ISSN 1982-5587

Articles

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

Perfil dos Estudantes Retidos na Universidade do Estado de Mato Grosso: análise de uma Instituição Pública Pantaneira

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

HOW TO CITE: MALANGE, F. C. V.; MESQUITA, M. P. Profile of Students Retained at the Universidade do Estado de Mato Grosso (State University of Mato Grosso): analysis of a Pantanal Public Institution. Revista IberoAmericana de Estudos em Educação, Araraquara, v. 19, esp. 3, e19455, 2024. eISSN: 19825587. DOI: https://doi.org/10.21723/riaee.v19i00.1945502

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.

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.

INTRODUCTION

Brazilian Higher Education experienced a boom in its number of vacancies, as well as an encouragement to occupy vacancies in private universities, in the 1990s and early 2000s. We can name the Programa de Financiamento Estudantil (Student Funding Program) (FIES), the Programa Universidade para Todos (University For All Program) (PROUNI), the Reestruturação

***Corresponding author:**
fmalange@unemat.br

Submitted: July 06, 2024
Reviewed: August 09, 2024
Approved: September 17, 2024

Financial support: nothing to declare.
Conflicts of interest: There are no conflicts of interest.

Ethics committee approval: Not applicable.
Data availability: No dataset or material is available online or for consultation. Study conducted at Programa de Pós Graduação em Educação, Universidade do Estado de Mato Grosso – UNEMAT, Cáceres (MT), Brasil.

e Expansão das Universidades Federais (Restructuring and Expansion of Federal Universities) (REUNI) and the creation of the Institutos Federais de Ciência e Tecnologia (Federal Institutes of Science and Technology) (IFET) as propellants of said increase. The policies for reservation of vacancies would also promote a greater access to higher education to those who historically have been marginalized from this educational level, shaping it up to be, at least in theory, a diverse and democratic environment.

Despite this expansion, the policies of permanence aren't implemented in the same proportion as the number of vacancies in Brazilian universities. Many students, for a variety of reasons, turn out to not complete their studies, as revealed by Nunes and Pereira (2019), between the years of 2006 and 2016, there was an important leap in the amount of students who hadn't graduated (69,06%), under the condition of retained or dropout students in brazilian higher education (Brasil, 1996).

The numbers pertaining to UNEMAT's vacancies, entrances and dropouts align themselves to the national scenario, and, in the case of the latter, can reach the alarming rate of 50% in some degrees (Hoffmann; Bitencourt, 2019; Nodari; Lima; Maciel, 2018).

The Censo da Educação Superior (Census of Higher Education) (CENSUP) from 2020 (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2020) revealed that, in the year of 2019, 21,786 students had ties to one of the on-campus undergraduate courses at UNEMAT. Of these, 2,120 graduated, 2,501 cut their ties to the institution and 2,455 had already surpassed the minimum regulatory time for course completion. The census also pointed out that the average duration of the courses in the institution was 4.5 years, whereas the graduated students spent an average of 5.57 years studying at the institution. A share of 810 of the students that cut their ties in 2019 were in the university for, at least, 4 years.

From these data, we learn that the Universidade do Estado de Mato Grosso (State University of Mato Grosso) has latent issues with retention, and that it would be reasonable to assume that the graduation rates of the university would be higher if the retentions could have been converted into graduations, or even, if the dropouts could have been converted into permanences, and later, future course completions.

We understand that it is not trivial to explain the reason behind these retentions, but learning about the profiles of those students who find themselves in this situation is an important step to better direct the management efforts, pedagogical policies and mitigation strategies for dropouts and/or retentions.

This research addresses the study of the retentions in a multi campi university, which offers different courses in a variety of fields of study, attending to dozens of cities in the state of Mato Grosso. By not worrying about the predictive character of the data, nor the explanatory character of the phenomenon, our focus was in analyzing the profile of the students that currently find themselves retained in the institution through data clusterization.

Data clusterization is based on analyzing a dataset, which, in our case, were the student records, and grouping them together according to their similarities. Amongst its advantages, as pointed out by Bertarelli Junior (2016), are the possibility of assessing dimensionality, identifying *outliers*[1], and suggesting hypotheses regarding the structure of the relations, besides being an important Data Mining technique.

