Análise de Perfis de Engajamento de Estudantes de Ensino a Distância

Pedro H. R. Macêdo[1], **Wylliams B. Santos**[1], **Alexandre M. A. Maciel**[1]

[1]Escola Politécnica – Universidade de Pernambuco (UPE)

phrm@ecomp.poli.br, wbs@upe.br, amam@ecomp.poli.br

Resumo. *O uso de ambientes virtuais de aprendizagem permite mais interação no ensino a distância, mas ao mesmo tempo faz com que os professores tenham dificuldade em identificar seu aluno. Esse trabalho tem como objetivo detectar os perfis de engajamento dos estudantes de ensino a distância a partir de técnicas de agrupamento através dos bancos de dados de um ambiente virtual de aprendizagem. O estudo utilizou 2 algoritmos de agrupamento. Os perfis encontrados na base utilizada possuem uma fraca relação entre eles, porém, é possível detectar os traços comportamentais dos estudantes nas plataformas de ensino a distância.*
Palavras-Chave: *ambiente virtual de aprendizagem, mineração de dados educacionais, engajamento, agrupamento.*

Abstract. *The use of virtual learning environments allows more interaction in distance learning, but at the same time makes it difficult for teachers to identify their students. This work aims to detect the engagement profiles of distance learning students using clustering techniques through the databases of a virtual learning environment. This study used 2 clustering algorithms. The profiles found in the base used have a weak relationship between them, but it is possible to detect the behavioral traits of students in distance learning platforms.*
Keywords: *virtual learning environment, educational data mining, engagement, clustering.*

1. Introdução

O ensino a distância proporciona aos estudantes acesso facilitado à educação, contudo é sabido que este tipo de modalidade traz consigo dificuldades que requerem um alto grau de comprometimento, e, não por acaso em muitos casos resultam em abandono e descrédito (Macedo, et al. 2019).

Casos extremos com esses não são imediatos e sim fruto de um processo gradativo de falta de engajamento. Segundo (Hayati, et al. 2016) o engajamento pode ser definido como a habilidade de se envolver e completar determinada tarefa. Já (Beer, et al. 2010) afirma que o engajamento está associado ao desempenho escolar e ao sucesso institucional. No ensino a distância a análise de perfis de engajamento é fundamental para a melhoria no processo de ensino-aprendizagem e consequentemente diminuição de abandono escolar (Macedo, et al. 2019).

Os Ambientes Virtuais de Aprendizagem, plataformas amplamente utilizadas no ensino a distância, possuem diversos recursos educacionais que armazenam em suas bases de dados inúmeras informações sobre a interação e o desempenho dos estudantes. Esta ampla disponibilidade de dados tem proporcionado o desenvolvimento de muitos

trabalhos a fim de analisar dados e propor soluções (Martins, e Ribeiro, 2018), (Beer, et al. 2010), (Abu Tair, e El-Halees, 2012) a fim de compreender melhor os perfis de engajamento.

Apesar dos relevantes resultados obtidos na literatura, há uma grande dificuldade em definir um consenso para quais atributos utilizar nos estudos visto que existe uma grande variedade de variáveis que são utilizadas para realizar a análise dos perfis de engajamento. Este trabalho tem como objetivo realizar um levantamento das principais variáveis e identificar os principais perfis de engajamento presentes nos ambientes virtuais de aprendizagem. (Martins, e Ribeiro, 2018).

2. Fundamentação Teórica

2.1. Tipos de Engajamento

Segundo (Kahu, e Nelson, 2018) o engajamento do estudante é entendido como o seu estado psicossocial, esse é dividido em três tipos: o cognitivo, o emocional e o comportamental. O engajamento cognitivo é relativo às ambições do estudante, ou ao investimento psicológico necessário para estudar aquele conteúdo. Já o engajamento emocional está relacionado com as reações emocionais do estudante com sua escola ou professores. E por último, o engajamento comportamental é medido a partir das ações do estudante, seja entregar atividades, seguir os prazos ou fazer questionamentos (Kahu, e Nelson, 2018).

