UNA APLICACIÓN DEL ANÁLISIS CLÚSTER A LA FORMACIÓN DE GRUPOS EN LA MATERIA DE MATEMÁTICAS EN UNA INSTITUCIÓN DE EDUCACIÓN PERSONALIZADA EN BOGOTÁ

DANIEL SEBASTIÁN BUITRAGO ARRIA

FUNDACIÓN UNIVERSITARIA LOS LIBERTADORES
DEPARTAMENTO DE CIENCIAS BÁSICAS
ESPECIALIZACIÓN EN ESTADÍSTICA APLICADA
BOGOTÁ, D.C. COLOMBIA
2016

UNA APLICACIÓN DEL ANÁLISIS CLÚSTER A LA FORMACIÓN DE GRUPOS EN LA MATERIA DE MATEMÁTICAS EN UNA INSTITUCIÓN DE EDUCACIÓN PERSONALIZADA EN BOGOTÁ

Presentado por:

DANIEL SEBASTIAN BUITRAGO ARRIA

Asesor:

JUAN CARLOS BORBON ARIAS

FUNDACIÓN UNIVERSITARIA LOS LIBERTADORES
DEPARTAMENTO DE CIENCIAS BÁSICAS
ESPECIALIZACIÓN EN ESTADÍSTICA APLICADA
BOGOTÁ, D.C. COLOMBIA
2016

NOTA DE ACEPTACIÓN

_____

_____

_____

_____

_____

_____

_____
Firma del presidente del jurado

_____
Firma del jurado

_____
Firma del jurado

Bogotá D.C., 30 de Julio del 2016

Resumen

A partir de la necesidad de conformar grupos para la materia de Matemáticas en una institución de educación personalizada en Bogotá, se aplica un análisis clúster para determinar los integrantes de cada grupo de acuerdo a las características de su desempeño en la materia en períodos anteriores y así sugerir una distribución del estudiantado para el período siguiente. Se encontraron 3 grupos para los estudiantes que ingresan a octavo grado, 3 grupos para los que ingresan a noveno y 4 grupos para los que ingresan a décimo grado con base en la similitud de las características mencionadas.

Palabras clave

Análisis clúster, grupos, educación, personalizada, desempeño.

Abstract

From the need of conforming groups for the Mathematics subject in a personalized education institution in Bogotá, a cluster analysis is applied in order to determine the elements on each group based on the performance of each student in the previous cut and in this way suggest the student distribution for the next period. The analysis suggested three groups for the students admitted for 8^{th} grade, three groups for the students admitted for 9^{th} grade and four groups for the students admitted for 10^{th} grade based on the similarity features presented.

Key words

Cluster analysis, groups, personalized, education, performance.

Tabla de contenido

Índice de tablas

Índice de ilustraciones

Introducción

Tándem es una institución de carácter privado ubicada en el nororiente de la ciudad de Bogotá, en la localidad de Usaquén. Su razón social es la educación semestralizada y personalizada para adultos. La población estudiantil está conformada por adolescentes y jóvenes en extra edad para realizar sus estudios de media básica y media vocacional, provenientes de familias de un sector socioeconómico alto. Su actividad económica se basa en el Decreto 3011 de 1997 como la nivelación y validación de la formación académica para personas con una edad mayor a la aceptada regularmente en la educación tradicional por años. De esta manera se presenta como una propuesta educativa a aquellas personas que deseen nivelar sus estudios o mejorar sus procesos escolares en las instituciones educativas tradicionales. El origen del capital de Tándem es privado y su representación legal está encabezada por la rectora y la vice rectora quienes, junto con la coordinadora, han diseñado toda la estructura organizacional, así como los procesos, mecanismos e instrumentos de planeación, evaluación, seguimiento y control para todas las instancias de la entidad.

La mayoría de los estudiantes de Tándem ha cursado algunos grados en los más prestigiosos y costosos colegios de la ciudad, pero debido a diferentes circunstancias, sus resultados académicos en éstos no han sido los esperados, por lo que han optado por acudir a una alternativa que, además de aportarle los contenidos de un buen programa académico en bachillerato, mejore sus hábitos de estudio.

