Identificação de perfis de interação de estudantes de educação a distância por meio de técnicas de agrupamentos

Jorge Luis C. Ramos[1], Laís Fernanda L. Santos[1], João C. Sedraz Silva[1], Rodrigo L. Rodrigues[2]

[1]Universidade Federal do Vale do São Francisco - Juazeiro - BA – Brasil
[2]Universidade Federal Rural de Pernambuco - Recife-PE, Brasil.

Abstract. *According to the Ministry of Education, in 2017, the number of students of Distance Education (EAD) was approximately 7.8 million. This large number of students results in a large volume of data, which is stored in virtual learning environments, which have databases with relevant records of the educational context. Thus, with the application of educational data mining techniques, this work identified profiles of students of courses offered in the distance modality by Universidade Federal do Vale do São Francisco. It was possible to observe the distribution of clusters with Low, Medium and High Interaction as the main profile of the students. From this, it was possible to extract new knowledge that could contribute to the improvement of this modality in the institutions.*

Resumo. *Segundo o Ministério da Educação, em 2017, o número de estudantes de Educação a Distância (EAD) foi de, aproximadamente, 7,8 milhões. Essa quantidade de alunos resulta em um grande volume de dados, sendo estes armazenados em ambientes virtuais de aprendizagem, os quais possuem bases de dados com registros relevantes do contexto educacional. Desse modo, com a aplicação de técnicas de mineração de dados educacionais, este trabalho identificou perfis dos estudantes de cursos ofertados na modalidade a distância pela Universidade Federal do Vale do São Francisco. Foi possível observar a distribuição dos clusters com Baixa, Média e Alta Interação como perfil principal dos alunos. A partir disso, foi possível extrair novos conhecimentos que poderão contribuir para o aperfeiçoamento dessa modalidade nas instituições.*

1. Introdução

A educação a distância (EAD) assumiu um importante papel na capacitação e na formação continuada das pessoas. A modalidade promove um acesso amplo à educação, alcançando pessoas e regiões antes desprovidas de processos educativos nos vários níveis de formação.

No Brasil, segundo o Censo da Associação Brasileira de Educação a Distância (ABED – *www.abed.org.br*), em 2017/2018, existiam 7.773.828 alunos matriculados na modalidade, nos mais diversos tipos de cursos (livres, graduação, pós entre outros). São indicadores como esses que reforçam a necessidade maior de pesquisas na área, a fim de proporcionar melhores instrumentos, métodos, processos e abordagens para a EAD.

A quantidade de dados na área educacional tem crescido exponencialmente e gerenciar esses dados é um dos maiores desafios das instituições [Romero; Ventura 2013]. Por meio da Mineração de Dados Educacionais, (do inglês *Educational Data Mining* - EDM), é possível criar e utilizar modelos com o intuito de descobrir novos conhecimentos sobre esses dados [Rabelo *et al*. 2017].

A identificação de perfis de alunos possibilita detectar tendências de evasão ou reprovação, além de oferecer subsídios para ganhos pedagógicos, de forma a melhorar o planejamento dos cursos ofertados na modalidade EAD [Kampff; Reategui; Lima 2008]. Conhecer o perfil dos estudantes da EAD é importante, pois pode revelar grupos distintos, para os quais podem ser necessárias ações diferenciadas, como um melhor acompanhamento e direcionamento dos procedimentos para motivar e engajar os alunos.

DOI: 10.5753/cbie.sbie.2020.932

Diante das considerações aqui firmadas, este estudo visa apresentar um processo para identificar perfis de estudantes EAD, por meio de mineração de dados, especificamente usando técnicas de clusterização. Assim, este trabalho foi guiado pela questão norteadora: é possível identificar diferentes perfis de estudantes EAD por meio da utilização de técnicas de agrupamento, a fim de formar grupos de estudantes com perfis similares entre si e heterogêneos entre os grupos?

2. Fundamentos teóricos

Com base na revisão de literatura, esta seção tem como propósito apresentar algumas das definições e técnicas que fundamentam este trabalho.

2.1. Interações em sistemas de gerenciamento da aprendizagem

Os sistemas de gerenciamento de aprendizagem, conhecidos como LMS, do inglês *Learning Management Systems*, têm como função gerenciar atividades de ensino e aprendizagem *online*. Os LMS distribuem o material didático e funcionam como meio para a comunicação síncrona e assíncrona, contribuindo para os processos de ensino e interatividade entre professores e alunos. Esses sistemas são formados por banco de dados e por módulos de gestão, publicação de conteúdo e interatividade [Haguenauer; Mussi; Cordeiro Filho 2009].

Os dados registrados no LMS sobre as interações de estudantes permitem analisar o comportamento do discente e identificar seu perfil, de modo a melhorar o ensino do professor e a aprendizagem do aluno [Dias; Gasparini; Kemczinski 2009].

