ISSN 1870-4069

Engagement and Reasons for Selecting University Studies Using the Clustering Technique

Erika Yunuen Morales Mateos[1], María Arely López Garrido[1], José Alberto Hernández Aguilar[2], Carlos Alberto Ochoa Ortiz[3], Arturo Corona Ferreira[1], Oscar Alberto González González[1]

[1] Universidad Juárez Autónoma de Tabasco Cunduacán, Mexico

[2] Universidad Autónoma del Estado de Morelos, Cuernavaca, Mexico

[3] Universidad Autónoma de Ciudad Juárez, Mexico

erika.morales@ujat.mx, arely.lopez@ujat.mx, jose_hernandez@uaem.mx, alberto.ochoa@uacj.mx, arturo.corona@ujat.mx,oscar.gonzalez@ujat.mx

Abstract. The objective of this work is to discover groupings of university students, characterized by their student engagement, motivation for career selection and academic performance, this through the application of a descriptive data mining technique called low cluster The k-means algorithm. The study was carried out in a university in southern Mexico with students belonging to careers in technology. The UWES-S instrument and a series of questions related to the reasons for choosing university studies were used to measure student engagement. A descriptive data analysis and a heat map were performed, which graphically represents the variables involved and their relationship, then the techniques necessary for the development of clustering were applied. In order to carry out these analyzes, the R language was used. As results, two clusters were found, cluster 1 is identified by students with a high level of student engagement, and the main reason for selecting the career is due to the possibility of work It offers, followed by the academic quality of the career, on the other hand, cluster 2 is characterized by a medium student engagement also by the possibility of work it offers followed by the aptitudes for studies.

Keywords: Engagement, career choice, heat map, cluster, k-means.

1 Introduction

University students can present different behavior in front of their university careers, different levels of student engagement, which is probably related to other variables such

as career selection reasons, to know if the career they study is the one they wanted to study. Students make the decision to select their career according to a series of parameters, such as if they have aptitudes for it, for the academic quality of the career, and even for the job opportunities offered by the career, among others.

These are elements that are interesting to know, as universities are concerned with offering quality programs to their students and terminal efficiency is an important factor, since the objective is for students to finish their careers in accordance with the guidelines established in the programs of studies. Therefore, in this study it is proposed to analyze student engagement, motivation for career selection and academic performance, in students of technology careers of a university in southern Mexico, so that groups of students with characteristics are identified similar among members of a group and different from other groups.

The identification of groups can be done through clustering, which is a data mining technique that organizes information into segments, has a great capacity for prediction as new data can be classified into existing groups, thanks to this It knows that they share common characteristics and behaviors, and in addition to each identified group, other techniques based on different algorithms can be applied, so that the data can be studied in a more efficient way [1].

In general terms, research has been carried out that measures student engagement with statistical techniques, offering interesting results when related to other variables, among these works are the following:

A work developed by Vizoso and Arias [2] determined the relationship between engagement, burnout, academic performance and career choice priority, in a group of students from the University of León from different careers, resulting in students who are pursuing the career as a first option have higher levels of engagement, that is, high levels of vigor, dedication and absorption, as well as greater academic performance, than those who are not studying the preferred career. As for the burnout that is a state of exhaustion and cynicism, both groups of students feel equally fatigued, it is manifested that it may be due to the academic stress of studying a university degree.

Carrasco and Martínez [3] conducted an investigation on the level of engagement and its implication with academic performance with university students of health sciences, found with respect to the level of engagement, of the 100% evaluated sample He found that 82.0% of university students are at the high level and 18.0% at the low level, which generally means that all students have adequate levels of engagement, although no relationship was found. Significantly with academic performance in general, the analyzes by faculties reported that there is a significant relationship in two faculties.

In a research carried out at the University of Malaga in Spain, on perceived stress, burnout and engagement in university students, the results obtained show that students have medium-low levels in cynicism, means in exhaustion, inadequacy, vigor, absorption and stress perceived and medium-high in dedication.

In general, as students perceive more stress, they also show greater exhaustion, cynicism and inadequacy and less vigor, dedication and absorption [4].

2 Proposed Materials and Methods

2.1 Description of the Data

This paper shows mainly results on student engagement and the reason for the selection of university studies that have students in technology careers, from a population of 141 students of a computer science faculty at a university in southern Mexico, who Willingly agreed to answer a questionnaire that includes the UWES-S and additional elements related to the investigation, the period of application of the survey was February-August 2015. The sample selected was not probabilistic, directed and for convenience [5].

