Identificação de Desigualdades Sociais a partir do desempenho dos alunos do Ensino Médio no ENEM 2019 utilizando Mineração de Dados

Vinicius A. Alves da Silva [1], Lorenza L. Oliveira Moreno[1] Luciana B. Gonçalves [1]
Stênio Sã R. Furtado Soares[1], Robson R. de Souza Júnior[2]

[1] Departamento de Ciência da Computação
Universidade Federal de Juiz de Fora, Juiz de Fora, MG - Brasil.

[2]Departamento de Ciências Humanas
Universidade do Estado de Minas Gerais, Unidade Barbacena, MG. Brasil

{viniciusalves,lorenza,lbrugiolo,ssoares}@ice.ufjf.br

robson.junior@uemg.br

Abstract. *This paper presents the use of Data Mining - DM techniques on data from high school graduates who performed the National High School Exam - ENEM in 2019 with a focus on identifying social inequalities from the analysis of performance in the exam. The use of Clustering and Association Rules algorithms allowed us to map the determining variables in the students' performance as well as to characterize two distinct groups based on the results obtained in the five evaluations that compose the exam.*

Resumo. *O presente trabalho apresenta o uso de técnicas de Mineração de Dados com foco na identificação de desigualdades sociais a partir da análise do desempenho dos estudantes concluintes do ensino médio que prestaram o Exame Nacional do Ensino Médio - ENEM no ano 2019. O uso de algoritmos de Clusterização e de Regras de Associação permitiu mapear as variáveis determinantes no desempenho dos estudantes bem como caracterizar dois grupos bem definidos a partir dos resultados obtidos nas cinco avaliações do exame.*

1. Introdução

O Exame Nacional do Ensino Médio (ENEM) significa, para milhares de alunos, uma oportunidade de acesso à educação superior gratuita e de qualidade que é ofertada por instituições públicas de educação superior através do Sistema de Seleção Unificada (SISU) e por meio da concessão de bolsas de estudo através do Programa Universidade para Todos (ProUni). Além disso, os seus resultados permitem o desenvolvimento de estudos e indicadores sobre a educação brasileira [INEP 2020]. Por isso, embora o ENEM não seja um exame de avaliação da qualidade da educação no país, a análise dos dados socioeconômicos dos candidatos permite identificar distorções sociais que podem nortear políticas públicas de ampliação de acesso ao ensino superior.

Ao inscreverem-se no exame, os participantes respondem a um questionário socioeconômico. O escopo deste trabalho foi definido de forma a identificar desigualdades sociais refletidas nos dados do ENEM 2019, considerando as respostas do questionário e o desempenho dos estudantes concluintes do Ensino Médio em escolas de Minas Gerais.

DOI: 10.5753/cbie.sbie.2020.72

A base de microdados do ENEM 2019 utilizada neste trabalho está disponível em [INEP 2020], tendo sido obtida no repositório online, extraída e preparada para análise. No processo de extração de conhecimento a partir dessa base de dados, utilizou-se duas técnicas de Mineração de Dados: (1) clusterização, onde aplicou-se o algoritmo *k-means* para diferentes valores de k tomando-se apenas os dados referentes às notas das provas dos estudantes; (2) mineração de Regras de Associação, onde aplicou-se algoritmo *Apriori*, que permite identificar elementos frequentes na base de dados e o nível de afinidade entre esses elementos. O objetivo deste trabalho foi caracterizar os grupos identificados no processo de clusterização e, a partir do conjunto de regras de associação extraídas da base de dados, identificar correlações entre variáveis relacionadas a aspectos socioeconômicos, o desempenho do estudante no exame e o grupo ao qual o estudante está inserido.

Este trabalho está organizado da seguinte maneira: a Seção 2 destaca publicações interessantes sobre o tema de mineração de dados e desempenho educacional; a Seção 3 apresenta a metodologia e algoritmos utilizados nesta pesquisa e fundamenta métricas de avaliação que são usadas na seção de resultados. Os resultados alcançados são discutidos na Seção 4 e, por último, conclusão e trabalhos futuros na Seção 5.

2. Trabalhos Relacionados

Dada a importância do ENEM na sociedade brasileira, diversos trabalhos sobre o tema foram escritos, sob diferentes perspectivas. Em [Lima et al. 2019] é apresentada uma revisão sistemática sobre análise de dados do ENEM e do Exame Nacional de Desempenho de Estudantes (ENADE) até 2016. O texto indica que os estudos feitos a partir dos dados desses exames têm escopos limitados, muitas vezes se restringindo apenas ao campo da estatística descritiva, ressaltando a importância de novas pesquisas abordando outras técnicas, como mineração de dados.

