ISSN 2477-9105
DOI: 10.47187/perf.v1i25.118

Perfiles

Número 25 Vol.1 (2021)

APLICACIÓN DEL ALGORITMO K-MEDOID PARA LA SEGMENTACIÓN DE LOS ALUMNOS INGRESANTES DE UNA UNIVERSIDAD.

Application of the k-medoid algorithm for the segmentation of entering students at a university.

Ledvir Ayrton Walter Chavez Valderrama*, Jesús Walter Salinas Flores

Universidad Nacional Agraria la Molina, Facultad de Economía y Planificación, Departamento de Estadística e Informática, Lima, Perú.

*lchavezvalderrama@gmail.com

Resumen

Actualmente, en el área de educación superior se ha vuelto indispensable la gestión de los datos para la toma de decisiones académicas y la mejora de los procesos educativos, para ello la analítica y estadística han sido llevados al ámbito tecnológico, donde prima la automatización de procesos y la gestión de grandes bases de datos a través de algoritmos de Machine Learning, uno de los más utilizados son los algoritmos clustering, cuyo propósito es agrupar datos por similitud. El presente estudio tuvo como objetivo encontrar tipos de estudiantes universitarios respecto a sus variables sociodemográficas, económicas y de rendimiento académico, utilizando el algoritmo K-medoid en datos de alumnos ingresantes a la Universidad Nacional Agraria La Molina de Lima, Perú. Se pudo determinar que los ingresantes en estudio se pueden segmentar en 3 grupos, cada uno con características propias, lo que permitirá impulsar cambios a favor de la calidad educativa y promover la renovación de los espacios de enseñanza de manera personalizada en torno al tipo de estudiante que la universidad gestiona.

Palabras Claves: Perfil del ingresante, algoritmos de agrupamiento, segmentación, K-medoid.

Abstract

Currently, in the area within higher education, data management has become essential for academic decision making and the improvement of educational processes. Analytics and statistics have been taken to the technological field, where the processes automation and the large databases management through Machine Learning algorithms are the most used, among which are the clustering algorithms, whose purpose is to group data by similarity. The objective of this study was to find types of university students with respect to their sociodemographic, economic and academic performance variables, using the K-medoid algorithm on data of students entering the Universidad Nacional Agraria La Molina in Lima, Peru. It was determined that the students under study can be segmented into 3 groups, each with its own characteristics, which will make it possible to promote changes in favor of educational quality and promote the renovation of teaching spaces in a personalized way around the type of student that the university manages.

Keywords: Admitted student profile, clustering algorithms, segmentation, K-medoid.

Fecha de recepción: 03-01-2021 **Fecha de aceptación:** 09-03-2021 **Fecha de publicación:** 31-05-2021

I. INTRODUCCIÓN

La preocupación de la comunidad educativa y los responsables de las políticas educativas en las instituciones de educación superiores gira en torno a mejorar la eficiencia académica y del entorno educativo, buscando prevenir problemas como la deserción universitaria que en el Perú anualmente puede

alcanzar el 30% de la cantidad de alumnos ingresantes (7), universitarios que por distintas razones como: problemas económicos, falta de vocación en la carrera profesional, falta de apoyo por parte de la universidad y plana universitaria (profesores/orientadores), expectativas defraudadas en la formación y bajo rendimiento académico dejan sus estudios superiores; para ello, es de suma importancia conocer al estudiante que se gestiona desde su inicio en la vida universitaria; conocer sus fortalezas y debilidades, ello permitirá al docente evaluar y proponer las mejores prácticas y metodologías que requieran sus estudiantes.

El objetivo principal de la investigación es lograr identificar cuáles son los distintos grupos de estudiantes que ingresan a una universidad. Se busca adicionalmente caracterizar cada uno de estos grupos y entender sus peculiaridades, conocimientos que promueven la sinergia de esfuerzos entre estudiante – docente, para que este último tenga información del tipo de ingresante que gestiona y con ello diseñe estrategias y renueve sus espacios de enseñanza de manera personalizada, aprovechando toda información del ámbito de la enseñanza (8).

II. MATERIALES Y MÉTODOS

La investigación fue realizada con los datos de los alumnos ingresantes de la Universidad Nacional Agraria La Molina (UNALM) en Lima, Perú durante los semestres 2015-I y 2015-II, los datos fueron obtenidos a partir de la vinculación entre las bases de datos de la Oficina de Estudios y Registros Académicos, del Centro de Admisión y Promoción y la Oficina de Bienestar Universitario y Asuntos Estudiantiles.