2.2 Utretch Work Engagement Scale for Students (UWES-S)

Salanova and Schaufeli [6] define engagement as "a positive psychological state characterized by high levels of energy and vigor, dedication and enthusiasm for work, as well as total absorption and concentration in work activity". Student engagement characterized by vigor, dedication and absorption is measured through the Utretch Work Engagement Scale for Students (UWES-S), which is composed of 17 items, measured through a Likert scale with values from zero to six .

Vigor is related to high levels of energy, effort, not fatigue easily, and persistence. Dedication refers to the meaning of studies, to feeling enthusiastic, proud, and inspired. Absorption is feeling happily immersed in their studies and presenting difficulty separating from them, feeling that time passes quickly and forgets everything around [7].

2.3 Circumstances of Choice of Studies

Four variables were used to know the reasons why university students selected their careers, later these variables were related to student engagement. These four variables of the choice of studies are: the educational value, academic quality of the career, aptitudes for studies and the possibility of work. For this, four questions that measure these variables were taken as a basis and are integrated in an answer, where 0 means nothing, 2 equals very little, 3 means little, 4 equals half-mind, 5 corresponds to enough, and 6 means a lot [8].

2.4 Analysis of the Data

A descriptive analysis was applied to know the characteristics of the study population, such as quantities and percentages of the variables gender, age, career and social stratum.

Subsequently, the minimum, maximum, average and standard deviation values of the variables engagement, vigor, dedication and absorption were calculated, as well as the variables related to the reason for choosing university studies.

A heat map is presented that shows through different colors and all the intensity of the relationship between variables, from a light tone to a high ratio, through yellow and orange colors to intense red that indicates a minimum relationship [9].

To know and describe characteristics of groups in the population, the clustering technique was applied with the variables involved in the study. With this technique, the data is grouped so that those belonging to the same group have similarities to each other and those belonging to different groups show significant differences [1]. The algorithm used to carry out the clustering was the k-means, where the number of clusters to be created must be specified in advance.

To determine the number of clusters that should be presented, the elbow method was used, where in a graph the point of the elbow is the place where there is a significant change and is the number of clusters that should be generated.

The R language was used to perform the different analyzes presented. R is a powerful software in the implementation of complex graphics and analysis, since it has a wide variety of packages for these tasks [9, 10].

Clustering

Clustering is a descriptive task that consists in obtaining "natural" groups from a data set. The data is grouped based on the principle of maximizing the similarity between the elements of a group by minimizing the similarity between the different groups. Clustering is related to summarization, where each group is considered as a summary of the elements that form them in order to accurately describe the data [11].

An algorithm widely used to group by partitioning is the k-means thanks to its simplicity, it is described in Algorithm 1. To use it you must specify the number of clusters that are going to be generated, it is the parameter k, for which you randomly select k elements, which represent the center or average of each cluster. Subsequently, each of the instances (example) is assigned to the center of the nearest cluster according to the Euclidean distance that separates it from it.

For each cluster the centroid of all its instances is calculated. These cen-troids are taken as the new centers of the clusters. The entire process is repeated with the new cluster centers. The iteration continues until the assignment of the same instances to the same clusters is repeated, since the central points have stabilized and will remain in variables after each iteration [1].

Algorithm 1. K-means [1].

Algorithm 1: K-means

Choose k examples that act as seeds (k number of clusters).

For each example, add example to the most similar class.

Calculate the centroid of each class, which become the new seeds.

If you do not reach a convergence criterion (for example, two iterations do not change the classifications of the examples), go back to step 2.

Table 1. Characteristics of the population study sample.

Variables	Values	N	%
Gender	Women (1)	54	38
	Men (2)	87	62
Age	18-19	35	25
	20-21	69	49
	22-23	37	26
Career	LIA (1)	51	36
	LSC (2)	71	50
	LTI (3)	12	9
	LT (4)	7	5
Social stratum	Low-Low (1)	17	12
	Low-High (2)	43	30
	Medium-Low (3)	68	47
	Medium-High (4)	11	7
	High-Low (5)	5	3
	High-High (6)	1	1

Table 2. Descriptive statistics of the population study sample.