Trabalhos que abordam o engajamento comportamental costumam analisar a boa conduta do estudante, se o estudante costuma seguir regras escolares, se entrega as atividades dentro ou fora do prazo, além do nível de interação do estudante com os professores e tutores (Fredricks, et al. 2004). Considerando que os AVAs registram diversas ações realizadas pelo estudante dentro de sua plataforma, (Amaral, et al. 2015), é possível utilizar técnicas de mineração de dados educacionais para detectar esse tipo de engajamento sem a necessidade de contato direto com o estudante.

2.2. Técnicas de Agrupamentos

A análise de grupos pode ser definida como a organização de um conjunto de objetos (normalmente representados por vetores de características, ou seja, pontos em um espaço multidimensional) em grupos baseada na similaridade entre eles. Intuitivamente, objetos pertencentes ao mesmo grupo são mais similares entre si do que a objetos pertencentes a grupos distintos, explica (FERRARI, e SILVA, 2017).

Segundo (FERRARI, e SILVA, 2017), o processo de agrupamento consiste em quatro fases: I) Pré-processamento dos dados, II) Definição das medidas de similaridades, III) Execução do método de agrupamento, e IV) Avaliação do agrupamento. A Figura 1 mostra o esquema geral do processo de agrupamento.

Figura 1. Esquema geral do processo de agrupamento

Na fase de pré-processamento são realizadas limpeza e transformação dos dados a fim de eliminar ruídos e inconsistências na base de dados. A definição da medida de similaridade está diretamente relacionada ao tipo do dado a ser analisado. No contexto deste trabalho que utiliza dados contínuos a medida utilizada foi a Distância Euclidiana visto que possui a propriedade de representar a distância física entre pontos em um espaço m-dimensional (Tan, et al. 2009).

Os métodos de agrupamento podem pertencer a uma das seguintes categorias: Particionamento, Hierárquico, Baseado em Densidade, Baseado em Grade e Baseado em Modelo (Han, et al. 2011). Neste trabalho foram utilizadas técnicas de particionamento K-means (MacQueen, et al. 1967), e hierárquicos Algoritmo de Agrupamento Aglomerativo (Gowda, e Krishna, 1978). Por fim, para avaliar a qualidade dos grupos gerados o Método do Cotovelo (*Elbow Method*) o qual é comumente utilizado para calcular a melhor quantidade de grupos que um algoritmo pode gerar em determinada base de dados (Marutho, et al. 2018), e o Coeficiente de Silhueta (*Silhouette Index*) uma métrica utilizada apenas para técnicas de particionamento, como o K-means, cujo objetivo é demonstrar o quão alocado um objeto está em seu grupo (Rousseeuw, 1987).

3. Levantamento das Variáveis de Engajamento

O processo escolhido para levantamento das variáveis de engajamento se inspirou em (Kitchenham, 2004) que é composto por 3 fases: I) Planejamento, onde são definidas as Perguntas de Pesquisa e Estratégias de busca, II) Condução da revisão onde são analisados os resultados e III) Documentação, onde a pesquisa é reportada e documentada.

Para realização da pesquisa foram definidas as seguintes perguntas:

- P1: Quais são as variáveis utilizadas para detectar o engajamento do estudante no ambiente virtual de aprendizagem?
- P2: Existem grupos de alunos com comportamentos parecidos em cursos diferentes?

Como estratégia de pesquisa foram escolhidas cinco ferramentas de busca: IEEE-Explorer, ACM Digital Library, Science Direct, Springer e Eric. Foi realizada uma busca automática apenas por artigos científicos revisados por pares. A busca foi realizada no título, palavras-chaves e resumo. A String de Busca utilizada foi a seguinte:

((”E-learning”OR ”E-learn”OR ”Intelligent Tutor System”OR ”Virtual Learning”OR ”Moodle”OR ”LSM”) AND (”Engagement”OR ”Participation”OR ”Communication”OR ”Post Frequency”OR ”Number of Posts”OR ”Grades”OR ”Absorption”))

Os critérios de inclusão (CI) são: CI01 - Somente estudos relacionados a EAD; CI02 - Estudos que apresentam informações sobre engajamento ou participação em plataformas EAD; Como critérios de exclusão (CE), foram definidos: CE01 - O artigo não aborda o EAD; CE02 - O artigo aborda Ensino Semipresencial ou Presencial; CE03 - Estudos que não respondem nenhuma das perguntas de pesquisa; CE04 - Estudos duplicados em relatórios sem informação adicional; CE05 - Não acessíveis pela Internet; CE06 - Documentos incompletos, estudos terciários, rascunhos, slides de apresentações e resumos estendidos; CE07 - Estudos que não foram publicados no período entre 2001 e 2019; e CE08 - Estudos que não foram escritos em Inglês ou Português.

