ISSN 1982-5587

Artigos teóricos

Perfil dos estudantes retidos na Universidade do Estado de Mato Grosso: análise de uma Instituição Pública Pantaneira

Profile of students retained at the Universidade do Estado de Mato Grosso (State University of Mato Grosso): analysis of a Pantanal Public Institution

Fernando Cezar Vieira Malange[1*], Marcos Paulo de Mesquita[2]

[1]Universidade do Estado de Mato Grosso (UNEMAT), Faculdade de Ciências Exatas e Tecnológicas, Programa de Pós-graduação em Educação (PPGEdu), Cáceres, MT, Brasil
[2]Universidade do Estado de Mato Grosso (UNEMAT), Faculdade de Ciências Exatas e Tecnológicas, Cáceres, MT, Brasil

COMO CITAR: MALANGE, F. C. V.; MESQUITA, M. P. Perfil dos estudantes retidos na Universidade do Estado de Mato Grosso: análise de uma Instituição Pública Pantaneira. Revista IberoAmericana de Estudos em Educação, Araraquara, v. 19, esp. 3, e19455, 2024. eISSN: 19825587. DOI: https://doi.org/10.21723/riaee.v19i00.1945501

Resumo

O presente trabalho traz uma abordagem baseada em clusterização (agrupamento) de dados educacionais como proposta alternativa para compreender o fenômeno da retenção nos cursos de graduação da Universidade do Estado de Mato Grosso (UNEMAT) analisando os diferentes perfis dos estudantes retidos na instituição. Neste Estudo de Caso, os sujeitos foram os estudantes retidos na UNEMAT no ano de 2019 dos cursos regulares de graduação. Foi considerado retido o estudante que não obtivera diploma, ainda que passado o tempo mínimo regulamentar de conclusão de curso. Dessa definição foi possível identificar no nosso banco de dados 2.169 acadêmicos nessa condição. Deste ponto, três perfis de retenção foram obtidos por meio do algoritmo de clusterização *K-Modes*, (técnica de Mineração de Dados) e com os resultados foi possível identificar um comportamento de persistência dos estudantes retidos da UNEMAT e fornecer insights importantes para a instituição sobre possíveis estratégias de intervenção e apoio aos estudantes em risco de evasão.

Palavras-chave: retenção na educação superior; clusterização; dados educacionais.

Abstract

This paper presents an approach based on clustering of educational data as an alternative proposal to understand the phenomenon of retention in undergraduate courses at the Universidade do Estado do Mato Grosso (State University of Mato Grosso) (UNEMAT) by analyzing the different profiles of students retained at the institution. In this Case Study, the subjects were students from on-campus undergraduate courses retained at UNEMAT in 2019. Students who did not obtain a diploma were considered retained, even after the minimum regulatory time for course completion. From this definition, it was possible to identify 2,169 students under this condition in our database. From this point, three retention profiles were obtained through the K-Modes clustering algorithm (Data Mining technique) and with the results, it was possible to identify a behavior of persistence from the retained students at UNEMAT and provide important insights for the institution on possible intervention and support strategies for students at risk of dropping out.

Keywords: retention in higher education; clustering; educational data.

INTRODUÇÃO

A Educação Superior brasileira experimentou tanto uma franca expansão na oferta de vagas na década de 1990 e no início dos nos 2000, quanto um fomento na ocupação de vagas nas universidades particulares. Como propulsores deste incremento podemos citar o

***Autor correspondente:**
fmalange@unemat.br

Submetido: Julho 06, 2024
Revisado: Agosto 09, 2024
Aprovado: Setembro 17, 2024

Fonte de financiamento: nada a declarar.
Conflitos de interesse: Não há conflitos de interesse.

Aprovação do comitê de ética: Não aplicável.
Disponibilidade de dados: Não há conjunto de dados ou material disponível on-line ou para consulta. O estudo foi realizado no Programa de Pós Graduação em Educação, Universidade do Estado de Mato Grosso (UNEMAT), Cáceres, MT, Brasil.

Programa de Financiamento Estudantil (FIES), o Programa Universidade para Todos (PROUNI), a Reestruturação e Expansão das Universidades Federais (REUNI), a criação dos Institutos Federais de Ciência e Tecnologia (IFET). As políticas de reserva de vagas vão promover também maior acesso ao ensino superior àqueles que historicamente estiveram marginalizados desse nível, configurando-o, pelo menos em tese, como um ambiente plural e democrático.