Variables	Mean	Standard Deviation	Minimum	Maximum
Engagement	4.32	0.94	1.60	6.00
Vigor	4.06	0.99	1.50	6.00
Dedication	4.77	0.98	1.80	6.00
Absorption	4.13	1.05	1.20	6.00
Formative value	4.22	1.24	1.00	6.00
Academic quality	4.23	1.52	1.00	6.00
Aptitudes	4.37	1.09	1.00	6.00
Possibility work	4.67	1.11	1.00	6.00
Average	8.15	0.69	6.70	9.80
Selected career	2.45	1.33	0.00	4.00

3 Results

Table 1 shows a sample of 141 students, 87 men and 54 women, the age ranges from 18-19, 20-21, 22-23, where the values of 35, 69 and 37 correspondingly, the population of students with ages in the range of 20-21 being greater. The courses considered are four, where the largest population is in LSC (Bachelor in Computer Systems) with 71, followed by LIA (Bachelor in Administrative Computer Science) with 51, LTI (Bachelor in Information Technology) with 12 and LT (Licensed in Telematics) with 7. The social stratum of the students is mostly medium-low with 68 students.

A descriptive analysis was carried out that shows the mean, standard deviation, minimum and maximum, for the variables presented in Table 2. The student

Table 3. Proposed scale to measure the UWES-S (Adapted from Schaufeli and Bakker [7]).

Valores	Engagement
Very low	score < 2.20
Low	2.20 ≤ score < 3.30
Medium	3.30 ≤ score < 4.70
High	4.70 ≤ score ≤ 6.00

engagement, presents a minimum of 1.60 and a maximum of 6.00, so as an average of 4.32. Of the dimensions of the student engagement, the dedication presents the highest average value with 4.77, minimum of 1.80, maximum of 6.00, followed by absorption with an average of 4.13, minimum of 1.20, maximum of 6.00, finally the vigor with a average of 4.06, minimum of 1.50, maximum of 6.00,

Regarding the career choice by, the training value, academic quality, aptitudes and work possibility presents in the average the values of 4.22, 4.23, 4.37, 4.67 correspondingly, as well as a minimum of 1.00 and a maximum of 6.00. As for the average, it has values of minimum 6.70, maximum of 9.80 and an average of 8.15. Another variable is the selected race where the average is 2.45, the minimum of 0 and the maximum of 4.00.

Table 3 shows an equivalence table for the categorical values according to their score, with established values the observations of the cases can be evaluated. Schaufeli and Bakker present more techniques to obtain the level of engagement to the UWES-S [7].

Using the proposed scale, it can be interpreted in a general way that the student engagement is 4.32, so that the students feel regularly engaged.

3.1 Heat Map of the Correlations

The correlation is a descriptive task that analyzes the percentage of similarity between the values of two numerical variables. The mathematical model with a correlation coefficient r is used, which takes values between 1 and -1.

The strongly correlated variables have a coefficient of 1 or -1 (positively or negatively), on the contrary if the value is 0 there is no correlation. These analyzes allow studying relationships between cause-effect attributes [1].

To find the possible relationships between variables, a heat map was used, since visualization is one of the first forms of data inspection.

The variables in the data set are shown in the heat map of Figure 1, the level of correlation is observed depending on the intensity of the color, the most intense color indicates a correlation closer to the value 0 and the lighter color indicates a correlation closest to 1.

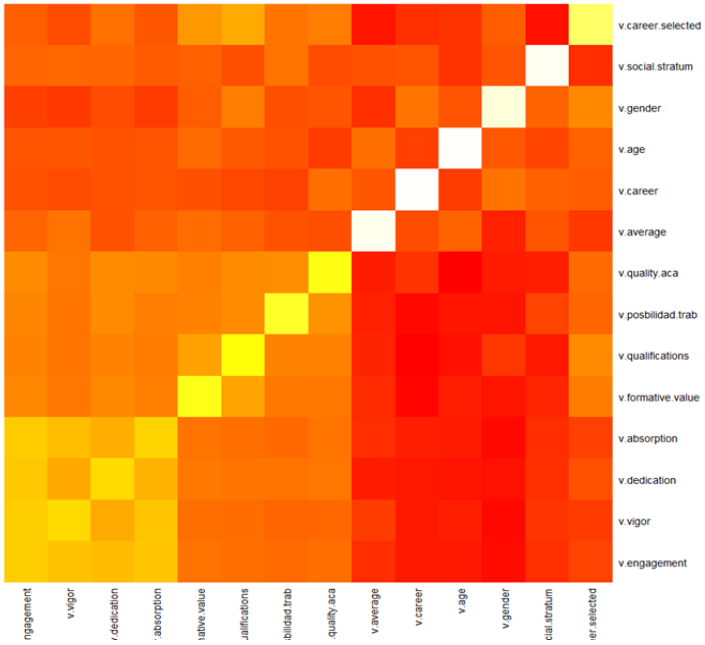

Fig. 1. Heat map of the correlations between attributes.