La población investigada fueron todos los alumnos ingresantes de la UNALM de las modalidades: Concurso Ordinario y Dos Primeros Puestos de Colegios de Educación Secundaria, con un total de 690 estudiantes. Las variables identificadas en la aplicación de ambas técnicas se muestran en la Tabla 1.

Variable	Descripción
Variables sociodemográficas	Tiempo transcurrido desde que terminó el 5to año de secundario e ingresó a la universidad, Edad del ingresante al momento del examen de admisión, Ubicación del colegio donde cursó el 5to año de secundaria (Lima o Provincia), Sexo del ingresante
Variables socioeducativas	Tipo de institución de procedencia (Privada o Pública)
Variables socioeconómicas	Aporte semestral asignado al ingresante
Variables de rendimiento en las áreas del conocimiento en la secundaria	Nota obtenida en el 5to año de secundaria en el área de Ciencia tecnología y Ambiente, en el área de Comunicación, en el área de Matemática, Nota promedio del último año de estudios
Variables de rendimiento en el examen de Admisión	Nota obtenida en los cursos de RM, RV, Matemática, Física, Química y Biología en el examen de admisión. Nota general obtenida en el examen de admisión. Si el alumno pertenece o no al tercio superior en la especialidad a la que ingresó.
Variables de elección en el ingreso a una carrera	Modalidad de ingreso a la universidad. Carrera a la que ingresó. Orden de elección que tuvo la carrera a la cual ingresó (1°, 2° o 3° opción)

Tabla 1. Determinar el número de clusters con el índice de Dunn

El tipo de investigación fue de carácter descriptivo, se identificó los grupos de ingresantes de la UNALM a través de la descripción de sus variables. El diseño de la investigación fue de carácter no experimental-transversal, ya que se contó con datos de los estudiantes que se recolectaron de diferentes fuentes. Para identificar los grupos se utilizó un algoritmo clustering que es un método exploratorio multivariado iterativo no supervisado (22, 23, 26) que describe el comportamiento de los objetos en grupos en la fase exploratoria de su investigación, ya que el resultado es exclusivo de los objetos incluidos en el análisis (27) de modo que el analista no asigna las clases previamente, es utilizado en varias áreas desde la década de 1960 (10, 24). El clustering clasifica los objetos, asignándolos en grupos internamente homogéneos, pero también heterogéneos entre ellos (4, 9, 17). Uno de los algoritmos clustering más utilizados y conocidos es el K-means (6, 15), técnica que distribuye los objetos a través del sistema de particiones en un número k de clusters previamente definido por el investigador (19, 13), sin embargo, este enfoque, tiene un inconveniente frente a la presencia de elementos con outliers (2, 12, 18) que pueden tener un efecto extremo en el análisis y provocar un agrupamiento inadecuado (3, 14, 20).

Frente a ello, se han desarrollado algoritmos más apropiados para lidiar con los valores atípicos (21).

Un algoritmo más robusto a los outliers y al ruido, que ocurren en un ambiente real sin control es el algoritmo K-medoid (5), el cual se basa en similitud (1). En lugar de utilizar la media convencional, se utiliza medoids para representar los clusters (16).

El medoid es un elemento del conjunto de datos y es el más centralizado del conjunto de datos. El algoritmo K-medoids inicia con la selección aleatoria de *k* elementos de datos como centros iniciales para representar los *k* clusters, los elementos restantes se incluyen en el grupo que tiene el medoid más cercano a ellos y posteriormente se determina un nuevo centro que puede representar mejor al grupo. En cada iteración, todos los elementos distintos a los centros se asignan nuevamente a los clusters que tienen el medoid más cercano, provocando que los medoids alteren su ubicación.

El algoritmo minimiza la suma de las distancias entre cada elemento de datos y su correspondiente medoid, este ciclo se repite hasta que ningún medoid cambie su colocación, esto marca el final del proceso y se tienen los clusters finales. La ubicación de cada centro puede cambiar en cada una de las $\frac{n!}{k!(n-k)!}$ iteraciones, así se encuentran los k clústers que representan n objetos de datos; el algoritmo fue diseñado para no depender del orden de las observaciones o una semilla inicial, debido a que prueba todas las posibles combinaciones, por lo que siempre converge en la misma solución.

