K-means na análise de características socioeconômicas de candidatos ao ensino superior.

Marília Magalhães Maia
Engenharia Elétrica
Universidade Federal Rural do Semi-Árido
Mossoró, Brasil
marilia.maia@alunos.ufersa.edu.br

Luiza Helena Felix de Andrade
Dept. de Ciências Naturais, Matemática e Estatística
Universidade Federal Rural do Semi-Árido
Mossoró, Brasil
luizafelix@ufersa.edu.br

Silvio Fernandes
Dept. de Computação
Universidade Federal Rural do Semi-Árido
Mossoró, Brasil
silvio@ufersa.edu.br

Resumo*—A necessidade de técnicas de aprendizado de máquinas tem sido evidenciada nas últimas décadas pela crescente manipulação e análise de dados para as mais variadas finalidades. Dentre os processos de aprendizado não supervisionado, o K-means é um dos algoritmos mais utilizados para agrupar dados sem classificação prévia e para diferentes finalidades. Diante disso, neste trabalho foram implementadas no K-means as proficiências de candidatos ao ensino superior no Exame Nacional do Ensino Médio (Enem), afim de analisar as características que os diferem na perspectiva das cotas de acesso ao ensino superior.

***Palavras-chaves*—Mineração de Dados; K-means; Cluster; Dados Socioeducacionais.**

I. INTRODUÇÃO

O aumento do volume de dados produzidos nas últimas décadas motiva os estudos dos processos de aprendizado de máquinas. Estes processos possibilitam a manipulação dos dados paras as mais diversas finalidades, variando de acordo com o interesse da análise e do tipo de dados trabalhados. Quando se trabalha com dados que não apresentam informações categóricas, aplica-se o processo de aprendizado não-supervisionado, caso contrário o processo aplicado é o supervisionado

Entre os não-supervisionados, para realizar um agrupamento sobre um conjunto de dados, o algoritmo K-means é um dos mais simples e populares. Esse agrupa os dados que mais se aproximam entre si, e para representá-los cria-se um centroide, de modo que um agrupamento (*cluster*) possua um centroide e os dados selecionados. O algoritmo tem como parâmetros a quantidade de agrupamentos, nomeada de K, a métrica que mede a similaridade dos dados e a inicialização aleatória. A métrica mais utilizada é a distância Euclidiana ao quadrado, devido a existência e unicidade do ponto que a minimiza e este mesmo ponto, sendo o centroide, descrito matematicamente pela média aritmética [1], [2]. A Fig. 1 apresenta um fluxograma demonstrando o funcionamento do K-means.

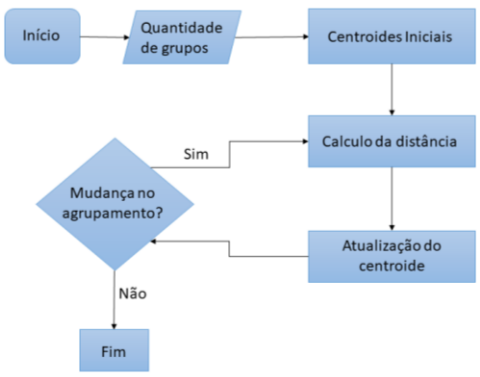

Fig. 1. Fluxograma do funcionamento do K-means.

A determinação aleatória dos K centroides é o passo inicial do algoritmo, sendo restrito à primeira iteração, onde K é definido de acordo com a quantidade desejada de grupos. Como forma de otimização os centroides iniciais são escolhidos aleatoriamente entre os dados que se deseja agrupar [3]. Dessa forma, a partir da segunda iteração, os centroides passam a ser determinados pela métrica, sendo definidos pelos pontos que minimizam a soma das distâncias aos *n* pontos pertencentes *cluster*. A variação da função distancia com a métrica escolhida, também varia a determinação dos centroides [3]. Além disso, ainda segundo [3], no passo seguinte, são calculadas as distâncias de todos os dados a todos os centroides, as quais são usadas para rotular cada dado como pertencendo ao agrupamento do centroide ao qual este dado tenha a menor distância. E por fim, o teste de parada do algoritmo realiza a paralização do mesmo ou permite o prosseguimento para a próxima iteração. A paralização ocorre quanto é atingida a quantidade máxima de iterações desejada ou quando não houver significativa mudança nos *clusters* entre a atual iteração e as anteriores. Caso não seja efetuada a paralização, todos os centroides são recalculados, como já citado e a nova iteração se inicia. Dessa forma, o resultado do algoritmo são os K *clusters*, sobre os quais é possível a realização de análises quanto as características que os agrupam