But the formation of clusters alone doesn't enable us to advance in the issues regarding retention. The term has been historically established, and the cluster analysis was made in light of these theories that attempt to handle the phenomenon's many facets.

Different approaches and perspectives have been used in the theoretical models of student retention in the last few decades, with an emphasis in the high rates of student dropouts and retentions in Higher Education. Many of which seem to be little-known by the educators, managers and policy makers in the educational institutions, due to their complexities. Tinto's theoretical model (Tinto, 1975), widely studied and used as a basis for several other proposals, argues that the retention process takes place through the interactions of the students with the systems that he will call institutional and academic systems, having as a foundation the commitments and goals of the students.

[1] *Outliers* are data that drastically differ from all others. In other words, an outlier is a figure that sets itself apart and can (and probably will) cause anomalies in the results obtained by means of algorithms and analytic systems.

From this emerged our research hypothesis that a database that contained different variables disregarding the model dimensions could be used to categorize the students retained at the institution.

The clusterization techniques help to identify groups of the retained students that share a profile, resemble each other. In this way, regarding the mitigation of the retention issue, better formulated policies, effective and occasional interventions, and strategic planning can be adopted by the institution.

METHOD

In general terms, the research was conceived as such:

1. Construction of the database based on CENSUP 2020 and later enrichment of the base with data sourced from UNEMAT's Sistema Integrado de Gestão de Atividades Acadêmicas (Integrated System of the Management of Academic Activities) (SIGAA);
2. Data processing and cleansing;
3. Selection of features (variables) aligned to the adopted theoretical framework of retention;
4. Data Clustering;
5. Analysis and presentation of the results.

In step 1, the construction of the database, the quality of the database in which the clustering algorithm operated was directly proportional to the quality of the final result. Thus, we extracted the data on UNEMAT available on CENSUP 2020 by creating our first structure. CENSUP provides approximately 100 variables per each student registered there. The variables reflect several characteristics of the students. The census also registers data from the institution, the courses and the faculty. Our main interest was the students' variables, but we also accessed the microdata of the courses in order to obtain the minimum time of their completion, as well as the places where they are offered. Then we filtered all of the records regarding UNEMAT, retrieving a total of 21,786 students. In this first phase, it was already possible to carry out the first exploratory analysis of the data and extract some general metrics of the institution, such as: graduation, retention and dropout rates, average duration of the courses, and average duration of ties with the university for students who graduated.

By the microdata of the courses contained in the census, we identified the minimum duration of the courses of all students on our data bank. At this point, our interest was to identify which of these were under the condition of 'retained'. From there, we filtered the students whose minimum regulatory time for course completion had already passed, without a graduation. Our filter came back with 2,455 students.

Once our data bank of retained students was constituted, we followed through by enriching it with the coeficiente de rendimento acadêmico (coefficient of academic performance) (CRA) of the students. This was made through consultations to UNEMAT's SIGAA. The consultation interface of SIGAA did not allow us to obtain the CRAs individually, so that we categorized in A, B and C the students with CRAs of, respectively, below 5, equal to 5 and below 7, and above 7 points (out of 10). We took into account the academic regulation of the institution that instructs that the students, in a given subject, with a grade-point average below 5 points are failed students; with a grade-point average between 5 and 7 points are subject to a Make-up Test; and, with a grade-point average above 7 points are students who passed.

The second stage consisted of data processing and cleaning, and aimed to standardize the data, discard incomplete registers and discretize the continuous variables.

The third stage was the selection of the variables (features). We emphasize that, within the set of variables present in our data bank, we selected those that, according to Vincent Tinto's theoretical model, could reflect some aspect related to retention. Chart 1 describes the variables that were used in the research.

The information contained on CENSUP assure a sufficient amount of variables that made it possible to apprehend the different dimensions of the students that were of our interest and, consequently, to the reasonableness of our proposal.

Chart 1. Descriptions of the Variables.