A despeito desta expansão, as políticas de permanência na universidade brasileira não são implementadas na mesma proporção da oferta de novas vagas. Muitos estudantes por razões diversas acabam por não concluir seus estudos, como revelado por Nunes e Pereira (2019), entre os anos de 2006 e 2016, houve um salto importante no número de não-concluintes (69,06%), nas condições de estudantes retidos ou evadidos no ensino superior do Brasil (Brasil, 1996).

Alinham-se ao cenário nacional os números da UNEMAT de vagas ofertadas, ingressos e evasão e, neste último, podendo em alguns cursos atingir o alarmante índice de 50% (Hofffmann; Bitencourt, 2019; Nodari; Lima; Maciel, 2018).

O Censo da Educação Superior (CENSUP) de 2020 (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2020) revelou que no ano de 2019, 21.786 estudantes possuíam vínculo com algum dos cursos presenciais de graduação da UNEMAT. Destes, 2.120 diplomaram, 2.501 se desvincularam da instituição e 2.455 já haviam ultrapassado o tempo regulamentar mínimo de integralização de seus cursos. O censo apontou também que o tempo médio de duração dos cursos da instituição fora 4,5 anos, ao passo que os estudantes diplomados estudaram em média 5,57 anos na instituição. Uma parcela de 810 dos estudantes que se desvincularam em 2019, estiveram na universidade por pelo menos 4 anos.

Desses dados, vemos que a Universidade do Estado de Mato Grosso tem questões de retenção latentes e que seria razoável supor que as taxas de diplomação da universidade seriam maiores se as retenções pudessem ser convertidas em diplomação, ou ainda, se as evasões pudessem ser convertidas em permanência e depois futura conclusão de curso.

Sabemos que não é trivial explicar o porquê dessas retenções, mas conhecer o perfil daqueles estudantes que se encontram nessa situação é um passo importante para melhor direcionar esforços de gestão, políticas pedagógicas e estratégias de mitigação de evasão e/ou retenção.

Esta pesquisa aborda o estudo de retenção em uma universidade multicampi, que oferece diferentes cursos nas mais diversas áreas do conhecimento, atendendo dezenas de cidades do estado de Mato Grosso. Ao não nos preocuparmos com o caráter preditivo dos dados, ou explicativo do fenômeno, nosso foco foi analisar os perfis dos estudantes que atualmente se encontram retidos na instituição por meio de clusterização de dados.

A clusterização de dados tem por base, analisar um conjunto de dados, que no nosso caso foram os registros de estudantes, e agrupá-los em função de suas similaridades. Entre as suas vantagens, como apontado por Betarelli Junior (2016) estão a possibilidade de avaliar a dimensionalidade, identificar *outliers*[1] e sugerir hipóteses acerca das estruturas de relações. Além de ser uma importante técnica exploratória de dados.

Mas tão somente a constituição dos *clusters* não nos habilita avançar nas questões da retenção. Historicamente, o termo tem sido constituído, e a análise dos *clusters* foi feita à luz dessas teorias que tentam dar conta das multifacetas do fenômeno.

Diferentes abordagens e perspectivas têm sido utilizadas nos modelos teóricos de retenção de estudantes nas últimas décadas com foco nas altas taxas de evasão e retenção de estudante na Educação Superior. Muitas delas parecem ser pouco conhecidas dos educadores, administradores e formuladores de políticas das instituições de ensino em função das suas complexidades. O modelo teórico de Tinto (1975), amplamente estudado e tomado como base para diversas outras propostas, argui que o processo de retenção se dá por meio das

[1] Os *outliers* são dados que se diferenciam drasticamente de todos os outros. Em outras palavras, um outlier é um valor que foge da normalidade e que pode (e provavelmente irá) causar anomalias nos resultados obtidos por meio de algoritmos e sistemas de análise.

interações dos estudantes com os sistemas que ele vai chamar de sistemas institucionais e acadêmicos tendo como base os compromissos e objetivos dos estudantes.

Daí emergiu nossa hipótese de pesquisa de que uma base de dados, que contivesse diferentes variáveis abstraindo as dimensões do modelo, poderia ser usada para categorizar os estudantes retidos na instituição.

As técnicas de clusterização auxiliam a identificação de grupos de estudantes retidos que comungam de um perfil, que se assemelham. Dessa forma, na mitigação do problema da retenção, políticas mais bem formuladas, intervenções pontuais e eficientes, e planejamento estratégico podem ser adotadas pela Instituição.

