Comparative study of unsupervised techniques of data mining for segmentation of students

Leticia Laura Ochoa, Mg[1], Karina Rosas Paredes, Mg[1], José Esquicha Tejada, Mg[1] [1]Universidad Católica de Santa María, Perú, llaura@ucsm.edu.pe, kparedes@ucsm.edu.pe, jesquicha@ucsm.edu.pe

Abstract- ***There are several clustering algorithms that yield different grouping results; thus, it is necessary to choose an algorithm that offers the best results for the segmentation of students. Herein, a comparative study of unsupervised data mining techniques is conducted for the segmentation of students according to their academic performance using algorithms such as If-means and PAM within partition clustering and methods such as Ward, single, complete, average, Mcquitty, median and centroid of hierarchical clustering agglomerative. Then, a data mining algorithm is chosen based on the best grouping quality that is obtained using internal measures, such as intra-cluster and inter-cluster distances, and the silhouette coefficient, thereby obtaining improved results with the partition-clustering technique If-means for the segmentation of students in three groups that can be used to reinforce student learning at the basic, intermediate, and advanced levels.***

Keywords– EDUCATIONAL DATA MINING, STUDENT SEGMENTATION, CLUSTERING, UNSUPERVISED TECHNIQUES.

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2017.1.1.115
ISBN: 978-0-9993443-0-9
ISSN: 2414-6390

Estudio Comparativo de Técnicas no Supervisadas de Minería de Datos para Segmentación de Alumnos

Leticia Laura Ochoa, Mg[1], Karina Rosas Paredes, Mg[1], José Esquicha Tejada, Mg[1]
[1]Universidad Católica de Santa María, Perú, llaura@ucsm.edu.pe, kparedes@ucsm.edu.pe, jesquicha@ucsm.edu.pe

***Resumen*– Existen varios algoritmos de clustering que generan diferentes resultados de agrupamiento, por lo que es necesario elegir el algoritmo que ofrezca mejores resultados para la segmentación de alumnos, en este trabajo se realiza un estudio comparativo de técnicas no supervisadas de minería de datos para la segmentación de alumnos según su rendimiento académico utilizando algoritmos de K-means y PAM dentro del clustering particional y métodos de ward, single, complete, average, mcquitty, median y centroid del clustering jerárquico aglomerativo, luego se elige el algoritmo de minería de datos con la que se obtiene mejor calidad de agrupamiento utilizando medidas internas como las distancias intra-cluster e inter-cluster, y el coeficiente de silueta, obteniendo mejores resultados con la técnica de clustering particional K-means para la segmentación de alumnos en tres grupos que puede ser utilizado para reforzar el aprendizaje de los alumnos en los niveles básico, intermedio y avanzado.**

***Palabras claves*– Minería de datos educacional, segmentación de alumnos, clustering, técnicas no supervisadas.**

I. Introducción

En el ámbito educativo es evidente la necesidad de disponer de sistemas de gestión que permitan tomar decisiones académicas y elaborar estrategias a partir del conocimiento oportuno, ya que esto no solo incide directamente sobre la funcionalidad de los departamentos académicos, u otras cuestiones internas, sino que también podrían incidir sobre actividades como las evaluaciones y acreditaciones de instituciones y carreras. Entre los problemas más complejos que enfrentan las instituciones de educación podemos mencionar: mejorar la calidad académica, disminuir la deserción y la reprobación, evitar el atraso estudiantil y los bajos índices de eficiencia relacionado con las tasas de graduación. Esto requiere gestionar estrategias y tomar medidas frente a estos acontecimientos; para ello es posible recurrir al proceso denominado Minería de Datos Educacional (MDE), es decir, la aplicación del proceso de Descubrimiento o Extracción de Conocimiento en Bases de Datos (KDD) en ámbito educativo [1].

Según [2] la Minería de Datos Educacional (MDE), es un campo de estudio dedicado a desarrollar métodos matemáticos para analizar datos provenientes de ambientes relacionados a la educación, y extraer la mayor cantidad de información para tratar de entender mejor a los estudiantes, profesores y actores relacionados, con el fin de mejorar los procesos educativos.

Digital Object Identifier (DOI): http://dx.doi.org/10.18687/LACCEI2017.1.1.115
ISBN: 978-0-9993443-0-9
ISSN: 2414-6390

La MDE permite responder preguntas sobre qué sabe realmente un estudiante y cómo está aprendiendo. De esta manera, la MDE permite descubrir información útil que ayuda a los docentes y responsables de las instituciones educativas en determinar la manera más pertinente para guiar a sus estudiantes, maximizando su aprendizaje [3]. Es el proceso de transformar los datos en bruto recolectados por los sistemas educativos en información útil que puede utilizarse para tomar decisiones y responder a preguntas de investigación [4][5]. Es la aplicación de la minería de datos en el ámbito de la educación, para comprender mejor el proceso de aprendizaje de los estudiantes y de su participación global en el proceso, con el objetivo de mejorar la calidad y rentabilidad del sistema educativo [1][4].

Las técnicas de minería de datos permiten llevar a cabo las tareas predictivas y descriptivas haciendo uso de algoritmos de minería de datos. Según el objetivo del análisis de los datos, los algoritmos utilizados se clasifican en supervisados y no supervisados [6]. Entre las técnicas más empleadas en el ámbito educativo están el agrupamiento automático (clustering), el análisis de asociaciones y patrones frecuentes en secuencias [4], que corresponden a las técnicas no supervisadas de minería de datos.

Clustering es una de las técnicas más útiles para descubrir conocimiento oculto en un conjunto de datos. En la actualidad el análisis de clustering en minería de datos ha jugado un rol muy importante en una amplia variedad de áreas tales como: reconocimiento de patrones, análisis de datos espaciales, procesamiento de imágenes, cómputo y multimedia, análisis médico, economía, bioinformática y biometría principalmente [7]. El análisis de clusters agrupa objetos basados solamente en la información encontrada en los datos que describen a los objetos y sus relaciones. El objetivo es que los objetos dentro de un grupo sean similares (o relacionados) entre sí y diferentes de (o no relacionados con) los objetos en otros grupos. A mayor similitud (u homogeneidad) dentro de un grupo y a mayor diferencia entre grupos, mejor o más distinto es el clustering [8]. En la literatura existen una gran cantidad de técnicas de clustering que varían de acuerdo a la arquitectura que utilizan [9].

En este trabajo se realiza un estudio comparativo de técnicas no supervisadas de minería de datos para segmentación de alumnos, que permita agrupar a los estudiantes según su rendimiento académico que pueda servir para tomar acciones de reforzamiento personalizada.

II. Materiales

Los materiales utilizados son:

1) *R:* Para utilizar el lenguaje de programación R en la aplicación de técnicas no supervisadas de minería de datos, así como la creación de gráficos. Se eligió R por ser el software más utilizado para minería de datos y ciencia de los datos según encuestas realizadas por [10][11].
2) *RStudio:* Entorno de desarrollo integrado (IDE) para R (lenguaje de programación). Incluye una consola, editor de sintaxis que apoya la ejecución de código, así como herramientas para el trazado, la depuración y la gestión del espacio de trabajo [12].
3) *Pentaho Data Integration:* Para desarrollar el proceso ETL (Extracción, Transformación y Carga), utilizado para la transformación y limpieza de los datos académicos.