VARIABLE	DESCRIPTION
RACE	Declared race of the student
AGE	Age (in years) of the student
COURSE_CINE	Cine code of the course of the student
CRA	Coefficient of academic performance of the student
COURSE_DEGREE	Describes whether the course is a bachelor's or a licenciatura[2]
PLACE_OFFER	Name of the place where the course is offered
STATUS	Describes if the student is enrolled, taking a break, or transferred
TIME	The time of the offered course
ENROLLMENT	The year and semester of the student's enrollment
COURSE_NAME	The name of the course
HS_TYPE	Type of former high school: public or private high school
NATIONALITY	Nationality of the student
FORM_ADMIN	Form of admission: by standardized testing or otherwise
MARITSTATUS	Marital status of the student
SEX	Sex of the student
PERCENTCH	Relationship between the completed workload and the total regulatory workload of the course

Source: Elaborated by the Authors.

The fourth stage was clustering, also known as grouping. The clustering technique belongs to the unsupervised methods of machine learning. We refer to machine learning as an interdisciplinary field, interfacing especially with statistical data analysis, artificial intelligence, pattern recognition and data visualization. Its main goal is to extract information from massive sources of data. These studies are called unsupervised because the observed data do not have any labels that identify them, leaving the discovery of such information to the method.

Some important definitions to understand the grouping technique are:

- ***Cluster***: Set of elements that share a degree of similarity with one another. Elements of the same group tend to be homogenous amongst themselves while heterogenous in relation to elements of another group.
- ***Centroid***: the central element of a cluster. This element is the "leader" of the group, as all of the elements belonging to a group have similar characteristics to those of the centroid, and consequently, to those of the other elements of the group.
- ***Distance***: describes how similar the points (elements) of a cluster are. The smaller the distance between the points, the more similar they are. Different techniques of clustering use different formulas to calculate the distance.

To carry out the clustering for this paper, we used the K-modes algorithm, which is a grouping algorithm created by Huang (1998). It is an alternative proposal to the grouping analysis when you work with categorical variables, or discretized continuous numerical variables. By means of mathematical formulas, it is possible to calculate the distance (usually, the Euclidean Distance) of each point in a group to its respective centroid. After the distances are calculated, the arithmetic averages of the distances are found and these figures come to be the figures of the observations of the centroid.

[2] A Licenciatura, in Brazil, is a type of undergraduate degree, similar to a Bachelor's, that grants the student the diploma to teach the chosen field at schools, which is typically not allowed with a Bachlor's.

RESULTS

Because it is an unsupervised technique, the *k* amount of clusters is not defined and/or known *a priori*. The formulation of the amount of groups was made by a technique called the "elbow method". This technique consists in plotting the value of the function cost produced by different *k* values; the *k* value where the cost has a significant drop is called the 'elbow' of the curve, which indicates that this value splits the data well. Through Figure 1, we observed that the curve cost begins to be smoothed from 3 groups, this point (the elbow of the curve) defined that this was the ideal number of clusters, given that the cost function (the total sum of the differences between the clusters) began to drop.

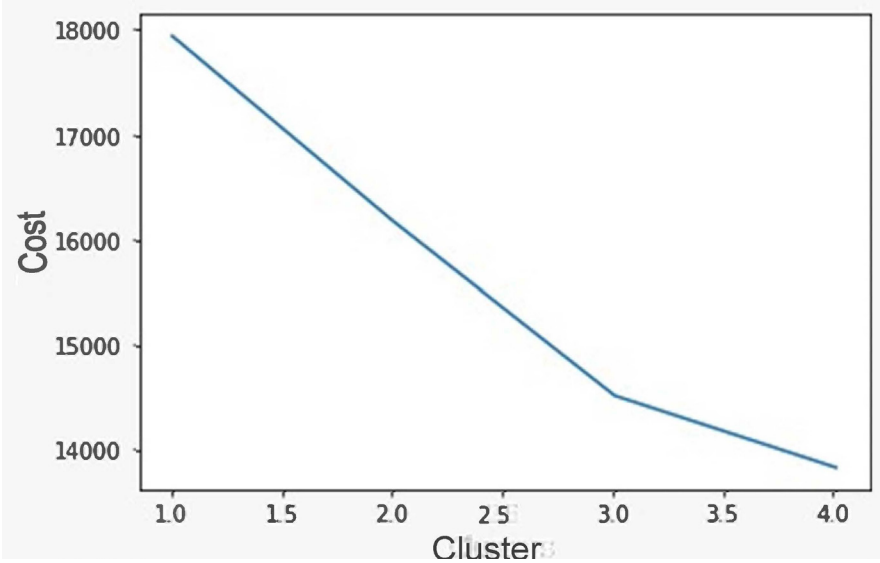

Figure 1. Curve of cost establishing the amount of Clusters.

