Tierra Nuestra
ISSN 2519-738X (En línea), ISSN 1818-4103 (Impresa)
https://revistas.lamolina.edu.pe/index.php/tnu

ARTÍCULO ORIGINAL – RESEARCH ARTICLE

© Los autores. Publicado por la Universidad Nacional Agraria La Molina
El artículo es de acceso abierto y está bajo la licencia CC BY
Recibido: 30/06/2021; Aceptado: 01/12/2021; Publicado: 30/12/2021
http://dx.doi.org/10.21704/.rtnv15i2.1825

Segmentación de los alumnos ingresantes a una universidad pública aplicando el algoritmo K-prototype

Segmentation of admitted student to a public university applying K-prototype algorithm

Ledvir Ayrton Walter Chávez Valderrama [1]*; Jesús Walter Salinas Flores[2]

* Autor de correspondencia: jsalinas@lamolina.edu.pe

RESUMEN

En la actualidad, el análisis de datos es una labor desafiante, especialmente en el campo de la educación, debido a que se realizan investigaciones profundas para conocer, entender y gestionar la diversidad de alumnos que ingresan a cada institución superior y con ello plantear estrategias educativas para mejorar el modelo de enseñanza – aprendizaje. El objetivo de este artículo fue caracterizar el perfil de los ingresantes de una universidad pública respecto a sus variables sociodemográficas, económicas y de rendimiento académico utilizando el algoritmo *K-prototypes*, para lo cual se utilizó datos de alumnos ingresados a la Universidad Nacional Agraria La Molina (Lima, Perú) recolectados a partir del examen de admisión, ficha del ingresante y su certificado de estudios escolares. Se pudo determinar que los ingresados en estudio se ajustan a 5 perfiles, cada uno con características propias, permitiendo agrupar a los ingresados con características similares, contribuyendo a la mejora de políticas de acompañamiento, impulsando cambios a favor de la calidad educativa y promoviendo la renovación de los espacios de enseñanza de manera personalizada en torno al perfil del alumno que la universidad gestiona.

Palabras clave: perfil del ingresado, algoritmos de agrupamiento, segmentación, K-prototype, calidad educativa.

ABSTRACT

Currently, data analysis is a challenging task, especially in the field of education, because in-depth research is carried out to know, understand and manage the diversity of students who enter for higher education institution and with it, to propose educational strategies to improve the teaching-learning model. The objective of this article was to characterize the pro|44file of admitted student of a public university with respect to their socio-demographic, economic and academic performance variables using *K-prototypes* algorithm. For this purpose, data from admitted student of La Molina National Agrarian University (Lima, Peru) were collected from the admission exam, the

Forma de citar el artículo (Formato APA):
Chávez, L., Salinas, J. (2021). Segmentación de los alumnos ingresados a una universidad pública aplicando el algoritmo K-prototype. *Tierra Nuestra*. 15(2), 10-21. http://dx.doi.org/10.21704/rtn.v15i2.1825.

[1] Universidad Nacional Agraria La Molina, 15024, Lima, Perú. lchavezvalderrama@gmail.com
[2] Universidad Nacional Agraria La Molina, 15024, Lima, Perú. jsalinas@lamolina.edu.pe

entrant's file and their school certificate. It was possible to determine that admitted student fits with 5 profiles, each one with its own characteristics, allowing the grouping of students with similar characteristics, contributing to the improvement of support policies, promoting changes in favor of educational quality and promoting the renovation of teaching spaces in a personalized way around the student profile that the university manages.

Keywords: semantic maps, graphic organizers, written texts, reading comprehension, comprehension levels.

1. Introducción

La gestión de datos se ha vuelto indispensable para la toma de decisiones, análisis y elaboración de estrategias a partir del conocimiento oportuno. La educación superior, un área relativamente nueva en este tema, se ha visto enfrentada a grandes retos bajo el impacto de la globalización, el crecimiento económico y la innovación tecnológica. Junto a ello, la preocupación por la mejora continua de la enseñanza-aprendizaje en las universidades incide directamente sobre la funcionalidad de los departamentos académicos. Por tal motivo, las instituciones de educación superior plantean diferentes políticas educativas como el asesoramiento de los alumnos desde el comienzo de la vida universitaria. Frente a esta situación, es vital el conocimiento de las características que poseen los ingresantes, reconocer sus fortalezas y debilidades y con ello direccionar en base a información, estrategias que promuevan la superación de distintas problemáticas que puedan impedir el normal desenvolvimiento académico en estos alumnos. Con este fin, la evaluación del desempeño académico del ingresado es básica y necesaria para controlar la progresión del rendimiento del alumno en una institución superior. Sobre la base de este tema crítico, la agrupación de alumnos en diferentes categorías, de acuerdo con los conocimientos adquiridos en el ámbito escolar y/o universitario, se ha convertido en una tarea compleja, ya que, con la segmentación tradicional de alumnos, basada solamente en sus puntajes promedios, es difícil obtener una visión completa y detallada del estado en el que se encuentra el rendimiento de los universitarios. Sin embargo, la analítica y la estadística han sido llevadas al ámbito tecnológico, donde prima la automatización de procesos y la gestión de grandes bases de datos a través de algoritmos de *Machine Learning* (procedimientos matemáticos que buscan descubrir la mejor solución para una situación dada) con lo cual se descubre información útil que ayuda a los docentes y responsables de las instituciones educativas a determinar la manera adecuada para guiar a sus alumnos, maximizando su aprendizaje y contribuyendo a la mejora de la calidad de la educación superior. Tal como menciona Arias (2015), quien logró caracterizar a ingresados de una carrera en una universidad pública, consiguiendo entender y comprender la situación socio-demográfica, socio-educativa y socio-económica de los alumnos.