Há 3 (três) tipos de interações na EAD: aluno-conteúdo, aluno-professor e aluno-aluno [Moore, 1989]. A interação aluno-aluno é importante à medida que permite o trabalho em equipe, a aprendizagem colaborativa, além de contribuir para o desenvolvimento do aspecto social do aluno. A interação aluno-conteúdo representa o contato do estudante com o material instrucional, como slides, textos, vídeos [Haguenauer; Mussi; Cordeiro Filho 2009]. A interação aluno-professor é o que define o processo educativo, já que o aprendizado ocorre a partir do compartilhamento de conhecimento entre esses atores [Silva; Navarro 2011].

As interações são fundamentais para a aprendizagem. Além de trabalhar o aspecto social, o estudante sabe que não está sozinho, pois mesmo estando fisicamente distante, há sempre alguém para tirar suas dúvidas e ajudá-lo no seu aprendizado [Grossi; Moraes; Brescia 2013].

2.2. Perfis educacionais

Traçar perfis educacionais consiste em identificar grupos de alunos, analisar quais características em comum apresentam e os que diferem de outros grupos. Com o estudo das interações, é possível observar quais características e tipos de perfis educacionais, para aquela amostra, que levam a uma reprovação ou aprovação, assim como possíveis causas de dificuldades no aprendizado [Oliveira et al. 2017].

A identificação de perfis estudantes pode revelar problemas de aprendizagem, o que contribui para auxiliar na prevenção desses problemas, além disso, contribui para a melhora do rendimento dos estudantes [Gomes 2010].

2.3 Mineração de dados por agrupamento (clusterização)

A Mineração de Dados Educacionais (EDM), é a área que aplica as tarefas típicas de Mineração de Dados na área educacional. Por meio da EDM, é possível dar um significado prático para os dados extraídos de sistemas educacionais, tornando dados brutos significativos para instituições educacionais [Romero; Ventura 2013]. Nela, as mesmas tarefas da mineração de dados tradicional são utilizadas (classificação, agrupamento, associação, dentre outros), porém com um enfoque específico em contextos educacionais.

A tarefa de agrupamento, caracterizada em mineração de dados como aprendizagem não supervisionada, agrupa os objetos de estudo baseado em similaridades comportamentais. Geralmente, utiliza-se a distância entre dois objetos para estabelecer os grupos. Agrupamentos (*clusters*) bons

devem ter alta homogeneidade interna e alta heterogeneidade externa, ou seja, objetos do mesmo grupo devem ser parecidos e devem apresentar uma diferença significativa em relação a outros grupos. Assim, pode-se sintetizar grandes volumes de dados para um agrupamento de informações concisas e importantes do conjunto, extraindo conhecimento e desenvolvendo hipóteses em cima desses dados [Linden 2009].

Comumente, processos de agrupamentos são divididos em hierárquicos e não hierárquicos. Nos hierárquicos, não é necessário definir previamente a quantidade de *clusters* desejados. Neles também é possível representar resultados por meio de dendogramas, que mostram graficamente e de forma hierárquica, o grau de semelhança entre os grupos [Ramos et al. 2016].

Já os não hierárquicos dependem de fatores que serão determinados pelo projetista, incluindo o número de clusters. Essa característica pode pesar negativamente nos resultados gerados, a depender das configurações utilizadas. Dessa forma, pelo menos para um estudo inicial dos dados, é mais vantajoso o tratamento hierárquico, já que o dendograma mostra, graficamente, como os dados se relacionam e quais as distâncias entre eles [Linden 2009].

2.4. Trabalhos Relacionados

Nesta seção, é apresentado um estudo de trabalhos relacionados ao tema abordado, atendendo a dois objetivos específicos: identificar as variáveis relevantes de interação dos alunos em ambientes virtuais de aprendizagem e selecionar técnicas para a descrição de perfis de estudantes da EAD.

No estudo de Kampff, Rategui e Lima (2008), estudantes com necessidades semelhantes foram associados a um mesmo grupo. Dessa forma, foi possível identificar perfis com riscos de evasão e/ou reprovação, assim como verificar as variáveis que influenciam nesses problemas, e, desta maneira, por meio de alertas no sistema, auxiliar a instituição a realizar quaisquer intervenções necessárias. A análise ocorreu não apenas considerando os dados gerados, mas também foi estudada a relação entre esses dados com o sexo, idade, polo e curso do aluno. O experimento teve uma amostra de 161 estudantes, de 5 áreas de conhecimento diferentes, de uma mesma disciplina, em duas turmas.

Pinheiro et al. (2014) analisaram os dados oriundos de um curso de 480 horas para formação de aproximadamente 1400 agentes de inclusão digital, oferecido pelo LMS Moodle. Com o intuito de encontrar perfis de estudantes e analisar os comportamentos deles, foi escolhido o algoritmo *K-Means*. A quantidade de grupos estabelecidos foi de cinco (k=5), equivalentes às classes excelente, bom, regular, insuficiente e a última contendo os *outliers*.