II. Metodología

El desarrollo del presente trabajo consta de 5 pasos:

1) *Recolección de datos académicos:* Los datos académicos fueron proporcionados por la Escuela Profesional de Ingeniería de Sistemas de la Universidad Católica de Santa María, los cuales fueron extraídos del Sistema Académico de Ingreso de Notas.
2) *Selección de registros y atributos:* Se seleccionaron los registros correspondientes a los alumnos de primer año ya que según investigaciones es donde se produce mayor deserción [13][14], por lo que se hace necesario un reforzamiento personalizado para mejorar su rendimiento académico. Los atributos seleccionados que se utilizaron para la segmentación son: Código de curso, nombre de la asignatura, código del alumno, nombre del alumno y promedio.
3) *Transformación y limpieza de datos:* Se utilizó la herramienta Pentaho para el proceso de transformación y limpieza de los datos.
4) *Selección y aplicación de técnicas y algoritmos de minería de datos:* Se seleccionaron y aplicaron los algoritmos de K-means y PAM dentro del clustering particional y métodos de ward, single, complete, average, mcquitty, median y centroid del clustering jerárquico aglomerativo para el estudio comparativo.
5) *Evaluación de los algoritmos de minería de datos:* Para evaluar las técnicas no supervisadas de minería de datos aplicadas para la segmentación de alumnos se realizó teniendo en cuenta las medidas internas como las distancias intra-cluster e inter-cluster, y el coeficiente de silueta.

III. Resultados y Discusión

En este trabajo se realizó el estudio comparativo de técnicas no supervisadas de minería de datos para segmentación de alumnos del primer año de la Escuela profesional de Ingeniería de Sistemas de la Universidad Católica de Santa María.

Luego de la recolección, selección, transformación y limpieza de los datos académicos, se utilizaron las variables que se muestran en la Tabla 1 para la segmentación académica.

TABLA 1
VARIABLES USADAS PARA LA SEGMENTACIÓN

VARIABLE	DESCRIPCIÓN
CODIGO	Código del Alumno
CALC_DIFE	Cálculo Diferencial
DESA_HUMA	Desarrollo Humano
COMU_ORAL_ESCR	Comunicación Oral y Escrita
FUND_PROG	Fundamentos de Programación
PROG_I	Programación I
ESTR_DISC_I	Estructuras Discretas I

A. Selección y Aplicación de Técnicas y Algoritmos de Minería de Datos

Para poder realizar la segmentación de alumnos se requiere utilizar técnicas de aprendizaje no supervisado de minería de datos que permitan realizar el clustering (agrupamiento).

Las técnicas de clustering seleccionadas a utilizar son:

1) Clustering jerárquico aglomerativo: Un método jerárquico crea una descomposición jerárquica de un conjunto de datos, formando un dendrograma (árbol) que divide recursivamente el conjunto de datos en conjuntos cada vez más pequeños [9]. Este enfoque de clustering se refiere a una colección de técnicas de agrupamiento estrechamente relacionadas que producen un agrupamiento jerárquico comenzando con cada punto como un cluster singleton (con un solo elemento) e iterativamente lo agrupa con los dos clusters más cercanos hasta que un único cluster que abarca a los todos los demás permanece [15].

El algoritmo de clustering aglomerativo es la técnica de clustering jerárquico más popular. Los algoritmos jerárquicos tradicionales utilizan una matriz de similitud o distancia [8]. Existen diferentes métodos de clustering jerárquico aglomerativo como ward, single, complete, average, mcquitty, median y centroid, según el cálculo de distancias entre clusters que se utiliza.

Utilizando el método ward, se observa a partir del dendrograma generado que se pueden formar varias agrupaciones de dos, tres o cuatro clusters como se muestra en la Fig. 1, Fig. 2 y Fig. 3 respectivamente:

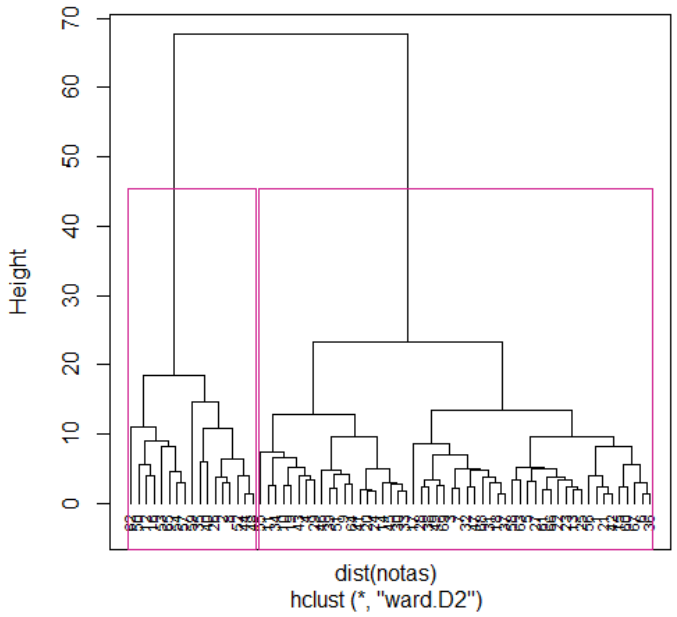

Fig. 1 Agrupación jerárquica de dos grupos.

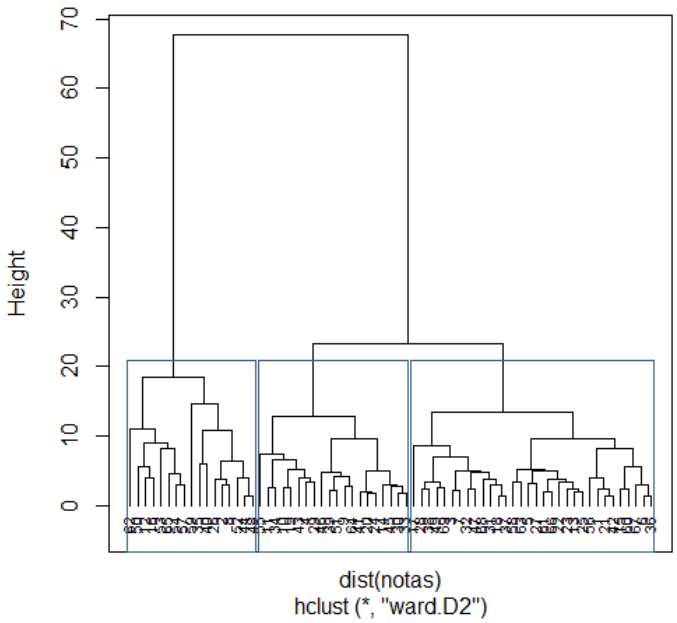

Fig. 2 Agrupación jerárquica de tres grupos.

