, A. (2011). Mineração de dados educacionais: Oportunidades para o brasil. *Revista Brasileira de Informática na Educação*, 19(02):03.

Bielschowsky, C. E. (2020). Tendências de precarização do ensino superior privado no brasil. *Revista Brasileira de Política e Administração da Educação*, 36(1):241–271.

BRASIL (2004). Lei nº 10.861, de 14 de abril de 2004. `http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2004/lei/l10.861.htm` [Acessado em 26-mar-2022].

BRASIL (2022). Instituto nacional de estudos e pesquisas educacionais anísio teixeira. `https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enade` [Acessado em 25-mar-2022].

Capelari, L. O. O. e Schwerz, A. L. (2021). O perfil socioeconômico dos concluintes de computação do sul do brasil. In *Computer on the Beach*, COTB'21, pages 133–140, Itajaí, SC, Brasil. Universidade do Vale do Itajaí.

Cavalcanti, I. T. d. N., Andrade, C. S. M., Tiryaki, G. F., e Costa, L. C. C. (2019). Desempenho acadêmico e o sistema de cotas no ensino superior: evidência empírica com dados da universidade federal da bahia. *Avaliação: Revista da Avaliação da Educação Superior*, 24(1).

Costa, E., Baker, R. S., Amorim, L., Magalhães, J., e Marinho, T. (2012). Mineração de dados educacionais: Conceitos, técnicas, ferramentas e aplicações. In *Anais da VIII Jornada de Atualização em Informática na Educação*, JAIE'12, Porto Alegre, RS, Brasil. SBC.

Cunha, R., Sales, C., e Santos, R. (2021). Análise automática com os microdados do enade para melhoria do ensino dos cursos de ciência da computação. In *Anais do XXIX Workshop sobre Educação em Computação*, WEI'21, pages 208–217, Porto Alegre, RS, Brasil. SBC.

Daly, A., Dekker, T., e Hess, S. (2016). Dummy coding vs effects coding for categorical variables: Clarifications and extensions. *Journal of Choice Modelling*, 21:36–41.

de Vasconcelos, L. G. e de Castro, M. M. B. (2020). Uso de técnicas de aprendizagem de máquina para análise dos fatores de sucesso de cursos do ensino superior no enade. In *Congresso Nacional de Educação*, CONEDU'20, pages 1868–1887, Campina Grande, PB, Brasil. Realize Editora.

Elisa Lovison, T. e Glasenapp, D. (2021). A escolaridade dos pais e a alfabetização dos estudantes. *Monumenta - Revista de Estudos Interdisciplinares*, 1(2):30–53.

Esteves de Moraes, A. G. e Belluzzo Júnior, W. (2014). O diferencial de desempenho escolar entre escolas públicas e privadas no brasil. *Nova Economia*, 24(2).

Fagundes, C., Luce, M., e Espinar, S. (2015). O desempenho acadêmico como indicador de qualidade da transicao ensino médio-educação superior. *Ensaio: Avaliação e Políticas Públicas em Educação*, 22(84):635–670.

Farias, M. R. S., Alves, F. d. S., e Farias, K. T. R. (2015). Desempenho acadêmico em métodos quantitativos nos cursos de ciências contábeis. *Enfoque: Reflexão Contábil*, 34(2):37–50.

Francelino, W. L. e Machado, L. S. (2020). Mineração de dados nos microdados enade computação. Relatório técnico, Faculdade Ânima Educação, Tubarão, SC, Brasil.

INEP (2022). Microdados do enade. https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enade.

Lima, A. M. S., Florez, A. Y. C., Lescano, A. I. A., de Oliveira Novaes, J. V., de Fatima Martins, N., Junior, C. T., de Sousa, E. P. M., Junior, J. F. R., e Cordeiro, R. L. F. (2020). Analysis of enem's attendants between 2012 and 2017 using a clustering approach. *Journal of Information and Data Management*, 11(2):115–130.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297. University of California Press.

Rezende, C. C. d. S., Cantarino, L. A. B., de Souza, P. F., Alves, T. O. M., e Campos, R. S. (2022). O impacto de aspectos socioeconômicos no desempenho de estudantes de sistemas de informação no enade. *Revista Brasileira de Informática na Educação*, 30:157–181.

Rocha, A. L. d. P., Leles, C. R., e Queiroz, M. G. (2018). Fatores associados ao desempenho académico de estudantes de nutrição no enade. *Revista Brasileira de Estudos Pedagógicos*, 99:74 – 94.

Rodrigues, E. e Gouveia, R. (2021). Técnicas de machine learning para predição do tempo de permanência na graduação no Âmbito do ensino superior público brasileiro. In *Anais do VI Congresso sobre Tecnologias na Educação*, Ctrl+e'21, pages 128–137, Porto Alegre, RS, Brasil. SBC.

Sampaio, B. e Guimarães, J. (2009). Diferenças de eficiência entre ensino público e privado no brasil. *Economia Aplicada*, 13(1):45–68.

Vautero, J., Silva, A., Marques, C., e Taveira, M. (2017). A influência da escolaridade dos pais no prestígio do curso universitário escolhido pelos filhos. *Psicologia: Revista da Associação Portuguesa Psicologia*, 31(2):155–158.

Vista, N. P. B., Barasuol, J. B., Figueiró, M. F., Chicon, P. M. M., e Ansuj, A. P. (2018). Análise de agrupamento hierárquico aplicada aos microdados do enade do curso de graduação em ciência da computação. *Revista Eletrônica Argentina-Brasil de Tecnologias da Informação e da Comunicação*, 1(8).

Webber, C. G., Zat, D., e Lima, M. F. W. P. (2013). Utilização de algoritmos de agrupamento na mineração de dados educacionais. *RENOTE*, 11(1).

Wu, J. (2012). *Advances in K-Means Clustering: A Data Mining Thinking*. Springer, Nelson Hall 3350.

Xu, D. e Tian, Y. (2015). A comprehensive survey of clustering algorithms. In *Annals of Data Science*, pages 165–193.