Após as buscas automáticas, foram encontrados 7.135 artigos. Para o primeiro filtro, foi checado o título, resumo e palavras-chaves dos artigos e aplicado os critérios de inclusão e exclusão, com isso, reduziu-se a amostra para 186 artigos. Após essa etapa,

todos os artigos foram lidos por completo e avaliados segundo o critério de inclusão e exclusão. Ao final 53 artigos foram selecionados o que resultou no levantamento de 69 variáveis de engajamento.

As variáveis consideradas importantes foram as que apresentaram cinco ou mais evidência nos artigos, são elas: Número de acessos à plataforma, Número total de postagens no fórum, Tempo na plataforma, Número de Atividades completas, Notas finais, Número de páginas visitadas, Número de visualizações por página, Número de respostas no fórum, Participação no fórum, Número de acessos ao fórum, Tempo gasto em atividades e Descrição do evento.

3.1. Metodologia Experimental

3.1.1. Descrição da base de dados

A base de dados utilizada neste trabalho foi gentilmente cedida pelo Núcleo de Ensino a Distância (NEAD) da Universidade de Pernambuco (UPE). Os dados foram gerados pela plataforma Moodle, a partir de quatro cursos de graduação que armazena um histórico de mais de 30 mil alunos entre o período de 2009 à 2016.

Para análise dos perfis de engajamento foram selecionados dois cursos com perfis de estudantes distintos (letras e biologia). No ambiente Moodle do NEAD foram encontradas todas as variáveis da tabela 2 contudo após a realização de uma análise descritiva observou-se que 8 dessas variáveis possuíam um baixo grau de dispersão nos dois cursos. Dessa forma foram selecionadas 5 variáveis para a realização do experimento, foram elas: Número de acessos ao fórum, Número de mensagens postadas no fórum pelo estudante, Número de atividades completas no prazo correto, Número de acessos ao AVA e Notas finais.

3.2. Pré-processamento

Cada curso apresenta valores diferentes relacionados à participação do estudante, sendo assim, não é possível considerar que uma alta participação em um curso represente alta participação nos outros cursos. Com o objetivo de analisar os valores apresentados considerando apenas o seu contexto, cada grupo foi comparado apenas aos outros grupos do curso no qual ele está contido, e cada variável foi analisada de forma a considerar apenas os valores do curso.

Para comparar os valores das variáveis: “Número de acessos ao fórum”, “Número de mensagens postadas no fórum” e “Número de acessos ao AVA” a seguinte equação foi formulada:

$$X = \frac{100Mv}{MGv}$$

Onde: v: Variável em questão Mv: Média aritmética da variável no grupo; MG: Média geral da variável naquele curso.

Com o uso dessa equação o valor de X foi categorizado aplicando as seguintes regras: Muito acima da média: X $>$ 190; Acima da média: X $>$ 125 e X $<=$ 190; Pouco acima da média: X$>$100 e X$<=$ 125; Média: X = 100; Pouco abaixo da média: X$>$75 e X$<$100; Abaixo da média: X $>$ 10 e X $<=$ 75; Muito abaixo da média: X$<=$ 10;

3.3. Instrumentação de Pesquisa

Para a implementação dos métodos de agrupamento escolhidos: K-means e Agrupamento Aglomerativo foi utilizada a biblioteca PyCaret, na versão 2.0. Essa é uma biblioteca que funciona no Python 3.6.

4. Resultados e Discussões

4.1. Curso de Letras

Utilizando o método do cotovelo para o algoritmo K-Means no curso de letras foi possível detectar que a quantidade de grupos ideal é de quatro grupos (Figura 2 (a)). Com o número ideal de grupos é possível calcular o coeficiente de silhueta que foi de aproximadamente 0,3 Figura (2 (b)). Esse valor representa que existe relação entre as variáveis em seus respectivos grupos, mas que é uma estrutura fraca.