Já [Alves et al. 2018], apresenta aplicações usando árvores de decisão e redes Bayesianas, para prever o desempenho dos alunos na prova de redação a partir de dados socioeconômicos. Em [Leoni and Sampaio 2017], a base do ENEM por Escola de 2015 é clusterizada em função do desempenho médio dos alunos e indicadores contextuais. Os resultados apontaram a formação de dois grupos, onde há padrões de desempenho equivalente entre escolas com indicadores similares.

Na linha metodológica deste trabalho em [Silva et al. 2014], é descrita a aplicação de mineração de regras de associação utilizando-se o algoritmo *apriori* em uma implementação através da ferramenta *RapidMiner 5.1*. Foram usados os dados do desempenho na prova e algumas questões do questionário socioeconômico de estudantes das capitais do sudeste. Os resultados apontaram que a renda familiar, o nível de escolaridade e o número de moradores da casa são fatores importantes no desempenho dos estudantes.

Por sua vez, [Gomes et al. 2017] discutem a utilização de algoritmos de Mineração de Dados na base do ENEM 2014 no âmbito da região Nordeste. Os autores concluíram que há uma relação entre a renda familiar e o desempenho dos estudantes, sobretudo para aqueles que cursaram o ensino básico em escolas públicas.

3. Técnicas de Mineração de Dados utilizadas

A descoberta de conhecimento em bases de dados é chamada na literatura de *Knowledge Discovery in Databases* (KDD), geralmente constituído pelas fases de seleção de dados,

limpeza, enriquecimento, transformação, mineração de dados e construção de conhecimento [Simon and Cazella 2017, Gomes et al. 2017]. Nesta seção, as etapas do processo de KDD na base de dados do ENEM 2019 são descritas.

3.1. Leitura, seleção e limpeza dos dados

A base do ENEM 2019, no formato csv, tem aproximadamente 3,12 GB de tamanho. Optou-se por utilizar a linguagem de programação *Python* e a biblioteca *Pandas* [Wes McKinney 2010] para leitura e tratamento da base de dados.

Como o escopo deste trabalho restringe-se aos concluintes do ensino médio do Estado de Minas Gerais, foi realizada uma seleção na base pelo campo referente ao código da unidade da federação da escola (CO_UF_ESC), filtrando os registros do referido estado. Dado que, no momento da inscrição, o preenchimento de dados da escola é requisitado apenas para alunos que declaram estar cursando o último ano do ensino médio, os registros que foram filtrados por estes dados podem ser considerados apenas dos concluintes.

Este recorte inicial gerou uma base com 108.173 registros. Como o objetivo deste trabalho é analisar o desempenho dos estudantes sob uma perspectiva socioeconômica, apenas registros devidamente preenchidos e não zerados foram considerados, o que resultou em uma base com 88.659 estudantes que efetivamente responderam as provas.

3.2. Transformação dos dados

Através de análise descritiva, chegou-se aos atributos socioeconômicos mais relevantes e possivelmente relacionados com o desempenho do estudante. As variáveis selecionadas foram: raça autodeclarada, tipo administrativo da escola, existência de computador pessoal (PC) no domicílio, nível de escolaridade da mãe e classe econômica, definida segundo a renda média familiar.

Para raça autodeclarada, aqueles que se afirmaram pretos, pardos, indígenas e amarelos foram tomados como Não-Branco (critério justificado em [Souza et al. 2010]), reduzindo o domínio a $\{$"Branco", "Não-Branco" ou "Não Declarado"$\}$. A variável "Quantidade de PCs no domicílio" foi tratada como binária $\{$"Não Possui" ou "Possui"$\}$.

Para a mineração de Regras de Associação, a variável "Escolaridade da mãe" foi tratada para indicar se a mãe do estudante contém o ensino "Médio Incompleto", "Médio Completo" ou a opção "Não Sabe". Contudo, na Seção 4.3, consideram-se os valores: "Fundamental Incompleto", "Fundamental Completo", "Médio", "Superior" e "Não Sabe"(registros de "Pós-Graduação" foram unificados como "Superior").