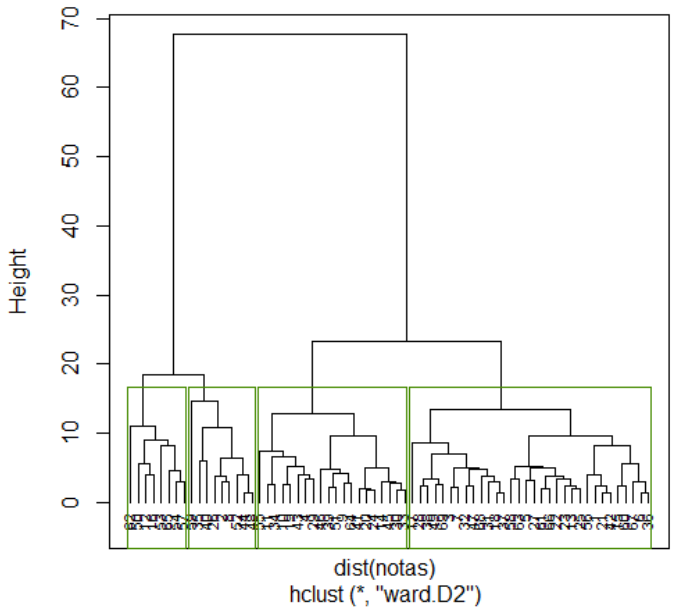

Fig. 3 Agrupación jerárquica de cuatro grupos.

El dendrograma es un gráfico usado en el procedimiento jerárquico que permite visualizar el proceso de agrupamiento de los cluster en los distintos pasos, formando un diagrama en árbol. Da una idea visual de la proximidad entre cluster y ayuda a decidir cuántos grupos formar [16]. Se eligió realizar la segmentación en tres grupos que pueda servir para el reforzamiento de alumnos en los niveles básico, intermedio y avanzado.

En la Fig. 4 se muestra un gráfico de barras construidos a partir de los centros de gravedad, los cuales son los promedios de cada cluster, en donde se puede apreciar que el Cluster 2 corresponde a los alumnos de bajo rendimiento, el Cluster 1 a los alumnos de rendimiento medio y el Cluster 3 a los alumnos de mejor rendimiento en los diferentes cursos.

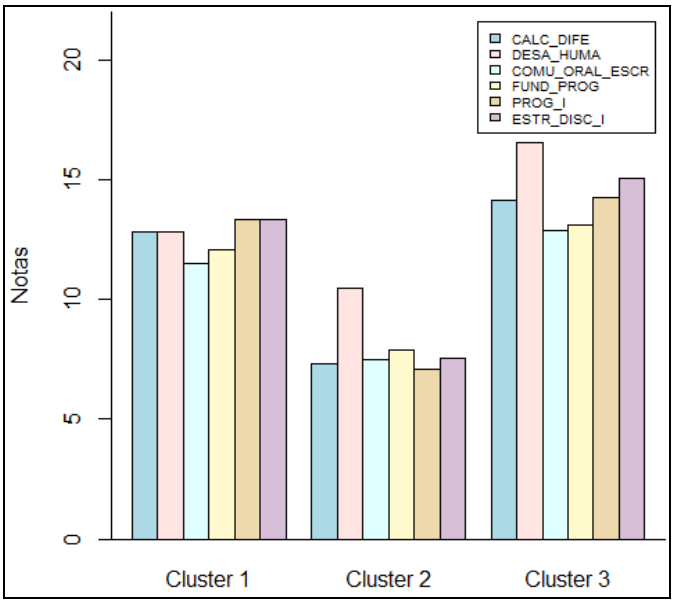

Fig. 4 Promedio de notas en los clusters - método ward.

Utilizando el método single (Agregación del salto mínimo), se puede observar a partir del dendrograma de la Fig. 5 que no es muy recomendable para la agrupación de tres clusters, ya que se forman dos grupos integrados solamente por un alumno.

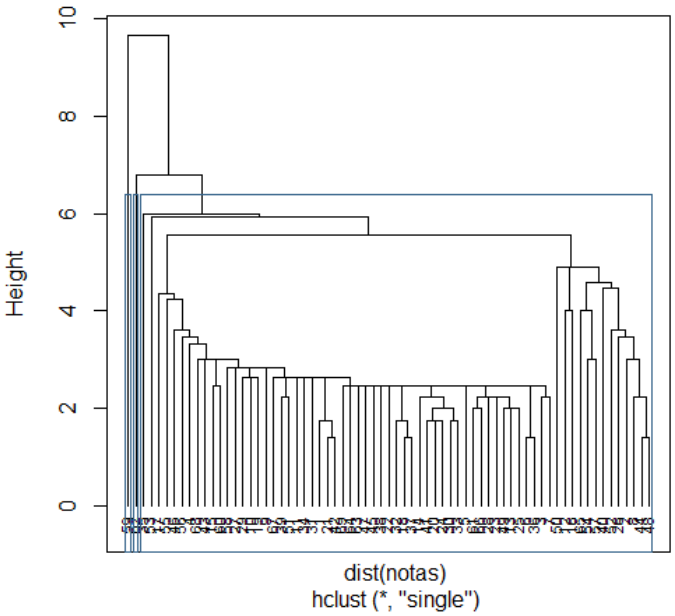

Fig. 5 Dendrograma utilizando el método "single".

Utilizando el método complete (Agregación del salto máximo), se generó el dendrograma que se muestra en la Fig. 6, el cual si se podría utilizar para la segmentación en tres grupos.

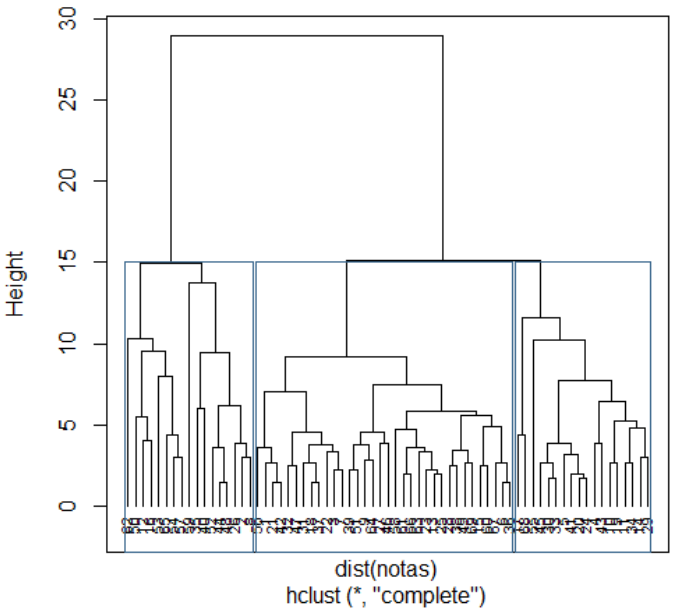

Fig. 6 Dendrograma utilizando el método "complete".