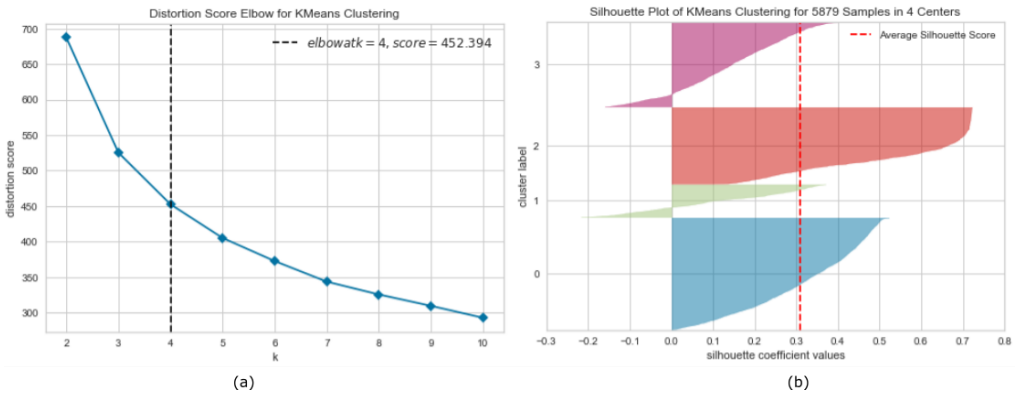

Figura 2. Numero de clusters para curso de letras usando K-means

Os perfis de engajamento identificados pelo K-Means apresentam uma forte relação de participação no fórum com o desempenho do estudante. Nenhum dos perfis encontrados apresentam uma média de 100% das atividades entregues. Os perfis são apresentados na Tabela 1.

	Grupo 1	Grupo 2	Grupo 3	Grupo 4
Acessos ao Fórum	Pouco acima da média	Muito acima da média	Abaixo da média	Pouco abaixo da média
Postagens no Fórum	Acima da média	Acima da média	Abaixo da média	Pouco abaixo da média
Acessos à Plataforma	Pouco acima da média	Muito acima da média	Abaixo da média	Abaixo da média
Desempenho Médio	7,03	6,78	0,84	3,5
Atividades Entregues	48,25%	47,50%	2,50%	26,25%

Tabela 1. Perfis do curso de letras gerados pelo K-Means

Utilizando o método do cotovelo para o algoritmo de Agrupamento Aglomerativo no curso de letras foi possível detectar que a quantidade de grupos ideal é de 4 grupos (Figura 3).

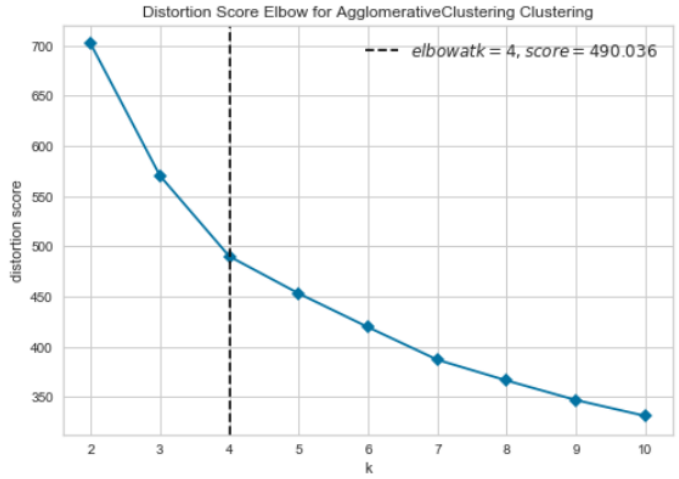

Figura 3. Numero de grupos para curso de letras usando Agrupamento Aglomerativo.

O Algoritmo de Agrupamento Aglomerativo apresentou dados semelhantes ao K-Means incluindo a quantidade de grupos. Contudo é possível analisar algumas pequenas diferenças nos perfis por exemplo o perfil com o menor desempenho não apresenta registros de atividades entregues no prazo. Os perfis são apresentados na Tabela 2.