MÉTODO

Em termos gerais a pesquisa foi concebida da seguinte forma:

1. Construção do banco de dados a partir do CENSUP 2020 e posterior enriquecimento da base com dados provenientes do Sistema Integrado de Gestão de Atividades Acadêmicas (SIGAA) da UNEMAT;
2. Tratamento e limpeza dos dados;
3. Seleção de *features* (variáveis) alinhadas ao modelo teórico de retenção adotado;
4. Clusterização dos Dados
5. Análise e Apresentação dos Resultados.

Na etapa 1, construção do banco de dados, a qualidade da base de dados na qual o algoritmo de clusterização operou foi diretamente proporcional à qualidade do resultado final. Assim, extraímos os dados da UNEMAT disponíveis no CENSUP 2020 criando nossa primeira estrutura. O CENSUP traz aproximadamente 100 variáveis para cada estudante ali registrado. As variáveis refletem diversas características dos estudantes. O censo também registra dados da instituição, dos cursos e do corpo docente. Nosso interesse maior foram as variáveis dos estudantes, mas acessamos também os microdados dos cursos para obter o tempo mínimo de integralização deles, assim como os locais onde são ofertados. Então filtramos todos os registros referentes à UNEMAT, recuperando um total de 21.786 estudantes. Nesta primeira fase já foi possível uma primeira análise exploratória dos dados e alguns indicadores gerais da instituição foram extraídos como: taxas de diplomação, de retenção e de evasão, tempo médio de duração dos cursos e tempo médio de vínculo com a instituição dos estudantes diplomados.

Pelos microdados dos cursos contidos no censo, identificamos a duração mínima dos cursos de todos os estudantes do nosso banco de dados. Neste ponto nosso interesse foi identificar quais estavam na condição de retidos. Disso, filtramos aqueles estudantes com tempo mínimo de integralização já transcorrido e sem diplomação. Nosso filtro retornou 2.455 estudantes.

Constituído, então, nosso banco de dados de estudantes retidos, seguimos enriquecendo-o com o coeficiente de rendimento acadêmico (CRA) dos estudantes. Isso foi feito por meio de consultas ao SIGAA da UNEMAT. A interface de consultas do SIGAA não nos permitiu obter individualmente o CRA, de modo que categorizamos em A, B, C aqueles estudantes com CRA, respectivamente, abaixo de 5, igual a 5 e inferior a 7, e acima de 7 pontos. Levamos em conta a normativa acadêmica da instituição que instrui que estudantes, numa dada disciplina, com média final abaixo de 5 pontos, são estudantes reprovados; com média final entre 5 e 7 pontos, são estudantes em Prova Final e, média acima de 7 pontos, são estudantes aprovados.

A segunda etapa, constituiu-se da limpeza e tratamento dos dados e teve como objetivo uniformizar os dados, descartar registros incompletos e discretizar as variáveis contínuas.

A terceira etapa foi a escolha das variáveis (*features)*. Salientando que do conjunto de variáveis presentes no nosso banco de dados, selecionamos aquelas que, segundo o modelo teórico de Vincent Tinto, poderiam refletir algum aspecto relacionado à retenção. O Quadro 1 descreve as variáveis que foram utilizadas na pesquisa.

Quadro 1. Descrição das Variáveis.

VARIÁVEL	DESCRIÇÃO
RACA	raça declarada do estudante
IDADE	idade em anos do estudante
CURSO_CINE	código CINE do curso do estudante
CRA	índice de rendimento acadêmico do estudante
CURSO_GRAU	descreve se o curso é bacharelado ou licenciatura
LOCAL_OFERTA	nome do local onde o curso é ofertado
STATUS	descreve se o estudante está trancado, transferido ou cursando
TURNO	o turno do curso
INGRESSO	o ano e o semestre de ingresso do estudante
NOME_CURSO	o nome do curso
TIPO_EM	tipo ensino médio: escola pública ou privada
NACIONALIDADE	nacionalidade dos estudante
FORMA_INGRESSO	tipo de ingresso: por vestibular, enem ou outra forma
ESTCIVIL	estado civil do estudante
SEXO	sexo do estudante
PERCENTCH	relação entre carga horária cumprida e carga horária total do curso

Fonte: Elaborado pelos Autores.

As informações presentes no CENSUP garantiram uma quantidade suficiente de variáveis que possibilitaram capturar as diferentes dimensões dos estudantes que eram de nosso interesse e, consequentemente, à razoabilidade da nossa proposta.

