Mapeamento de Desempenho de Concluintes de Computação no Nordeste: Uma Abordagem de Clusterização

Caio C. F. Silva[1], Tarcísio B. Costa[2], Victor H. S. Oliveira[1]

[1]Departamento de Estatística e Informática, [2]Departamento de Computação
Universidade Federal Rural de Pernambuco (UFRPE) – Recife – PE – Brasil

{caio.fsilva, tarcisio.bcosta, victor.henriqued}@ufrpe.br

Abstract. *Educational Data Mining is a field that has historically developed solutions for various challenges in education. However, the effective implementation of these solutions often requires adaptation to different regional and institutional realities. In this context, this study aims to carry out a socioeconomic clustering approach using public data made available by INEP, focusing on the Northeast region of Brazil. Based on the KDD methodology, the K-Means clustering algorithm was applied, with the optimal number of clusters determined using the Elbow Method, in order to identify latent patterns in the data. The results indicate the presence of three predominant profiles among graduates, with significant distinctions in terms of overall score, family income, educational background, and access to support policies. The findings contribute to a more contextualized understanding of the educational landscape in the Northeast and offer valuable insights for the formulation of targeted public policies, as well as for future studies in higher education.*

Resumo. *A Mineração de Dados Educacionais é uma área que, historicamente, desenvolve soluções para diversos problemas presentes na educação. No entanto, o pleno desenvolvimento dessas soluções exige adaptações às diferentes realidades regionais e institucionais. Nesse sentido, este trabalho tem como objetivo realizar uma abordagem baseada em agrupamento socioeconômico, utilizando dados públicos disponibilizados pelo INEP referentes à região Nordeste do Brasil. A partir da metodologia KDD, empregou-se o algoritmo de clusterização K-Means, com definição do número ideal de agrupamentos por meio do Método do Cotovelo, visando identificar padrões latentes nos dados. Os resultados indicam a existência de três perfis predominantes entre os concluintes, com distinções relevantes quanto à nota geral, renda, escolaridade e acesso a políticas de apoio. Os achados contribuem para uma compreensão mais contextualizada da realidade educacional nordestina, oferecendo subsídios à formulação de políticas públicas mais direcionadas, bem como a futuros estudos voltados ao ensino superior.*

1. Introdução

Ao observar o ecossistema educacional brasileiro, múltiplos obstáculos são identificados no caminho para a universalização do ensino e a valorização/respeito aos direitos dos discentes e docentes no país, como o analfabetismo, as desigualdades socioculturais e as diferentes condições de docentes e infraestrutura. Devido a estes e outros fatores, a educação brasileira, principalmente no ensino superior, enfrenta altas taxas de evasão (Barroso, 2022) e reduções no crescimento de matrículas (Neves, 2012).

Dentre as estratégias existentes para o combate aos desafios enfrentados pelas Instituições de Ensino Superior (IES) brasileiras, a Mineração de Dados Educacionais surge como um conjunto de ferramentas tecnológicas para compreender e melhorar o

processo de aprendizagem através da análise e transformação de dados em informação (Romero, 2020; dos Santos, 2021). Abrangendo várias soluções para a área, a EDM demonstra ser um método eficiente para a construção de um ciclo dinâmico e contínuo de feedback entre ferramenta e gestor acadêmico (Silva, 2014), facilitando o desenvolvimento de políticas para melhorar índices acadêmicos e combater a evasão estudantil (Marques, 2023).

Entretanto, também existem desafios para a implementação de soluções de EDM na educação brasileira, sejam elas de cunho estrutural como a falta de independência estatística dos dados (Baker, 2011), ou a necessidade de adaptar as soluções para o contexto da área ou da instituição (Zapparolli, 2017). Dentre outros, estes obstáculos dificultam a utilização simples e prática de EDM com os dados coletados, exigindo etapas extras para o desenvolvimento pleno de soluções tecnológicas.

Dado este contexto, este trabalho busca realizar uma análise socioeconômica dos perfis estudantis, especificamente da região Nordeste, através de um agrupamento de vários dados públicos disponibilizados pelo INEP. Espera-se, através do processo KDD (*Knowledge Discovery in Databases*), obter observações e *insights* da região, como questões de infraestrutura através do censo e notas como conceito ENADE, a fim de contextualizar a situação socioeconômica da região e fornecer estas informações à comunidade acadêmica para a elaboração de futuras políticas e sistemas, além de fornecer a metodologia para replicar o estudo em outras regiões.

2. Trabalhos Relacionados

Esta seção apresenta os trabalhos relacionados ao estudo relacionado. Estes trabalhos tratam da aplicação de técnicas de mineração de dados sobre bases educacionais com o propósito de descobrir conhecimentos úteis que contribuam para a melhoria da qualidade do ensino. Para isso, consideram-se variáveis relacionadas ao perfil socioeconômico dos estudantes, bem como dados institucionais, empregando-se abordagens de aprendizado de máquina supervisionado e não supervisionado.