A nota do estudante, média simples das notas das provas do ENEM nas cinco áreas - Linguagens e Códigos, Ciências Humanas, Ciências da natureza, Matemática e Redação, é representada pela variável "Nota Média". Para aplicação do algoritmo *apriori*, essa variável foi discretizada utilizando-se o método *equal width Binning*, que divide os valores do domínio em 4 faixas de mesma largura $w = (maiorValor - menorValor)/n = 132,61$, gerando as seguintes faixas para a "Nota Média": abaixo de $452,9$; entre $452,9$ e $585,6$; entre $585,6$ e $718,2$; e maior ou igual a $718,2$.

No passo seguinte, o processo de clusterização, foram consideradas as notas das provas de cada inscrito. Os valores foram normalizados usando o método *quantile_transform*, com o número de quantis igual a 5. Neste método os atributos são transfor-

mados em uma distribuição uniforme, espalhando os valores dos atributos mais frequentes, reduzindo o impacto de outliers.

3.3. Clusterização de dados

Dado um conjunto C com n objetos, o problema da k-clusterização consiste em dividi-lo em k subconjuntos disjuntos, chamados *clusters*, baseando-se na similaridade entre seus objetos. Neste trabalho, o algoritmo de clusterização escolhido foi o *k-means*, devido à sua simplicidade no uso e à sua escalabilidade em função do número de itens da base. Detalhes do algoritmo podem ser encontrados em [Leoni and Sampaio 2017]. Optou-se por utilizar a versão disponível na biblioteca *sci-kit learn* [Pedregosa et al. 2011], muito usada em aplicações acadêmicas e comerciais. No intuito de identificar o número adequado de clusters, utiliza-se o Índice Silhueta, descrito em [Rousseeuw 1987]. O valor de Silhueta, no intervalo $[-1, 1]$, indica a qualidade do agrupamento, sendo que, quanto mais próximo de 1, maior o indicativo de que os objetos estão bem agrupados. Foram testados outros algoritmos (*DBSCAN*, *Aglomerative Clustering* e *Birch*, da mesma biblioteca), que não se mostraram viáveis devido ao tamanho da base e ao elevado consumo de memória RAM.

3.4. Mineração de regras de associação

A técnica de mineração de regras de associação é usada para encontrar afinidade entre itens de uma base de dados. Um item corresponde a cada variável que pode descrever um objeto, como Raça="Branco"ou EscolaridadeMãe="NãoSabe". O algoritmo *Apriori* é usado na literatura para extrair regras de associação a partir da popularidade dos itens. Regras de Associação são relações de implicação $(A \rightarrow B)$, onde lê-se A (antecedente) implica em B (consequente). Numa regra, A e B são *itemsets*, subconjuntos de itens da base de dados, obrigatoriamente disjuntos e com ao menos um elemento.

O *toolkit Orange* [Demšar et al. 2013] é uma ferramenta de Mineração de Dados que usa programação visual, muito utilizada para exploração de dados. Para este trabalho, foi utilizado o algoritmo *apriori* disponibilizado no *add-on Orange3-Associate*. As vantagens desta implementação são a facilidade de uso, uma vez que a ferramenta não exige um formato específico para entrada dos dados.

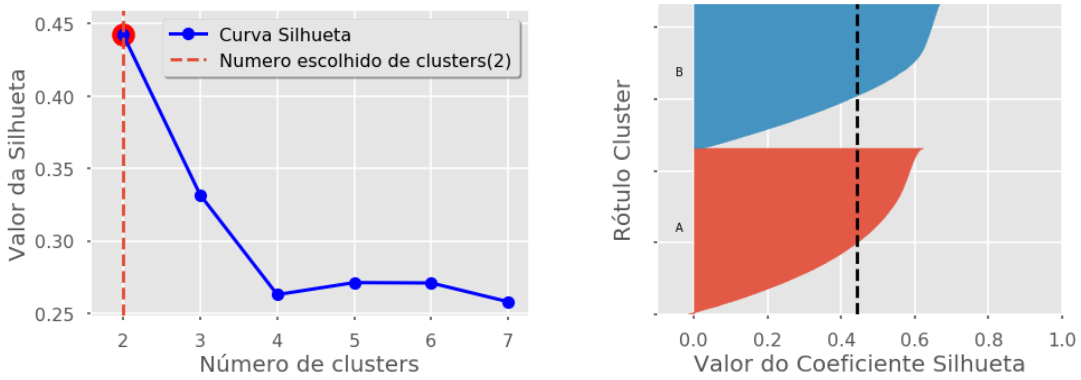
Diferentes métricas norteiam o algoritmo *apriori*. O suporte indica o quanto um *itemset* B é frequente na base de dados, sendo dado pela razão entre o número de registros que contém o *itemset* e o total de registros da base, $Suporte(B) = \frac{Registros\ Contendo\ B}{Total\ Registros}$. Caso o suporte refira-se a uma regra de associação $(A \rightarrow B)$, o numerador deve indicar o número de registros que contém tanto A quanto B. A confiança, dada por $Confianca(A \rightarrow B) = \frac{Registros\ contendo\ A\ e\ B}{Registros\ contendo\ A}$, representa a chance de B ocorrer em um registro, sabendo-se que A ocorre a priori. Já a métrica $Lift(A \rightarrow B) = \frac{Suporte(A \rightarrow B)}{Suporte(A) * Suporte(B)}$, refere-se ao aumento na proporção em que B ocorre quando A ocorre a priori, levando em conta a popularidade do *itemset* B. Nos casos em que os valores de *Lift* são inferiores a 1, as respectivas regras não têm significância. Para cada regra de associação, o algoritmo *Apriori* calcula os valores destas três métricas. Observando-as, é possível estimar a relevância de cada regra de associação encontrada no processo de mineração. Mais detalhes em [Agrawal et al. 1994].