En la Fig. 7 se muestra un gráfico de barras construido a partir de los centros de gravedad de los clusters, de forma similar al método ward se puede apreciar que el Cluster 2 corresponde a los alumnos de bajo rendimiento, el Cluster 1 a los alumnos de rendimiento medio y el Cluster 3 a los alumnos de mejor rendimiento en los diferentes cursos.

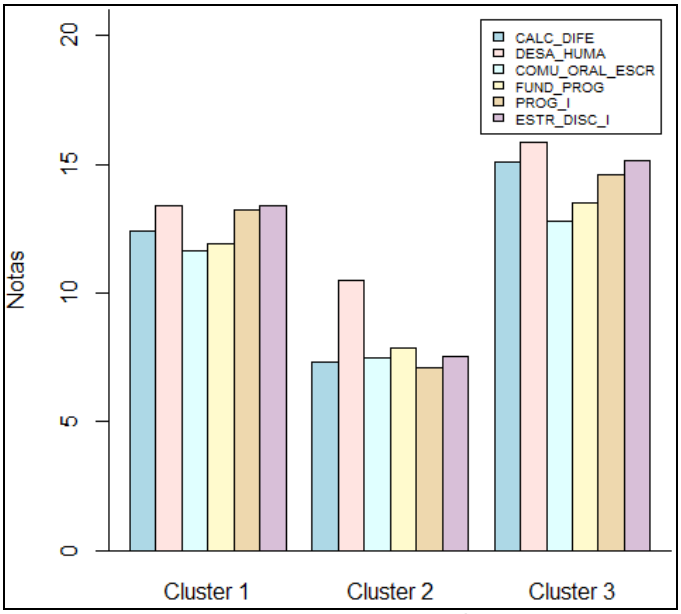

Fig. 7 Promedio de notas en los clusters - método complete.

Utilizando el método average, mcquitty, median y centroid del algortimo de clustering jerárquico aglomerativo, a partir de sus dendrogramas se formaban grupos de un solo elemento de forma similar al método single para la

segmentación de alumnos por lo que no se consideraron en esta sección.

2) K-means: Esta técnica está basada en el clustering particional que intenta encontrar un número de clusters (K) especificados por el usuario, los cuales son representados por sus centroides. El algoritmo básico se describe a continuación: Primero se eligen K centroides iniciales, donde K es un parámetro especificado por el usuario y corresponde al número de clusters deseados. Cada punto es asignado a su centroide más cercano y cada colección de puntos asignado a un centroide representa un cluster. El centroide de cada cluster se actualiza basado en la asignación de puntos al cluster. Se repiten los pasos de asignación y actualización hasta que los puntos dentro del cluster no cambien, o equivalentemente, hasta que los centroides dejen de cambiar [15].

La ventaja de K-means es ser un algoritmo simple, efectivo para pequeñas y medianas cantidades de datos. Utiliza el promedio para representar los centros de los clusters. K-means permite especificar de forma inicial la cantidad de clusters o grupos deseados.

Se realizaron pruebas de segmentación en tres grupos utilizando los diferentes algoritmos disponibles en K-means: Hartigan-Wong, Lloyd, Forgy y MacQueen, dando los mismos resultados para los datos probados, por lo que se utilizó el algoritmo por defecto "Hartigan-Wong", cuyos resultados se muestran en la Fig. 8, donde se puede apreciar la cantidad de alumnos por cada cluster que son 17, 32 y 20. También se puede observar los promedios de cada cluster en los diferentes cursos, se puede notar por ejemplo que el cluster 1 corresponde a los alumnos de bajo promedio.

```
> grupos_Kmeans<-kmeans(notas, 3,iter.max=50, nstart=50)
> grupos_Kmeans
K-means clustering with 3 clusters of sizes 17, 32, 20

Cluster means:
  CALC_DIFE DESA_HUMA COMU_ORAL_ESCR FUND_PROG    PROG_I ESTR_DISC_I
1  7.294118  10.47059       7.470588  7.882353  7.117647    7.529412
2 12.625000  12.96875      11.531250 11.968750 13.218750   13.250000
3 14.500000  16.35000      12.850000 13.300000 14.500000   15.200000

Clustering vector:
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
 2  1  2  3  2  2  2  1  3  3  3  1  2  3  2  1  3  2  3  3  2  2  2
24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46
 3  2  1  2  2  3  3  2  2  3  3  1  2  2  2  3  1  3  2  3  1  3  2
47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69
 2  1  2  1  3  1  1  1  3  2  1  2  1  2  2  1  2  3  1  2  2  2  2

Within cluster sum of squares by cluster:
[1] 571.7647 340.8750 276.5000
 (between_SS / total_SS =  68.5 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"
[5] "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
```

Fig. 8 Resultados de la agrupación utilizando K-means.

En la Fig. 9 se muestra un gráfico de barras construidos a partir de los promedios de los clusters generados por el algoritmo K-means para la segmentación en tres grupos, en donde se puede apreciar que el Cluster 1 corresponde a los alumnos de bajo promedio, el Cluster 2 a los alumnos de promedio regular y el Cluster 3 a los alumnos de alto promedio en los diferentes cursos.

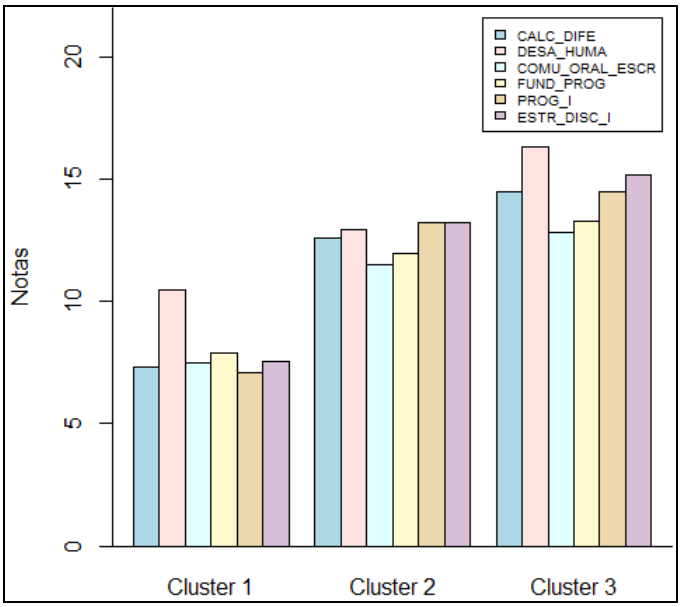

Fig. 9 Promedio de notas en los clusters - Kmeans.

3) PAM: PAM (Partitioning Around Medoids) es un algoritmo típico de agrupación K-medoides. Aborda el problema de una manera iterativa. Al igual que el algoritmo K-means, los objetos representativos iniciales (llamados semillas) son elegidos arbitrariamente. Consideramos si la sustitución de un objeto representativo por un objeto representativo mejoraría la calidad de la agrupación. Todos los posibles reemplazos son probados. El proceso iterativo de reemplazar objetos representativos por otros objetos continúa hasta que la calidad de la agrupación resultante no puede ser mejorada por ningún reemplazo. Esta calidad se mide mediante una función de coste de la disimilitud media entre un objeto y el objeto representativo de su agrupación [17].