No estudo de Lima et al. (2020), aplicou-se o algoritmo de agrupamento K-means aos microdados do ENEM no período de 2012 a 2017, com o objetivo de identificar padrões de desempenho entre os participantes a partir das notas obtidas em cada área do conhecimento e na redação. O trabalho foi conduzido com rigoroso pré-processamento para a padronização dos dados e prevenção de viés estatístico. Como resultado, foram formados três grupos — baixo, médio e alto desempenho — e observou-se a predominância de estudantes de escolas públicas nos grupos de menor desempenho, enquanto os de maior desempenho apresentaram participação mais equilibrada entre os tipos de escola.

De forma semelhante, Figueiró et al. (2018) aplicaram análise de agrupamento por clusterização hierárquica aos microdados de 2014 para o curso de Ciência da Computação, formando grupos das IES do Consórcio das Universidades Comunitárias Gaúchas. Considerando as notas brutas de formação geral, conhecimento específico e geral, foram identificados quatro clusters que evidenciaram diferenças de desempenho entre as universidades comunitárias, fornecendo subsídios para estratégias de melhoria da qualidade do ensino.

Silva, Hoed e Saraiva (2019) utilizou uma técnica não supervisionada de regras de associação aplicada aos microdados do Enade 2017 para cursos de Computação, a fim de descobrir fatores que se relacionam ao desempenho dos concluintes. As regras geradas evidenciaram influência conjunta de variáveis socioeconômicas e institucionais, oferecendo indícios de que desigualdades de origem e contexto acadêmico repercutem nas notas dos estudantes.

A literatura indica que a análise dos microdados apresenta caráter multifatorial, de modo que vários autores investigaram determinantes de desempenho sob perspectivas diversas. Observa-se contudo, que os estudos da área de computação concentram-se, em geral, em conjuntos restritos de instituições. Carece-se de investigações que contemplem de maneira sistemática outras regiões do país. Nesse contexto, é proposto o uso das bases disponíveis para todos os cursos da área de computação da Região Nordeste, contemplando os nove estados que a compõem.

3. Métodos

Neste trabalho foi empregado o método Knowledge Discovery in Databases – KDD [Frawley et al. 1992] para conduzir o estudo quantitativo dos microdados do ciclo de computação. Esse modelo organiza a extração de conhecimento não-trivial e potencialmente valioso em grandes bases de dados, articulando etapas de preparação das informações, aplicação de técnicas analíticas para mineração dos dados e por fim, a avaliação dos resultados obtidos.

3.1 Pré-processamento

O pré-processamento de dados é responsável por efetuar a limpeza – corrigindo inconsistências e removendo redundâncias – para preparação dos dados. Como resultado, gerou-se um conjunto de dados com características dos cursos superiores da área de computação no Brasil na região Nordeste, processadas a partir dos microdados do ENADE[1].

3.1.1. Seleção dos Dados

Na etapa de seleção dos dados, levou-se em consideração a reformulação introduzida pelo INEP em 2022 para atender à Lei Geral de Proteção de Dados. Em razão disso, as respostas do questionário socioeconômico e demais informações passaram a ser distribuídas e aleatorizadas em arquivos distintos. Desse modo, tornou-se inviável rastrear o conjunto completo de respostas de um mesmo concluinte, sendo possível inferir apenas características agregadas no nível do curso de graduação[2].

Para abranger três ciclos completos do Enade — cujas avaliações de cursos de tecnologia ocorrem em intervalos trienais — selecionaram-se, no pré-processamento, os microdados das edições de 2014, 2017 e 2021, assegurando a amplitude temporal necessária à análise. Ressalte-se ainda que a edição de 2020 foi realizada apenas em 2021 em razão da pandemia COVID-19, e que o questionário socioeconômico sofreu

[1] Microdados - ENADE
[2] Adequação à LGPD dos microdados do ENEM/ENADE

alterações ao longo dos anos; as bases mais antigas apresentaram discrepâncias significativas em comparação às edições recentes e, por esse motivo, foram descartadas. Nas edições mantidas dos últimos três triênios, identificou-se um conjunto de variáveis comuns que possibilitaram a comparação entre ciclos, representadas na Tabela 1 com suas respectivas categorias.

Atributo	Categoria
Idade; sexo; ano de conclusão do ensino médio; ano de início da graduação; estado civil; raça; renda total da família; situação financeira; situação de trabalho; bolsa de estudo ou financiamento; bolsa acadêmica; auxílio permanência; escolaridade do pai; escolaridade da mãe; políticas de ação afirmativa; tipo de escola do ensino médio; família com curso superior	Socioeconômico
Código do curso; código da IES; categoria administrativa; organização acadêmica; área de enquadramento; modalidade de ensino; região; município; turno da graduação	Curso
Nota geral; tipo de presença.	Prova

Tabela 1. Atributos selecionados do ENADE

3.1.2. Limpeza dos Dados

A limpeza dos microdados iniciou-se pela identificação dos cursos efetivamente pertencentes à área de Computação. Para isso, recorreu-se ao código oficial de enquadramento utilizado pelo Enade, selecionando as sete categorias que abrangem bacharelados, licenciaturas e tecnológicos (Tabela 2).

