ISSN: 2319-0124
josif 2023
15ª JORNADA CIENTÍFICA E TECNOLÓGICA
12º SIMPÓSIO DE PÓS-GRADUAÇÃO
DO IFSULDEMINAS
REALIZAÇÃO INSTITUTO FEDERAL Sul de Minas Gerais
APOIO FAPEMIG CNPq

Análise de Dados e Clusterização Utilizando K-means: Um relato de experiência

Laura F. C. FERREIRA[1]; Guilherme O. RANGEL[2]; Juliete A. R. COSTA[3]

RESUMO

Este trabalho relata a experiência de estudantes do curso Técnico em Informática Integrado ao Ensino Médio em primeiro contato com pesquisa cientifica na área de aprendizado de máquina. A partir da utilização da biblioteca *scikit-learn* disponível em linguagem de programação Python, foi aplicado o algoritmo K-Means para analisar quão bem o algoritmo consegue realizar agrupamento e, consequentemente, a relação entre variáveis da base de dados. Os resultados parciais obtidos fornecem informações valiosas para análise dos dados e limitações da amostra de dados analisada.

Palavras-chave: Análise de Dados; Visualização; Métricas de Avaliação.

1. INTRODUÇÃO

A análise de dados tem se tornado uma ferramenta essencial em diversas áreas de pesquisa e tomada de decisões (KUČAK; JURIČIĆ; ĐAMBIĆ, 2018). Nesse contexto, técnicas de clusterização têm sido amplamente utilizadas para identificar padrões e agrupamentos em conjuntos de dados complexos. Este trabalho apresenta uma abordagem utilizando o algoritmo K-means para realizar a clusterização em um conjunto de dados gerados em ambientes educacionais com o objetivo de analisar os agrupamentos inerentes aos dados e identificar relações significativas entre variáveis.

O algoritmo K-means é uma técnica de aprendizado não supervisionado amplamente utilizada, que visa agrupar dados em clusters com base em suas similaridades. Ele é um método iterativo que busca minimizar a soma dos quadrados das distâncias entre os pontos de um cluster e o centroide desse cluster (AGGARWAL, 2015).

2. MATERIAL E MÉTODOS

O método utilizado para análise dos dados educacionais foi dividido em três etapas como ilustrado pela Figura 1.

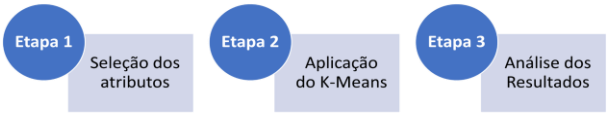

Figura 1 - Método para Análise dos Dados

[1]Bolsista PIBIC-EM/CNPq, IFSULDEMINAS – *Campus* Avançado Carmo de Minas. E-mail: laura.fernanda@alunos.ifsuldeminas.edu.br
[2]Bolsista PIBIC-EM/CNPq, IFSULDEMINAS – *Campus* Avançado Carmo de Minas. E-mail: guilherme.rangel@alunos.ifsuldeminas.edu.br
[3]Professora EBTT, IFSULDEMINAS – *Campus* Avançado Carmo de Minas. E-mail: juliete.costa@ifsuldeminas.edu.br

Na primeira etapa aconteceu a seleção de atributos para realizar a aplicação do algoritmo de clusterização, visto que a maioria destes algoritmos trabalham apenas com dados numéricos. Desta forma, de uma base de dados educacional original com 15 atributos e 3187 registros de sessões de acesso ao sistema (COSTA; DORÇA; ARAÚJO, 2020), foram selecionados apenas atributos numéricos que não tinham valores faltantes, são eles:

- *dispositivo*: tipo de dispositivo utilizado pelo estudante para acessar o sistema;
- *largura_de_banda*: largura de banda utilizada durante o acesso;
- *tempo_de_acesso*: tempo de acesso da sessão do estudante;
- *abertura_aula*: quantidade de vezes que o estudante abriu uma aula na sessão;
- *visualização_de_slide*: quantidade de vezes que o estudante visualiza um slide;
- *mudança_modelo*: quantidade de vezes que o usuário mudou a visualização do modelo estudante dentro da plataforma;
- *colaboração*: quantidade de colaborações do estudante durante a sessão, como, comentário, marcação de estrela para uma aula, dentre outras.
- *resposta_quiz*: quantidade de vezes que o estudante respondeu ao quis;
- *resposta_quiz_correta*: quantidade de acertos de quizzes do estudante dentro da sessão

A segunda etapa do processo de análise de dados foi a aplicação do algoritmo K-Means utilizando um código desenvolvido em Python com suporte da biblioteca de aprendizagem de máquina Scikit-learn[4] (BARANWAL; BAGWE; M, 2019) sobre os dados selecionados na etapa 1, com o objetivo de analisar correspondências entre as variáveis. Finalmente, a etapa 3 deste processo foi a análise dos resultados obtidos pelo algoritmo e considerações sobre a pesquisa e trabalhos futuros.

A implementação prática dessa técnica é realizada neste trabalho utilizando a linguagem de programação Python com suporte da biblioteca de aprendizagem de máquina Scikit-learn (BARANWAL; BAGWE; M, 2019) para realizar a análise dos dados.

3. RELATO DE EXPERIÊNCIA

A fim de verificar o melhor resultado da medida de avaliação do algoritmo K-Means, foi utilizada a medida coeficiente de silhueta que basicamente verifica quão bem os dados estão agrupados dentro de um cluster e como é possível analisar na Figura 2, quando se tem 2 clusters o coeficiente de silhueta alcança o valor mais próximo de 1.0, o que indica que os dados destes dois clusters estão agrupados de acordo com a sua similaridade próxima de 77%.