K-medoides usa medianas en lugar de promedios para representar a los clusters. PAM funciona de manera eficiente para pequeños conjuntos de datos.

En la Fig. 10 se muestra los resultados obtenidos con el algoritmo PAM para la segmentación de tres grupos, en la parte superior se muestran los medoides, las cuales son las medianas de los clusters que representan los grupos, se puede notar que el segundo cluster corresponde a los alumnos de bajas notas en los diferentes cursos.

```
> grupos_PAM <- pam(notas,3)
> grupos_PAM
Medoids:
   ID CALC_DIFE DESA_HUMA COMU_ORAL_ESCR FUND_PROG PROG_I ESTR_DISC_I
25 25        12        14             11        11     13          14
8   8         9        11              9         8      7           7
45 45        15        15             13        14     14          14
Clustering vector:
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
 1  2  3  3  3  1  1  2  1  3  3  2  1  3  1  2  3  3  3  3  1  1  1
24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46
 3  1  2  1  1  3  3  3  1  3  3  2  1  3  1  3  2  3  1  3  2  3  1
47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69
 1  2  1  2  1  2  2  2  3  1  2  1  2  1  1  2  1  1  2  1  1  3  1
objective function:
   build     swap
4.061587 4.061587

Available components:
 [1] "medoids"    "id.med"     "clustering" "objective"  "isolation"
 [6] "clusinfo"   "silinfo"    "diss"       "call"       "data"
```

Fig. 10 Resultados de la agrupación utilizando PAM.

En la Fig. 11 se muestra un gráfico de barras construidos a partir de los medoides (medianas) que representan a los clusters generados por el algoritmo PAM para la segmentación en tres grupos. Se puede apreciar que el Cluster 2 corresponde a los alumnos de bajas notas, el Cluster 1 a los alumnos de notas regulares y el Cluster 3 a los alumnos de altas calificaciones en los diferentes cursos.

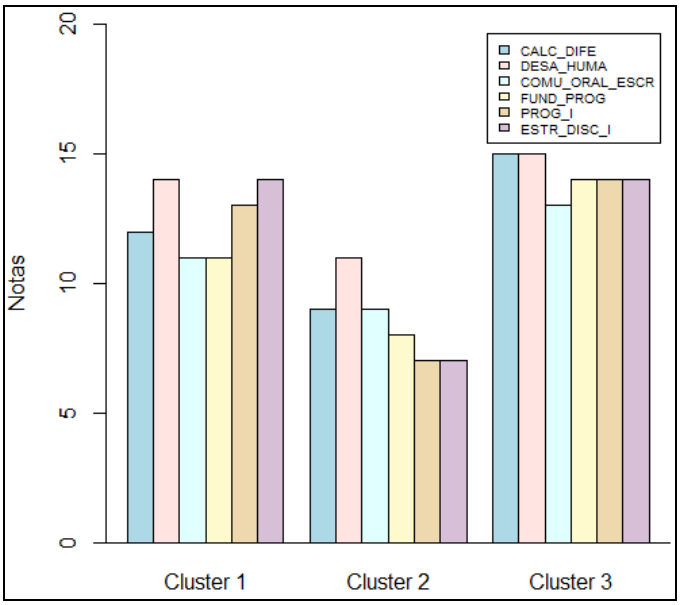

Fig. 11 Mediana de notas en los clusters – PAM.

El clustering jerárquico aglomerativo y los algoritmos K-means y PAM del clustering particional se pueden utilizar para la segmentación de alumnos ya que funcionan de manera eficiente para pequeños conjuntos de datos, también pueden ser utilizados para medianas cantidades de datos. Además permiten definir la cantidad de grupos que se desea para la segmentación.

B. Evaluación de los Algoritmos de Minería de Datos

Para la segmentación de alumnos se utilizaron algoritmos de clustering, por lo que se debe evaluar la calidad de los grupos formados.

Un buen método de agrupamiento producirá clusters de alta calidad [17] si presenta: (a) Alta similitud intra-cluster: Cohesiva dentro de los clusters, (b) Baja similitud inter-cluster: Distintivo entre clusters.

Algunas de las razones de evaluar los clusters según [8] son: (a) Establecer el número correcto de clusters, (b) Para comparar los algoritmos de clustering, (c) Comparar dos conjuntos de clusters para determinar cuál es el mejor.

Las medidas no supervisadas de validación de clusters se dividen a menudo en dos clases: las medidas de cohesión del cluster, que determinan la similitud entre los objetos de un cluster y las medidas de separación del cluster, que determina la diferencia o separación entre los clusters [8].

Según [18] el coeficiente silueta (silhouette) mide cuan buena es la asignación de un elemento o dato a su cluster. Para esto compara las distancias de este elemento respecto a todos los demás elementos del cluster al que pertenece, contra las distancias respecto a los clusters vecinos. El coeficiente silueta del elemento i se denota s(i).

- Si $s(i) \approx -1$, el dato i está mal agrupado
- Si $s(i) \approx 0$, el dato i está entre dos clusters
- Si $s(i) \approx 1$, el dato i está bien agrupado

Es por esta razón que la evaluación de las técnicas no supervisadas de minería de datos aplicadas para la segmentación de alumnos se realizó teniendo en cuenta: (a) Las distancias intra-cluster e inter-cluster, y (b) El coeficiente de silueta.

1) Distancias intra-cluster e inter-cluster: Se debe seleccionar una técnica que minimice la distancia intra-cluster (cohesión) y maximice la distancia inter-cluster (separación).

Según [8] la cohesión Within-cluster Sum of Squares (WSS) se mide por la suma de los cuadrados intra-cluster:

$$WSS = \sum_{i} \sum_{x \in C_i} (x - m_i)^2 \qquad (1)$$

Donde:

- i es el identificador del cluster.
- m_i corresponde al promedio del cluster
- x es el punto de datos que pertenece al grupo C_i

Y la separación Between-cluster Sum of Squares (BSS) se mide por la suma de los cuadrados inter-cluster:

$$BSS = \sum_{i} |C_i| (m - m_i)^2 \qquad (2)$$

Donde:

- $|C_i|$ es el tamaño del cluster i
- m_i corresponde al promedio del cluster
- m corresponde al promedio total

En la Tabla 2 se muestra los resultados obtenidos de las distancias intra-cluster por cada método del clustering jerárquico aglomerativo, luego a partir del Total WSS se generó el gráfico de la Fig. 12 para su interpretación.