A quarta etapa foi a clusterização, também conhecido como agrupamento. A técnica de clusterização pertence aos métodos não supervisionados de aprendizagem de máquina. Referimo-nos à aprendizagem de máquina como sendo uma área interdisciplinar, interfaceando-se sobretudo com a análise estatística de dados, inteligência artificial, reconhecimento de padrões e visualização de dados. Seu objetivo é extrair informações de fontes volumosas de dados. São chamados de não supervisionados porque os dados observados não possuem nenhum rótulo que o identifica, cabendo ao método descobrir tal informação.

Algumas definições importantes para entender a técnica de agrupamento são:

- ***Cluster***: também chamado de grupo, de agrupamento ou de aglomerado – Conjunto de elementos que compartilham um grau de similaridade entre si. Elementos de um mesmo grupo tendem a ser homogêneos entre si ao passo que heterogêneos em relação aos elementos de outro grupo.
- ***Centróide***: elemento central de um *cluster*. Este elemento é o "líder" do grupo, todos os elementos pertences a um grupo possuem características similares às do centróide, e consequentemente aos dos demais elementos do grupo.
- ***Distância***: descreve quão semelhantes os pontos (elementos) de um *cluster* são. Quanto menor a distância entre os pontos, mais semelhantes eles são. Diferentes técnicas de clusterização, utilizam diferentes fórmulas para o cálculo de distância.

Para realização da clusterização, utilizamos para este trabalho o algoritmo *K-modes,* que é um algoritmo de agrupamento criado por Huang (1998). É uma proposta alternativa para análise de agrupamento quando se trabalha com variáveis categóricas, ou variáveis numéricas contínuas discretizadas. Por meio de fórmulas matemáticas, é possível calcular a distância (geralmente a distância euclideana) de cada ponto de um grupo ao seu respectivo centróide. Após as distâncias serem calculadas, as médias aritméticas das distâncias são encontradas e esses valores passam a ser os valores das observações do centróide.

RESULTADOS

Por se tratar de uma técnica não supervisionada, a quantidade *k* de *clusters* (agrupamentos) não é definida e/ou conhecida *a priori*. A definição da quantidade de grupos foi feita por meio de uma técnica chamada "regra do joelho". Esta técnica consiste em plotar o valor da função custo produzida por diferentes valores de *k;* o valor de *k* onde o custo tem uma queda significativa é chamado de 'joelho' da curva, o qual indica que esse valor divide bem os dados. Pela Figura 1, observamos que a partir de 3 grupos a curva de custo começa a ser suavizada, esse ponto (joelho da curva) definiu que este foi o número ideal de clusters já que a função custo (a soma total das diferenças entre os clusters) começou a diminuir.

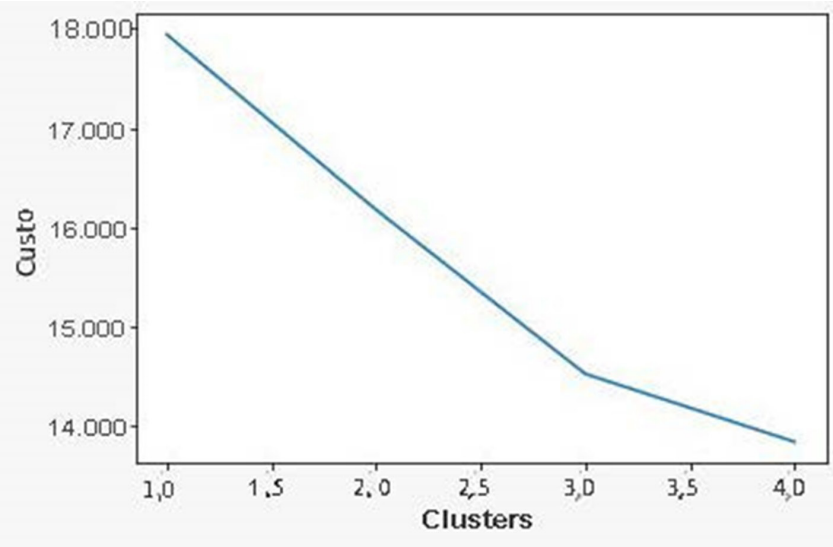

Figura 1. Curva de custo determinando o número de Clusters.

Considerando então 3 a quantidade ideal de clusters, aplicamos o algoritmo *K-Modes* e como resultado, nosso conjunto de 2.169 registros foi particionado em três grupos, nomeados de *cluster 1*, *cluster 2* e *cluster 3*, contendo 984, 715 e 470 estudantes respectivamente.

O Quadro 2 apresenta os centróides destes 3 *clusters*. Lembrando que cada centróide sumariza os valores mais frequentes (a moda) de cada variável do grupo. Pelos centróides, uma primeira caracterização dos grupos pode ser feita.