Silva et al. (2015) propuseram agrupamentos de alunos com comportamentos semelhantes no Moodle, de forma a analisar essas interações, e, consequentemente, prever o desempenho de 200 alunos, distribuídos em 6 polos. Esses alunos estavam matriculados em uma disciplina com duração de 4 semanas e, ao final do curso, 161 foram aprovados, 15 reprovados por nota e 24 evadiram. A estratégia adotada para realizar o agrupamento foi utilizar o método de Ward e a distância euclidiana, dividindo os dados em 4 grupos.

Lira et al. (2016) extraíram do ambiente Moodle informações de 3195 estudantes, divididos em 248 cursos, em um intervalo de um ano. Para o agrupamento, foi utilizado o algoritmo *K-Means*, que dividiu a base em 5 grupos: muito bom, bom, regular, ruim e muito ruim. Após a definição desses grupos, foram realizadas classificações com os algoritmos *J48, JRip, SimpleCart e Random Forest* para se obter a acurácia no modelo. O melhor resultado foi obtido com o *Random Forest*, 98,27% de instâncias classificadas corretamente, e o pior com o *JRip*, com 96,95%. Os autores ainda destacam a grande quantidade de alunos que possuem desempenho insatisfatório ou que evadem na EAD, ressaltando assim a importância desse tipo de estudo para apontar sugestões para melhorar essa situação.

Os experimentos de Souza et al. (2016) utilizaram o Moodle para extrair informações de 2227 alunos de 10 cursos de graduação distintos. O trabalho teve intuito de analisar não só o comportamento dos alunos, mas também avaliar os 38 tutores destes cursos. Para o agrupamento, utilizou-se o algoritmo *K-Means*, criando três grupos: baixo desempenho, desempenho moderado ou

bom desempenho. Diferente dos outros trabalhos, neste a instituição pôde acompanhar o desempenho do aluno em tempo real, através de um núcleo de apresentação de dados.

A partir dessa análise, percebeu-se que o algoritmo *K-Means* é predominante nos estudos sobre agrupamentos e que os atributos como número de acessos ao LMS, número de acessos ao fórum, número de postagens em fóruns, número de atividades realizadas, quantidade de mensagens enviadas e recebidas e nota final, são os atributos mais usados para estabelecimento dos perfis dos alunos.

Outra informação importante obtida na análise dos trabalhos, é que predominou o processo de *Knowledge Discovery in Databases* (KDD) como método de extração de conhecimento dos dados. O KDD inclui atividades que são compostas basicamente, por cinco etapas: Seleção dos dados, Pré-processamento, Formatação, Mineração de Dados e Interpretação dos resultados [Fayyad, 1996].

3. Percurso Metodológico

Existem vários processos utilizados para padronizar e detalhar as etapas da descoberta de conhecimentos. O percurso metodológico escolhido para este trabalho é uma adaptação do KDD para contextos educacionais, conforme descrito por Romero e Ventura (2013). Na Figura 1, pode-se observar essas etapas.

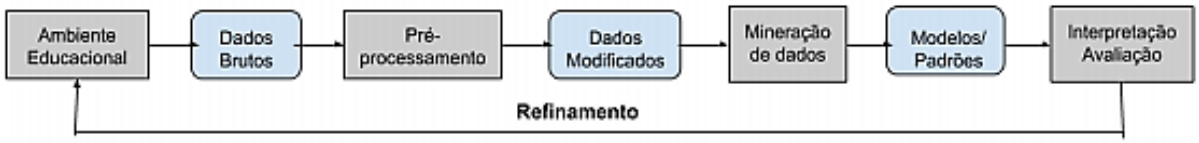

Figura 1: Etapas do processo de KDD em ambientes educacionais. Fonte: Adaptado de Romero e Ventura (2013).

Este processo começa com a coleta ou escolha da base para o estudo do ambiente educacional. Para este estudo, foi utilizada a base de dados do LMS Moodle que está em uso pela Secretaria de Educação a Distância (SEAD) de dois cursos de graduação na Universidade Federal do Vale do São Francisco (UNIVASF), com uma base que contém todos os registros das interações dos alunos dentro da plataforma, no período de 2013 até 2018, nos cursos de Administração Pública e Pedagogia, conforme quantitativos listados no Quadro 1.

Após a coleta, na seleção de dados, foram analisadas as informações extraídas, que são relevantes para o estudo, de modo a diminuir o volume de dados.

Quadro 1: Descrição da base de dados. Fonte: Autores (2019).

BASES	ALUNOS MATRICULADOS	DISCIPLINAS	Nº REGISTROS
Administração Pública	273	41	8.199
Pedagogia	142	39	3.855

Utilizando a análise das interações dos alunos no ambiente e a verificação do ganho de informação, por meio de bibliotecas da linguagem de programação *Python*, foram definidos os atributos considerados relevantes para o estudo.

Os dados brutos obtidos necessitaram de limpeza e pré-processamento (fusão de dados heterogêneos, tratamento de dados faltosos e incorretos, conversão de dados, seleção de recursos entre outros). Esta fase requereu a utilização de técnicas clássicas da mineração.