TABLA 2
DISTANCIAS INTRA-CLUSTER – CLUSTERING JERÁRQUICO

Método	Cluster 1	Cluster 2	Cluster 3	Total WSS
Ward	377.88	571.76	261.80	1211.44
Single	3377.82	0	0	3377.82
Complete	374.26	571.76	275.94	1221.96
Average	911.58	388.40	46.50	1346.48
Mcquitty	911.58	461.31	0	1372.89
Median	3313.64	0	0	3313.64
Centroid	911.58	461.31	0	1372.89

En la Fig. 12 se puede observar que la menor distancia intra-cluster, que representa mayor cohesión de la agrupación, se obtiene con el método Ward (1211.44).

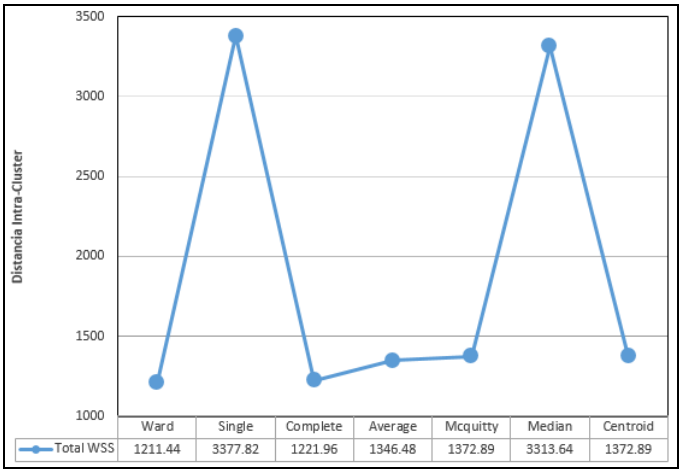

Fig. 12 Distancias intra-cluster – clustering jerárquico.

En la Tabla 3 se muestra los resultados obtenidos de las distancias inter-cluster por cada método del clustering jerárquico aglomerativo.

TABLA 3
DISTANCIAS INTER-CLUSTER – CLUSTERING JERÁRQUICO

Método	Cluster 1	Cluster 2	Cluster 3	Total BSS
Ward	4.66	101.33	34.30	2557.69
Single	0.12	172.84	210.22	391.31
Complete	3.48	101.33	39.24	2547.16
Average	10.83	109.99	104.78	2422.65
Mcquitty	10.83	103.76	172.84	2396.24
Median	0.02	289.61	164.81	455.49
Centroid	10.83	103.76	172.84	2396.24

En la Fig. 13 se puede observar que la mayor distancia inter-cluster, que representa mayor separación entre los clusters formados, se obtiene con el método Ward, cuya distancia es 2557.69.

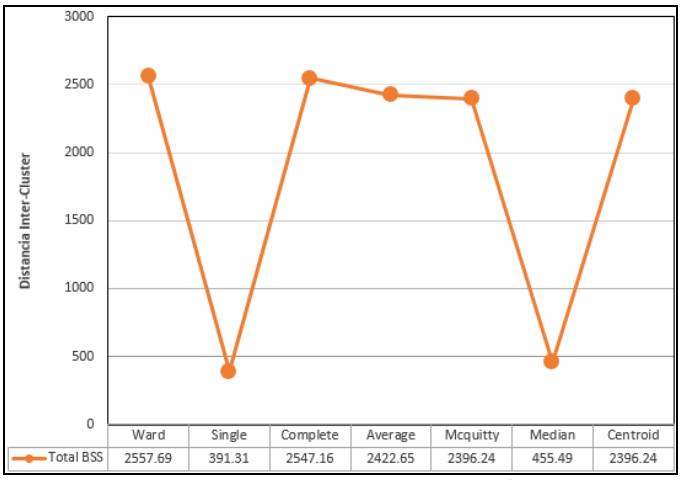

Fig. 13 Distancias inter-cluster – clustering jerárquico.

Dentro de los métodos de clustering jerárquico aglomerativo se elige el método de Ward, ya que los grupos formados según las distancias halladas tienen mayor similitud dentro del grupo y mayor diferencia entre los diferentes grupos.

En la Tabla 4 y Tabla 5 se muestran las distancias intra-cluster e inter-cluster respectivamente para el algoritmo de agrupación K-means.

TABLA 4
DISTANCIAS INTRA-CLUSTER – KMEANS

Cluster 1	Cluster 2	Cluster 3	Total WSS
571.76	340.88	276.50	1189.14

TABLA 5
DISTANCIAS INTER-CLUSTER – KMEANS

Cluster 1	Cluster 2	Cluster 3	Total BSS
101.33	3.49	37.29	2579.99

En la Tabla 6 y Tabla 7 se muestran las distancias intra-cluster e inter-cluster respectivamente para el algoritmo PAM.

TABLA 6
DISTANCIAS INTRA-CLUSTER – PAM

Cluster 1	Cluster 2	Cluster 3	Total WSS
306.14	571.76	351.13	1229.03

TABLA 7
DISTANCIAS INTER-CLUSTER – PAM

Cluster 1	Cluster 2	Cluster 3	Total BSS
2.52	101.33	32.36	2540.10

En la Tabla 8 se muestra el resumen de las distancias intra-cluster e inter-cluster del método ward del clustering jerárquico aglomerativo y algoritmos k-means y PAM del clustering particional.

TABLA 8
DISTANCIAS INTRA-CLUSTER E INTER-CLUSTER

Algoritmos	Total WSS	Total BSS
Ward	1211.44	2557.69
Kmeans	1189.14	2579.99
PAM	1229.03	2540.10

En la Fig. 14 se puede observar que la menor distancia intra-cluster, que representa mayor cohesión y similitud de la agrupación, se obtiene con el algoritmo Kmeans.

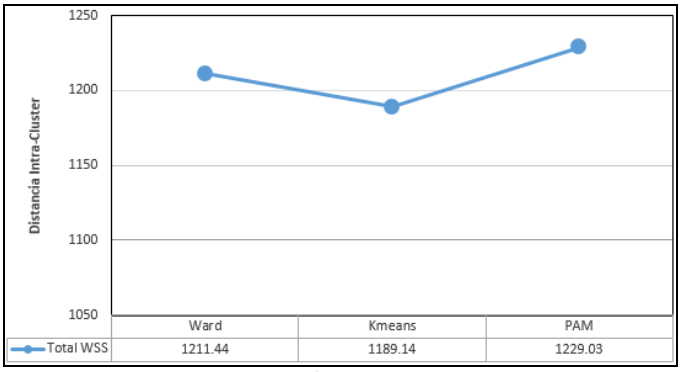

Fig. 14 Comparación distancia intra-cluster.

En la Fig. 15 se puede observar que la mayor distancia inter-cluster, que representa mayor separación o diferencia entre los clusters formados, se obtiene con el algoritmo Kmeans.