Precisamente, su misión determina “Apoyar académicamente a los estudiantes a través de una institución educativa formal de adultos y fortalecer sus procesos de pensamiento por medio de una educación personalizada, que responda a sus necesidades particulares, mejorando los hábitos de estudio y creando una metodología que le permita alcanzar los resultados esperados de acuerdo a sus capacidades.” (Samper & Olarte, 2009)

Tándem cuenta con toda la información académica y actitudinal posible sobre el estudiante que ingresa y el desarrollo de sus procesos de aprendizaje es monitoreado en cada materia de manera periódica a través de reportes directos por parte de cada profesor en fechas establecidas en 3 ocasiones cada 7 semanas o mediante correos electrónicos cada vez que se requiera. Sin embargo, a partir del segundo semestre de 2013, la institución decidió continuar con la metodología personalizada a través de aulas de máximo 3 estudiantes con el fin de minimizar costos. Con esta decisión se genera el problema de cómo organizar dichos grupos, debido a que cada estudiante era asignado de manera individual a un profesor. Se sugiere que dicha clasificación sea basada en el desempeño inmediatamente anterior en la materia, y en la información adicional que se tenga de éste, lo que concuerda con la metodología llevada por Luan (2002). El presente trabajo lleva a cabo una propuesta de dicha clasificación para la materia de Matemáticas con base en el Análisis Clúster para los estudiantes de 7, 8 y 9 grados teniendo en cuenta las directrices señaladas para que, con base en una agrupación bajo similitud de desempeños, se puedan decidir los grupos de máximo 3 estudiantes de la manera más homogénea posible.

Formulación del problema

La pregunta esencial que se busca responder con este trabajo de acuerdo a lo anteriormente expuesto es:

¿Cómo debe ser la distribución de los estudiantes en grupos para la materia de Matemáticas para el siguiente período lectivo en el colegio Tándem con base en su desempeño inmediatamente anterior en la materia?

En sentido estricto se busca entonces elaborar una propuesta clasificatoria diseñada con base en un análisis clúster que haya tenido como entrada las variables de “nota definitiva” y “porcentaje del programa visto”.

Objetivos

Objetivo general

Determinar los grupos de estudiantes más afines de acuerdo a las características de desempeño de nota definitiva, y porcentaje del programa visto.

Objetivos específicos

- Describir la población de estudiantes en términos de las variables: “nota definitiva”, “nivel” y “porcentaje del programa visto”.

- Aplicar un análisis clúster que permita determinar una propuesta de clasificación de los estudiantes en los determinados grupos.

Justificación

Como ya se mencionó, la principal motivación para realizar este estudio parte de la necesidad de agrupar las clases impartidas por grupos de 3 estudiantes en las distintas materias. En cada una de ellas existen estudiantes aventajados o con un bagaje o interés mucho mayor relativamente, que hace que los procesos y competencias alcanzadas en una determinada materia sean de un nivel de profundidad mucho mayor. De aquí que, agrupar aleatoriamente a los estudiantes puede resultar en situaciones en las que un estudiante aventajado se encuentre en el mismo grupo que uno no tan aventajado que originará uno de dos resultados: o el aventajado no desarrolla plenamente su potencial, o el no tan aventajado perderá fácilmente el hilo de la clase y probablemente reprobará o desertará. Como la filosofía de los centros de educación personalizada procuran, al contrario del proceso de normalización y homogenización de los colegios tradicionales, potenciar las habilidades y talentos particulares de sus estudiantes destacados, y al mismo tiempo reforzar aquellos componentes donde presenten dificultades, se busca entonces conformar grupos lo más homogéneos posibles en cuanto a las habilidades manifiestas hacia la materia en términos de evidencias cuantitativas en su proceso educativo, recopiladas por los mismos profesores.