4. Resultados e Discussões

Esta seção discorre sobre os resultados obtidos no processo de clusterização, apresenta as regras de associação e traz ainda uma análise descritiva dos dados. Ao longo da seção, a "Nota Média"é usada como medida de desempenho do aluno.

4.1. Grupos identificados na base

Para decidir qual o melhor número de clusters, o *k-means* foi testado com valores de k variando de 2 a 10 sendo o melhor valor de Silhueta encontrado para $k = 2$, conforme indica o gráfico da Figura 1(a). Os valores se estabilizam menores que 0,275 para $k \geq 4$. O gráfico da Figura 1(b) mostra o valor de silhueta para cada elemento conforme o cluster, com destaque para o valor do índice silhueta encontrado. Nestes testes, o número máximo de iterações foi fixado em 1.200.



(a) Silhueta em função de K (b) Índice Silhueta

Figura 1. Valor de Silhueta dado número de clusters e distribuição da Silhueta para K=2

Após a clusterização, os rótulos "*Cluster A*" e "*Cluster B*" foram adicionados à base de dados, para identificar o posicionamento de cada registro. Do total de estudantes, 46.419 foram rotulados como "Cluster A" e 40.240 como "Cluster B". A Figura 2 apresenta características da distribuição das notas dos registros de cada grupo nas provas do ENEM.

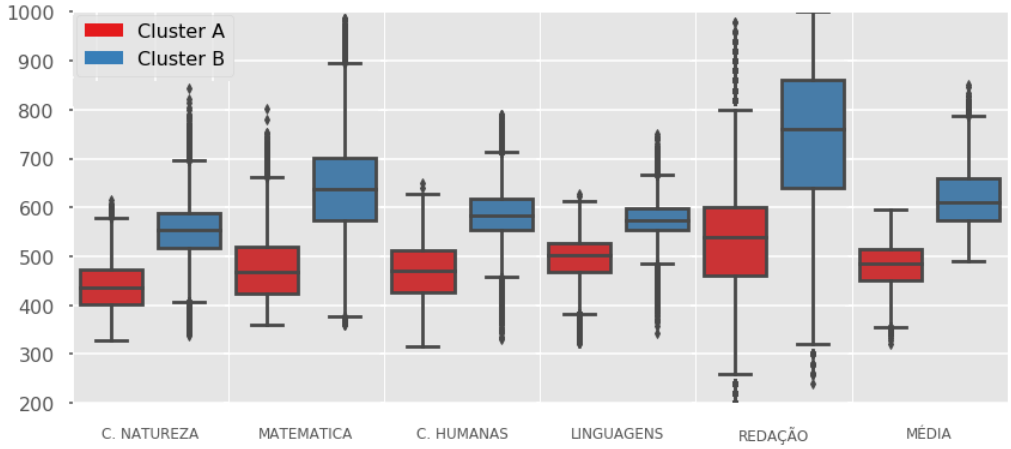


Figura 2. Distribuição das notas para os clusters resultantes

Observando a Figura 2, verifica-se que os participantes do Cluster B apresentam desempenho superior em todas as provas. Em qualquer delas, o segundo quartil do Cluster B é maior que o terceiro quartil do Cluster A, ou seja, a nota de 75% dos alunos do Cluster B é estritamente maior que a de 75% dos estudantes do Cluster A. Na comparação da Média, o valor mínimo do Cluster B é muito próximo à mediana do Cluster A, confirmando a diferença significativa na distribuição desta variável entre os dois grupos.