O presente artigo trata-se do desenvolvimento de um algoritmo K-means e uma aplicação, a fim de demonstrar o funcionamento do mesmo e a possibilidade de discutir problemas sociais através de procedimentos computacionais.

A aplicação implementada é uma análise dos microdados [4], disponibilizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), sobre dos candidatos ao ensino superior que realizaram o Exame Nacional do Ensino Médio (Enem) em 2018. O conjunto de dados é constituído pelas respostas do questionário socioeconômico aplicado no momento da inscrição no exame, as notas dos alunos nas 5 áreas do conhecimento avaliadas, além da situação quanto a presença dos candidatos nos dias de realização das provas ou se foram desclassificados, sem os dados pessoais dos candidatos. Entretanto, a análise será realizada sobre características especificas fornecidas no microdados, descritas mais à frente.

O Enem, desenvolvido pelo Ministério da Educação (MEC), através do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) como ferramenta para "avaliar o desempenho do estudante ao fim da escolaridade básica" [5], teve, também, como finalidade o acesso à educação superior de instituições públicas e privadas, que fizeram uso das notas dos candidatos no processo seletivo. Em 2009, passou por reformulação, tornando-se a principal porta de acesso as instituições de ensino superior públicas e o

processo de seleção unificado das universidades públicas federais.

Entretanto, como tentativa de diminuir os impactos causados pela desigualdade socioeconômica a uma parcela da sociedade, em 2012, foi sancionada a Lei n° 12.711/2012, a Lei da Cotas, que prevê a reserva de 50% das vagas das instituições públicas federais de ensino superior para estudantes de escolas públicas, subdividindo-se metade-metade entre aqueles que possuem renda *per capita* até 1,5 salários-mínimos e aqueles que não possuem, internamente a essa subdivisão ainda há uma outra, quanto a cor/raça dos candidatos, entre aqueles que são pardos, pretos ou indígenas e que não são, sendo essa divisão proporcional aos percentuais das etnias nas unidades da Federação, de acordo com o último censo do Instituto Brasileiro de Geografia e Estatística (IBGE)[6]. A Fig. 2 apresenta um fluxograma que demonstra as divisões das cotas. Diante disso, a analise realizada sobre o agrupamento será baseada nas características abordadas na lei.

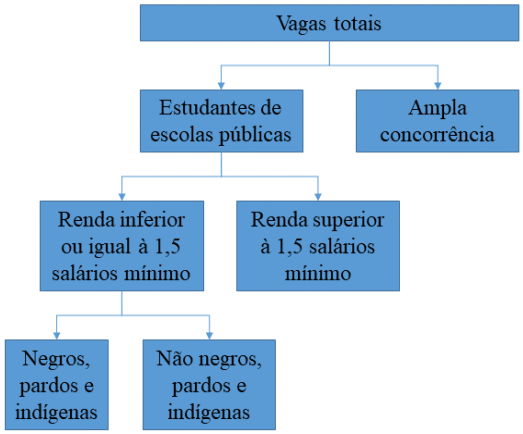

Fig. 2. Fluxograma demonstrativo das divisões das cotas.

II. Metodologia

Neste trabalho, através de um Jupyter notebook, optou-se por implementar o algoritmo K-means, completamente ao invés de utilizar de bibliotecas como Scikit Learn [7]. O intuito dessa implementação é fazer também experimentações com diferentes métricas de agrupamento ou otimizações como paralelismo para aumentar a eficiência.