Diversos estudios, como los de Feldman, Goncalves, Chacón-Puignau, Zaragoza, & Pablo, (2008) y Plazas, Penso, & López, (2006) por mencionar algunos, han establecido una relación entre el bienestar social de un estudiante y su rendimiento académico; por lo que lograr el desarrollo del potencial de cada estudiante a su medida repercute de manera importante en la meta de mejorar su calidad de vida, y esta es en parte la misión de las instituciones de educación personalizada. La educación personalizada no es una modalidad nueva en Colombia; sin embargo, las instituciones educativas que las ofrecen sí lo son, así como las problemáticas que plantean. No es de extrañar por tanto, que difícilmente se encuentren estudios sobre la población estudiantil en este tipo de escolaridad en el país y mucho menos sobre clasificación en términos de desempeño escolar.

Marco de referencia

Los estudios sobre planeación educativa con base en análisis estadísticos provienen en su mayoría de la Psicología (Renninger & Wozniak, 1985). Debido a esto, las herramientas utilizadas y la tipología de los estudios giraban en torno a aquellos vistos en las facultades de Psicología: en su mayor parte estudios de correlación o asociación de variables, o influencia de factores en la atención, motivación o aprendizaje utilizando análisis de varianza o regresión. Sin embargo, con avances en los métodos computacionales y algorítmicos, el inventario de herramientas disponibles aumentó, diversificando a su vez la tipología de estudios posibles. Tal como lo relata Luan (2002), se ha puesto a disposición de los investigadores todo un repertorio de técnicas que permiten *administrar el conocimiento* sobre la educación, y recopilar, analizar y tomar decisiones con grandes volúmenes de información. Existen por otro lado diversos estudios que sostienen que el mejor predictor del desempeño de un estudiante es su rendimiento anterior (Repáraz, Tourón, & Villanueva, 1990), por lo que se puede encontrar numerosa evidencia a favor de que el rendimiento académico previo es una variable determinante a la hora de categorizar a un estudiante.

En esta dirección se puede encontrar el trabajo de Rico & Habana (2012), en donde se utilizó un modelo de regresión logística para predecir el rendimiento académico en una clase con base en los resultados inmediatamente anteriores. Este trabajo utilizó únicamente como variables explicativas las notas y así mismo la variable a predecir era una nota.

Por otro lado, la investigación de Ullmer (2012) es una de las primeras en incluir otro tipo de variables para pronósticos en el ámbito educativo. El autor consideró factores como juicios valorativos verbales y género (entre otros) para predecir el desempeño en una clase introductoria de Economía. A pesar de encontrar que el género no era estadísticamente significativo en la predicción, sí lo fue el juicio valorativo dado por un evaluador calificado.

A la hora de clasificar individuos, la técnica de Análisis Clúster ha sido aplicada exitosamente en investigaciones sobre educación por autores como Antonenko, Toy, & Niederhauser (2012) para trazar perfiles de comportamiento en plataformas de aprendizaje online. Pero fue el trabajo de Organista-Sandoval & Henríquez-Ritchie (2012) el que ha utilizado esta técnica de la manera más similar al objetivo del presente trabajo al proponerse clasificar estudiantes de nuevo ingreso a una universidad pública con base a variables de desempeño académico inmediatamente anteriores.

Por supuesto, no es de esperar que exista un trabajo completamente similar, ya que las circunstancias que rodean la problemática originada en el presente es poco común, por lo que los desarrollos que más se asemejan, son aquellos utilizados en la clasificación de estudiantes en categorías de acuerdo a su desempeño.

Siguiendo la metodología desarrollada por Organista-Sandoval & Henríquez-Ritchie (2012), se tomarán como variables de clasificación aquellas que den cuenta de una u otra manera del desempeño del estudiante en dicha materia. En este caso particular, la institución ha recopilado tres medidas para cada estudiante: Su "nota definitiva", que es una asignación numérica de 1 a 10 (donde 10 es el más alto) del nivel de dominio de las temáticas vistas durante el período lectivo; "el nivel", que es un juicio valorativo dado por el docente a cargo que tiene tres niveles (alto, medio o bajo) dependiendo del nivel de profundidad y complejidad al que pueden llevarse las temáticas de esa materia en particular con ese estudiante en particular; y finalmente el "porcentaje del programa visto", que refleja de manera cuantitativa la velocidad a la que el estudiante pudo avanzar en las temáticas, expresando en porcentaje la parte total abarcada por éste durante el período en cuestión. Estas serán las variables base a partir de las cuales se determinará el número de grupos de 3 estudiantes para el siguiente nivel.