Analisando os dados socioeconômicos dos registros do Cluster A, foi possível observar que $94,51\%$ deles, 43.869 estudantes, são oriundos de escolas estaduais (valor que representa 65% do total de estudantes da rede estadual). Além disso, 66% dos membros do Cluster A se declararam negros, pardos, amarelos ou indígenas (não-brancos); $52,59\%$ afirmaram ter PC em casa e 71% pertencem a Classe E, indicando que a família tem renda mensal inferior a R$\$2.000,00$ mensais. Sobre a escolaridade da mãe, $51,92\%$ dos candidatos afirmaram que a mãe não completou o Ensino Médio.

Quanto à composição Cluster B, $56,85\%$ são estudantes de escolas estaduais, $29,89\%$ de escolas particulares, $11,75\%$ de escolas federais e $1,51\%$ de escolas municipais. Embora mais da metade dos alunos pertença à rede estadual, verifica-se que estão no Cluster B 89% dos alunos de escolas particulares, 88% do total de alunos das escolas federais e apenas 35% dos alunos de escolas estaduais. Neste grupo, 51% se declararam Brancos, 79.65% relataram possuir PC e 70% têm mãe com pelo menos o Ensino Médio completo. Em relação à renda mensal, $38,54\%$ dos alunos são da Classe E, $29,34\%$ da D e $22,45\%$ da classe C, sendo os 10% restantes das classes A e B.

4.2. Regras de Associação

A mineração de regras de associação teve como objetivo caracterizar os clusters a partir das variáveis socioeconômicas. O processo de mineração foi executado considerando-se toda a base de dados e, também, os registros cada cluster isoladamente. Devido ao espaço limitado, neste texto, são apresentadas apenas algumas das regras de associação obtidas em cada execução [1]. Os valores das métricas, que permitem validar e entender melhor o comportamento das regras de associação, foram limitados da seguinte forma: suporte mínimo de 20%, confiança de 70% e lift maior que 1.

Ao executar o algoritmo com todos os registros da base, foram geradas 80 regras de associação, das quais 8 encontram-se na Tabela 1(a). A primeira regra indica que, com 99,9% de confiança, os alunos com médias entre 585,6 a 718,21 estão no Cluster B. De acordo com a regra 2, os alunos com média entre 452,99 e 585,6, com 72% de confiança, encontram-se no Cluster A (há alunos do Cluster B com valores de média neste intervalo).

Considerando informações originadas do questionário socioeconômico, entre os alunos do Cluster A, $94,5\%$ vem de escolas estaduais (regra 3). Entre os estudantes de escolas estaduais que se autodeclararam negros, pardos, amarelos ou indígenas (não-brancos), $70,1\%$ estão no Cluster A (regra 4). Além disso, o Cluster A contém $73,6\%$ dos estudantes da classe E cujas mães não concluíram o ensino médio (regra 5). A regra 6 indica que os estudantes do Cluster B que possuem PC em casa, com $75,7\%$ de confiança, têm mãe que, ao menos, completou o ensino médio. Entre o grupo majoritário dos que possuem PC, encontram-se $79,9\%$ dos indivíduos do Cluster B (regra 7); já aqueles que não possuem encontram-se no Cluster A (com $72,9\%$ de confiança), vide regra 8.

[1] Em `https://bit.ly/33f3QgJ`, encontra-se um relatório com todas regras de associação geradas.