Código	Área de Enquadramento	Grau
72	Análise e Desenvolvimento de Sistemas	Tecnológo
79	Redes de Computadores	Tecnológo
4004	Ciência da Computação	Bacharelado
4005	Ciência da Computação (Licenciatura)	Licenciatura
4006	Sistemas de Informação	Bacharelado
6409	Gestão da Tecnologia da Informação	Tecnológo
5809	Engenharia de Computação	Bacharelado

Tabela 2. Áreas de Enquadramento do ENADE para cursos de computação

Essa escolha evita ambiguidade na denominação dos cursos e independe de variações textuais nos arquivos. Vale registrar que algumas categorias não aparecem em todas as edições analisadas: Gestão da TI surge apenas a partir de 2017, enquanto Engenharia de Computação deixa de ser listada em 2021, fato refletido nos totais anuais da Tabela 1. Na sequência, foi conduzida uma análise exploratória dos dados, com o objetivo de identificar padrões e orientar a formulação de questões de pesquisa pertinentes ao estudo.

Para assegurar a integridade da análise, foi necessário realizar o alinhamento e a ordenação das informações entre os diferentes arquivos disponibilizados, uma vez que apresentavam variações na estrutura e no posicionamento das variáveis. Esse processo teve como base o código dos cursos, permitindo a consolidação das respostas associadas aos concluintes, ainda que sem um mapeamento individual completo. Tal estratégia possibilitou uma correspondência mais coerente entre os dados, favorecendo uma análise mais precisa e representativa. Além disso, adotou-se um critério de filtragem que considerou apenas os concluintes válidos — ou seja, aqueles cuja participação na prova e no questionário foi validada pelo INEP —, minimizando o risco de enviesamento nas interpretações estatísticas.

Ano	Antes do Pré-Processamento			Após o Pré-processamento		
	Concluintes	**Concluintes de TI**	**Atributos**	**Concluintes de TI**	**Concluintes Válidos**	**Atributos**
2014	481.718	51.774	150	8.285	6.669	30
2017	537.358	42.533	148	7.129	5.675	30
2021	489.866	55.277	159	7.564	5.651	30

Tabela 3. Comparativo dos microdados durante o processamento

Após a aplicação de etapas de saneamento, os dados foram significativamente reduzidos, refletindo um recorte mais específico e coerente com os objetivos da pesquisa. Conforme apresentado na Tabela 3, o número total de concluintes inicialmente disponíveis em cada edição do Enade sofreu sucessivas reduções: da população geral para os concluintes dos cursos de TI, e, em seguida, para os concluintes válidos da região Nordeste, com presença e notas reconhecidas pelo INEP. Paralelamente, o número de atributos também foi uniformizado entre os anos, resultando em um conjunto final composto por 30 variáveis comuns e compatíveis entre os três ciclos analisados.

3.1.3. Transformação

O processo de transformação dos dados foi realizado visando à padronização e à viabilidade analítica das informações. Inicialmente, procedeu-se à filtragem de registros válidos, de modo a excluir entradas com informações incompletas ou inconsistentes que pudessem comprometer a qualidade da análise. Em seguida, algumas colunas foram

padronizadas e uniformizadas quanto à nomenclatura e ao formato, garantindo maior consistência estrutural.

As variáveis categóricas foram mapeadas por meio de dicionários específicos, construídos manualmente para converter as alternativas em valores textuais padronizados. Para lidar com a alta granularidade de determinadas questões — especialmente aquelas relacionadas a tipos de bolsas, auxílios e escolaridade —, realizou-se uma redução da granularidade, sem comprometer o significado analítico, por meio do agrupamento de respostas similares em categorias mais representativas.

Adicionalmente, os valores ausentes foram tratados com base na frequência relativa de cada categoria em cada ano, assegurando coerência estatística na imputação. Posteriormente, as variáveis categóricas foram transformadas por meio da técnica de One-Hot Encoding, empregada para converter variáveis nominais em uma representação binária adequada aos algoritmos de aprendizado não supervisionado. Essa etapa resultou na expansão do número de atributos, passando de aproximadamente 28–30 colunas para cerca de 110 variáveis após a codificação.

No que diz respeito aos atributos numéricos, foi aplicada a padronização utilizando o método de Z-score, especificamente sobre as variáveis "idade" e "nota geral" dos participantes. Essa normalização visou equilibrar a influência desses campos nos algoritmos de agrupamento, prevenindo distorções provocadas por escalas discrepantes.