Para evaluar cuán diferente son dos observaciones de tipo mixto X e Y con m atributos, donde se tiene p atributos numéricos y m-p atributos categóricos, el algoritmo calcula la disimilitud (11), como:

$$d(X,Y) = \sum_{j=1}^{p}(x_j - y_j)^2 + \gamma \sum_{j=p+1}^{m}\delta(x_j, y_j) \quad (1)$$

donde el primer término es la medida de distancia euclidiana al cuadrado en los atributos numéricos y el segundo es la medida de disimilitud de coincidencia simple en los atributos categóricos, siendo $\delta(x_j, y_j)=0$ para $x_j=y_j$ y $\delta(x_j, y_j)=1$ para $x_j \neq y_j$, γ es un peso para atributos categóricos, introducido para evitar favorecer cualquier tipo de atributo. Un cálculo estimado de γ es de la siguiente manera:

$$\gamma = \frac{Promedio\ (Varianza\ o\ desviación\ estándar\ de\ las\ variables\ numéricas)}{Promedio(Heurística\ para\ variables\ categóricas)} \quad (2)$$

donde la heurística para variables categóricas se calcula como: $1- \sum_{j=p+1}^{m} p_j^2$ o $1-max(p_j)$ con $j = p+1,\ldots,m$; siendo p_j la proporción de la categoría j en la variable cualitativa. La solución para encontrar el mejor algoritmo clustering y el número óptimo de conglomerados *k* se llama generalmente validez del cluster. Para esta investigación, se utilizó el Índice de validación de Dunn (25), cuyo objetivo es identificar un conjunto de clústeres que sean compactos, con una varianza pequeña entre los miembros del clúster, y que éstos estén bien separados de los miembros de otros clústeres. Un valor más alto del índice de Dunn indica un mejor rendimiento del algoritmo de clustering, tiene un valor entre cero e infinito.

III. RESULTADOS

Para aplicar el algoritmo K-medoid es necesario conocer a priori el número de clusters k a formarse. En este caso, se utilizó el índice de validación interna de Dunn, calculándolo de manera iterativa cambiando el número de cluster y el valor de semilla inicial, el valor de k seleccionado fue aquel que permitió obtener el índice de Dunn más alto.

N° Cluster											
2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
0.1	0.16	0.13	0.13	0.1	0.15	0.11	0.11	0.11	0.08	0.08	0.08

Tabla 2. Determinar el número de clusters con el índice de Dunn

Se observa en el Tabla 2 que al aplicar el algoritmo K-medoid los valores del índice de validación interna de Dunn óptimos fue 0.16 por lo que el número clusters óptimo es *k=3*.
Analizando los resultados obtenidos, se realizó la Tabla 3 de resumen general para caracterizar cada uno de los grupos de ingresantes 2015 de la UNALM. Asimismo, se obtuvo que el 36% de los ingresantes pertenecen al cluster 1, el 42% al cluster 2 y el 22% al cluster 3.

Denominación	Cluster1	Cluster2	Cluster3
Años_colegio_admisión	↓	→	↑
Edad_admisión	↓	→	↑
Aporte_Semestral	↑	→	↓
CTA_Colegio	→	↑	↓
COM_Colegio	→	↑	↓
MAT_Colegio	→	↑	→
Nota_Colegio	→	↑	↓
RM_Admisión	↑	→	↓
RV_Admisión	↑	→	↓
MAT_Admisión	↑	↓	→
FIS_Admisión	↑	↓	→
QUI_Admisión	↑	↓	→
BIO_Admisión	↑	↓	→
Nota_Admisión	↑	↓	→
Dept_Colegio	Lima y provincia	Lima y provincia	Lima
Sexo	Masculino	Femenino	Masculino
Tipo_Colegio	Pública	Privada y Pública	Pública
Tercio_Superior_Esp	Si	No	No
Modalidad	Concurso Ordinario	Concurso Ordinario y Dos Primeros Puestos de Colegios de Educación	Concurso Ordinario
Elección_Esp_Ingreso	Primera	Segunda o Tercera	Segunda o Tercera

Tabla 3. Resumen general de los segmentos formados

Se observó que el cluster con mayor porcentaje de alumnos ingresantes fue el 2 con 42%, dado los resultados obtenidos los ingresantes se clasificaron en:

Cluster		Características
1	Ingresante previsto	Se caracterizan por evidenciar conocimientos previstos o esperados al ingresar a la universidad, ya que en su mayoría mostraron tener un alto rendimiento en el examen de admisión con desempeño académico medio en el colegio, en su mayoría ocuparon el tercio superior en su carrera e ingresaron a la especialidad que eligieron su primera opción, se les fue asignado un aporte semestral mayor al promedio dada su situación socioeconómica.
2	Ingresante en proceso	Se caracterizan por estar en camino a lograr conocimientos previstos o esperados al ingresar a la universidad por lo cual requieren acompañamiento durante un tiempo razonable para alcanzarlo, en su mayoría mostraron tener un desempeño académico muy bueno en el colegio pero no suficiente para afrontar el examen de admisión ya que alcanzaron un rendimiento entre regular y bajo en este, en su mayoría no ocuparon el tercio superior en su carrera e ingresaron a la especialidad que eligieron como su segunda o tercera opción, se les fue asignado un aporte semestral igual al promedio dada su situación socioeconómica.
3	Ingresante en inicio	Se caracterizan por estar empezando a desarrollar conocimientos previstos o esperados al ingresar a la universidad por lo cual necesita mayor tiempo de acompañamiento e intervención del consejero de acuerdo con su ritmo y estilo de aprendizaje para alcanzarlo, en su mayoría mostraron tener un desempeño académico malo en el colegio y alcanzaron un rendimiento entre regular y bajo en el examen de admisión, en su mayoría no ocuparon el tercio superior en su carrera e ingresaron a la especialidad que eligieron como su segunda o tercera opción, se les fue asignado un aporte semestral menor al promedio dada su situación socioeconómica.

Tabla 4. Características de cada segmento

Con el fin de validar los resultados obtenidos de la segmentación se cruzó esta información con el promedio ponderado acumulado de los alumnos que obtuvieron al término de su primer año de estudios superiores, ya que en este periodo los universitarios llevan cursos generales que buscan reforzar sus conocimientos adquiridos antes de ingresar a la universidad. Para el análisis se clasificó el promedio ponderado acumulado como: EXCELENTE: notas ente 16.5 y 20, BUENO: notas entre 12,5 y 16.5,

REGULAR: notas entre 10,5 y 12.5 o MALO: notas entre 0 y 10.5. Se observó que más de la mitad de los ingresantes que tuvieron un promedio ponderado acumulado en su primer año de estudios EXCELENTE se encuentran en el cluster 1, los alumnos BUENOS y REGULARES se encuentran en su mayoría en el cluster 2, mientras que los alumnos MALOS se encuentran agrupados en el cluster 3. Validando así lo mencionado anteriormente.

Todo esto permite entender que los clusters 2 y 3, son los perfiles de ingresantes que deben ser atendidos con prioridad por autoridades pertinentes dentro de la institución, a través de diversas estrategias educativas, apoyo económico y orientación con el fin de que a futuro no tengan bajo rendimiento académico, retraso en sus estudios, dilatación del tiempo de estudio, deserción, entre otros.

V. CONCLUSIONES

Al aplicar el algoritmo de clustering K-medoid, es posible agrupar a los ingresantes de una universidad pública respecto a sus variables socioeconómicas, demográficas y de rendimiento educativo, se pudo identificar 3 tipos de ingresantes cada uno con características diferentes, se denominaron:

Ingresante previsto, Ingresante en proceso e Ingresante en inicio; este último dado sus características necesita mayor tiempo de acompañamiento e intervención del consejero de acuerdo con su ritmo y estilo de aprendizaje frente a los otros segmentos, por otro lado el Ingresante previsto puede ser considerado el grupo de ingresantes con mejores características para la universidad.

VI. AGRADECIMIENTOS

Los autores desean agradecer al personal de la Universidad Nacional Agraria La Molina
por las facilidades brindadas en la recopilación de la información de las bases de datos de cada oficina.