Considering then 3 as the ideal number of clusters, we applied the K-Modes algorithm and as a result, our set of 2,169 registers was partitioned into three groups, named *cluster 1*, *cluster 2* and *cluster 3*, containing 984, 715 and 470 students respectively.

Chart 2 presents the centroids of these 3 clusters. Keep in mind that each centroid summarizes the most frequent values (the mode) of each variable of the group. Through the centroids, a first characterization of the groups could be done.

By analyzing the behavior of these variables in each of the groups, we were able to describe the profile

retained at the institution.

Cluster 1, the largest amongst them, contains 984 retained students. 58.43% of these students study courses in the broad field of Business and Law (CINE 4). The bachelor courses represent a considerable portion of this group (88.1%). A little over half of the students in group 1 declare themselves to be of mixed race (50.3%). They focus on the campus of Sinop (24.08%), Tangará de Serra (16.36%) and Cáceres (13.21%). Students taking night classes represent 62.4% of the group. The vast majority of these students come from public schools and enrolled to the university via ENEM (58.84%). The year of enrollment of most of the students was 2015 (48.17%). In relation to the sex of these students there isn't a significant difference, the group is equally distributed amongst both sexes. A little over 16.16% of these students didn't achieve 50% of their minimum regulatory workload for course completion, whereas 52.44% of them have over 80% of it completed. In terms of academic performance, 50.3% of the students have a coefficient equal to or greater than 5 and below 7. This indicates that half of these students have been subjected to make-up tests; 33.02% of them have a CRA equal to or greater than 7, indicating passing without make-up exams, and 20.12% of these students have been experiencing failings.

The second largest group of retained students, *cluster 2*, is composed of 715 students, Here, 87.97% of them are in courses of the board field of Education (CINE 1); the Licenciaturas

Chart 2. Centroids of the Clusters.

VARIABLE	*CLUSTER*		
	1	**2**	**3**
RACE	Mixed	N/A	White
COURSE_CINE	0413A01	0114B01	0811A04
COURSE_CINE1	4	1	8
CRA	B	C	B
DEGREE	1	2	1
PLACE	Sinop	Cáceres	Tangará da Serra
STATUS	1	1	1
TIME	Nightime	Nightime	Fulltime
COURSE_NAME	Business	Biological Sciences	Agronomy
HSTYPE	Public	Public	Public
FORM_ADMIN	ENEM[3]	Vestibular[4]	Vestibular
MARITSTATUS	Single	Single	Single
SEX	F	F	M
AGE RANGE	20-25	20-25	20-25
ENROLLMENT	2015	2014	2014
PERCENTCH	0.9-1	0.8-0.9	0.8-0.9

comprise 88% of these students; 45.31% of the students don't disclose their race. In this cluster, the Cáceres campus is the one that concentrates the largest part of these students (49.51%); in this group the night classes predominate (75.52%). Almost the totality of these students are alumni of public schools (93.85%). The most frequent method of enrollment of the students in this cluster is the Vestibular (58.32%). In this cluster the female sex shows up a little more often, 66.30%. The years of enrollment of these students are concentrated in 2014 and 2015 (60.7%). The completed regulatory workload of 34.27% of the students in *cluster 2* is between 80% and 90%, 24.33% have already completed more than 90% of their workload and a sum of 15.24% still has not achieved over 50% of their workload. On academic performance, 53.14% exhibit a coefficient equal to or superior then 7, which indicates passing without make-up tests. There are 31.46% of the students with a CRA in the range between 5 and 7, and 15.3% exhibit a coefficient below 5.