Ao final do processo de transformação, foram gerados três conjuntos de dados distintos, com os três anos unificados em cada um deles: (i) um dataset com os dados mapeados (ENADE mapeado); (ii) um dataset com as variáveis categóricas já convertidas via One-Hot Encoding; e (iii) um dataset com os dados categóricos convertidos e os atributos numéricos normalizados via Z-score. Este último foi utilizado como base para a aplicação do algoritmo de clusterização e para a execução do Método do Cotovelo, abordado em maior detalhe na seção seguinte.

4. Metodologia

Nesta seção, serão descritas as técnicas e ferramentas específicas utilizadas ao longo do trabalho durante as diferentes etapas de processamento e análise, tais como as particularidades na abordagem e implementação das mesmas dado o contexto do trabalho.

4.1. K-Means

O *K-means* é uma técnica de agrupamento de dados, que usa como método a partição, a fim de dividir o conjunto em k grupos de objetos semelhantes entre si, onde k é um número pré-definido (MURPHY, 2012). Isso é realizado para que a cada possível atualização, o centróide seja atualizado para refinação da qualidade dos grupos.

Nesse aspecto, o algoritmo consiste em um método de agrupamento pelo qual é utilizado a partição de dados não rotulados, isto é, realiza-se a busca e reconhecimento de similaridades entre dados a fim de agrupá-los conforme a definição de um argumento

k, sendo esse o número de clusters. Para tal abordagem, o método faz uso da distância euclidiana para obter a distância entre dois pontos no espaço, que é calculada pela soma da raiz quadrada da diferença das coordenadas desses pontos, de forma que os grupos sejam homogêneos, isto é, fazer com que os dados existentes dentro de cada cluster sejam mais parecidos entre si. Isso é efetuado a partir do conceito de centróide, em que é definido um ponto central de cada grupo com base em medidas estatísticas, como a média, a medida que a cada iteração o ponto central é atualizado até que um critério de parada seja atingido.

O uso dessa abordagem permite que sejam identificados padrões de forma que se torne possível realizar a segmentação e a compreensão das estruturas nos conjuntos de dados. Para os experimentos realizados neste trabalho, os alunos são representados através de características contidas nos microdados. Alunos considerados similares são aqueles no qual possuem características que se assemelham ao espaço dimensional listado na Tabela 1.

4.2. Método do Cotovelo

O algoritmo de aprendizado *K-Means* requer a definição do número de possíveis grupos de classificação, conhecidos como clusters. Abordagens heurísticas podem ser utilizadas para escolher a quantidade de agrupamentos, como o Método do Cotovelo (*Elbow Method)*. Esse método consiste em calcular as distâncias dos pontos em relação aos centróides, para diferentes valores de k, e representar os resultados em um gráfico, no qual a curvatura significativa na curva indica o número ideal de clusters para o conjunto de dados em análise (BAILEY, 1975).

Para determinar esse número, existem duas métricas comumente utilizadas: a distorção (*within-cluster sum of squares*) e a inércia (*between-cluster sum of squares*). A distorção é calculada pela soma das distâncias quadráticas dos pontos aos seus respectivos centróides dentro de cada cluster (KING, 2000). Em outras palavras, ela mede o quão compactos estão os clusters e quanto os pontos estão próximos dos centróides – por meio da distância Euclidiana. Já a inércia é obtida somando as distâncias quadráticas entre os centróides dos clusters e o centróide global dos dados, ponderadas pela quantidade de pontos em cada cluster. Dessa forma, a inércia mede o quão separados estão os clusters entre si.

Após o cálculo utilizando as métricas de distorção ou inércia, é elaborado um gráfico para diferentes valores de k, onde é possível observar uma curva que se assemelha ao formato de um cotovelo. O ponto de cotovelo representa o número ideal de clusters, pois é nesse ponto que ocorre um equilíbrio entre a redução da distorção e a manutenção da separação entre os clusters (ALDEFENDER et al, 1984). Em outras palavras, k determina o valor limite em que a composição de outros clusters não contribui significativamente para a melhoria do modelo de agrupamento.

A Figura 1 apresenta os resultados da aplicação do método para cada uma das edições do Enade. Observa-se que, em todos os anos, há uma inflexão acentuada nas curvas entre $k = 2$ e $k = 3$, indicando que os ganhos na redução da soma dos erros quadráticos deixam de ser expressivos a partir desse ponto. Essa característica sugere que a escolha de dois ou três agrupamentos tende a representar uma divisão mais

eficiente dos dados, evitando segmentações excessivas que pouco agregam em termos de explicabilidade[3].