De posse dos atributos escolhidos, a próxima etapa consistiu no pré-processamento de dados, que tratou, por exemplo, dos *missing values* e *outliers*, de forma a assegurar a qualidade e confiabilidade dos dados. Assim, informações inconsistentes, ausentes e errôneas foram tratadas.

Na etapa de transformação, os dados foram ajustados para poderem ser utilizados pelo algoritmo escolhido, a partir de discretizações e normatizações necessárias.

Na etapa de mineração, o conhecimento foi então extraído. Como o objetivo do trabalho foi identificar grupos de estudantes e analisá-los, a tarefa escolhida foi o agrupamento, em seus dois

principais métodos: o hierárquico, como forma de subsidiar a escolha da quantidade de grupos e o não hierárquico, para realizar as devidas análises dos perfis nos grupos obtidos, conforme descrito por Ramos et al. (2016). A escolha do *K-Means* como método não hierárquico se deu por ser o mesmo bastante consolidado em estudos com dados educacionais, permitindo assim comparações ou adoção de estratégias similares e também com fácil implementação e parametrização por meio de bibliotecas.

Finalmente, o último passo é a interpretação e a avaliação dos resultados obtidos, onde o conhecimento é evidenciado e relatado. Lembrando, que por ser um processo iterativo, pode-se, em qualquer etapa, retornar para realizar quaisquer ajustes na base de dados e realizar novo processamento dessa base [Ramos et al., 2016].

O conhecimento adquirido a partir da EDM sobre os dados gerados pelos alunos e instrutores pode vir a fornecer rápidas e importantes compreensões acerca do desempenho, da motivação e do nível de participação dos alunos no curso. Este conhecimento pode sugerir mudanças no curso, intervenções significativas na metodologia ou mesmo um contato individual com alunos desmotivados ou com baixa interação.

4. Resultados e Discussões

A escolha das variáveis de interação foi baseada no levantamento dos trabalhos relacionados. Vale salientar que a ID de cada variável segue os scripts de outra coleta de dados, realizada por Ramos (2016a) em outro cenário educacional também na EAD e, como algumas variáveis foram reutilizadas, optou-se por manter os identificadores utilizados no trabalho citado. O Quadro 2 descreve essas variáveis.

Quadro 2: Atributos escolhidos para o experimento. Fonte: Ramos (2016).

ID	Descrição da variável escolhida
VAR01	Quantidade de postagens do aluno em fóruns por disciplina
VAR04	Quantidade de mensagens enviadas pelo aluno no semestre
VAR05	Quantidade de mensagens recebidas pelo aluno no semestre
VAR13	Quantidade de acesso por turno, 13(a) manhã, 13(b) tarde e 13(c) noite, no semestre
VAR16	Quantidade de colegas diferentes que o aluno enviou mensagens no semestre
VAR18	Quantidade de acessos do aluno ao ambiente no semestre
VAR21	Quantidade de mensagens recebidas de colegas no *LMS* no semestre
VAR22	Quantidade de mensagens enviadas para outros colegas no semestre
VAR31	Quantidade de acessos aos fóruns por disciplina

A coleta dessas informações foi realizada no software *MySQL Workbench*, utilizando consultas na linguagem SQL. Basicamente, os ID do estudante juntamente com o ID da disciplina formaram uma chave primária que serviu para coletar a interatividade de cada instância nas variáveis escolhidas. Destaca-se que o que é denominado de instância, representa o conjunto de variáveis que refletem o comportamento de um aluno em determinada disciplina, já que o mesmo aluno pode apresentar diferentes comportamentos em diferentes disciplinas.

4.1. Agrupamento Hierárquico

Essa etapa foi realizada com o auxílio de programa criado na linguagem *Python*, utilizando o Método de Ward com os parâmetros *default* da biblioteca Pandas para este tipo de agrupamento. O objetivo foi, através da observação do dendograma, estimar um possível melhor "k" para realizar a clusterização não hierárquica. A partir dos gráficos gerados, pôde-se obter uma linha de corte horizontal que cruzava os eixos de cada grupo, possibilitando uma escolha apropriada para o número de grupos.

Mesmo na perspectiva da realização de clusterização não hierárquica para fins de determinação dos perfis de cada grupo de estudante formado, escolha essa por ser a mais encontrada nos trabalhos relacionados (Seção 2.4), esta clusterização hierárquica serviu para indicar uma tomada de decisão acerca do número de *clusters* que iria ser usado como parâmetro para a outra técnica de

agrupamento escolhida para análise final.

Para cada curso, foram obtidos os dendogramas representativos dos grupos formados, conforme mostrados nas Figuras 2 e 3, nas quais foi observável que um possível valor para o parâmetro Número de *Clusters* (k) poderia ser k=3, já que contemplaria grupos sem uma grande discrepância de tamanhos e com boa representatividade em relação aos dados.

Em ambos os gráficos do tipo dendogramas, o eixo x corresponde aos dados normalizados de cada instância e o eixo y corresponde à distância euclidiana calculada para cada ponto. Estudantes com variáveis com valores mais similares tendem a ficar em uma mesma ramificação, que por sua vez, são ligadas em outras ramificações similares a partir de certas distâncias euclidianas próximas. Para o curso de Administração Pública, os *clusters* ficaram assim como mostrados na Figura 2.