Cluster 3 comprises 470 students. Of these, 35.32% declare themselves to be white, and 30.42%, mixed. The courses of the broad field of Agriculture, Forestry, Fishery and Veterinary (CINE 8) concentrate 50.21% of the students, whereas 22.13% of the students find themselves in courses of the broad field of Engineering, Manufacturing and Construction (CINE 7). All of the students of this group are in bachelor's courses. Most of them are in the campus of Tangará de Serra (24.04%), followed by Nova Xavantina (16.4%), and Cáceres (14.9%). These students are essentially in full-time courses (mornings and afternoons) (86.4%), they are also alumni of public schools (80.85%) and 61.45% of them enrolled via Vestibular. The male sex is the most frequent in this group (69.8%). 54.04% enrolled in the year of 2014. The minimum regulatory workload already completed by 24.5% of the students in *cluster* 3 is between 80% and 90%, 23.83% have already completed more than 90% of their workload and a sum of 18.72% still has not achieved more than 50% of their workload. On academic performance, 54.9% of the students still have a CRA in the range between 5 and 7, 28.08% of them have a CRA below 5, and 17.02% of the students display a CRA equal to or superior then 7.

[3] ENEM refers to Exame Nacional do Ensino Médio (National High School Exam), a type of standardized testing that is accepted for all university admissions in Brazil.

[4] Vestibular refers to a different type of testing for university admissions in Brazil, one that is specific to each institution.

Given these considerations, we can then summarize the profile of these students as:

- *Cluster 1* is characterized by students concentrated in bachelors within the broad field of Law and Business; who have been for 5.2 years in average at the university, have completed over 90% of their minimum regulatory workload, enrolled via ENEM, study at nightime, and have a CRA between 5 and 7.
- *Cluster 2* is characterized by students of Licenciaturas within the field of Education, mostly female, with a CRA superior then 7, have been in the university for 5.4 years on average, enrolled via Vestibular and study on the nightime, with a completed minimum regulatory workload between 80% and 90%.
- *Cluster 3* is characterized by students in full-time courses, all bachelors. They have been in the institution for 5.95 years on years on average, are mostly male, enrolled via Vestibular, have a CRA between 5 and 7 and have already completed over 80% of their minimum regulatory workload.

DISCUSSION

We reiterate that the clustering technique does not allow us to explain a fact. The method sought to group the students on the basis of the similarity amongst them. These groups represent, in practical terms, the extracted profiles. In any case, the behavior of some variables help us to understand and discuss these profiles. Some variables feature a very particular frequency distribution, whereas others do not distribute themselves so differently.

The variables RACE and AGE RANGE distribute themselves similarly on the three clusters. They reveal in actuality the profile of the students at UNEMAT and the clusters capture these characteristics: youth of 25 years old on average and of mixed race.

The variable CINE is an interesting one, as it displays higher occurrences on *cluster 1* and *cluster 2*. From this, it can be speculated that the students retained at UNEMAT are concentrated within the broad field of Business and Law and the broad field of Education. Vescovi (2020), by investigating predictive methods for evasion, points out that the qualitative variable containing the "name of the course" was of the greatest importance in all research models. The author argues that the great importance of this variable probably occurs due to the difference amongst the profiles of the students who attend different courses. This leads to another issue regarding the democratization of the access to higher education, which is, according to Knop (2020), its stratification. In other terms, people originating from higher social strata, with good initial studies, have a higher probability of enrolling in courses of greater prestige and financial returns and a higher chance of graduating, and the people on the opposite side, enroll in courses of lower financial returns and are more prone to evasion or facing challenges for their completion.

The variable SEX characterizes *clusters 2* and *3* well. On *cluster 2,* there are, as the large majority, students of the female sex, whereas on *cluster 3*, the male sex is more present.

Through COURSE DEGREE, it was observed that the *cluster*s *1* and *3* are characterized by bachelors', and *cluster 2*, by Licenciaturas. It is noteworthy that 100% of the academics on *cluster 3* are from bachelor courses.

The variable TIME displays a very particular distribution on *cluster 3*. This is the group of retained students that in most cases are in full time courses.

The analysis of the CRA and completed workload informs us of an interesting characteristic of the retained students. It allowed us to identify what Tinto calls high academic completion. The students of the three *cluster*s, in their vast majority, display good grades (CRA B and C), have already completed more than 80% of the workload of their courses and indeed persist in the institution. This persistence hypothesis is corroborated by the fact that the graduations have been occurring, on average, within 5.7 years, whereas the average duration of the courses at UNEMAT is of 4.5 years.