Ao analisarmos o comportamento dessas variáveis em cada um dos grupos conseguimos descrever o perfil dos alunos retidos na instituição.

O *cluster 1*, o maior deles, contém 984 estudantes retidos. 58.43% desses estudantes estão nos cursos da grande área de Negócios, Administração e Direito (CINE 4). Os cursos de bacharelado representam uma porção considerável deste grupo (88.1%). Um pouco mais da metade dos estudantes do grupo 1 declaram-se pardos (50.30%). Eles concentram-se no campus de Sinop (24.08%), Tangará de Serra (16.36%) e Cáceres (13.21%). O turno noturno representa 62.40% desses estudantes. A grande maioria desses alunos são oriundos da escola pública (88.71%) e ingressaram na universidade via ENEM (58.84%). O ano de ingresso da maioria dos estudantes foi 2015 (48.17%). Em relação ao sexo não se tem uma diferença significativa, o grupo está igualmente distribuído entre ambos. Pouco mais 16.16% desses estudantes não atingiram 50% de sua carga horária mínima de integralização de curso, ao passo que 52.44% deles já têm mais 80% concluída. Em termos de desempenho acadêmico, 50.3% dos estudantes possui um coeficiente de rendimento maior ou igual a 5 e menor que 7. Isso indica que quase a metade desses alunos têm se submetido a exames finais; 33.02% deles têm um CRA igual ou superior a 7, indicando aprovações sem exames finais e 20.12% desses alunos têm experienciado reprovações.

O segundo maior grupo de estudantes retidos, o *cluster 2* é formado por 715 estudantes. Aqui, 87.97% deles estão em cursos da grande área de Educação (CINE 1); As licenciaturas

Quadro 2. Centróides dos Clusters.

VARIÁVEL	*CLUSTER*		
	1	**2**	**3**
RACA	Parda	N/A	Branca
CURSO_CINE	0413A01	0114B01	0811A04
CURSO_CINE1	4	1	8
CRA	B	C	B
GRAU	1	2	1
LOCAL	Sinop	Cáceres	Tangará da Serra
STATUS	1	1	1
TURNO	Noturno	Noturno	Integral
CURSO_NOME	Administração	Ciências Biológicas	Agronomia
TIPOEM	Pública	Pública	Pública
TIPOINGRESSO	ENEM	Vestibular	Vestibular
ESTCIVIL	Solteiro(a)	Solteiro(a)	Solteiro(a)
SEXO	Feminino	Feminino	Masculino
FAIXAETARIA	20-25	20-25	20-25
INGRESSO	2015	2014	2014
CHPERCENT	0.9-1	0.8-0.9	0.8-0.9

compreendem 88% destes estudantes; 45.31% dos estudantes deste grupo não declaram cor. Neste *cluster*, o campus de Cáceres é o que concentra a maior parte desses estudantes (49.51%); neste grupo predominam os cursos noturnos (75.52%). Quase que a totalidade são estudantes egressos de escola pública (93.85%). O tipo de ingresso mais frequente dos estudantes deste *cluster* é o vestibular (58.32%). Neste *cluster* o sexo feminino aparece com um pouco mais de frequência, 66.30%. Os anos de ingresso desses estudantes concentram-se em 2014 e 2015 (60.7%). A carga horária integralizada de 34.27% dos estudantes do *cluster 2* está entre 80 e 90%, 24.33% já concluíram mais de 90% de seus créditos e um montante de 15.24% ainda não atingiu mais que 50% de seus créditos. Sobre o desempenho acadêmico, 53.14% apresentam um coeficiente de rendimento acadêmico igual ou superior a 7, o que indica aprovações sem exames finais. Com CRA no intervalo de 5 e 7, tem-se 31.46% dos estudantes e 15.3% deles apresentam um coeficiente de rendimento acadêmico inferior a 5.