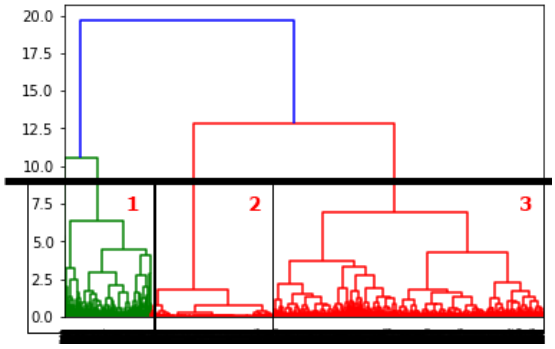

Figura 2: Dendograma do curso de Administração para 3 clusters. Fonte: Autores (2019).

Inicialmente, poderia ter sido feita a opção por dois grandes *clusters*, identificados no gráfico nas cores verde e vermelho. Nesse caso, teríamos um corte horizontal das linhas azuis, a partir do valor 12.5 no eixo y. Entretanto, os dados em vermelho possuem uma primeira subdivisão representativa, o que sugeria a necessidade de dividi-lo em dois *clusters* menores, o que assim foi feito, a partir da linha de corte horizontal no gráfico.

Observa-se que o *cluster* 1 reuniu dados com maiores distâncias euclidianas, o que indica que, alguns dados apresentam uma maior discrepância em relação aos demais, mesmo estando em um mesmo grupo. Uma possível razão disso poderia ser as características das variáveis coletadas para cada instância, onde algumas delas têm uma variabilidade bem maior que as demais. Inversamente, o *cluster* 2 apresenta os dados com menor distância euclidiana, sendo, portanto, mais homogêneo em relação aos demais. Por fim, o *cluster* 3, com maior número de instâncias entre os grupos formados, poderia ser subdividido em dois outros *clusters*, mas optou-se por mantê-lo integralmente, para fins de utilização nas análises posteriores.

Para o curso de Pedagogia, os *clusters* ficaram distribuídos conforme a Figura 3.

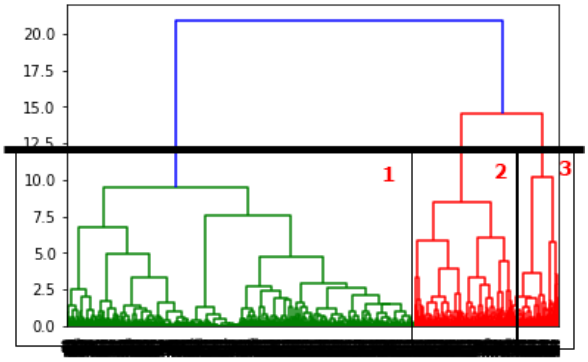

Figura 3: Dendograma do curso de Pedagogia para 3 clusters. Fonte: Autores (2019).

A mesma observação sobre os dois grandes *clusters* do gráfico do outro curso pode ser aplicada também neste caso. Observa-se que o *cluster* 1 agrupou mais da metade dos dados,

destacando-se em relação aos demais. Apesar do tamanho, boa parte dos seus elementos mantêm uma distância menor entre os centros, diferentes dos demais grupos. Os *clusters* 2 e 3 foram obtidos a partir da subdivisão de um *cluster* maior, a fim de se manter o padrão de 3 grupos para cada base.

4.2. Agrupamento Não Hierárquico

Após definido o número de 3 *clusters* e utilizando o algoritmo *K-Means*, com os parâmetros *default* da biblioteca Pandas para o algoritmo, foi obtida a distribuição dos dados nos 3 *clusters*, com a adição de uma coluna "*ClusterID*" que rotulou cada instância em cada uma das bases usadas. A Tabela 1 mostra a distribuição das instâncias em cada um dos grupos.

Tabela 1: Distribuição das instâncias nos grupos

	Número de instâncias		
	***Cluster* 0**	***Cluster* 1**	***Cluster* 2**
Administração	3408	3721	1070
Pedagogia	2393	437	1031

Essa divisão foi realizada com base no *K-Means*, onde foram calculados 3 centroides para cada variável e cada instância foi classificada de acordo com o *cluster* no qual era mais próxima, de forma a agrupar dados semelhantes.

4.3. Análise dos *clusters* por cursos

A partir da separação dos dados em cada *cluster*, foram obtidas estatísticas descritivas básicas de todas as variáveis nos cursos. Assim, foi possível compreender e descrever melhor os grupos formados e então associar cada um desses grupos a um perfil predominante dos estudantes da EAD em cada grupo. Para essa análise, foram considerados os valores originais das variáveis em cada instância (sem a normalização), acrescidos da ID do *cluster* ao qual pertence cada registro.