	Grupo 1	Grupo 2	Grupo 3	Grupo 4
Acessos ao Fórum	Pouco acima da média	Muito acima da média	Abaixo da média	Abaixo da média
Postagens no Fórum	Acima da média	Acima da média	Pouco abaixo da média	Abaixo da média
Acessos à Plataforma	Pouco acima da média	Muito acima da média	Abaixo da média	Abaixo da média
Desempenho Médio	6,83	6,77	2,84	1,5
Atividades Entregues	47,75%	47,75%	29,25%	00,00%

Tabela 2. Perfis do curso de letras gerados pelo Agrupamento Aglomerativo

4.2. Curso de Biologia

Utilizando o método do cotovelo para o algoritmo K-Means no curso de biologia foi possível detectar que a quantidade de grupos ideal é de cinco grupos (Figura 4 (a)). Com o número ideal de grupos é possível calcular o coeficiente de silhueta que foi de aproximadamente 0,32 Figura (4 (b)). Esse valor representa que existe relação entre as variáveis em seus respectivos grupos, mas que é uma estrutura fraca.

Os perfis de engajamento identificados pelo K-Means apresentam uma forte relação entre a frequência de acessos e a quantidade de atividades entregues. Foi possível identificar que os perfis com alta participação no fórum apresentam também os melhores desempenho. Os perfis são apresentados na Tabela 3.

Assim como no curso de Letras, o Algoritmo de Agrupamento Aglomerativo apresentou dados semelhantes ao K-Means. É possível analisar algumas pequenas diferenças na quantidade de acessos ao fórum ou quantidade de atividades entregues por curso. Os perfis são apresentados na Tabela 4.

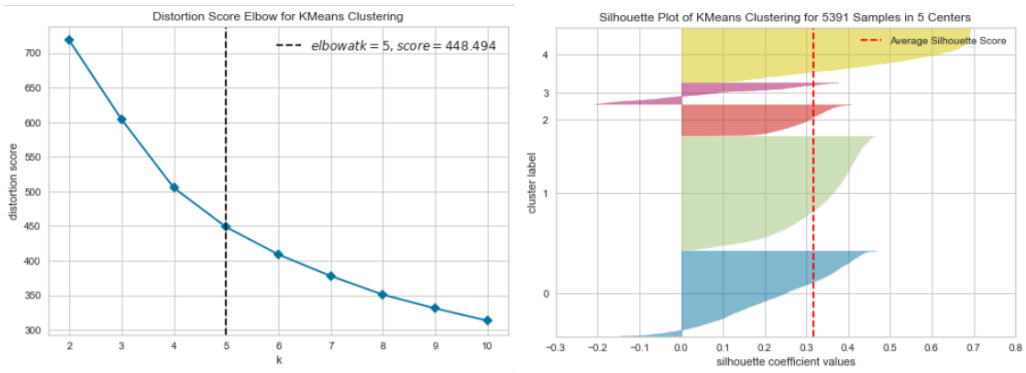

Figura 4. Numero de clusters para curso de biologia usando K-means.

‘cm	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
‘cm	Acessos acima da média Fórum	Pouco abaixo da média	Acima da média	Muito acima da média	Abaixo da média
‘cm	Postagens no Fórum	Pouco acima da média	Pouco acima da média	Acima da média	Abaixo da média
‘cm	Acessos à Plataforma	abaixo da média	Acima da média	Muito acima da média	Abaixo da média
‘cm	Desempenho Médio	7,29	6,22	6,64	1,17
‘cm	Atividades Entregues	65,60%	100,00%	61,60%	2,60%