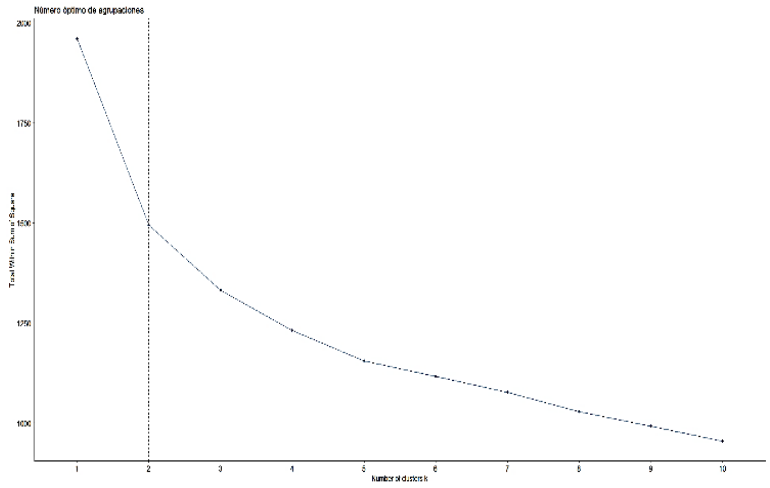

Fig. 1. Elbow technique to determine the optimal number of clusters.

Figure 1 shows that the most related variables are those found in the lower left with lighter colors, student engagement, vigor, dedication, absorption, formative value, academic quality, skills and possibility. work; on the contrary, the less related are found in the lower right, a box with darker shades (red) is observed.

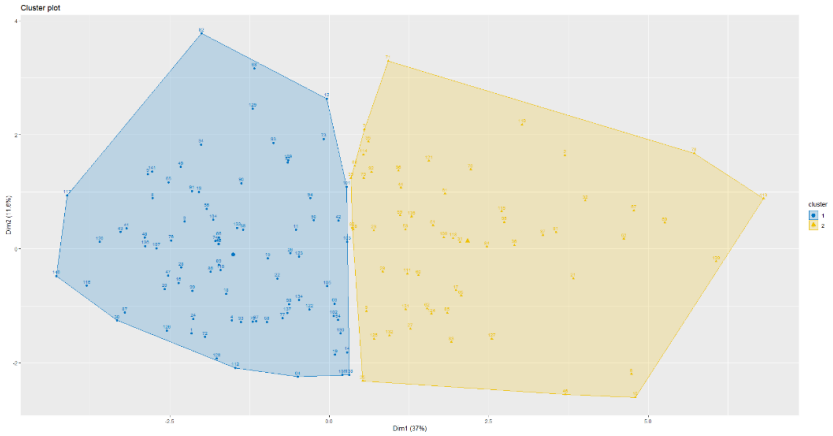

Fig. 2. Cluster graph obtained with the k-means algorithm.

3.2 Clustering

Clustering is a technique that allows you to find common properties that group a set of data. It seeks to maximize the distance of the instances to the grouping to which they do not belong, minimizing the distance of the own [1].

To carry out this analysis, the optimal number of clusters was first determined, the elbow technique that determines the moment when increasing the number of clusters does not imply a substantial measure in a quality measure was applied. Figure 2 shows that according to the elbow technique the optimal number of clusters is two.

Subsequently, the clustering of two groups was created, using the k-means algorithm, Table 4 shows the results,

Cluster 1 corresponds to the student engagement with a high level with 4.88, as well as its dimensions of vigor, dedication and absorption, with high values, 4.59, 5.34, 4.73 respectively; As for the variables that determine the reason for choosing studies, he leads the possibility of work offered by the career with 5.15, followed by the academic quality of the career with 4.91, very close to the aptitudes for studies with 4.90, subsequently by the formative value of the career with 4.81. Additionally, they present approximately an academic average of 8.21, they are from the LSC career, their age is 21, it is made up of men and women, the social stratum is medium-low and the career they study was the one they wanted to study.

Cluster 2 corresponds to the student engagement with a medium level with 3.51, as well as its dimensions of vigor, dedication and absorption, with average values, 3.30, 3.96 and 3.28 respectively; For the variables that determine the reason for the choice of studies, it leads the possibility of work offered by the career with 3.98, followed by the aptitudes for the studies with 3.62, later by the formative value of the career with 3.37 and the academic quality of the race with 3.25. As for the additional variables, this group has an academic average of 8.10, they are mainly from the LSC and LIA, their

Table 4. Characteristics of the obtained cluster.