4.3.1 *Clusters* de Administração Pública

Para verificar a semelhança entre os *clusters*, uma das abordagens escolhidas para este estudo foi a análise de estatísticas descritivas das variáveis coletadas. Foi observada como essas variáveis se comportam por *cluster*, conforme a média e mediana de cada uma. A mediana foi usada como complemento às análises por se tratar de um indicador que não é suscetível a erros por conta de valores muito discrepantes dos demais (*outliers*). Como o resultado de ambas as análises não evidenciou diferenças nos resultados, por limitação do espaço deste estudo, serão apresentados e analisados os dados obtidos com as médias.

A Tabela 2 apresenta os dados das médias das variáveis nos três *clusters* obtidos. O Gráfico de densidade com esses dados são apresentadas na Figura 4.

Tabela 2: Médias das variáveis nos *clusters* do curso de Administração Pública.

CLUSTER	**VAR01**	**VAR04**	**VAR05**	**VAR13a**	**VAR13b**	**VAR13c**	**VAR16**	**VAR18**	**VAR21**	**VAR22**	**VAR31**
Cluster0	0,57	4,42	39,50	7,60	9,91	10,09	0,28	28,18	0,94	0,54	3,33
Cluster1	2,59	35,46	107,34	33,76	47,30	53,83	1,84	138,78	6,78	5,82	16,37
Cluster2	4,82	94,69	168,07	80,62	101,17	105,14	4,40	297,22	30,20	31,65	37,86

Cabe destacar, como forma auxiliar na compreensão do gráfico, a variável de maior valor absoluto nos *clusters* 1 e 2 (VAR18), corresponde à *quantidade média de acessos do aluno ao ambiente no semestre*, justificando assim a sua amplitude.

Já as variáveis com menores valores absolutos foram a VAR01 (Quantidade média de postagens do aluno em fóruns por disciplina) e VAR16 (Quantidade média de colegas que o aluno enviou mensagens pelo ambiente, no semestre).

As possíveis explicações para esse baixo número, é que nem toda disciplina tinha fórum de discussão habilitado ou então quando tinha, podia não ser avaliativo, não despertando o interesse do aluno em postar. A baixa quantidade de mensagens também pode ser explicada pelo uso de outras

formas de comunicação fora do ambiente, como e-mail, aplicativos de mensagens, redes sociais entre outros.

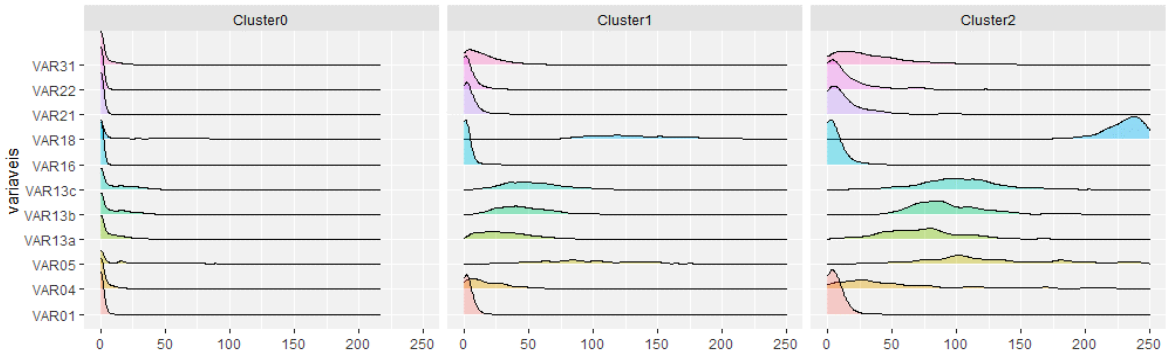

Figura 4: Gráfico de densidade das médias das variáveis em cada *cluster*, do curso de Administração Pública. Fonte: Autores (2020).

Também foi evidenciado uma predominância do acesso ao ambiente nos turnos noturno e vespertino. Isso pode caracterizar um perfil de alunos que exercem uma atividade matutina.

Ao analisarmos os dados das médias e medianas das variáveis nos *clusters*, percebe-se que os grupos podem ser descritos, de forma macro, da seguinte forma:

Cluster* 0 (3.408 Instâncias) -** Representado pelos estudantes com valores mais baixos em todas as variáveis coletadas. Ou seja, são aqueles com menor interação nos fóruns, que enviam e recebem menos mensagens e também os que menos acessaram o ambiente virtual durante o curso. Nesse *cluster*, foram alocadas cerca de 41,5% do total das instâncias do curso. A esse perfil foi dado o nome de ***Grupo de Baixa Interação.

Cluster* 1 (3.721 Instâncias) -** É o maior grupo e é representado pelos estudantes com valores intermediários nas variáveis coletadas, com cerca de 45,4% do total das instâncias do curso. A esse perfil foi dado o nome de ***Grupo de Média Interação.