O algoritmo faz uso das bibliotecas Pandas [8] e Numpy [9] e recebe como parâmetro um *DataFrame*, a quantidade de *cluster*s desejados e quantidade máxima de iterações. Internamente, são adicionadas 2 colunas ao *DataFrame*, uma com a identificação do *cluster* atual e outra com a distância calculada, essas guardam as informações produzidas, sendo atualizadas a cada iteração, na etapa do cálculo das distâncias e formação do agrupamento, e facilita a atualização dos centroides independente da métrica escolhida.

Para a aplicação optou-se por trabalhar com os microdados [4] referentes à aplicação do Enem em 2018, devido à ausência da informação se o aluno frequentou, ou não, escolas públicas durante todo seu ensino médio em microdados mais atuais. Tendo em vista que, essa característica classifica se o aluno pode usufruir das cotas ou não.

Para a implementação dos dados, primordialmente, fez-se uma amostragem para viabilizar implementação dos dados. Para isso, com nível de confiança de 95% e erro amostral de 2%, foi encontrado 1537 como o ótimo para a representação dos 5.513.662 inscritos no exame em 2018. Após a seleção aleatória dos 1537 inscritos, aplicou-se uma amostragem, selecionando apenas aqueles que estavam presentes nos dois dias de aplicação das provas, não foram desclassificados e poderiam concorrer a vagas para ingressar no ensino superior, obtendo-se 919 candidatos que irão constituir, apenas com suas notas nas 5 áreas do conhecimento avaliadas, o *DataFrame* implementado. Dessa forma, o algoritmo foi posto em operação as proficiências dos candidatos, com o valor de K igual a 2, resultando 2 agrupamentos, e restrito a quantidade de 100 iterações máximas.

Após o agrupamento resultante do algoritmo, coletou-se as informações socioeconômicas dos candidatos referentes a cor/raça que os alunos se autodeclaram, variando as respostas entre, não declarado, branca, preta, parda, amarela e indígena, a renda familiar mensal, informando intervalos de valores proporcionais ao salário mínimo do ano, iniciando com 0 reais, até rendas superiores a 20 salários mínimos. Além disso, foram buscadas também as questões que se referem à quantidade de pessoas que moram na residência do candidato, variando a respostas de 1 a 20, para que fosse realizado o cálculo da renda *per capita*. E a última informação socioeconômica coletada foi o tipo de escola que frequentou o ensino médio, fornecendo a indicação como possível usuário das cotas apenas aqueles que responderam “somente em escola pública”. Diante disso, tornou-se viável a análise dos agrupamentos quanto as características de classificação para as cotas.

III. Resultados

A divisão dos dados resultante da implementação do K-means apresentam o *cluster* 0 constituído por 471 alunos e o *cluster* 1 com 448. Essa distribuição pode ser visualizada na Fig. 3, sendo o *cluster* 0 representado pela cor azul e o *cluster* 1 pela cor laranja.

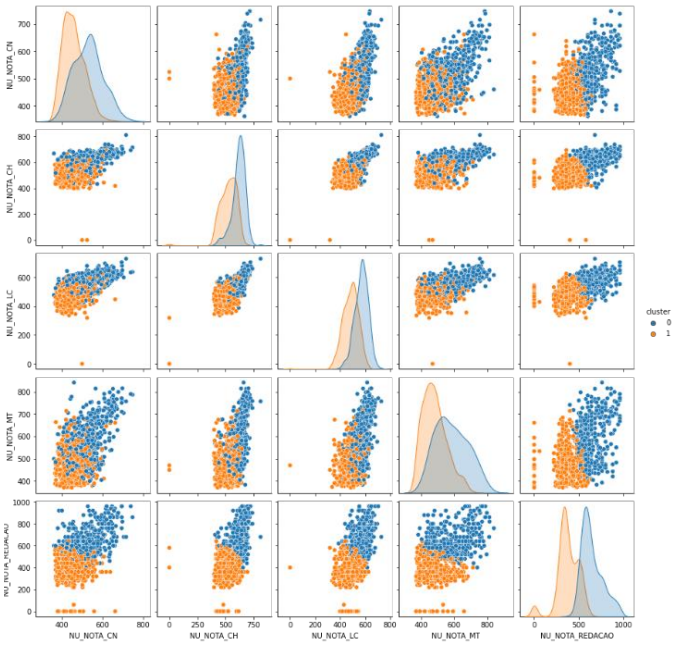

Fig. 3. Agrupamento resultante do K-means com as notas dos candidatos.