Una técnica muy utilizada es el análisis *clustering* con el cual es posible segmentar a los alumnos y descubrir características claves de su rendimiento; asimismo, utilizar esas características para predicciones futuras y empezar una atención personalizada del profesor hacia el alumno desde su ingreso a la universidad.

1. Marco teórico

Análisis clustering

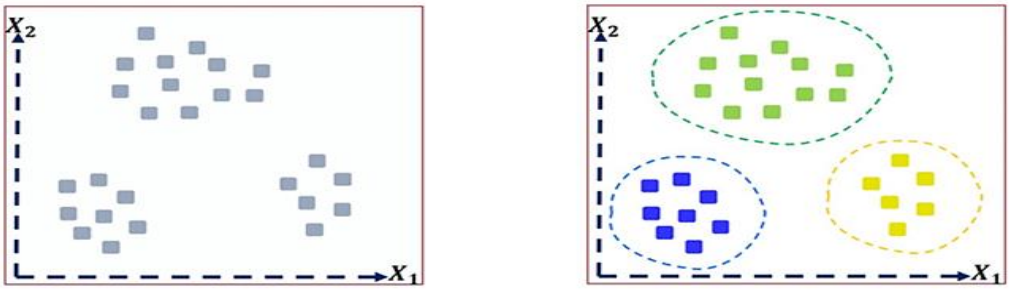
Rai y Singh (2010) señalaron que el análisis *clustering* representa a los datos en segmentos, provocando la pérdida de ciertos detalles finos, pero logrando la simplificación de información en base a análisis matemáticos, estadísticos y numéricos. Desde una perspectiva de técnicas de aprendizaje, el análisis *clustering* corresponde al mundo de patrones ocultos. La búsqueda de segmentos es un aprendizaje no supervisado y el sistema resultante representa un concepto de datos. La figura 1 ilustra un ejemplo de un resultado del proceso de *clustering* dadas las variables X_1 y X_2.

Adicionalmente, Kaur y Kaur (2013) definieron que un *cluster* es una colección de objetos similares entre sí y distintos a los objetos de otros *clusters*. Un

buen algoritmo de agrupamiento es capaz de identificar segmentos independientemente de la distribución de las observaciones. Otros requisitos de los algoritmos de agrupamiento son la escalabilidad, la capacidad de tratar con datos ruidosos y la insensibilidad al orden de los registros de entrada.

Figura 1

Resultado del proceso clustering



Datos sin agrupar

Datos después del proceso de agrupamiento

Algoritmo K-means

MacQueen (1967), señaló que el algoritmo *K-Means clustering* es un método no supervisado de agrupamiento particional o no jerárquico utilizado para segmentar un determinado conjunto de datos en k grupos, donde k representa el número de grupos especificados previamente por el analista. En la agrupación *K-Means*, cada *cluster* está representado por su centro o centroide que corresponde a la media de los puntos asignados al *cluster*.

La idea básica detrás de la agrupación *K-means* consiste en definir grupos, donde los objetos se representan con vectores reales de d dimensiones $d(x_1, x_2, \dots, x_n)$ y el algoritmo construye k grupos donde se minimiza la distancias de los objetos, dentro de cada *cluster* $C = \{C_1, C_2, \dots, C_k\}$ a su centroide μ_i. El algoritmo estándar es el algoritmo de Hartigan y Wong (1979) donde se define la variación intragrupo de un *cluster* como la suma de distancias al cuadrado entre los elementos y el correspondiente centroide:

$$W(C_k) = \sum_{x_j \in C_i} (x_j - \mu_i)^2 \qquad (1)$$

Cada observación x_j se asigna recursivamente a un grupo determinado, de tal manera, que la suma de los cuadrados de la distancia de observación a su centroide minimice la variación intragrupo (conocido como variación total dentro del grupo). Finalmente, se define la variación intragrupo de la segmentación como:

$$Var.intra = \sum_{i=1}^{k} \sum_{x_j \in C_i} (x_j - \mu_i)^2 \qquad (2)$$

La suma total de las variaciones intragrupos mide la concentración de la agrupación y se desea que sea la mínima posible.

Algoritmo K-modes

Según Huang (1998), el algoritmo de agrupamiento *K-Modes* es una extensión del algoritmo *K-Means*. Sin embargo, el proceso de agrupamiento de *K-Means* estándar no se puede aplicar a datos categóricos debido a la función de distancias entre observaciones y el uso de centroides para representar los centros de conglomerados. Para usar *K-Means* en la agrupación de datos categóricos, se tendría que convertir cada categoría única a un atributo binario ficticio y usar 0 o 1 para indicar el valor categórico ausente o presente en un registro de datos. Este enfoque no es adecuado para datos categóricos de alta dimensión.