Tabela 1. Regras de Associação

#	Antecedente → Consequente	Suporte	Conf.	*Lift*
(a) Regras considerando toda a base				
1	mediaNota=585,6 - 718,21 → Cluster=B	28,5%	99,9%	2,152
2	mediaNota=452.99 - 585.6 → Cluster=A	39.4%	72,0%	1,344
3	Cluster=A → ADM_ESC=Estadual	50,6%	94,5%	1,227
4	RACA=NãoBranco, ADM_ESC=Estadual → Cluster=A	34,2%	70,1%	1,309
5	EstudoMae=Medio_Inc, Classe=E → Cluster=A	22,4%	73,6%	1,373
6	TemPC=Sim. Cluster=B → EstudoMae=Medio_Comp.	28,0%	75,7%	1,334
7	Cluster=B → TemPC=Sim	37,0%	79,6%	1,222
8	TemPC=Não → Cluster=A	25,4%	72,9%	1,361
(b) Regras considerando apenas o Cluster A				
1	RACA=NãoBranco → Classe=E	50%	75%	1,051
2	Classe=E → RACA=Não-Branco	50%	75%	1,051
3	EstudoMae=Medio_Comp. → mediaNota=585,6 - 718,21	34%	78%	1,062
4	mediaNota=585,6 - 718,21 → Classe=E	20%	78%	1,099
(c) Regras considerando apenas o Cluster B				
1	NotaMedia=452,99 - 585,6 → ADM_ESC=Estadual	27%	84%	1,470
2	ADM_ESC=Particular → NotaMedia=585,6 - 781,21	25%	75%	1,226

Considerando apenas os dados dos estudantes do Cluster A, quatro das regras de associação obtidas são destacadas na Tabela 1(b). As duas primeiras apresentam uma relação forte da autodeclaração da raça e a classe econômica, indicando que 75% dos autodeclarados como não-brancos são da classe E e vice-versa. Já a regra 3, indica que 78% dos estudantes do Cluster A cujas mães concluíram pelo menos o ensino médio têm nota média entre $585,6$ e $718,21$ (as maiores notas do cluster), sendo que isso ocorre em 34% dos registros do Cluster A. Além disso, a regra 4 mostra que 78% dos indivíduos do Cluster A que têm desempenho médio de $585,6$ a $718,21$ são da classe E.

Por outro lado, tomando-se apenas os dados dos estudantes do Cluster B, observa-se uma variação maior das faixas de notas médias. A regra 1 referente ao Cluster B na Tabela 1(c) mostra que 27% do total de registros do Cluster B, apresentaram nota média na faixa baixa ($452,99$ a $585,6$) e estudam em escolas estaduais. E, entre os estudantes do Cluster B que tiveram este rendimento, 84% são de escolas estaduais. Além disso, 75% dos estudantes de escolas particulares atingiram desempenho na faixa ($585,6$ a $781,21$).

4.3. Variáveis que evidenciam as desigualdades

Visto que a maioria dos inscritos são de Escolas Estaduais ou da Classe econômica E, as regras de associação, que se baseiam na frequência dos itens, ficam saturadas por essas variáveis. Nesta seção, são detalhadas outras correlações entre as variáveis socioeconômicas e o desempenho dos estudantes.

A Figura 3 apresenta, para cada faixa de renda, o percentual dos estudantes em cada cluster. Nesta figura fica claro que, quanto maior a renda familiar, maior a probabilidade deste estudante se encaixar no Cluster B, onde estão os alunos com melhor desempenho. Destaca-se que 68,22% do total estudantes possuem renda entre 998,00 e 2495,00 reais.

Na Figura 4(a), é possível observar a distribuição da média das notas dos alunos

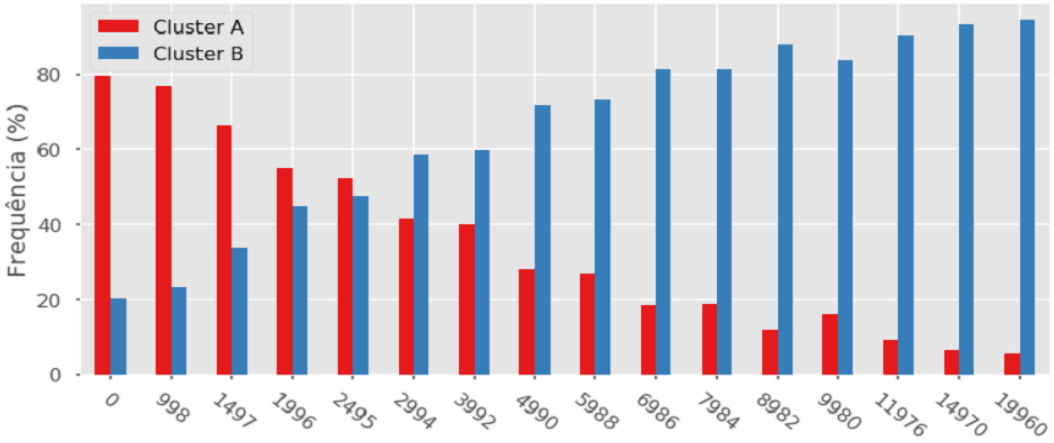
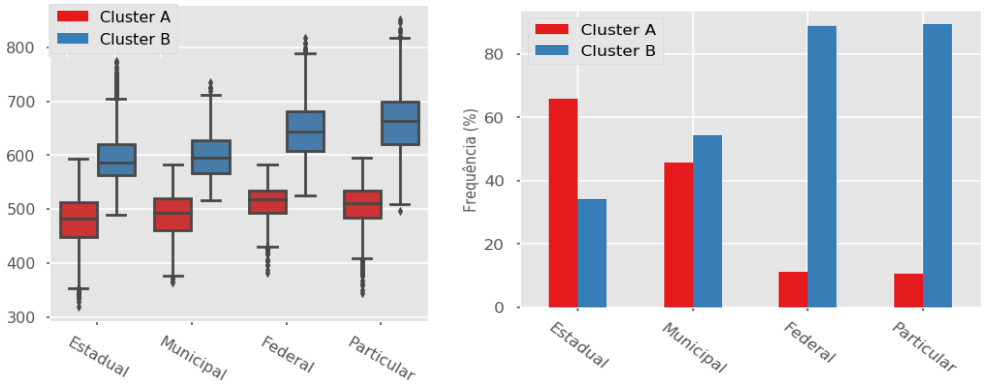