O *cluster 3* compreende 470 estudantes. Desses, 35.32% declaram-se brancos e 30.42%, pardos. Os cursos da grande área de Agricultura, Silvicultura, Pesca e Veterinária (CINE 8) concentram 50.21% dos estudantes enquanto que 22.13% dos estudantes encontram-se nos cursos da grande área de Engenharia, Produção e Construção (CINE 7). Todos os discentes desse grupo estão nos cursos de bacharelado. A maioria deles estão no campus de Tangará de Serra (24.04%), seguido por 16.4% (Nova Xavantina) e Cáceres (14.9%). Estes estudantes estão essencialmente em cursos integrais (86.4%), também são egressos de escola pública (80.85%) e 61.45% deles ingressaram via Vestibular. O sexo masculino é o mais frequente neste grupo (69.8%). 54.04% ingressaram no ano de 2014. A carga horária mínima de curso integralizada de 24.5% dos estudantes do *cluster* 3 está entre 80 e 90%, 23.83% já concluíram mais de 90% de seus créditos e um montante de 18.72% ainda não atingiram mais que 50% de seus créditos. Sobre o desempenho acadêmico, 54.9% dos estudantes têm um CRA no intervalo de 5 e 7, 28.08% deles possuem um CRA inferior a 5 e 17.02% dos estudantes apresentam um CRA igual ou superior a 7.

Feitas essas considerações, podemos então sumarizar os perfis desses estudantes como:

- O *cluster 1* é caracterizado por estudantes que se concentram nos bacharelados da área de Direito, Administração e Negócios; o grau estão em média há 5,2 anos na universidade, já concluíram mais de 90% de sua carga-horária mínima, ingressaram via ENEM, estudam no turno noturno e possuem um CRA entre 5 e 7.
- O *cluster 2* é caracterizado por estudantes das licenciaturas da área de Educação, a maioria do sexo feminino, com CRA superior a 7, estão na universidade em média 5,4 anos, ingressaram via vestibular e estão no turno noturno com uma carga horária mínima cumprida entre 80 e 90%.
- O *cluster 3* é caracterizado por estudantes de cursos de período integral. Todos em bacharelados. Estão na instituição em média 5,95 anos, são na maioria de sexo masculino, ingressaram via vestibular, têm um CRA entre 5 e 7 e já cumpriram mais de 80% de sua carga horária mínima de integralização de curso.

DISCUSSÃO

Reiteramos que a técnica de clusterização não nos permite explicar um fato. O método buscou agrupar os estudantes em função da maior similaridade entre eles. Esses grupos representam, em termos práticos, os perfis extraídos. De qualquer modo, o comportamento de algumas variáveis nos ajuda a entender e discutir esses perfis. Algumas variáveis apresentam uma distribuição de frequência bem particular enquanto outras já não se distribuem tão diferentemente.

As variáveis RAÇA e FAIXA ETÁRIA distribuem-se de forma semelhante nos três *clusters*. Elas revelam na verdade o perfil do estudante da UNEMAT e os *clusters* captam estas características: jovens com idade média de 25 anos e pardos.

A variável CINE é uma variável interessante, ela apresenta as maiores frequências no *cluster 1* e *cluster 2*. Disso, pode-se especular que os estudantes retidos na UNEMAT concentram-se nos cursos da grande área de Negócios, Administração e Direito e da grande área Educação. Vescovi (2020) ao investigar métodos preditivos para evasão, ressalta que a variável qualitativa contendo o "nome do curso", foi a de maior importância em todos os modelos investigados. O autor argumenta que a elevada importância desta variável provavelmente ocorre devido à diferença entre os perfis de estudantes que frequentam cursos distintos. Isto levanta uma outra questão acerca da democratização do ensino superior que é, segundo Knop (2020) sua estratificação. Em outros termos, indivíduos oriundos de estratos sociais mais elevados, com boa formação inicial tem maior probabilidade de ingressar em cursos de maior prestígio e retorno financeiro e maior chance de diplomação, os indivíduos do lado oposto, ingressam em cursos de menor retorno financeiro e estão mais propensos a evadir ou enfrentar dificuldades para conclusão.

A variável SEXO caracteriza bem os *clusters 2* e *3*. No *cluster 2,* estão, em grande maioria, estudantes do sexo feminino, enquanto no *cluster 3*, o sexo masculino é mais presente.

Pelo GRAU DOS CURSOS observou-se que os *cluster*s *1* e *3* são caracterizados pelos bacharelados e o *cluster 2* pelas licenciaturas. Ressalta-se que 100% dos acadêmicos do *cluster 3* são dos cursos de bacharelado.

A variável TURNO apresenta uma distribuição bem particular no *cluster 3*. Esse é o grupo de estudantes retidos que na maioria dos casos estão em cursos de período integral.

A análise do CRA e do percentual de carga horária cumprida nos revela uma característica interessante dos estudantes retidos. Ela nos permitiu identificar o que Tinto chama de alta integração acadêmica. Os estudantes dos três *cluster*s, em sua grande maioria, apresentam boas notas (CRA B e C), já cumpriram mais 80% da carga horária de seus cursos e de fato persistem na instituição. Corrobora com essa hipótese de persistência, o fato de que as diplomações têm ocorrido, em média, em 5,7 anos ao passo que a duração média dos cursos na UNEMAT é de 4,5 anos.