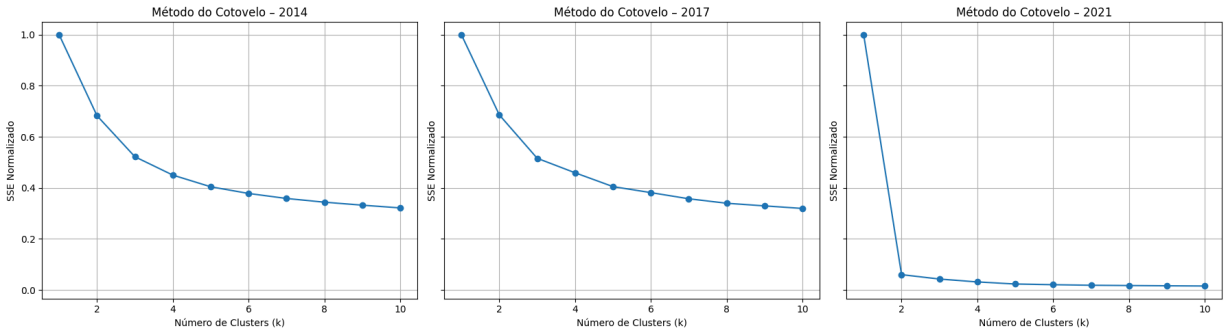

Figura 1. Método do Cotovelo aplicado ao ENADE 2014, 2017 e 2021

A Figura 2 complementa a análise apresentada na Figura 1 ao trazer uma visualização comparativa da soma dos erros quadráticos normalizada para cada valor de k entre os anos de 2014, 2017 e 2021. Percebe-se que, apesar de pequenas variações entre os anos, o comportamento das curvas é bastante semelhante, reforçando a tendência de estabilização da SSE a partir de $k = 3$.

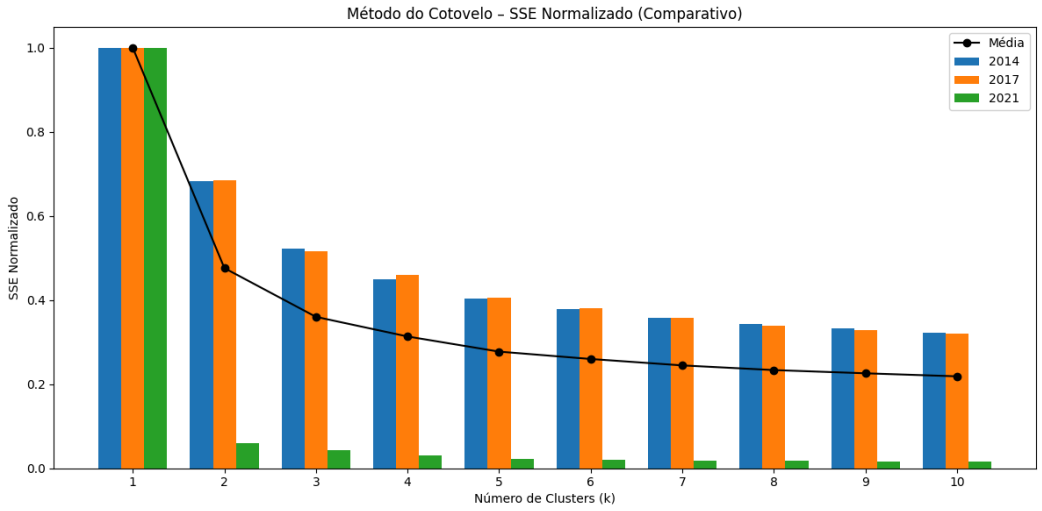

Figura 2. Comparativo do Método do Cotovelo para cada ano

5. Resultados e Discussões

Nesta seção apresenta-se os resultados provenientes da aplicação do K-Means utilizando $k = 3$ conforme foi encontrado anteriormente. Estes agrupamentos passaram por diferentes análises baseadas em outros fatores presentes nos dados, através da geração de gráficos que serão apresentados a seguir. Acompanhando os gráficos, foram realizadas observações e análises sobre as informações obtidas.

A Figura 3 ilustra uma visão geral da situação estudantil da área, composta por mais de 15 mil estudantes. Foram obtidos três perfis claros de desempenho, bem balanceados em número, e com notas globalmente diferenciadas. O panorama organiza

[3] GitHub

os grupos de discentes por desempenho, nota e ao passar dos anos. Observa-se que os grupos estão bem alinhados à nota geral como métrica (obtida ao comparar com o desvio padrão, onde o quanto mais próximo de 4, maior é a nota do estudante), indicando que as classes do modelo estão, num geral, bem representadas, com notas globalmente diferenciadas por grupo e onde o grupo majoritário é o de estudantes com baixo desempenho acadêmico. Apesar disso, ao analisar a distribuição de estudantes ao longo dos anos, pode-se notar uma tendência de redução nos estudantes de baixo rendimento e aumentos nos grupos de alto e médio rendimento. Essa inversão sugere efeitos positivos das políticas públicas de melhoria educacional implementadas na última década.