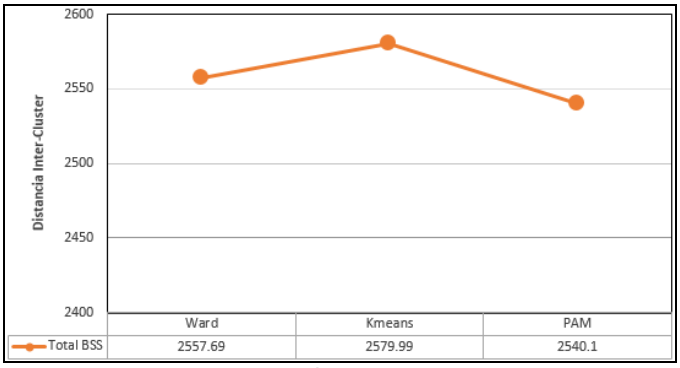

Fig. 15 Comparación distancia inter-cluster.

Para la segmentación de alumnos se elige el algoritmo kmeans por ser la técnica con menor distancia intra-cluster y mayor distancia inter-cluster, que representa mayor homogeneidad dentro del grupo y mayor diferencia entre los diferentes grupos formados.

2) Coeficiente de silueta: Se utiliza el coeficiente de silueta (silhouette) para medir la calidad de los clusters. Típicamente el coeficiente silueta está entre 0 y 1. Mientras más cercano a 1 es mejor. Si es negativo representa una mala agrupación.

En las Fig. 16 se muestra una representación gráfica de los coeficientes de silueta obtenidas para tres clusters por el método jerárquico aglomerativo Ward, algoritmo Kmeans y k-medoids (PAM), en donde presenta una mejor estructura de grupos el algoritmo de Kmeans con un coeficiente de silueta promedio de 0.30 a comparación de los algoritmos Ward y PAM que tienen 0.29 y 0.25 como coeficientes de silueta promedio respectivamente. También se puede notar que el algoritmo k-medoids (PAM) para el cluster 3 que contiene 23 alumnos no es una buena agrupación ya que tiene un coeficiente de silueta promedio bajo (0.14) y además para varios de los alumnos de ese cluster se tiene coeficientes de silueta negativos, en el caso del método Ward también se puede observar en el cluster 3 coeficientes de silueta negativos para algunos de los alumnos.

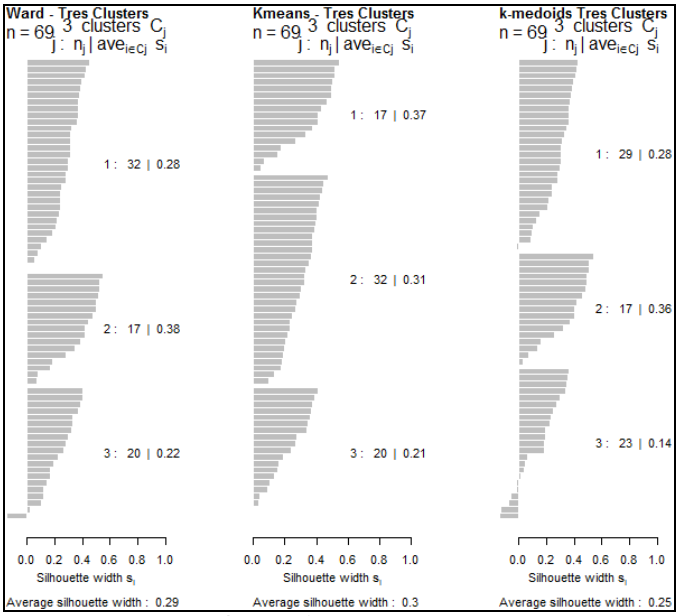

Fig. 16 Comparación de coeficientes de silueta para tres clusters.

Se elige el algoritmo Kmeans para la segmentación académica en tres grupos para el reforzamiento de los alumnos en los niveles básico, intermedio y avanzado, ya que con esta técnica de clustering se obtiene grupos de mejor calidad con coeficientes de silueta positivos para los tres clusters y con menor distancia intra-cluster, que representa mayor similitud dentro del grupo, y mayor distancia inter-cluster, lo cual quiere decir que presenta mayor diferencia entre los diferentes grupos.

En la Fig. 17 se muestra un análisis de componentes principales, donde se aprecian los tres clusters formados por el

algoritmo Kmeans, donde el cluster 1 representa a los alumnos de RENDIMIENTO BAJO ya está más alejado de los cursos, así como el cluster 3 a los alumnos de RENDIMIENTO ALTO, ya que está más cerca del círculo de correlaciones de los cursos. Se puede observar también que los cursos CALC_DIFE (Cálculo Diferencial) y ESTR_DISC_I (Estructuras Discretas I) están altamente correlacionadas así como FUND_PROG (Fundamentos de Programación) y PROG_I (Programación I).

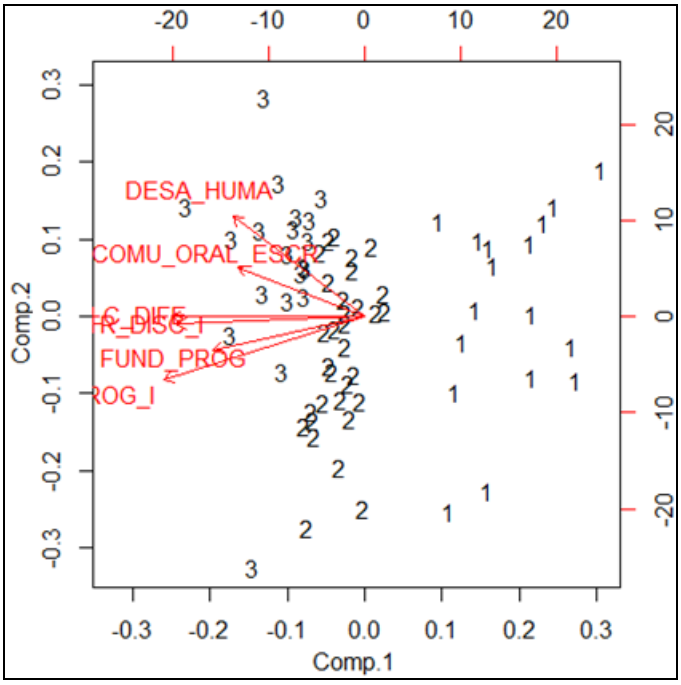

Fig. 17 Componentes principales.

En la Fig. 18 se observa gráficamente que el 25% de la cantidad de alumnos pertenecen al Cluster 1 que son los alumnos con rendimiento bajo, el 46% corresponde al Cluster 2 que son los alumnos de rendimiento medio y el 29% representa al Cluster 3 que son los alumnos con rendimiento alto.

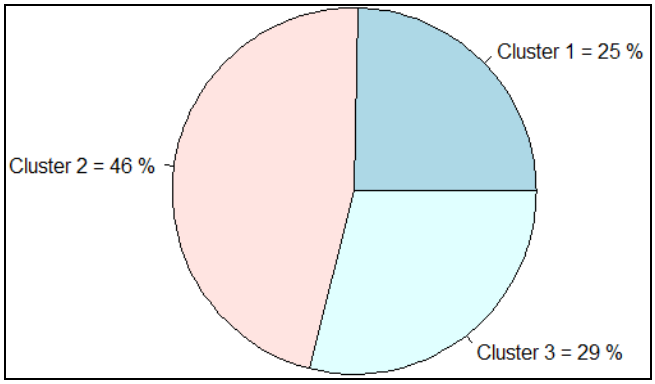

Fig. 18 Porcentajes de cantidad de alumnos en cada cluster.