Entretanto, 356 estudantes retidos (16.42%) ainda não cumpriram mais de 50% de suas cargas horárias. Uma investigação dos coeficientes de rendimento acadêmico destes estudantes nos revela que no *cluster 1*, 126 apresentam CRA inferior a 5, 26 estudantes com CRA entre 5 e 7 e

7 deles apresentam CRA superior a 7; no *cluster 2*, 85 tem CRA inferior a 5, 19 apresentam CRA entre 5 e 7, e 5 estudantes têm CRA superior a 7; no *cluster 3*, 79 estudantes têm CRA inferior a 5, 8 com CRA inferior a 7 e apenas 1 estudante tem CRA superior a 1. O que identificamos então é uma baixa integração acadêmica desses 356 indivíduos, uma vez que nos 3 *cluster*s, essa parcela de estudantes está longe de integralizar seus créditos para diplomação, e ainda enfrentam problemas de notas. Pela Teoria de Tinto esses alunos teriam grande probabilidade de evadirem.

Tinto (1993) destaca também a importância dos atributos pessoais do estudante antes de entrar na instituição de ensino superior. Embora não tenhamos adicionados todos os atributos possíveis em no nosso modelo, os atributos TIPO DE ENSINO MÉDIO e TURNO poderiam nos dizer algo a respeito. Estudantes egressos de ensino médio de escola pública são predominantes em todos os *clusters*, no entanto, no *cluster 1* e *cluster 2* a maioria dos estudantes estão em cursos noturnos o que pode implicar na necessidade de conciliar trabalho com estudos. Nesta mesma direção, Brandão (2018) também discute que o aumento da retenção para os cursos noturnos possui as mesmas causas elencadas para a evasão: a necessidade de conciliar estudo com trabalho, a maior deficiência dos alunos que estudam à noite, o menor envolvimento do aluno noturno nas atividades extracurriculares da instituição de ensino, entre outras.

Podemos pautar algumas especulações advindas da análise desses grupos: Como o fenômeno de estratificação ocorre (e se ocorre) na instituição?; Como os cursos da área de Educação, em sua maioria noturnos, apresentam um elevando número de estudantes retidos ainda que com um alto coeficiente acadêmico? Teriam as diretrizes nacionais que determinam cargas-horárias mínimas dos bacharelados e das licenciaturas algum impacto sobre as diplomações além do prazo regulamentar?

Agora encontramos um norte. Sabemos quem são os estudantes retidos, sabemos que seus perfis podem ser agrupados, sabemos as características mais relevantes desses perfis.

CONCLUSÃO

Buscamos compreender os fatores que levam à retenção de estudantes na Universidade do Estado de Mato Grosso ao identificar e analisar os diferentes perfis dos estudantes retidos. Por meio da clusterização dos dados, foi possível agrupar os estudantes com características semelhantes, permitindo uma visão mais clara e detalhada dos padrões de retenção. Essa abordagem possibilitou a identificação de variáveis relevantes, como o curso de graduação, tempo de permanência na universidade, forma de ingresso, desempenho acadêmico, entre outros, que podem influenciar na retenção dos estudantes.

Com os perfis constituídos e apresentados, um próximo passo, não alcançado pelo escopo deste trabalho, é uma investigação pontual e específica sobre as reais causas sobre os diferentes perfis.

Os desafios de estudar retenção, dadas a complexidade e a multidimensionalidade deste tema estiveram presentes neste trabalho. A falta de um consenso na conceituação do termo, as diferentes métricas de mensuração das taxas de retenção, o estreitamento desse fenômeno com os de permanência e evasão, tudo isso resume um pouco o que foram os desafios.

Desafios a parte, nosso trabalho ao trazer de forma inédita a técnica de clusterização, e mais especificamente pelo algoritmo *K-Modes*, para estudos de retenção na Educação Superior se mostrou aplicável no estudo de caso da UNEMAT, podendo facilmente ser aplicada em qualquer instituição.