El enfoque *K-modes* modifica el proceso *K-means* estándar para agrupar datos categóricos reemplazando la función de distancia con una medida de disimilitud para datos categóricos, utilizando modas para representar los centros de los *clusters* y las modas se

actualizan con los valores categóricos más frecuentes en cada iteración del proceso de agrupamiento. Estas modificaciones garantizan que el proceso de agrupación converja a un resultado mínimo local y se mantenga la eficiencia del proceso de agrupamiento.

La disimilitud para datos categóricos se define como el total de falta de correspondencia (desajuste) entre las categorías de atributos correspondientes entre dos observaciones. Cuanto menor sea el número de desajustes, más similares serán los dos objetos. Esta medida a menudo se denomina coincidencia simple, y se presenta en la siguiente ecuación:

$$d(X,Y) = \sum_{j=1}^{m} \delta(x_j, y_j) \quad (3)$$

Donde: $\delta(x_j, y_j) = \begin{cases} 0 & (x_j = y_j) \\ 1 & (x_j \neq y_j) \end{cases}$

Siendo x_j e y_j dos observaciones del conjunto de datos descritos por m atributos.

Para agrupar un conjunto de datos categóricos en k grupos, el proceso de agrupación de este algoritmo consiste en los siguientes pasos:

El primer paso es seleccionar aleatoriamente k observaciones como los centros de los *clusters* iniciales. Luego, se calculan las distancias entre cada objeto y los centroides; se asigna el objeto al grupo cuyo centro tiene la distancia más corta al objeto, a este paso se denomina "etapa de asignación de *cluster*"; se repite este procedimiento hasta que todos los objetos estén asignados a los *clusters*. La asignación de *cluster* y la "actualización de los centros" se repiten iterativamente hasta que las asignaciones de las observaciones dejen de cambiar (hasta lograr la convergencia).

Como es esencialmente similar al algoritmo *K-means*, el algoritmo de agrupamiento de *K-modes*, posee las mismas propiedades, es decir, es eficiente para agrupar grandes datos categóricos y también produce resultados de clustering localmente óptimo, que dependen de los centros iniciales.

Algoritmo K-prototypes

Según Huang (1998), los algoritmos *K-means* y *K-modes* se integran en el algoritmo *K-prototype* que se utiliza para agrupar elementos de tipo mixto. El algoritmo *K-prototype* es muy útil porque los elementos encontrados con frecuencia en las bases de datos del mundo real son de tipo mixto. La disimilitud entre dos elementos de tipo mixto X e Y se puede medir con la siguiente ecuación:

$$d(X,Y) = \sum_{j=1}^{m_r} (x_j{}^r - y_j{}^r)^2 + \gamma \sum_{j=m_r+1}^{m_c} \delta(x_i{}^c, y_j{}^c) \quad (4)$$

Donde:

- El primer término es la medida de distancia euclidiana al cuadrado en los atributos numéricos.
- El segundo término es la medida de disimilitud de coincidencia simple en los atributos categóricos.
- $\delta(x_j, y_j) = 0$ para $x_j = y_j$ y $\delta(x_j, y_j) = 1$ para $x_j \neq y_j$, $x_j{}^r$ y $y_j{}^r$ son valores de atributos numéricos, mientras que $x_j{}^c$ y $y_j{}^c$ son valores de atributos categóricos.
- m_r y m_c son las cantidades de atributos numéricos y categóricos respectivamente.
- γ es un peso para atributos categóricos, introducido para evitar favorecer cualquier tipo de atributo.

La elección de γ depende de la distribución de atributos numéricos, en términos generales, γ se relaciona con σ, la desviación estándar promedio de los atributos numéricos. En la práctica, σ puede usarse como una guía para determinar γ. Según, Huang (1998) un γ adecuado se encuentra entre 1/3 σ y 2/3 σ para los conjuntos de datos, el cálculo estimado de γ es de la siguiente manera:

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

donde la heurística para variables categóricas se calcula usando:

$$1 - \sum_i p_i^2 \text{ o } 1 - max_i\, p_i$$

siendo p_i la proporción de la categoría i en la variable cualitativa.

El algoritmo *K-prototype* se describe en los siguientes pasos:

- Paso 1: Seleccionar *k prototypes* iniciales de un conjunto de datos, uno para cada grupo.
- Paso 2: Asignar cada elemento a un *cluster* cuyo *prototype* sea el más cercano a él de acuerdo con la

ecuación (4). Luego, actualizar el *prototype* en cada *cluster* después de cada asignación.

- Paso 3: Después de que todos los elementos hayan sido asignados a un *cluster*, volver a probar la disimilitud de los objetos con los *prototypes* actuales. Si se encuentra un elemento tal que su *prototype* más cercano pertenece a otro *cluster* en lugar de su actual, reasignar el elemento al otro *cluster* y actualizar los prototypes de ambos grupos.
- Paso 4: Repetir el paso 3 hasta que ningún elemento haya cambiado de *cluster*.