Referencias

1. Arora P, Virmani D, Varshney S. Analysis of K-Means and K-Medoids Algorithm for Big Data. Procedia Computer Science. 2016; 78: 507-512. Disponible en: https://bit.ly/2s5X9xy.
2. Adams J, Hayunga D, Mansi S, Reeb D, Verardi V. Identifying and treating outliers in finance. Financial Management. 2019; 48(2): 345–384. Disponible en: https://cutt.ly/9hawFSN.
3. Acock A. A gentle introduction to Stata. 4th ed. College Station: Stata Press. 2014
4. Aggarwal C. An introduction to cluster analysis. In C. Aggarwal, C. Reddy (Eds.). Data clustering: Algorithms and applications (pp. 1-28). New York: CRC Press. 2014.
5. Bhat A. K-medoids clustering using partitioning aroud medoids for performing face recognition. International Journal of Soft computing, Mathematics and Control. 2014; 3(3): 1-12. Disponible en: https://cutt.ly/Hharh0H.
6. Boehmke B, Greenwell B. K-means Clustering. In Hands-On Machine Learning with R (pp. 399–416). 1st ed. New York: CRC Press. 2014. Disponible en: https://cutt.ly/KhaqBcJ.
7. Castro M. Factor principal que determina la deserción de los estudiantes del primer y segundo ciclo de una universidad privada de lima - campus lima centro, durante el periodo 2018 I – II [Tesis de maestría]. Perú: Universidad Tecnológica del Perú; 2019. Disponible en: https://n9.cl/157s.
8. Eckert K, Suénaga R. Aplicación de técnicas de Minería de datos al análisis de situación y comportamiento académico de alumnos de la UGD. In XV Worshop de Investigadores en Ciencias de la Computación. Argentina. 2013. Disponible en: https://bit.ly/2QSvppC.
9. Everitt B, Hothorn, T. Cluster analysis. In B. Everitt, T. Hothorn, An Introduction to Applied Multivariate Analysis with R (pp. 163–200). 1st ed. New York: CRC Press. 2011.
10. Fávero L, Belfiore P. Análise de agrupamentos. In Manual de análise de dados: Estatística e modelagem multivariada com Excel, SPSS e Stata (pp. 309–378). 1st ed. São Paulo: GEN. 2017.
11. Huang Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery. 1998; 2: 283 - 304. Disponible en: https://bit.ly/2FMUgoH.
12. Hair J, Black W, Babin B, Anderson R. Multivariate data analysis. 8th ed. Ireland: Cengage Learning EMEA. 2018
13. Hartigan J, Wong M. Algorithm AS 136: A K-means clustering algorithm. Journal of the Royal Statistical Society. 1979; 28(1): 100-108. Disponible en: https://bit.ly/30jLpV1.

14. Irizarry R, Love M. Data analysis for the life sciences with R. 1st ed. United Kingdom: Chapman and Hall/CRC. 2016.
15. Janssen A, Wan P. K-means clustering of extremes. Electronic Journal of Statistics. 2020; 14(1): 1211–1233. Disponible en: https://cutt.ly/ihaupE6.
16. Kaufman L, Rousseeuw P. Partitioning around medoids (Program PAM). In Finding groups in data: An introduction to cluster analysis (pp. 68–125). 1st ed. New York: Wiley-Interscience. 1990.
17. Ketchen D, Shook C. The application of cluster analysis in strategic management research: An analysis and critique. Strategic Management Journal. 1996; 17(6): 441–458. Disponible en: https://cutt.ly/Whaq1Kh.
18. Loperfido N. Kurtosis-based projection pursuit for outlier detection in financial time series. The European Journal of Finance. 2020; 26(2–3); 142–164. Disponible en: https://cutt.ly/dhaq0Oc.
19. MacQueen J. Some methods for classification and análisis of multivariate observations. Proceedings of the Berkeley symposium on mathematical statistics and probability. 1967; 1: 281–297. Disponible en: https://cutt.ly/YhaubYD.
20. Malhotra N. Marketing research: An applied orientation. 7th ed. New York: Pearson. 2018.
21. Pandey P, Singh I. Comparision between K-mean clustering and improved K-mean clustering. International Journal of Computer Applications. 2016; 146(13): 39–42. Disponible en: https://cutt.ly/Shaq3uw.
22. Rai P, Singh S. A Survey of Clustering Techniques. International Journal of Computer Applications. 2010; 7(12): 1-5. Disponible en: https://cutt.ly/OhauJpX.
23. Raulji G. A Review on Fuzzy C-Mean Clustering Algorithm. International Journal of Modern Trends in Engineering and Research. 2014; 2(2): 751-754. Disponible en: https://bit.ly/2FSxewM.
24. Scoltock J. A survey of the literature of cluster analysis. The Computer Journal. 1982; 25(1), 130–134. Disponible en: https://cutt.ly/Ghaq8Rg.
25. Vallejo D. Clustering de documentos con restricciones de tamaño [Tesis de maestría]. España: Universidad Politécnica de Valencia; 2015. Disponible en: https://n9.cl/r2mjx.
26. Velmurugan T, Santhanam T. A comparative analysis between K-medoids and fuzzy C-means clustering algorithms for statistically distributed data points. Journal of Theoretical and Applied Information Technolog. 2011; 27: 19-29. Disponible en: https://bit.ly/3867V6o.
27. Wang W, Zhang Y. On fuzzy cluster validity indices. Fuzzy Sets and Systems. 2007; 158(19): 2095-2117. Disponible en: https://cutt.ly/DhaifXB.