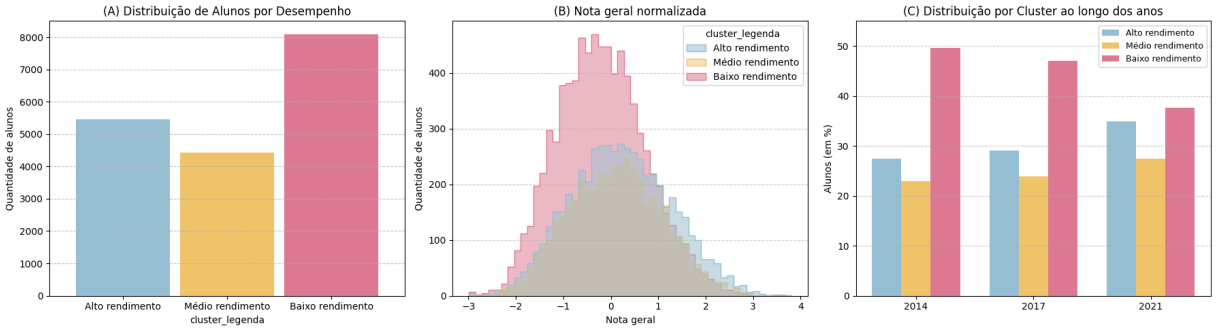

Figura 3. Visão geral dos agrupamentos: (a) quantidade de alunos por cluster; (b) distribuição da nota geral normalizada; e (c) divisão de alunos por grupo/ano.

Realizando uma observação mais aprofundada ao separar os grupos de aluno por curso/área, encontramos mais informações úteis para o direcionamento de políticas educacionais, como a maior concentração de discentes de baixo desempenho acadêmico nos cursos de Sistemas de Informação (BSI) e Análise e Desenvolvimento de Sistemas (ADS) – evidenciados na Figura 4 – enquanto o curso de Ciências da Computação (CC) apresenta uma quantidade maior de discentes de alto desempenho, subindo a média da área de TI num geral, mas também abriga expressivos contingentes de alunos de médio e baixo desempenho. Os demais cursos têm participação numérica mais modesta, Licenciatura em Computação, por sua vez, exibe a menor proporção de discentes de baixo rendimento, mas conta com um universo amostral reduzido.

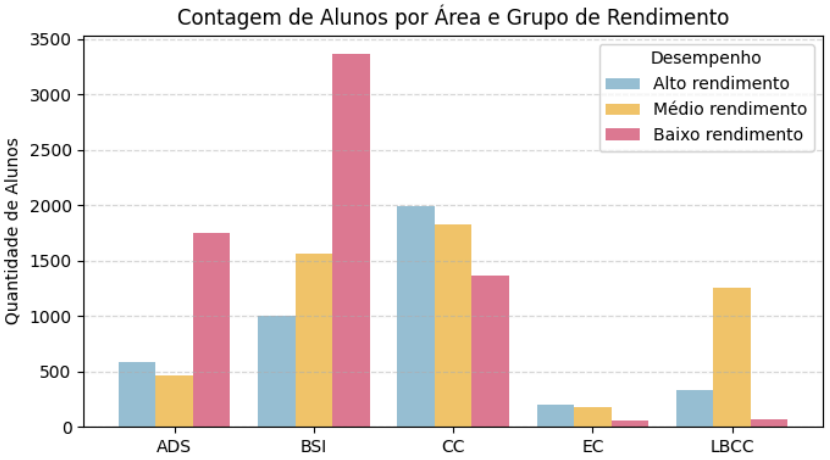

Figura 4. Comparação entre as áreas de enquadramento dos alunos.

De forma semelhante, a Figura 5 evidencia disparidades regionais ao retratar a distribuição dos perfis geograficamente por estados do Nordeste. Piauí, Rio Grande do Norte e Paraíba lideram em proporção de alunos de alto rendimento (54%, 50% e 42%, respectivamente), enquanto Bahia, Ceará e Pernambuco concentram as maiores fatias de baixo rendimento – 56%, 53% e 51%. Chama atenção que esses três últimos estados também respondem pelo maior contingente absoluto de estudantes na amostra, o que sugere um possível efeito cascata: turmas maiores podem dificultar o acompanhamento individualizado, reduzindo o desempenho médio. Esse padrão atrelado a estudos mais aprofundados pode apontar para desigualdades estruturais e reforça a necessidade de políticas focalizadas. Investigar variáveis como tamanho de turma, gasto educacional por aluno, aspectos da infraestrutura das IES pode ajudar a confirmar essa hipótese e direcionar uma conclusão mais fundamentada.