Figura 3. Distribuição dos participantes de cada cluster por renda

conforme o tipo de escola. Observando-se apenas o Cluster B, nota-se uma similaridade de desempenho entre estudantes de escolas públicas estaduais e municipais nos dois primeiros quartis, e ainda, uma maior semelhança de comportamento quando se comparam os resultados para estudantes da rede federal com os das escolas privadas. A mediana destas últimas categorias é maior que o terceiro quartil das escolas estaduais e municipais. Já a Figura 4(b) mostra, em porcentagem, a participação de cada tipo de escola nos clusters. É nítido que a maioria dos alunos das Escolas Federais e Particulares estão no Cluster B.



((a)) Nota Média por Adm. da Escola ((b)) População nos clusters por Adm da Escola

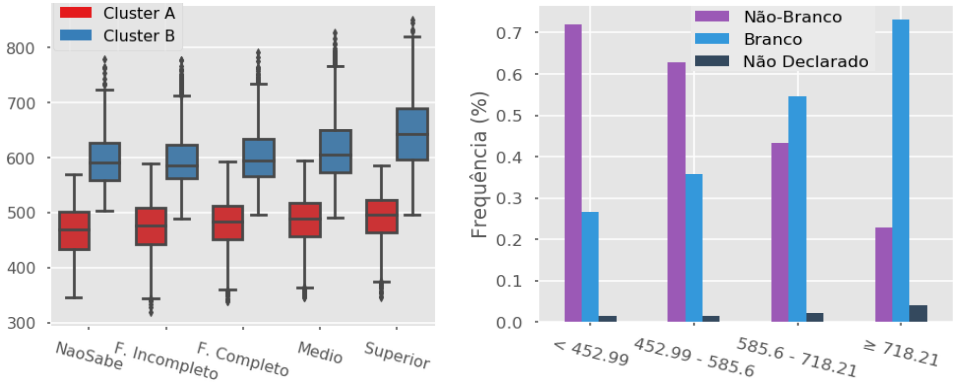
Figura 4. Distribuição da nota média e entre clusters de acordo com tipo de escola

Sobre a escolaridade da mãe, na Figura 5(a), verifica-se a distribuição da "Nota Média"pelos vários níveis de escolaridade. Convém destacar que a mediana da categoria referente ao ensino superior no Cluster B é próxima ao terceiro quartil do ensino médio para o mesmo cluster. Ou seja, quase 50% dos inscritos do Cluster B com mãe que cursou o Ensino Superior tem nota maior que 75% dos demais de cada uma das outras categorias.

Tanto para o Tipo Administrativo da Escola na Figura 4(a) quanto para a escolaridade da mãe na figura 5(a), no Cluster A nota-se a uma diferença menor na distribuição das notas pelos diversos tipos de categoria, ao contrário do Cluster B. Indicando de que outros aspectos socioeconômicos se sobressaem como determinantes.

Já a Figura 5(b) traz a informação sobre raça autodeclarada, onde é possível obser-

var, para cada intervalo de notas, a distribuição proporcional dos estudantes. Verifica-se facilmente um predomínio nas faixas iniciais por estudantes autodeclarados Não-brancos, enquanto a distribuição se altera à medida que a média aumenta, chegando-se à total inversão para as faixas das maiores médias.