[4] *https://scikit-learn.org/stable/*

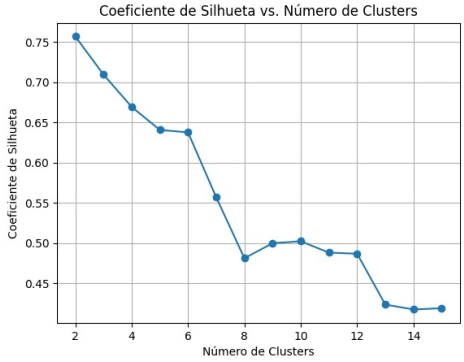

Figura 2 - Variação do Coeficiente de Silhueta X Número de Clusters

A aplicação do algoritmo K-Means sobre dados destacou pontos interessantes sobre as relações das variáveis. Ao analisar algumas variáveis dentro dos clusters gerados, observa-se que as variáveis “largura de banda” e “tempo de acesso” estão fortemente correlacionadas (veja Figura 3), indicando que alunos com internet melhor acessam mais o sistema. Além disso, foi observado que os alunos que possuem uma internet melhor e mais tempo de acesso possuem mais colaboração dentro da ferramenta, sugerindo que a relação entre esses dois dados tem um efeito positivo.

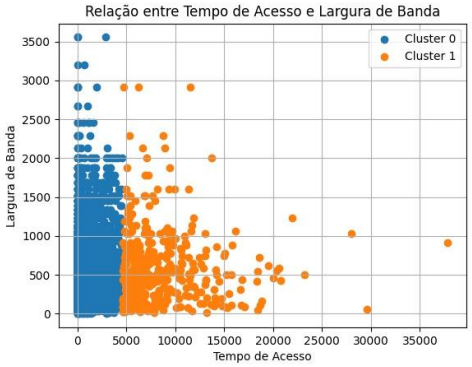

Figura 3 - Relação entre os atributos "largura de banda" e "tempo de acesso" para 2 clusters.

Observa-se também que a variável “*resposta_quiz*”, é maior entre alunos que possuem mais respostas certas nos exercícios (Figura 4). Essas tendências podem ser explicadas por fatores específicos e demandam uma investigação mais detalhada. A análise dos dados também revelou algumas limitações do estudo, como o tamanho da amostra e a heterogeneidade dos participantes. Esses fatores podem ter influenciado os resultados e devem ser considerados em análises futuras.

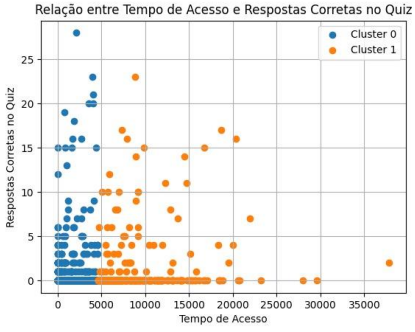

Figura 4 - Relação entre os atributos "tempo de acesso" e "respostas corretas aos quizzes" com 2 clusters.

4. CONSIDERAÇÕES FINAIS

Importante destacar que o objetivo deste trabalho foi destacar o relato da experiência de estudantes do ensino médio técnico integrado em informática em um primeiro contato com pesquisa cientifica na área de análise de dados e ao encerrar este estudo, pode-se afirmar que os resultados obtidos por meio da análise proporcionam de pontos interessantes obre a relação entre as diferentes condições investigadas. As evidências coletadas sugerem a presença de influências significativas que afetam os elementos estudados, destacando a importância de intervenções específicas.

Embora observe-se resultados promissores, é essencial reconhecer as limitações inerentes a este estudo, como o tamanho da amostra e possíveis variáveis não controladas. Essas limitações destacam a necessidade de investigações mais abrangentes e rigorosas para validar e expandir nossas descobertas. O trabalho realizado foi de grande importância para os estudantes, pois foi possível visualizar como utilizar uma linguagem de programação para analisar dados, algo totalmente diferente do que é visto no curso técnico. Como trabalhos futuros deste projeto pretende-se aplicar o algoritmo K-Means com outras configurações e outros algoritmos de agrupamento.

AGRADECIMENTOS

Agradecemos à Pró-Reitoria de Pesquisa, Pós-Graduação e Inovação (PPPI) e ao Programa Institucional de Bolsas de Iniciação Científica no Ensino Médio (PIBIC-EM) do IFSULDEMINAS

REFERÊNCIAS

AGGARWAL, Charu C. **Data mining: the textbook**. [*S. l.*]: Springer, 2015.

BARANWAL, Astha; BAGWE, Bhagyashree R.; M, Vanitha. Machine Learning in Python. [*s. l.*], v. 12, p. 128–154, 2019.

COSTA, Juliete A. R.; DORÇA, Fabiano A.; ARAÚJO, Rafael D. Avaliação do Comportamento de Estudantes em um Ambiente Educacional Ubíquo. [*s. l.*], n. Cbie, p. 182–191, 2020.

KUČAK, Danijel; JURIČIĆ, Vedran; ĐAMBIĆ, Goran. Machine learning in education - A survey of current research trends. **Annals of DAAAM and Proceedings of the International DAAAM Symposium**, [*s. l.*], v. 29, n. 1, p. 0406–0410, 2018.