Validación del clustering

Según Wang y Zhang (2007) el análisis *clustering* tiene como objetivo identificar grupos de objetos similares, por lo tanto, ayuda a descubrir la distribución de patrones y correlaciones interesantes en grandes conjuntos de datos. Sin embargo, el análisis *clustering* es un método no supervisado y en la mayoría de los casos, el usuario no tendrá ningún conocimiento previo sobre el número de grupos en el que se está separando el conjunto de datos, ni el algoritmo más adecuado para estos. Para llegar a ello, es necesario dividir conglomerados dispersos en dos o más *clusters* compactos, por lo contrario, si son pequeños, pueden combinarse más de un clúster separado.

La solución para encontrar el mejor algoritmo *clustering* y el número óptimo de conglomerados *k* se llama generalmente validez del *cluster*. Una vez que la partición se obtiene mediante un método de agrupación, la función de validez **cuantifica la precisión** de la estructura del conjunto de datos; para datos bidimensionales, los usuarios pueden verificar visualmente la validez de los resultados. Sin embargo, en el caso de grandes conjuntos de datos multidimensionales, la visualización es imposible. Por lo tanto, el objetivo de la validez del clúster es encontrar *k* conglomerados óptimos con el mejor algoritmo *clustering* buscando alcanzar una mejor precisión en la descripción de la estructura de datos multidimensionales.

La determinación del valor del parámetro *k* (el número de *clusters*) es importante cuando se aplica algoritmos *clustering*, ya que la elección inadecuada de este genera particiones que no reflejan el agrupamiento deseado de los datos.

Por ejemplo, la figura 2 (a) presenta un conjunto de datos donde se observa que este tiene tres grupos desde el ángulo visual. Sin embargo, si se considera un algoritmo de agrupamiento con el valor del parámetro *k* =4 para dividir el conjunto de datos, el resultado del proceso de agrupación sería el esquema de agrupación presentado en figura 2 (b). En el ejemplo, el algoritmo de agrupación encontró los cuatro mejores *clusters* en los que nuestro conjunto de datos podría ser particionado. Sin embargo, este no es el particionamiento óptimo para el conjunto de datos considerado.

Figura 2

Comparación del número de clusters

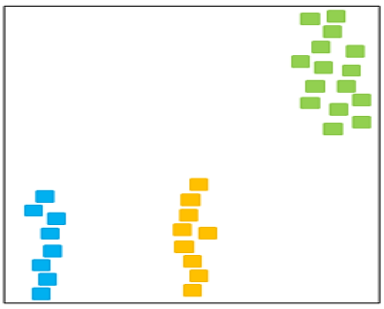


Fig. 2(a). Un conjunto de datos que consta de tres grupos

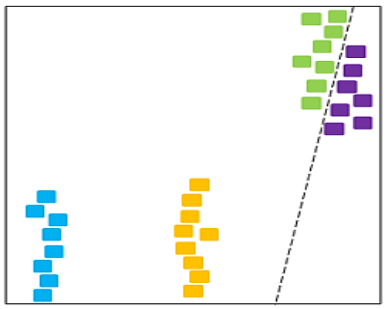


Fig. 2(b). Los resultados de la aplicación de algoritmos *clustering* cuando se piden cuatro grupos.

La partición obtenida por el algoritmo de agrupamiento en la figura 2(b) representa incorrectamente la estructura del conjunto de datos, es decir, no encaja bien en el conjunto de datos. El agrupamiento óptimo para el conjunto de datos será un esquema con tres *clusters*; como consecuencia, si se asigna un valor incorrecto a los parámetros del algoritmo de agrupación, el método obtendrá como resultado un esquema de partición que no es óptimo para el conjunto de datos específico, lo que conduciría a tomar decisiones incorrectas

Medidas de validación interna

Para verificar que los algoritmos de *clustering* agrupen objetos similares en un mismo *cluster* y objetos disímiles en diferentes *clusters*, se utilizan indicadores de validación interna, medidas que reflejan la cohesión y separación de las particiones de *cluster*.

La **cohesión** cuantifica el grado de proximidad del miembro de cada *cluster* a los otros miembros del mismo *cluster*; con ello se evalúa la homogeneidad dentro del *cluster* o varianza intragrupo; mientras que la **separación** entre *clusters* cuantifica el grado de heterogeneidad entre conglomerados o varianza intergrupal (usualmente midiendo la distancia entre los centroides). Ambas métricas permiten validar los resultados que se obtienen de una segmentación.

Índice de validación de Davies Bouldin

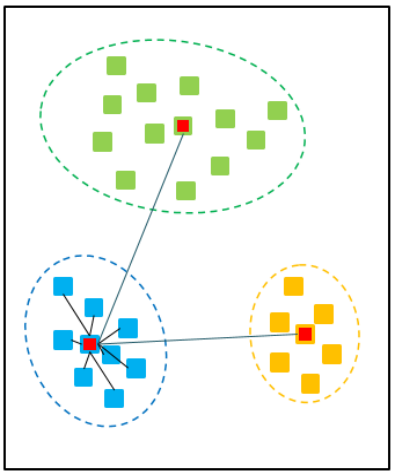
Según Chun (2012) el índice de Davies Bouldin es una métrica para evaluar el buen funcionamiento de los algoritmos de *clustering*. La fórmula de este índice se muestra en la siguiente ecuación:

$$DB = \frac{1}{k}\sum_{i=1}^{k} máx_{i\neq j}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right) \quad (5)$$

donde k es el número de *clusters*, c_i denota el centro del clúster, σ_i es la distancia media de todos los elementos del clúster i al centro c_i, y $d(c_i, c_j)$ es la distancia entre los centroides c_i y c_j, la idea de estas distancias se ve en la figura 3. El objetivo de los algoritmos de *clustering* es producir agrupamientos con baja distancia dentro del mismo clúster, y altas distancias entre los *clusters.* Por lo tanto, el máximo valor de este índice $máx_{i\neq j}\left(\frac{\sigma_i+\sigma_j}{d(c_i,c_j)}\right)$ representa el peor caso para el *cluster* i. La solución óptima es aquella que tiene el índice de Davies Bouldin más bajo.

Figura 3

Distancias utilizadas para el cálculo del índice de Davies Bouldin



2. Metodología

La técnica fue aplicada a los alumnos ingresantes de la Universidad Nacional Agraria La Molina (UNALM) en Lima, Perú de los semestres 2015-I y 2015-II. La información necesaria fue obtenida 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 (OBUAE).

La población investigada estuvo conformada por todos los alumnos ingresados de la UNALM en las modalidades: concurso ordinario y dos primeros puestos de colegios de educación secundaria en los semestres 2015-I y 2015-II con un total de 690 alumnos. El procesamiento de los datos se realizó utilizando el paquete clustMixType del software R.

Identificación de variables

Las variables identificadas en la aplicación fueron:

Variable	Denominación	Descripción
Variables socio-demográficas	Años_colegio_admisión	Tiempo transcurrido desde que terminó el 5to año de secundaria e ingresó a la universidad
	Edad_admisión	Edad del ingresante al momento del examen de admisión
	Dept_Colegio	Ubicación del colegio donde cursó el 5to año de secundaria (Lima o provincia)
	Sexo	Sexo del ingresado
Variables socioeducativas	Tipo_colegio	Tipo de institución de procedencia (Privada o Pública)
Variables socioeconómicas	Aporte_Semestral	Aporte semestral asignado al ingresante (pago de matrícula que realiza el alumno según su nivel socio-económico)
Variables de rendimiento en las áreas del conocimiento en la secundaria	CTA_Colegio	Nota obtenida en el 5to año de secundaria en el área de Ciencia, Tecnología y Ambiente
	COM_Colegio	Nota obtenida en el 5to año de secundaria en el área de Comunicación
	MAT_Colegio	Nota obtenida en el 5to año de secundaria en el área de Matemática
	Nota_Colegio	Nota promedio del último año de estudios
Variables de rendimiento en el examen de Admisión	RM_Admisión	Nota obtenida en el área de Razonamiento Matemático en el examen de admisión
	RV_Admisión	Nota obtenida en el área de Razonamiento Verbal en el examen de admisión
	MAT_Admisión	Nota obtenida en el área de Matemática en el examen de admisión
	FIS_Admisión	Nota obtenida en el área de Física en el examen de admisión
	QUI_Admisión	Nota obtenida en el área de Química en el examen de admisión
	BIO_Admisión	Nota obtenida en el área de Biología en el examen de admisión
	Nota_Admisión	Nota general obtenida en el examen de admisión
	Tercio_Superior_ESP	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	Modalidad de ingreso a la universidad
	Especialidad	Carrera a la que ingresó (12 carreras)
	Elección_ESP_Ingreso	Orden de elección que tuvo la carrera a la cual ingresó (1°, 2° o 3° opción)

El tipo de investigación fue de carácter descriptivo; se identificó y determinó el perfil del ingresado de la UNALM, a través de, la descripción de sus variables sociodemográficas, socioeducativas, socioeconómicas, rendimiento en las áreas del conocimiento en la secundaria, rendimiento en el examen de admisión y de elección en el ingreso a una carrera.

El diseño de investigación fue de carácter no experimental-transversal, ya que se contó con datos de alumnos, que fueron recolectados a partir de los archivos personales tomados de la ficha socioeconómica administradas por el personal de asistencia social de la OBUAE, también se recogió información de los registros del rendimiento en las áreas del conocimiento en la secundaria y las notas de los exámenes de admisión, los cuales son depositados oficialmente en el Centro de Admisión y Promoción.

Instrumento de colecta de datos

Examen de admisión

Uno de los instrumentos empleados en la obtención de los datos requeridos para la investigación fue el examen de admisión elaborado por el Comité Permanente de Admisión. Este instrumento tuvo un tiempo de aplicación de aproximadamente tres horas, 100 preguntas con cinco alterativas, donde solo hay una respuesta correcta. Las preguntas están distribuidas en nueve cursos o áreas de la siguiente forma: Razonamiento Matemático (l4), Razonamiento Verbal (20), Aritmética (8), Álgebra (6), Geometría (6), Trigonometría (4), Física (l4), Química (14) y Biología (l4). Cada pregunta bien contestada tuvo un valor de 1.00 punto, sin contestar 0.00 puntos y mal contestada -0.25 puntos.