A Fig. 3 apresenta a relação entre as notas das 5 áreas do conhecimento avaliadas pelo exame duas a duas e na diagonal principal a distribuição da coordenada por ela mesma. Sendo,

NU_NOTA_CN, NU_NOTA_CH, NU_NOTA_LC, NU_NOTA_MT e NU_NOTA_REDACAO as notas dos candidatos nas provas das ciências da natureza, ciências humanas, linguagens e códigos, matemática e redação, respectivamente, variando entre valores de 0 a 1000.

Mediante a análise da Fig. 3 é possível observar que o *cluster* 1, encontrando-se majoritariamente a esquerda e próximo aos eixos, o que permite classificar seus integrantes como aquele que possuem desempenho menor, em relação aos demais estudantes, os quais se encontram no *cluster* 0, localizado mais distante dos eixos e a direita dos gráficos.

A partir do agrupamento resultante, o estudo sobre das informações categóricas tornou-se viável, ao se contabilizar a quantidade de candidatos em cada *cluster* que satisfaz as características avaliadas para a adesão as cotas. A Fig. 4 apresenta esta contabilidade.

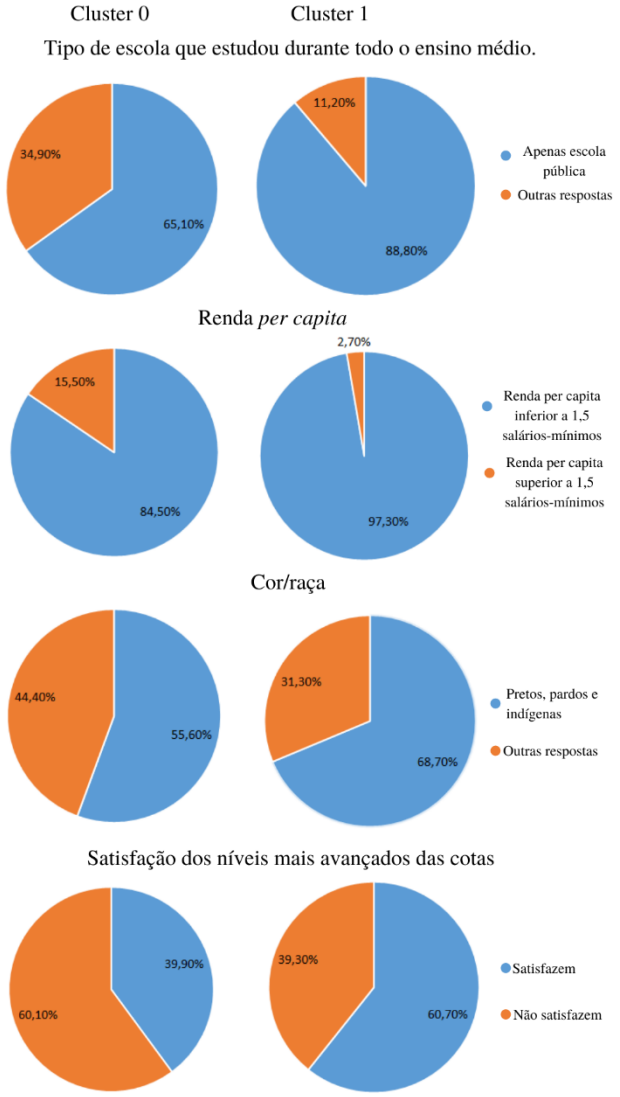

Fig. 4. Frequência das características relacionadas às cotas.