However, 356 retained students (16.42%) still haven't completed more than 50% of their workloads. An investigation on the coefficients of academic performance of these students informs us that, on *cluster 1*, 126 display a CRA inferior to 5, 26 students with a CRA between 5 and 7, and 7 of them display a CRA superior to 7; on *cluster 2*, 85 have a CRA inferior to 5,

19 display a CRA between 5 and 7, and 5 students have a CRA superior to 7; on *cluster 3*, 79 students have a CRA inferior to 5, 8 with a CRA inferior to 7, and only 1 student has a CRA superior to 7. What we identified then is a low academic completion of these 356 people, given that, on the 3 *cluster*s, this portion of students is far from completing their credits for graduation, and still suffer from grade issues. By Tinto's theory, these students would have a high probability of evasion.

Tinto (1993) also highlights the importance of personal attributes of the student before entering the institution of higher education. Although we have not added all of the possible attributes to our model, the attributes HIGH SCHOOL TYPE and TIME could tell us something about it. Students coming from public high schools are dominant in all of the *clusters*, however, on *cluster 1* and *cluster 2*, most of the students are in night time courses, which may imply the need to balance work and studies. In this same direction, Brandão (2018) also discusses that the increase in retention of night courses has the same causes listed for evasion: the need to balance work and studies, the greatest deficiency of the students that study at night, the lower engagement of the night students with extracurricular activities of the institution, among others.

We can propose some speculation arising from the analysis of these groups: How does the stratification phenomenon occurs (and if it occurs) on the institution?; How do the courses in the field of Education, in their majority, night courses, feature a high number of retained students, though with a high academic coefficient? Would the national guidelines that determine minimum workloads for undergraduate degrees have any impact on the graduations beyond the regulatory deadline?

Now we have found a direction. We know who the retained students are, we know that their profiles can be grouped, and we know the most relevant characteristics of these profiles.

FINAL THOUGHTS

We have sought to understand the factors that may lead to the retention of students in the Universidade do Estado de Mato Grosso (State University of Mato Grosso) by identifying and analyzing the different profiles of the retained students. By means of data clustering, it was possible to group the students with similar characteristics, allowing for a clearer and more detailed vision of the retention patterns. This approach enabled the identification of relevant variables, such as the undergraduate course, length of stay at the university, method of enrollment, academic performance, among others, which may influence student retention.

With the established and presented profiles, a next step, not reached by the scope of this work, is a specific investigation of the real causes interfering with the different profiles.

The challenges of studying retention, given the complexity and multidimensionality of the issue, were present in this paper. The lack of consensus in the term's conceptualization, the different metrics for measuring the retention rates, the narrowing between this phenomenon and those of permanence and evasion, all of these briefly summarize what the challenges were.

Challenges aside, our work, by bringing the clustering technique in an unprecedented manner, and more specifically, through the K-Modes algorithm, to the studies of retention in Higher education, was shown to be applicable to the Case Study at UNEMAT, and may be easily applied to any other institution.

Through the exploratory analysis of the clusters, it was possible to understand the phenomenon of retention in the undergraduate (on-campus) courses of the Universidade do Estado do Mato Grosso (State University of Mato Grosso). The groups generated by the clustering technique demonstrate that the students retained at UNEMAT have a high level academic completion, which implies the expectation of graduation, even if on a deadline exceeding the minimum time regulated by the Pedagogical Projects of the Courses. Despite this generalization, the profiles showcase particularities among each group, and thus, may offer important insights on possible strategies of intervention and support to the students at risk of evasion. Understanding the specific characteristics and needs of each group of retained students can assist the implementation of targeted and personalized actions in order to promote the permanence and academic success of these students.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the Research Group on Access and Permanence in Higher Education - GPAPES/UNEMAT for their contribution to the theoretical framework constructed.

REFERENCES

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.

Author contributions

FCVM: Research advisor and main author of the article. MPM: Research executor and co-author of the article.

Editor: Prof. Dr. José Luís Bizelli
Executive Editor for Latin America: Prof. Dr. Vilmar Alves Pereira