Tabela 3. Perfis do curso de biologia gerados pelo K-Means

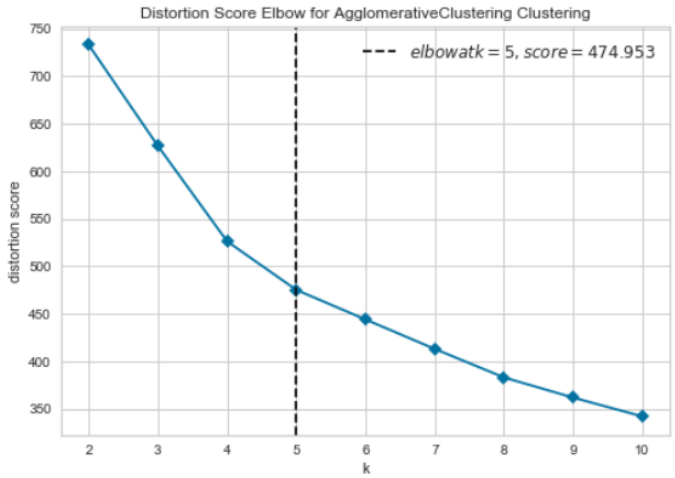

Figura 5. Numero de clusters para curso de biologia usando Algoritmo de Agrupamento Aglomerativo.

4.3. Discussões

As perguntas levantadas para essa pesquisa, foram utilizadas como guia para a análise dos resultados.

	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
Acessos ao Fórum	Pouco abaixo da média	Pouco abaixo da média	Muito acima da média	Abaixo da média	Pouco acima da média
Postagens no Fórum	Pouco abaixo da média	Pouco acima da média	Acima da média	Abaixo da média	Pouco acima da média
Acessos à Plataforma	Abaixo da média	Pouco abaixo da média	Muito acima da média	Abaixo da média	Pouco acima da média
Desempenho Médio	4,30	7,15	6,67	1,13	6,21
Atividades Entregues	26,70%	66,30%	67,35%	35,30%	100,00%

Tabela 4. Perfis do curso de biologia gerados pelo Algoritmo Aglomerativo

P1: Quais as variáveis utilizadas para detectar o engajamento do estudante no ambiente virtual de aprendizagem?

A partir do levantamento de dados feito para esse trabalho foi possível identificar 12 variáveis que são utilizadas para medir o engajamento dos estudantes, sendo que 5 delas recebem destaque por ter valores significantes e representativos na elaboração dos perfis, são elas: “Desempenho”, “Quantidade de acessos à plataforma”, “Atividades entregues no prazo”, “Quantidade de postagens no fórum” e “Quantidade de visualização ao fórum”.

As outras variáveis da base não apresentaram valores significativos e têm baixa correlação quando analisados. Porém, essas variáveis não devem ser descartadas, pois a análise do perfil de engajamento do estudante depende de como o curso é executado, e por isso são variáveis que devem ser consideradas em cursos de diferentes instituições de ensino.

P2: Existem grupos de alunos com comportamentos parecidos em cursos diferentes?

Os cursos aqui avaliados apresentaram média diferente de perfis, existindo 4 grupos no curso de letras e 5 no de biologia. Porém, nos perfis encontrados foi possível identificar que existem pontos de semelhança entre eles. Considerando o desempenho médio como um fator de saída e comparação, podemos notar que:

- O perfil que mantém o maior desempenho médio em todos os cursos é aquele perfil que apresenta alta participação no fórum, porém não a maior, alta frequência de acessos à plataforma, mas que não entregue 100% das atividades dentro do prazo.
- O perfil com menor desempenho médio em todos os cursos é aquele que apresenta a menor participação no fórum, tem um número muito baixo de acessos à plataforma, algo próximo de 10% da quantidade média de acessos do perfil com maior frequência de acessos, e costuma apresentar poucas atividades dentro do prazo, em alguma das vezes, nenhuma atividade.
- O perfil com maior quantidade de acessos à plataforma e maior participação no fórum, apresenta desempenho médio de aproximadamente 7 pontos, tendo uma média pouco menor que a média do perfil com o maior desempenho. Esse perfil

costuma apresentar 100% das atividades entregues dentro do prazo.

Os dois cursos apresentaram perfis com essas características. Sendo assim, é possível afirmar que existem perfis que apresentam o mesmo tipo de comportamento em cursos diferentes.

5. Conclusão e Trabalhos Futuros

Foi possível identificar que as variáveis encontradas não podem ser utilizadas para medir o comportamento do estudante de forma separada. As variáveis apresentadas podem fornecer valores diferentes de acordo com o curso e o AVA utilizado, portanto, é essencial destacar o problema que será analisado.