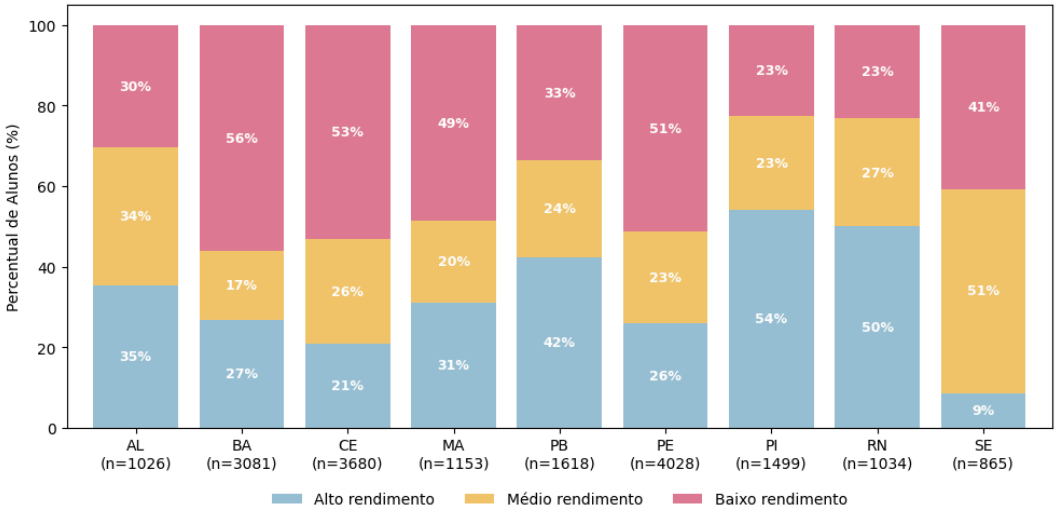

Figura 5. Distribuição dos agrupamentos entre os Estados do Nordeste.

Também evidenciado pela Figura 6, constatamos que a evolução dos três grupos ao longo do tempo corrobora com a visão geral de que há uma queda na quantidade de estudantes de baixo rendimento e um aumento na população dos outros grupos, tendência a qual é observada quase que consistentemente em todas as faixas de renda familiar.

Vale constatar que, apesar de terem a maior concentração de estudantes do grupo de baixo rendimento, aqueles de menor renda familiar (considerados de classe baixa) também apresentam as maiores quedas de quantidade de estudantes nesse grupo e o aumento no desempenho médio e alto, indicando que caso a redução esteja relacionada a políticas públicas anteriormente implementadas, elas estão beneficiando principalmente os grupos de maior necessidade econômica. Enquanto isso, os recortes com maiores rendas familiares tendem à distribuição uniforme de todos os três grupos.

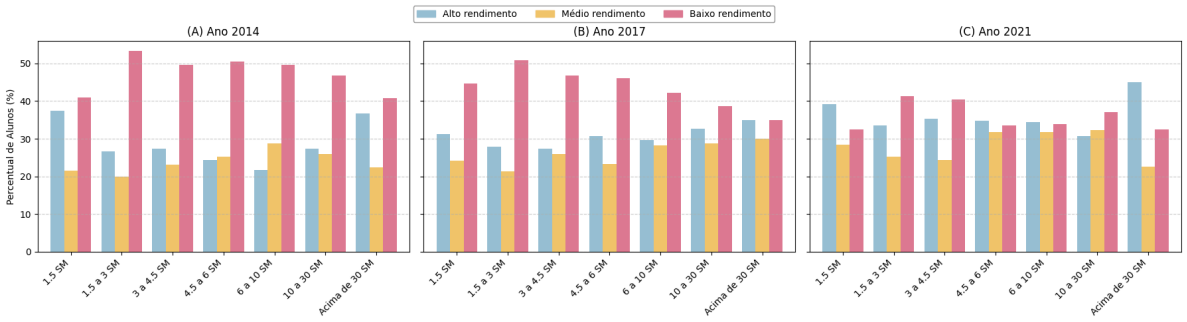

Figura 6. Evolução temporal da distribuição percentual de alunos por faixa de renda e desempenho acadêmico.

6. Conclusão

Ao realizar a análise socioeconômica dos perfis estudantis na área de computação do Nordeste nos anos de 2014, 2017 e 2021, conduzida pela aplicação do processo KDD e com o uso do algoritmo K-Means, foi possível a criação de três grupos de alunos, agrupados por desempenho acadêmico, que proporcionaram uma base estruturada para uma análise mais precisa. Com este conjunto de ferramentas e recursos, foi possível identificar padrões na relação entre o desempenho acadêmico e fatores institucionais, geográficos e econômicos.

Os resultados revelam uma predominância de estudantes com baixo desempenho dentro das universidades, embora também evidenciam uma tendência positiva ao longo dos anos, principalmente nos grupos de classe econômica mais baixa. Esta tendência pode estar atrelada a políticas públicas de apoio e inclusão estudantil anteriormente implementadas, indicando efeitos concretos na melhoria das notas dos alunos.