Os gráficos da Fig, 4 apresentam o *cluster* 1 como sendo aquele com percentuais mais elevados do que o *cluster* 0, quanto a frequência dos candidatos que estudaram apenas em escola pública durante o ensino médio, possuem renda *per capita* inferir a 1,5 salários-mínimos e são pretos, pardos e indígenas, demonstrando a relação entre as notas e as características citadas. Os estudantes com desempenho menor serem majoritariamente provenientes do ensino público é um resultado encontrado também por [10], o qual apresenta como resultado a afirmação da existência do agrupamento natural entre dos alunos por modalidade educacional frequentada. Além disso, é valido ressaltar que os autodeclarados pretos, pardos e indígenas, que são mais de 97% dos integrantes do *cluster* 1, possuem a taxa de analfabetismo em sua população mais que dobrada em relação a população branca, segundo [11], além de apresentarem a mesma relação entre as taxas de distribuição de renda e condições de moradia das pessoas a baixo da linha da pobreza.

Dessa forma, sendo todas essas características relevantes para a satisfação ou não pelas cotas e observando os últimos dois gráficos apresentados na Fig. 4, que expressam os percentuais dos candidatos que possuem, simultaneamente, todas as condições satisfeitas para usufruir dos níveis mais avançados das cotas, pode-se afirmar que as informações classificatórias são as notas e a predominância dos candidatos que satisfazem os critérios para adesão total as cotas.

Entretanto, é valido salientar que como o algoritmo K-means tem com um de seus parâmetros a inicialização aleatória, os resultados são variáveis, mantendo-se a relação encontrada, mas com valores diferentes, se mantidos os dados de entrada implementados nesta aplicação.

IV. Conclusão

Apresentou-se neste artigo o desenvolvimento do algoritmo de agrupamento, K-means e sua aplicação sobre os dados do Enem, disponibilizados pelo Inep, afim de realizar uma análise sobre as características que dividem os candidatos a ingressarem em instituições de ensino superior através dos seus desempenhos no exame. Esses dados foram divididos em 2 grupos pelo algoritmo, baseando-se apenas pelas notas nas 5 competências avaliadas pelo exame. Com os *cluster*s formados foi possível observar a relação das notas com as questões socioeconômicas dos candidatos.

Como apresentado em [11] e [12], a parcela da sociedade negra, parda e indígena sofre de forma mais acentuada a desigualdade socioeconômica existente no Brasil e como afirmado em [10], este fato recai diretamente sobre a situação educacional desse grupo, como demonstrado nos resultados encontrados, reafirmando sua coerência quanto a informação classificatória encontrada.

Trabalhos futuros incluem um estudo sobre as características socioeducacionais, a partir dos microdados sobre o Enem ao longo tempo utilizando o K-means e a implementação de outras métricas de similaridade baseadas em distribuição de probabilidades e uso do K-means em outras aplicações.

V. Referência

[1] A. K. Jain, “Data clustering: 50 years beyond K-means”, in *Pattern Recognit. Lett.*, vol. 31, nº 8, 2010, p. 651–666.

[2] S. P. Lloyd, “Least squares quantization in {PCM}. Special issue on quantization”, in *IEEE Trans. Inf. Theory*, vol. 28, nº 2, 1982, p. 129–137.

[3] M. K. Matte, “Impacto do Uso da Desigualdade Triangular para Acelerar o Algoritmo k-Means”, in

Centro Universitario Campo Limpo Paulista, 2020.

[4] INEP, “microdados_Enem2018,” in https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem, 2020.

[5] Mec, “Historico”, in http://inep.gov.br/enem/historico, 2019.

[6] BRASIL, “PL 3627/2004”, in https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=254614, 2004.

[7] A. G. et. al Fabian Pedregosa, Gaël Varoquaux, “Scikit-learn,” in http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html, 2011.

[8] W. McKinney, “Panda”, *{P}roceedings of the 9th {P}ython in {S}cience {C}onference*, https://pandas.pydata.org/, 2010.

[9] S. Berg *et al.*, “Numpy”, in https://numpy.org/, 2005.

[10] R. C. Leoni e N. A. de S. Sampaio, “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”, in *Cad. do IME - Série Estatística*, vol. 42, nº 0, 2017.

[11] IBGE, “Desigualdades sociais por cor ou raça no Brasil”, *Estud. e Pesqui. Informações Demográficas e Socioeconômicas*, vol. 41, 2019, p. 1–12.

[12] Inep, “O item cor ou raça no Censo Escolar da educação básica”, 2015, p. 10.