A mineração de dados apresentou resultados promissores quanto a utilização das variáveis encontradas. Sendo possível encontrar as variáveis mais representativas dentro da base de dados. Com as técnicas utilizadas para a mineração de dados foi possível detectar grupos comportamentais dos estudantes, o que representa seus perfis de engajamento.

Os perfis apontados pelo K-Means e pelo Algoritmo Aglomerativo apresentam semelhanças em sua grande maioria. Porém, é necessário a análise de outras técnicas de agrupamento para se obter resultados com melhor fator de agrupamento, visto que os algoritmos utilizados não apresentaram uma boa relação entre os grupos indicando que é possível detectar os perfis de engajamento, mas alguns indivíduos apresentam características encontradas em mais de um perfil.

Ao analisar os perfis encontrados nos cursos foi possível comprovar que os estudantes com comportamento mais participativo tendem a ter melhores resultados. Também foi possível detectar a importância da comunicação dentro do AVA no desempenho do estudante, visto que os participantes com melhores desempenhos apresentavam alta participação nos fóruns.

Como trabalhos futuros, pretende-se utilizar outras técnicas de agrupamento, como o Fuzzy C-Means, com o objetivo de obter grupos com melhores índices de alocação por itens. Também é pretendido analisar como as variáveis encontradas nessa pesquisa se comportam em outros cursos de outras instituições de ensino.

Finalmente, acredita-se que esta pesquisa pode apoiar estudos que avaliam o ensino à distância e usam técnicas de mineração de dados educacionais para avaliar o comportamento dos estudantes, focando principalmente em facilitar o trabalho do cientista de dados na seleção de variáveis, uma vez que as bases de dados dos AVAS são extensas e complexas.

6. Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Agradecimentos ao Núcleo de Ensino a Distância (NEAD) da Universidade de Pernambuco (UPE) por fornecer a base de dados para a realização dessa pesquisa.

Referências

Abu Tair, M. M. e El-Halees, A. M. (2012). Mining educational data to improve students' performance: a case study. *Mining educational data to improve students' performance: a case study*, 2(2).

Amaral, Y., Maciel, A., e Rodrigues, R. (2015). Development of a virtual assistant for alerts and notifications in a learning environment. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 26, page 742.

Beer, C., Clark, K., Jones, D., et al. (2010). Indicators of engagement. *Curriculum, technology & transformation for an unknown future. Proceedings ascilite Sydney*, pages 75–86.

FERRARI, D. G. e SILVA, L. N. D. C. (2017). *Introdução a mineração de dados*. Saraiva Educação SA.

Fredricks, J. A., Blumenfeld, P. C., e Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of educational research*, 74(1):59–109.

Gowda, K. C. e Krishna, G. (1978). Agglomerative clustering using the concept of mutual nearest neighbourhood. *Pattern recognition*, 10(2):105–112.

Han, J., Pei, J., e Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hayati, H., Tahiri, J. S., Idrissi, M. K., e Bennani, S. (2016). Classification system of learners engagement within massive open online courses. In *2016 4th IEEE International Colloquium on Information Science and Technology (CiSt)*, pages 527–530. IEEE.

Kahu, E. R. e Nelson, K. (2018). Student engagement in the educational interface: understanding the mechanisms of student success. *Higher Education Research & Development*, 37(1):58–71.

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26.

Macedo, M., Santana, C., Siqueira, H., Rodrigues, R. L., Ramos, J. L. C., Silva, J. C. S., Maciel, A. M. A., e Bastos-Filho, C. J. (2019). Investigation of college dropout with the fuzzy c-means algorithm. In *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*, volume 2161, pages 187–189. IEEE.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Martins, L. M. e Ribeiro, J. L. D. (2018). Os fatores de engajamento do estudante na modalidade de ensino a distância. *Revista Gestão Universitária na América Latina-GUAL*, 11(2):249–273.

Marutho, D., Handaka, S. H., Wijaya, E., et al. (2018). The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. In *2018 International Seminar on Application for Technology of Information and Communication*, pages 533–538. IEEE.

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.

Tan, P.-N., Steinbach, M., e Kumar, V. (2009). *Introdução ao datamining: mineração de dados*. Ciência Moderna.