Cluster* 2 (1.070 Instâncias) -** Representado pelos estudantes com os maiores médias e medianos em todas as variáveis coletadas, representando aproximadamente 13,1% das instâncias do curso. São, portanto, os alunos que mais interagem a partir de fóruns, acessam mais o ambiente e trocam mais mensagens seja em números totais tanto em números totais quanto em número de diferentes colegas com quem interagiu. A esse perfil foi dado o nome de ***Grupo de Alta Interação.

4.3.2 *Clusters* de Pedagogia

De forma análoga ao desenvolvido para o curso de administração pública, os dados de pedagogia foram agrupados em 3 *clusters*. A Tabela 3 apresenta os dados das médias das variáveis nos três *clusters* obtidos. O Gráfico é apresentado na Figura 5.

Tabela 3: Médias das variáveis nos *clusters* do curso de Pedagogia.

CLUSTER	VAR01	VAR04	VAR05	VAR13a	VAR13b	VAR13c	VAR16	VAR18	VAR21	VAR22	VAR31
Cluster0	2,48	13,85	48,03	13,50	23,74	33,03	1,43	72,36	4,02	2,65	15,49
Cluster1	6,10	89,77	127,53	38,57	59,69	76,05	27,48	179,36	56,97	65,83	57,23
Cluster2	4,81	38,84	85,68	44,64	69,27	105,07	2,27	226,34	7,00	6,58	44,87

A distribuição das instâncias em cada *cluster* apresenta características quase semelhantes à distribuição do curso de administração pública, com 1 *cluster* predominando em relação aos demais (*Cluster* 1). A diferença, em particular, refere-se às variáveis relacionadas ao acesso do aluno ao ambiente (13a, 13b, 13c e 18), que registraram valores abaixo do *Cluster* 2, embora ainda assim significativos. Uma possível causa é uma diferença menor verificada entre valores de outras variáveis dos dois *clusters* (01, 05 e 31). O um estudo sobre a importância das variáveis em cada *cluster* poderia confirmar essa causa ou apontar outra.

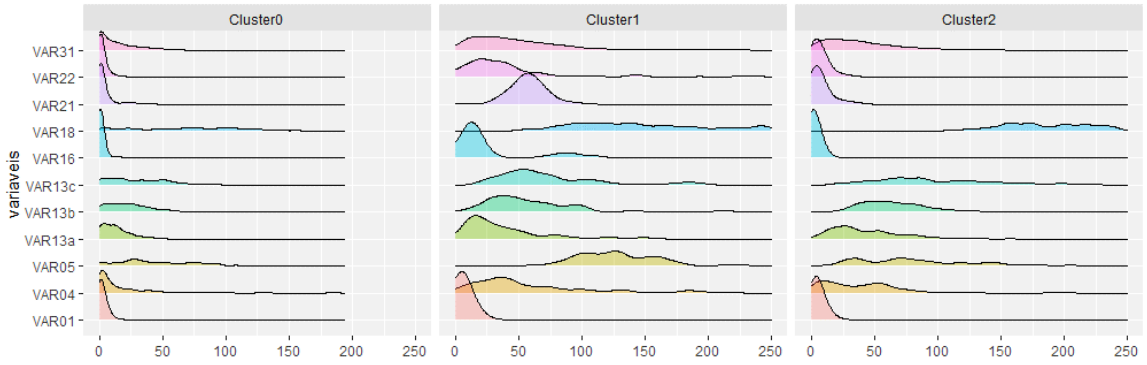

Figura 22: Gráfico de densidade das médias das variáveis em cada *cluster*, do curso de Pedagogia. FONTE: Autores (2020).

Ao analisarmos os dados das médias e medianas das variáveis nos *clusters*, percebe-se que os grupos podem ser descritos, de forma macro, da seguinte forma:

***Cluster* 0 (2.393 Instâncias) -** O maior *cluster* formado, com quase 62% das instâncias, mas que representa os estudantes com valores mais baixos em todas as variáveis coletadas (**Baixa Interação**). Isso indica que a maioria dos alunos do curso possuía esse perfil o que pode refletir, por exemplo, em baixos desempenhos ou mesmo em uma maior evasão.

***Cluster* 1 (437 Instâncias) -** É o menor grupo, com 11,3% das instâncias, mas inclui os estudantes com os maiores valores nas variáveis acesso e envio de postagens em fóruns e mensagens (**Alta Interação**).

***Cluster* 2 (1.031 Instâncias) -** Representado pelos estudantes com valores intermediários nas variáveis coletadas (**Média Interação**), excetuando-se pelas variáveis de acesso.

5. Considerações Finais e Trabalhos Futuros

O uso de técnicas de mineração de dados em contextos educacionais é um campo de pesquisa interdisciplinar emergente. A EDM é focada no desenvolvimento de métodos para explorar os diversos tipos de dados provenientes de ambientes educacionais. Seu objetivo é entender melhor como os alunos aprendem e identificar as configurações em que aprendem, para melhorar os resultados e obter conhecimentos e explicar os fenômenos educacionais.

Dessa forma, este trabalho buscou, a partir de técnicas de EDM, conhecer os perfis dos estudantes de graduação na EAD na UNIVASF utilizando dados das interações no ambiente virtual em uso, foram descobertas possíveis variações nos perfis em relação aos diferentes cursos onde os alunos são vinculados.