((a)) Nota por Escolaridade da Mãe ((b)) Autodeclaração de raça por nota média

Figura 5. "Escolaridade da Mãe"e "Raça Autodeclarada"

5. Conclusão

O acesso universal a uma educação de qualidade é condição necessária para a redução das desigualdades sociais em qualquer país. Neste sentido, a análise do desempenho escolar sob o prisma socioeconômico constitui um importante instrumento para trazer à tona discussões acerca das distorções sociais que contribuem na manutenção das desigualdades.

Este trabalho utilizou técnicas de clusterização, mineração de regras de associação e estatística descritiva para identificar quais variáveis socioeconômicas apresentam correlação com o desempenho de alunos do Estado de Minas Gerias, concluintes do ensino médio, na prova do ENEM de 2019.

Os resultados da clusterização permitiram identificar dois grupos bem característicos: um com os estudantes com notas, em geral, mais baixas (Cluster A) e outro com as notas mais altas (Cluster B). Apesar do agrupamento basear-se somente em atributos referentes às notas, no Cluster A existe um predomínio de alunos com características socioeconômicas similares entre si, destacando-se a educação em rede estadual de ensino e a baixa renda familiar. Por outro lado, no cluster de alunos com notas mais altas, observa-se uma diferença expressiva no desempenho de alunos das redes particular e federal, em relação aos demais, mesmo sendo os últimos a maioria entre estudantes do Cluster B.

O estudo possibilitou ainda verificar a semelhança no desempenho médio de estudantes oriundos de escolas da rede federal em relação aos da rede privada. Diante disso, duas questões se apresentam: a necessidade de que se discuta o modelo de gestão destas escolas, no sentido de replicá-lo nas demais redes públicas de ensino; e a necessidade de se aprofundar a análise dos aspectos socioeconômicos destes alunos, de forma a verificar similaridades e diferenças entre estudantes de escolas federais e de escolas particulares.

Em trabalhos futuros, pretende-se complementar a base do ENEM utilizando dados de outras bases, como a do censo escolar, o que permitiria constatar outras regras

de associação e relacionar o desempenho do aluno às características da escola, e assim entender quais os principais fatores que permitem ao aluno um melhor desempenho.

Referências

Agrawal, R., Srikant, R., et al. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499.

Alves, R. D., Cechinel, C., and Queiroga, E. (2018). Predição do desempenho de matemática e suas tecnologias do enem utilizando técnicas de mineração de dados. In *Workshops do Congresso Brasileiro de Informática na Educação*, volume 7, page 469.

Demšar, J. et al. (2013). Orange: Data mining toolbox in python. *Journal of Machine Learning Research*, 14:2349–2353.

Gomes, T., Gouveia, R., and Batista, M. (2017). Dados educacionais abertos: associações em dados dos inscritos do exame nacional do ensino médio. In *Anais do Workshop de Informática na Escola*, volume 23, page 895.

INEP (2020). Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - Microdados do ENEM 2019. `http://portal.inep.gov.br/web/guest/microdados`. Online: acessado 03 Julho 2020.

Leoni, R. C. and Sampaio, N. (2017). Desempenho das escolas públicas e privadas da região do vale do paraíba: Uma aplicação da técnica de agrupamentos kmeans com base nas variáveis do enem 2015. *Cadernos do IME-Série Estatística*, 42:31.

Lima, P. d. S. N., Ambrósio, A. P. L., Ferreira, D. J., and Brancher, J. D. (2019). Análise de dados do enade e enem: uma revisão sistemática da literatura. *Avaliação: Revista da Avaliação da Educação Superior*, 24(1):89–107.

Pedregosa, F. et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Silva, L. A., Morino, A. H., and Sato, T. M. C. (2014). Prática de mineraçao de dados no exame nacional do ensino médio. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, volume 3, page 651.

Simon, A. and Cazella, S. (2017). Mineração de dados educacionais nos resultados do enem de 2015. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, volume 6, page 754.

Souza, P. F. d., Ribeiro, C. A. C., and Carvalhaes, F. (2010). Desigualdade de oportunidades no brasil: considerações sobre classe, educação e raça. *Revista Brasileira de Ciências Sociais*, 25(73):77–100.

Wes McKinney (2010). Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings 9th Python in Science Conference*, pages 56 – 61.