ISBN 978-989-96995-3-3 (2019-06-07)

ANÁLISE DE CLUSTERS PARA SEGMENTAÇÃO DE ESTUDANTES NUMA INSTITUIÇÃO DE ENSINO SUPERIOR

CLUSTERS ANALYSIS FOR SEGMENTATION OF STUDENTS IN AN HIGHER EDUCATION INSTITUTION

Pedro Sobreiro[1,3]; Domingos Martinho[2,3]
[1]Instituto Politécnico de Santarém; [2]ISLA Santarém; [3]CEPESE, Porto
sobreiro@esdrm.ipsantarem.pt; domingos.martinho@islasantarem.pt;

Resumo

A segmentação do mercado é um tema importante para os administradores das instituições de ensino superior. A segmentação dos alunos permite a diferenciação e a definição de ações personalizadas de acordo com cada segmento e pode ser realizada recorrendo a dados existentes de alunos para serem posteriormente utilizados no desenvolvimento de ações de comunicação ou para realização de um acompanhamento interno diferenciado.
A metodologia utilizada para realizarmos a segmentação dos alunos (n=280) recorreu à análise de clusters utilizando o algoritmo k-means disponível na biblioteca scikit. O k-means é um algoritmo não supervisionado para a determinação dos clusters, que requer que o investigador determine à priori o número de clusters pretendidos, utilizando uma aproximação iterativa calculando o centro ótimo de cada cluster. A identificação do número de clusters foi baseada no método *elbow*, que utiliza o pressuposto de que o número de clusters ótimo é aquele em que adicionando mais clusters não reduz significativamente a variância entre clusters. Depois de obtivermos cada cluster realizamos a sua caraterização utilizando as variáveis existentes para termos uma melhor compreensão dos dados.
Os resultados obtidos permitiram identificar três clusters, onde obtivemos no cluster um 89 alunos, cluster dois 16 alunos e cluster três 175 alunos. Para facilitar a compreensão dos resultados obtidos realizamos a redução das variáveis existentes através de do *Principal Components Analysis*, uma redução de dimensões para podemos projetar os dados num espaço dimensional menor de duas dimensões, num gráfico de dispersão x,y. Realizamos a caraterização (média±desvio padrão) das variáveis idade, ano, estado civil e sexo.
Os resultados obtidos evidenciam que nos clusters um, dois e três as médias de idades são aproximadamente iguais 28,29 e 31, o estado civil é maioritariamente solteiros com 80%, 81% e 75% e o sexo feminino representa 49%, 51% e 50% respetivamente.
Os resultados conseguidos não são elucidativos considerando os indicadores obtidos em cada cluster. Para podermos retirar melhores conclusões deveriam ser incluídas mais variáveis, como cursos frequentados, resultados obtidos na frequência do curso e aumentar a amostra. Um aspeto que poderia ter sido equacionado seria a normalização dos dados, reduzindo impacto de variáveis em escalas diferentes na determinação do número de clusters. Por último seria interessante explorar as diferenças entre os alunos nos clusters existentes realizando a análise das variáveis existentes.

Palavras-chave: *Análise de clusters, alunos ensino superior, segmentação.*

Abstract

Market segmentation is a crucial issue for administrators of higher education institutions. Segmentation of students allows the differentiation and definition of customized actions according to each segment and can be performed using existing data of students to be used later in the development of communication actions or to perform differentiated internal monitoring.
The methodology used to segment the students (n = 280) resorted to the analysis of clusters using the k-means algorithm available in the scikit library. The k-means is an unsupervised algorithm for the determination of clusters, which requires that the researcher determines a priori the number of clusters desired, using an iterative approach calculating the optimal center of each cluster. The identification of the number of clusters was based on the elbow method, which uses

the assumption that the optimal number of clusters is one in which adding more clusters does not significantly reduce the variance between clusters. After obtaining each cluster, we perform its characterization using the existing variables to have a better understanding of the data.
The results obtained allowed to identify three clusters, where we obtained in the cluster one 89 students, cluster two 16 students and cluster three 175 students. In order to facilitate the understanding of the obtained results we reduced the existing variables through Principal Components Analysis, a reduction of dimensions so we can represent the data in a smaller dimension space of two dimensions, using a scatter plot x, y. We performed the characterization (mean ± standard deviation) of the variables age, year, marital status and sex.
The results show that in the clusters one, two and three the mean ages are approximately the same 28,29 and 31, the marital status is mostly unmarried with 80%, 81% and 75% and the female sex represents 49%, 51 % and 50% respectively.
The results obtained are not elucidative considering the indicators obtained in each cluster. In order to obtain better conclusions, more variables, such as courses attended, results obtained in the frequency of the course and increase of the sample should be included. One aspect that could have been considered could be the normalization of the data, reducing the impact of variables in different scales in determining the number of clusters. Finally, it would be interesting to explore the differences among students in existing clusters by performing the analysis of existing variables.

Keywords: *Cluster analysis, higher education students, segmentation.*