A utilização de 3 *clusters* possibilitou a definição de 3 tipos de interação: a baixa, média e alta. Foi possível identificar as médias e medianas bem definidas diferenciando as interações.

Como sugestões, alguns trabalhos futuros podem ser desenvolvidos, tais como:

- Fazer análises específicas por disciplina, a fim de perceber indícios de práticas pedagógicas inovadoras e motivadoras por parte dos professores.
- Usar técnicas de aprendizagem supervisionadas, como a classificação ou regressão, para predizer comportamentos ou outras características dos alunos.
- Desenvolver uma solução que implemente a coleta e análise de dados de forma automática, com a exibição direta dos grupos formados, sinalização de alunos com dificuldades ou baixa interação.
- Discutir com especialistas da Universidade, a necessidade de agregar ou retirar variáveis, dentro do contexto acadêmico da EAD local.

Espera-se que os resultados obtidos possam subsidiar gestores, professores, tutores e alunos nos procedimentos e tomadas de decisão que possam fortalecer a modalidade EAD.

Referências

DIAS, C. C. L.; GASPARINI, I.; KEMCZINSK, A. Identificação dos estilos cognitivos de aprendizagem através da interação em um Ambiente EAD. In: XVII Workshop sobre Educação em Informática (WEI), p. 489-498, 2009.

FAYYAD, U. *et al*. The KDD process of extracting useful knowledge from volumes of data. Communications of the ACM. p. 27- 34, 1996.

GOMES, C. M. A. Perfis de estudantes e a relação entre abordagens de aprendizagem e rendimento escolar. Psico, Porto Alegre, v. 41, n. 4, p.503-509, 2010.

GROSSI, M. G. R.; MORAES, A. L.; BRESCIA, A. T. Interatividade em ambientes virtuais de aprendizagem no processo de ensino e aprendizagem na educação a distância. Rev. Arquivo Bras. Educ., v.1, n.1, p. 75-92, 2013

HAGUENAUER, C.J.; MUSSI, M.V.; CORDEIRO FILHO, F. Ambientes virtuais de aprendizagem: definições e singularidades. Revista Educaonline, Rio de Janeiro (RJ), v. 3, n. 2, p. 1-23, 2009.

KAMPFF, A. J. C.; REATEGUI. E. B; LIMA, J. V. Mineração de dados educacionais para a construção de alertas em ambientes virtuais de aprendizagem como apoio à prática docente. RENOTE, v. 6, n. 1, 2008.

LINDEN, R. Técnicas de agrupamento. Revista de Sistema da Informação da FSMA, n. 4, p. 18-36, 2009.

LIRA, K. C. *et al*. Utilizando mineração de dados e sistemas multiagentes na análise da evasão em educação a distância por meio do perfil dos alunos. In: Encontro Nacional de Inteligência Artificial e Computacional, 2016.

MOORE, Michael G. Three types of interaction. 1989.

OLIVEIRA, M. *et al*. Mapeamento automático de perfis de estudantes em métricas de software para análise de aprendizagem de programação. In Simpósio Brasileiro de Informática na Educação-SBIE, v. 28, p.1337, 2017.

PINHEIRO, Márcia F. *et al*. Identificação de grupos de alunos em ambiente virtual de aprendizagem: Uma estratégia de análise de log baseada em clusterização. In: Anais dos Workshops do Congresso Brasileiro de Informática na Educação. 2014. p. 582.

RABELO, D.S.S. *et al*. Utilização de técnicas de mineração de dados educacionais para predição de desempenho de alunos EaD em ambientes virtuais de aprendizagem. Anais do Simpósio Brasileiro de Informática na Educação (SBIE), Recife-PE, 2017.

RAMOS, J. L. C *et al*. A comparative study between clustering methods in educational data mining. IEEE *Latin America Transactions*, v. 14, n. 8, p. 3755-3761, 2016.

RAMOS, J. L. C. Uma abordagem preditiva da evasão na educação a distância a partir dos construtos da distância transacional. Tese de Doutorado, CIn-UFPE. 2016a.

ROMERO, Cristobal; VENTURA, Sebastian. Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, v. 3, n. 1, p. 12-27, 2013.

SILVA, O. G.; NAVARRO, E. C. A relação professor: aluno no processo ensino-aprendizagem. Rev eletrônica Univar [Internet]. 2011 3 (8): 95-100.

SILVA, Ricardo *et al*. Mineração de dados educacionais na análise das interações dos alunos em um Ambiente Virtual de Aprendizagem. In: (Simpósio Brasileiro de Informática na Educação-SBIE). 2015. p. 1197.

SOUZA, R. *et al*. Um Ambiente Inteligente de Avaliação de Comportamentos de Tutores e Turmas no Ambiente Virtual de Aprendizagem Moodle. In Anais dos Workshops do Congresso Brasileiro de Informática na Educação. vol. 5, n. 1, p. 417, 2016.