Ficha del ingresante

La ficha del ingresante es un documento elaborado por la OBUAE como parte de la información que es manejada por el Departamento de Asuntos Estudiantiles. Este documento fue completado por los ingresados a la UNALM juntamente con los padres si son menores de edad o dependientes económicamente. Esta base de datos se encontró en formato físico y contenía la información básica que identifica al ingresado por apellidos y nombres, sexo, fecha de nacimiento, edad de ingreso, colegio o institución educativa de procedencia, nivel educativo de los padres, condición laboral de los padres, la situación laboral de los padres y del ingresado, la constitución de la familia, el ingreso total familiar.

Certificado de estudio escolar

El certificado de estudio escolar de los ingresantes es un documento que certifica y registra las notas de los cursos en los últimos años de educación secundaria. El documento es almacenado por la OBUAE como parte de la información que es administrada por el Departamento de Asuntos Estudiantiles en cada folder personal de los ingresados.

3. Resultados

Para aplicar el algoritmo *K- prototype* es necesario conocer a priori el número de *clusters* (*k*) a formarse. En este caso, se utilizó el índice de validación interna de Davies-Bouldin, calculado de manera iterativa cambiando el número de *cluster* y el valor aleatorio inicial; el valor de *k* seleccionado fue aquel que permitió obtener el índice de validación interna óptimo.

Se observa en la tabla N° 1 y la figura N° 4 que al aplicar el algoritmo *clustering K- prototype* el valor del índice de validación interna de Davies Bouldin óptimo es de 1.86 por lo que el número de *clusters* óptimo es K=5 con valor aleatorio inicial=10, visualizando la variabilidad en sus resultados por la naturaleza del algoritmo.

Analizando los resultados obtenidos, se elaboró una tabla de resumen general para caracterizar cada *cluster* por variable de los ingresados del año 2015 en la UNALM, esto se puede visualizar en la tabla 2. Por otro lado, en la figura 5 se muestra la tabla de distribución de observaciones por *cluster*.

Tabla 1

Determinación del número de clusters con el índice de Davies-Bouldin

N° aleatorio inicial	N° Cluster											
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
0	1.96	2.14	2.15	2.04	2.17	2.19	2.59	2.67	2.35	2.67	2.46	2.25
10	2.26	2.14	2.11	1.86	1.91	2.35	2.28	2.40	2.41	2.27	2.20	2.46
20	1.98	2.43	2.34	2.52	2.21	2.76	2.48	2.50	2.87	2.27	2.29	2.56
30	2.26	2.21	2.46	2.20	2.11	2.15	2.23	2.25	2.58	2.49	2.39	2.37
40	1.98	2.44	2.17	2.03	2.45	2.48	2.47	2.53	2.40	2.50	2.31	2.34
50	2.25	3.09	2.71	2.50	2.31	2.51	2.38	2.38	2.52	2.26	2.46	2.26
60	2.26	2.26	2.41	2.40	2.60	2.49	2.29	2.41	2.41	2.30	2.32	2.50
70	1.96	1.91	2.53	2.34	2.28	2.47	2.37	2.16	2.17	2.24	2.12	4.07
80	1.96	2.20	2.32	2.51	2.82	2.07	2.05	2.21	2.26	2.22	2.36	5.00
90	1.98	2.27	2.70	2.18	2.49	2.33	2.11	2.41	2.47	2.40	2.38	4.80
100	1.96	2.64	2.78	2.12	2.25	2.63	2.81	2.26	2.75	2.69	2.30	4.58

Tabla 2

*Caracterización de los **clusters***

Denominación	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
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 y provincia	Lima	Lima
Sexo	Femenino	Masculino	Femenino y Masculino	Masculino	Femenino
Tipo_Colegio	Pública	Privada	Privada	Pública	Privada
Tercio_Superior_Esp	No	No	Si	No	No
Modalidad	Concurso Ordinario	Concurso Ordinario	Concurso Ordinario	Concurso Ordinario	Concurso Ordinario y Dos Primeros Puestos de Colegios de Educación
Elección_Esp_Ingreso	Segunda o Tercera	Primera o Segunda	Primera	Segunda o Tercera	Segunda
Especialidad	Agronomía, Pesquería, Estadística Informática	Biología, Economía, Gestión Empresarial, Industrias Alimentarias, Ingeniería Agrícola, Meteorología	Biología, Ciencias Forestales, Ingeniería Ambiental	Pesquería, Zootecnia, Estadística Informática, Economía, Gestión Empresarial	Agronomía, Zootecnia, Pesquería