Por outro lado, constatou-se uma concentração alta de estudantes no grupo de baixo desempenho acadêmico em cursos e estados com maior quantidade total de alunos, sugerindo a presença de problemas estruturais nas instituições mais populosas, possivelmente relacionados a infraestrutura limitada e sobrecarga docente.

Estes resultados reforçam os benefícios da implementação de ferramentas de EDM para suporte à decisão. Ao descrever a metodologia deste trabalho, espera-se que este trabalho não apenas facilite o entendimento da situação acadêmica do Nordeste na área da tecnologia, mas que a análise possa ser realizada em diferentes contextos educacionais e contribua com a gestão educacional baseada em evidências.

Por fim, além da aplicação da metodologia em diferentes contextos educacionais, recomenda-se expandir o escopo da análise através de outras variáveis específicas dos dados das regiões, e a implementação de outros métodos de clusterização em comparação ao K-Means para reforçar a decisão ao dividir os grupos. Assim, espera-se que a compreensão de fatores que influenciam o desempenho estudantil seja facilitada, e que este trabalho auxilie a criação de estratégias eficazes ao combate das desigualdades educacionais.

Referências

ALDEFENDER, M. S.; BLASHFIELD, R. K. **Cluster Analysis**. California: Sage Publications, 1984.

BAILEY, K. D. Cluster Analysis. **Sociological Methodology**, v. 6, n. 1, p. 59-128, 1975.

BAKER, Ryan; ISOTANI, Seiji; CARVALHO, Adriana. **Mineração de Dados Educacionais: Oportunidades para o Brasil.** *Revista Brasileira de Informática na Educação*, v. 19, n. 2, p. 03, 2011.

BARROSO, Paula Cristina Freitas *et al.* **Fatores de Evasão no Ensino Superior: uma Revisão de Literatura**. *Psicologia Escolar e Educacional*, v. 26, e228736, 2022.

FIGUEIRÓ, M. F.; VISTA, N. P. B.; BARASUOL, J. B.; CHICON, P. M. M.; ANSUJ, A. P. **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*, Rio Grande do Sul, Brasil, 2018.

FRAWLEY, W. J.; PIATETSKY-SHAPIRO, G.; MATHEUS, C. J. **Knowledge Discovery in Databases: An Overview.** *AI Magazine*, v. 13, n. 3, p. 57, 1992.

INEP. *Microdados do Enade*. 2023. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enade. Acesso em: 13 jul. 2025.

KING, Elliot. **Data Warehousing and Data Mining: Implementing Strategic Knowledge Management.** Computer Technology Research, 2000.

LIMA, A. M. S.; FLOREZ, A. Y. C.; LESCANO, A. I. A.; DE OLIVEIRA NOVAES, J. V.; DE FA-TIMA MARTINS, N.; JUNIOR, C. T.; DE SOUSA, E. P. M.; JUNIOR, J. F. R.; CORDEIRO, R. L. F. **Analysis of ENEM's attendants between 2012 and 2017 using a clustering approach.** *Journal of Information and Data Management*, v. 11, n. 2, p. 115-130, 2020.

MARQUES, Ebony *et al.* **Sabia: uma plataforma para auxiliar a gestão baseada em evidências nas instituições de ensino superior.** In: *Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil* (WAPLA). SBC, 2023.

MURPHY, Kevin. **Machine Learning: A Probabilistic Perspective.** Massachusetts: The MIT Press, 2012.

NEVES, Clarissa Eckert Baeta. **Ensino Superior No Brasil: Expansão, Diversificação e Inclusão.** Trabalho apresentado no *Congresso da LASA*, São Francisco, Califórnia, 2012.

ROMERO, Cristobal; VENTURA, Sebastian. **Educational data mining and learning analytics: an updated survey.** *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, v. 10, n. 3, e1355, 2020.

SANTOS, Vitor Hugo Barbosa dos; SARAIVA, Daniel Victor; OLIVEIRA, Carina Teixeira de. **Uma análise de trabalhos de mineração de dados educacionais no contexto da evasão escolar.** In: *Simpósio Brasileiro de Informática na Educação* (SBIE). SBC, 2021.

SILVA, A.; HOED, R.; SARAIVA, P. **Análise do Desempenho dos Alunos de Cursos Superiores em Computação no ENADE – Uma Abordagem usando Mineração de Dados**. In: *Conferência Ibero-Americana*, 2019. p. 207-214.

SILVA, Leandro A.; SILVA, Luciano. **Fundamentos de mineração de dados educacionais.** *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, v. 3, n. 1, 2014.

ZAPPAROLLI, Luciana *et al.* **Aplicando Técnicas de Business Intelligence e Learning Analytics em Ambientes Virtuais de Aprendizagem.** *Simpósio Brasileiro de Informática na Educação*, v. 28, n. 1, p. 536-545, 2017.