Pela análise exploratória dos *clusters* foi possível compreender o fenômeno de retenção nos cursos de graduação (modalidade presencial) da Universidade do Mato Grosso. Os 3 grupos gerados pela técnica de clusterização evidenciam que os estudantes retidos na UNEMAT têm uma alta integração acadêmica que implica expectativa de diplomação, ainda que no prazo superior ao mínimo regulamentado nos Projetos Pedagógicos dos Cursos. Mesmo nessa generalização, os perfis demonstram particularidades entre cada grupo, assim, podendo fornecer insights importantes sobre possíveis estratégias de intervenção e apoio aos

estudantes em risco de evasão. Compreender as características e necessidades específicas de cada grupo de estudantes retidos pode auxiliar na implementação de ações direcionadas e personalizadas para promover a permanência e o sucesso acadêmico desses alunos.

AGRADECIMENTOS

Agradecimentos ao Grupo de Pesquisa sobre Acesso e Permanência na Educação Superior-GPAPES/UNEMAT pela contribuição ao referencial teórico construído.

REFERÊNCIAS

BETARELLI JUNIOR, A. A. Análise de Agrupamentos (Clusters). NOTA DE AULA DO PROGRAMA DE POS GRADUACAO EM ECONOMIA APLICADA, 2016, Juiz de Fora. **Apresentação**. Juiz de Fora: UFJF, 2016. Disponível em: https://www2.ufjf.br/lates//files/2016/12/Conte%C3%Bado-5-%E2%80%93-A_An%C3%A1lise-de-cluster-AA.pdf. Acesso em: 01 jul. 2020.

BRANDAO, J. dos S. **O impacto da evasão e retenção sobre o financiamento de universidades federais brasileiras:** um estudo a partir do indicador aluno equivalente. Dissertação de Mestrado – Fundação Universidade Federal do Tocantins, Palmas, 2018.

BRASIL. Ministério da Educação. **Comissão Especial de Estudos sobre Evasão nas Universidades Públicas Brasileiras. Diplomação, Retenção e Evasão nos Cursos de Graduação em Instituições de Ensino Superior Públicas. ANDIFES/ABRUEM/SESu/MEC**. Brasília, DF: Ministério da Educação, 1996.

HOFFFMANN, E.; BITENCOURT, L. P. A evasão discente nas licenciaturas de matemática presenciais da UNEMAT (2011 a 2015) e as políticas de combate a essa evasão. In: ANAIS ENCONTRO NACIONAL DE EDUCAÇÃO MATEMÁTICA, 13., 2019, UNEMAT. **Anais [...]**. UNEMAT, 2019.

HUANG, Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery, v. 2, p. 283–304, 1998.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. **Censo da Educação Superior (2000-2019)**. São José dos Campos: INPE, 2020. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-da-educacao-superior. Acesso em: 6 mar. 2024.

Knop, M. N. H. **Retenção e resiliência no Ensino Superior brasileiro:** determinantes das chances de conclusão. 2020. Tese (Doutorado em Sociologia) – Universidade de Brasília, Brasília, 2020.

NODARI, D. E.; LIMA, E. G. S.; MACIEL, C. E. O desempenho dos estudantes no Vestibular e a permanência nos cursos de graduação da UNEMAT. **Avaliação**: Revista da Avaliação da Educação Superior (Campinas), Sorocaba, v. 23, n. 2, p. 312-329, 2018. DOI: http://doi.org/10.1590/s1414-40772018000200003.

Nunes, S. I.; Pereira, F. A. Retenção no Ensino Superior: Reflexões a partir da produção acadêmica. In: CONGRESSO NACIONAL DE EDUCAÇAO – CONEDU, 6., 2019. Campina Grande. **Anais [...]**. Campina Grande: Realize Editora, 2019. p. 1-12. Disponível em: https://editorarealize.com.br/artigo/visualizar/62537. Acesso em: 1 jun. 2023.

Tinto, V. Dropout from higher education: a theoretical synthesis of recent research. **Review of Educational Research**, Washington, v. 45, n. 1, p. 89-125, 1975. DOI: http://doi.org/10.3102/00346543045001089.

Tinto, V. **Leaving college:** rethinking the causes and cures of student attrition. London: University of Chicago Press, 1993.

Vescovi, P. V. S. **Análise Preditiva na detectação de Evasão de Alunos no Ensino Superior Privado Brasileiro:** abordagem de algorítimos de aprendizado de máquina com base nas perspectiva acadêmicas, financeiras, geográficas e socioeconômicas. 2020. Dissertação (Mestrado em Gestão para a Competitividade) – Fundação Getúlio Vargas, São Paulo, 2020.

Contribuições dos autores

FCVM: Orientador da pesquisa e autor principal do artigo. MPM: Executor da pesquisa e coautor do artigo.

Editor: Prof. Dr. José Luís Bizelli
Editor Executivo para América Latina: Prof. Dr. Vilmar Alves Pereira