Figura 4

Determinación del número de clusters con el índice de Davies Bouldin

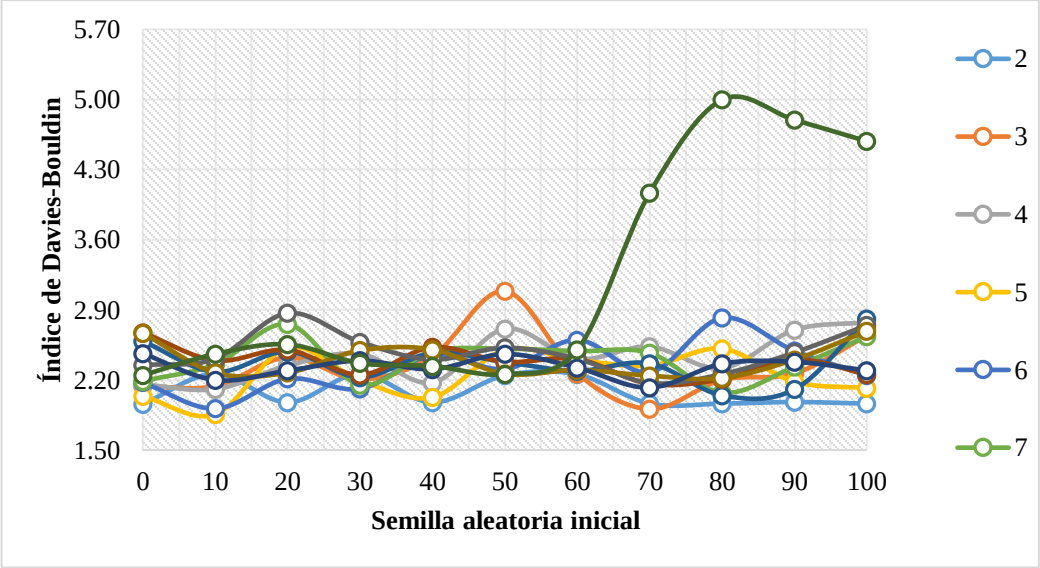
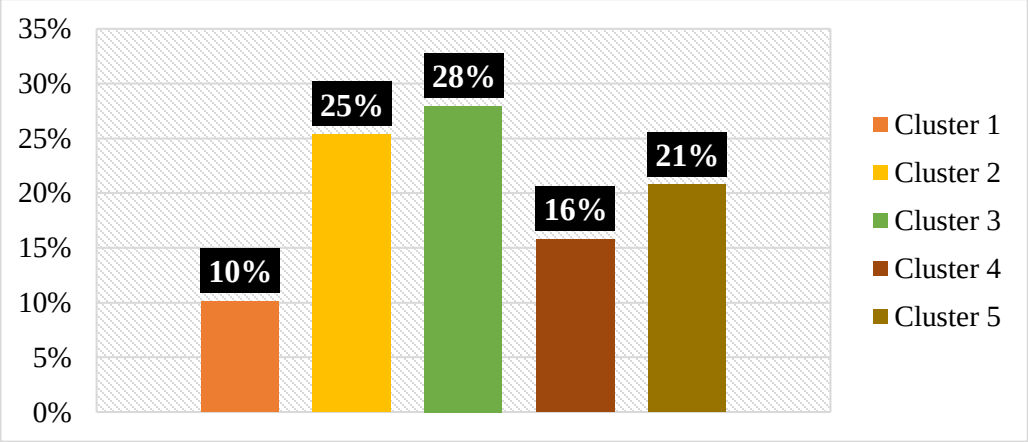


Figura 5

Tabla de distribución de observaciones por cluster



4. Discusión

Se observó que el *cluster* con mayor porcentaje de alumnos ingresados fue el 3 con 28%, los cuales tenían un alto rendimiento académico en el colegio y obtuvieron puntajes altos en el examen de admisión. Dado los resultados obtenidos, los ingresantes se clasificaron en:

- Ingresante destacado: son aquellos alumnos que se agruparon en el *cluster 3*, los cuales ingresaron poco tiempo después de terminar su educación secundaria, estos alumnos se caracterizan por evidenciar conocimientos destacados, ya que mostraron tener un alto rendimiento en el examen de admisión y venían con un desempeño académico

alto del colegio. Dada su situación socioeconómica, le asignaron un aporte semestral mayor al promedio. Los alumnos que tienen este perfil, en su mayoría ocuparon el tercio superior en su carrera e ingresaron en su mayoría a la carrera que eligieron como primera opción por la modalidad Concurso Ordinario; por lo general son alumnos que terminaron sus estudios en un colegio privado.