Variables	Cluster 1	Cluster 2
v.engagement	4.88	3.51
v.vigor	4.59	3.30
v.dedication	5.34	3.96
v.absorption	4.73	3.28
v.formative.value	4.81	3.37
v.qualifications	4.90	3.62
v.posbilidad.trab	5.15	3.98
v.quality.aca	4.91	3.25
v.average	8.21	8.10
v.career	1.89	1.72
v.age	20.79	20.75
v.gender	0.63	0.58
v.social.stratum	2.71	2.39
v.career.selected	2.95	1.74

age is 21 years, it is made up of men and women, the social stratum is low-high and the Career they studied was the one they wanted to study to a small extent.

Figure 3 shows the grouping obtained with the k-means algorithm, it is observed that the resulting clusters are not overlapping and all instances were classified.

4 Conclusions

In this work, an analysis has been developed that involves variables related to studies, that is, student engagement, knowing the reasons for career selection, knowing if the career they study is what they wanted, academic performance and other personal identification variables. The R language was used for the implementation of data analysis and representation techniques. Descriptive statistics indicated that for the general population it has a engagement of 4.32 considered medium, dedication of 4.77, followed by absorption with 4.13, and vigor 4.06, an average of academic performance of 8.15 and the selected career where the average is 2.45 That means medium-mind.

In the heat map, the relationship between variables was identified, the ones that present the most relationship are the dimensions of student engagement with those of reasons for the selection of university studies. Finally, a data mining technique called clustering was applied using the k-means algorithm, where two groups were identified, group 1 presents high characteristics of student engagement unlike group 2 that presents a medium engagement; both groups present as a reason for choosing studies the

possibility of work offered by the career, at different levels, group 1 quite, group 2 medium; In addition, group 1 selected their career fairly and group 2 a little, both groups have an average of about 8.

In these results in group 2 despite the average level of engagement, since it was not the career selected at a high level, maintains a favorable academic average. It is proposed to continue studying both identified groups, with other techniques to learn more about population behavior, as well as to replicate this study in other careers to know if there are differences given the nature of the same.

References

1. García, J., Molina, J.M., Berlanga, A., Patricio, M.A., Bustamante, A.L., Padilla, W.R.: Ciencia de datos, técnicas analíticas y aprendizaje estadístico. Alfaomega (2019)
2. Vizoso, C.M., Arias, O.: Engagement, burnout y rendimiento académico en estudiantes universitarios y su relación con la prioridad en la elección de la carrera. Revista de Psicología y Educación, 11(1), pp. 45–60 (2016)
3. Carrasco, M.A., Martínez, C.: Nivel de engagement y su implicancia en el rendimiento académico en estudiantes universitarios de ciencias de la salud de la Unheval-Huänuco. Revista Boletín Redipe, 8(2), pp. 131–139 (2019)
4. Vallejo, M., Aja, J., Plaza, J.J.: Estrés percibido en estudiantes universitarios: influencia del burnout y del engagement académico. International Journal of Educational Research and Innovation (IJERI), 9, pp. 220–236 (2018)
5. Hernández, R., Fernández, C., Baptista, M.: Metodología de la investigación. McGraw-Hill Interamericana (2010)
6. Salanova, M., Schaufeli, W.B.: El Engagement de los empleados un reto emergente para la dirección de recursos humanos. Estudios Financieros, (261), pp. 109–138 (2004)
7. Schaufeli, W.B., Bakker, A.: Utrecht work engagement scale (UWES). Escala de Engagement en el trabajo de Utrecht, Ocuppational Health Psychology Unit: Utrech University (2003)
8. Artunduaga, M.: Cuestionario sobre rendimiento académico y deserción en la universidad. Universidad Complutense de Madrid Facultad de Ciencias de la Educación Departamento de Métodos de Investigación y Diagnóstico en Educación (MIDE) (2005)
9. Guisande, C., Vaamonde, A.: Gráficos estadísticos y mapas con R. Ediciones Díaz de Santos (2013)
10. Development Core Team: A language and environment for statistical computing. R Foundation for Statistical Computing (2019)
11. Hernández, J., Ramírez, M.J., Ferri, C.: Introducción a la minería de datos. Editorial Pearson Educación (2004)