En la Fig. 19 se muestra los tres grupos obtenidos con el clustering Kmeans, de color rojo están representados los alumnos de rendimiento alto, de color verde los alumnos de rendimiento medio y de color negro los alumnos de rendimiento bajo.

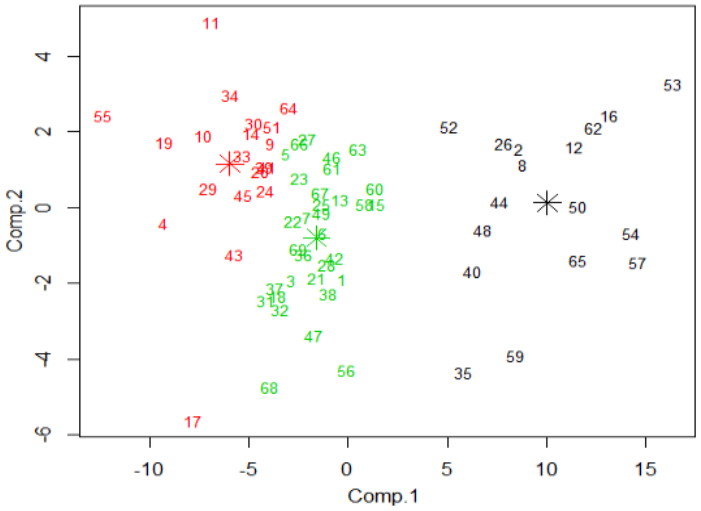

Fig. 21 Clustering Kmeans.

Con la segmentación de alumnos se puede mejorar el proceso de enseñanza-aprendizaje a través de cursos de nivelación personalizada.

Las entidades educativas con la segmentación académica utilizando técnicas y algoritmos de clustering de minería de datos pueden tener un mejor conocimiento de las características y rendimiento de sus alumnos, que les permita crear grupos de estudiantes y reforzar su aprendizaje mediante una enseñanza personalizada, pudiendo así disminuir la deserción y hacerlos más rentables en el tiempo.

IV. Conclusiones

Se realizó un estudio comparativo de técnicas no supervisadas de minería de datos para segmentación de alumnos en tres agrupaciones correspondientes a rendimiento BAJO, MEDIO y ALTO utilizando algoritmos de K-means y PAM dentro del clustering particional y métodos de clustering jerárquico aglomerativo, a partir de sus datos académicos, obteniendo grupos de mejor calidad, con menor distancia intra-cluster y mayor distancia inter-cluster con el algoritmo K-means, que representa mayor homogeneidad dentro del grupo y mayor diferencia entre los grupos. También se utilizó el coeficiente de silueta para medir la calidad de las agrupaciones formadas por las técnicas utilizadas que permitieron seleccionar el algoritmo K-means por ser la técnica de clustering con la que se obtuvo un mayor coeficiente de silueta promedio representando grupos de mejor calidad para la segmentación académica en tres grupos que

puede ser utilizado para el reforzamiento de aprendizaje de los alumnos en los niveles básico, intermedio y avanzado.

Como trabajo futuro se recomienda investigar y comparar otras técnicas no supervisadas de minería de datos para la segmentación como son agrupaciones basadas en densidad, algoritmos basados en redes neuronales, lógica difusa, algoritmos híbridos.

REFERENCIAS

[1] K. Eckert and R. Suénaga, "Aplicación de técnicas de minería de datos al análisis de situación y comportamiento académico de alumnos de la UGD," in *Proc. XV Workshop de Investigadores en Ciencias de la Computación*, RedUNCI, 2013, pp. 92-96.

[2] J. Molen, "Minería de datos educacionales: modelos de predicción del desempeño escolar en alumnos de enseñanza básica," Tesis para optar al título de ingeniero matemático, Universidad de Chile, Santiago, Chile, 2013.

[3] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert Systems with Applications*, vol. 33, no. 1, pp. 135-146, 2007.

[4] Investigación u-Learning, EDM http://ingenieriaeducacion.blogspot.pe/2010/06/edm.html, Revisado en Diciembre del 2016

[5] C. Heiner, R. Baker and K. Yacef, in *Proc. of the Workshop on Educational Data Mining at the 8th International Conference on Intelligent Tutoring Systems* (ITS 2006), Jhongli, Taiwan, 2006.

[6] S. Weiss and N. Indurkhya, *Predictive Data Mining: A Practical Guide*, San Francisco, CA, EEUU: Morgan Kaufmann, 1998.

[7] J. Han and M. Kamber, *Data Mining. Concepts and Techniques*, San Francisco, CA, EEUU: Morgan Kaufmann, 2006.

[8] P. Tan, M. Steinbach and V. Kumar, *Introduction to Data Mining*, New York, EEUU: Pearson Education, 2006.

[9] A. Jain, M. Murty and P. Flynn, "Data clustering: a review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, 1999.

[10] KDnuggets, "Analytics, Data Mining, Data Science software/tools used in the past 12 months," http://www.kdnuggets.com/polls/2015/analytics-data-mining-data-science-software-used.html, Revisado en Enero del 2016.

[11] KDnuggets, "R, Python Duel As Top Analytics, Data Science software – KDnuggets 2016 Software Poll Results," http://www.kdnuggets.com/2016/06/r-python-top-analytics-data-mining-data-science-software.html, Revisado en Noviembre del 2016.

[12] RStudio, "Take control of your R code," https://www.rstudio.com/products/rstudio/, Revisado en Enero del 2016.

[13] E. Himmel, "Modelos de análisis de la deserción estudiantil en la educación superior," *Revista Calidad en la Educación*. Consejo Superior de Educación. Ministerio de Educación, Chile, no. 17, pp. 91-108, 2002.

[14] R. Timarán and J. Jiménez, "Extracción de perfiles de deserción estudiantil en la institución universitaria CESMAG," *Investigium IRE*, vol. 6, no 1, pp. 30-44, 2015.

[15] C. Flores, "Exigencias de calidad de suministro en base a densidad de consumo mediante técnicas de minería de datos," Memoria para optar al título de ingeniero civil electricista, Universidad de Chile, Santiago, Chile, 2014.

[16] M. Vargas, Análisis cluster, http://www.ugr.es/~mvargas/1.acluster.pdf, Revisado en Mayo del 2016.

[17] J. Han, M. Kamber and J. Pei, *Data Mining: Concepts and Techniques*, *3rd ed.*, San Francisco, CA, EEUU: Morgan Kaufmann, 2011.

[18] Bioinformática, "Introducción al clustering en bioinformática," http://genoma.unsam.edu.ar/trac/docencia/wiki/Bioinformatica/Guias/DataMining, Revisado en Mayo del 2016.