- Ingresante con logro esperado: son aquellos alumnos que se agruparon en el *cluster 2*, los cuales ingresaron poco tiempo después de terminar su educación secundaria, estos alumnos se caracterizan por evidenciar conocimientos esperados, ya que mostraron tener un rendimiento medio en el examen de admisión, en algunas áreas con alto rendimiento y venían con un desempeño académico promedio del colegio. Dada su situación socioeconómica, le asignaron un aporte semestral mayor al promedio. Los alumnos que tienen este perfil, en su mayoría no ocuparon el tercio superior en su carrera e ingresaron mayormente a la carrera que eligieron como primera o segunda opción por la modalidad Concurso Ordinario; y por lo general son varones que terminaron sus estudios en un colegio privado.
- Ingresante regular: son aquellos alumnos que se agruparon en el *cluster 4*, los cuales ingresaron después de un periodo largo de tiempo de terminar su educación secundaria, estos alumnos se caracterizan por evidenciar conocimientos regulares, ya que mostraron tener un rendimiento regular en el examen de admisión, en algunas áreas con bajo rendimiento y con un desempeño académico bajo del colegio. Dada su situación socioeconómica le asignaron un aporte semestral menor al promedio. Los alumnos que tienen este perfil, en su mayoría no ocuparon el tercio superior en su carrera e ingresaron mayormente a la carrera que eligieron como segunda o tercera opción por la modalidad Concurso Ordinario, por lo general son varones que terminaron sus estudios en un colegio nacional.
- Ingresante en proceso: son aquellos alumnos que se agruparon en el *cluster 1*, los cuales ingresaron después de un tiempo regular de tiempo de terminar su educación secundaria, estos alumnos se caracterizan por evidenciar conocimientos en proceso, ya que mostraron tener un rendimiento regular en el examen de admisión, en algunas áreas con bajo rendimiento y venían con un desempeño académico alto en el colegio. Dada su situación socioeconómica le asignaron un aporte semestral menor al promedio, los alumnos que tienen este perfil mayormente no ocuparon el tercio superior en su carrera e ingresaron en su mayoría a la carrera que eligieron como segunda o tercera opción por la modalidad Concurso Ordinario, por lo general son mujeres que terminaron sus estudios en un colegio nacional.
- Ingresante en inicio: son aquellos alumnos que se agruparon en el *cluster 5*, los cuales ingresaron poco tiempo después de terminar su educación secundaria. Estos alumnos se caracterizan por evidenciar conocimientos en inicio, ya que mostraron tener un rendimiento bajo en el examen de admisión, sin embargo, venían con un desempeño académico alto en el colegio. Dada su situación socioeconómica, le asignaron un aporte semestral igual al promedio. Los alumnos que tienen este perfil mayormente no ocuparon el tercio superior en su carrera e ingresaron en su mayoría a la carrera que eligieron como segunda opción por la modalidad Concurso Ordinario y Dos Primeros Puestos de Colegios de Educación Secundaria, por lo general son mujeres que terminaron sus estudios en un colegio privado.

5. Conclusiones

Al aplicar el algoritmo de segmentación *K-prototype* es posible tener una visión completa y detallada de los tipos de alumnos que ingresan a una universidad con base a sus variables socioeconómicas, demográficas y de rendimiento educativo, permitiendo descubrir información útil que ayuda a los docentes y responsables de la institución educativa a determinar la manera más adecuada para guiar a sus alumnos, maximizando su aprendizaje de manera más personalizada y contribuyendo a la mejora de la calidad de la educación superior. A diferencia de otros algoritmos *cluster* como el jerárquico o el de partición, el *K-prototype* permite trabajar con datos mixtos.

Con el algoritmo *K-prototype*, es posible caracterizar el perfil de los ingresados de una universidad respecto a sus variables socioeconómicas, demográficas y de rendimiento educativo. Se pudo identificar 5 tipos de ingresados cada uno con características diferentes, los cuales se denominaron:

ingresante destacado, ingresante con logro esperado, ingresante regular, Ingresado 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 perfiles. Por otro lado, el ingresante destacado puede ser considerado el perfil ideal o deseado de los alumnos ingresantes a la universidad.

Estas características permiten identificar y entender la diversidad existente en el perfil del ingresante, las cuales deben ser atendidas por las autoridades de la institución, a través de diversas estrategias educativas: cursos de nivelación, apoyo económico y orientación con el fin de evitar en el futuro el bajo rendimiento académico, deserción estudiantil, dilatación del tiempo de estudio o retraso, entre otros.

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.

Conflictos de intereses

Los autores firmantes del presente trabajo de investigación declaran no tener ningún potencial conflicto de interés personal o económico con otras personas u organizaciones que puedan influir indebidamente con el presente manuscrito.

Contribuciones de los autores

Preparación y ejecución; Desarrollo de la metodología; Concepción y diseño; Edición del artículo; Supervisión del estudio: L-CV, J-SF.

Referencias bibliográficas

Arias, J. (2015). El perfil de ingreso en el rendimiento académico inicial de los estudiantes de la carrera de Agronomía de la Universidad Nacional Agraria La Molina, años 2011 a 2012. [Tesis de doctorado]. Universidad Nacional de Educación Enrique Guzmán y Valle.

Chun, L. (2012). Diseño e Implementación de algoritmos aproximados de clustering balanceado en PSO. [Tesis de maestría]. Universidad de Chile.

Hartigan, J y Wong, M. (1979). Algorithm AS 136: A K-Means clustering algorithm. Journal of the Royal Statistical Society, 28(1), 100-108. https://bit.ly/30jLpV1 .

Huang, Z. (1998). Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery, 2, 283 - 304. https://bit.ly/2FMUgoH .

Kaur, A; Kaur, N. (2013). Survey Paper on Clustering Techniques. International Journal of Science, Engineering and Technology Research (IJSETR), 2(4), 803-806. https://bit.ly/3a3Z72B .

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Berkeley Symposium on Mathematical Statistics and Probability, 281-297. https://bit.ly/384ZY1g .

Rai, P y Singh, S. (2010). A Survey of Clustering Techniques. International Journal of Computer Applications, 7(12), 1-5. https://bit.ly/30m6AWp .

Wang, W y Zhang, Y. (2007). On fuzzy cluster validity indices. Fuzzy Sets and Systems. Fuzzy Sets and Systems, 158(19), 2095-2117. https://n9.cl/rl9xy