Debido a que se prestableció un número de elementos en cada clúster (3 estudiantes en cada uno), el análisis pertinente es el de las *k*-medias. Esta técnica busca una partición de *n* individuos en k grupos o clústeres $G_1, G_2, \dots, G_k$ con base en un conjunto multivariado de datos, donde cada G_i denota el conjunto de n_i individuos en el *i*-ésimo grupo, con *k* dado. De esta manera, se busca minimizar la suma de cuadrados intragrupal (maximizando a su vez la intergrupal) sobre todas las variables involucradas. Esto es, minimizar la función:

$$SCIG = \sum_{j=1}^{q} \sum_{l=1}^{k} \sum_{i \in G_l} \left(x_{ij} - \left(\overline{x_j} \right)^{(l)} \right)^2$$

Donde $\left(\overline{x_j} \right)^{(l)} = \frac{1}{n_i} \sum_{i \in G_l} x_{ij}$ es la media de la variable *j* de los individuos en el grupo G_l. Debido a la complejidad de los cálculos para encontrar una solución exacta, se remite a métodos algorítmicos y numéricos que puedan arrojar una valoración aproximada de la misma pero que sin embargo, no garantizan la minimalidad. Uno de los algoritmos más utilizados es el de Steinley (2003), que se trata de un tipo de optimización por iteración y se puede describir mediante los siguientes pasos:

1. Se inicia con una partición aleatoria de la población en *k* grupos con base a k centros seleccionados de manera aleatoria sobre el hiper-volumen que contiene el conjunto de datos.
2. Se asigna cada individuo al clúster a cuyo centro se tenga la menor distancia.
3. Se recalcula la posición de los centros de los clústeres con base en los grupos formados.
4. Si con estos nuevos centros no hubo cambio en la pertenencia de un individuo a un clúster, se conserva esta distribución. De lo contrario se repite desde el paso 2.

Esta técnica presenta tres inconvenientes: En primer lugar, no es invariante en cuanto a escala; Esto implica que, usar los datos mismos o alguna estandarización de estos puede cambiar significativamente los resultados. En segundo lugar, el análisis supone una estructura esférica en los datos (debido a que utiliza una distancia euclidiana) y así mismo la forma geométrica del clúster, ignorando otras posibles formas de éste. Finalmente, es sensible a las condiciones iniciales; Puede suceder que la distribución inicial elegida pueda afectar la convergencia del algoritmo y mostrar resultados distintos para distribuciones iniciales distintas (i.e., la SCIG está convergiendo a un mínimo local y no global). Sin embargo, este abordaje permite su implementación mediante softwares estadísticos con una complejidad de orden lineal, logrando una respuesta razonablemente buena a la situación.

Marco metodológico

Tipo de estudio

El estudio es de tipo descriptivo. Se busca realizar inicialmente un análisis exploratorio de los datos que permita encontrar y describir las variables relevantes en la clasificación más adecuada de las unidades experimentales con base en las características de interés: el desempeño académico de la materia de Matemáticas.

Método

Para este trabajo se emplea un método estadístico para el análisis de los datos.

Población y muestra

La institución Tándem cuenta en estos momentos con dos instalaciones separadas. La primera se llama la casa AB y la otra CD. La primera, para el año 2013 es el espacio de formación para los estudiantes que se encuentran cursando los ciclos equivalentes a 7º, 8º y 9º. El interés del estudio se basa en esta población de estudiantes, ya que se necesita conocer la forma en cómo se distribuirán en cada materia para realizar la correspondiente planeación logística de los salones en las casas.

Descripción de la población

La población objetivo consta de 36 estudiantes (21 hombres, 15 mujeres) con edades entre los 13 y los 17 años, todos pertenecientes a sectores socioeconómicos altos y residentes en la zona del distrito capital. Debido a la delimitación espacial de la institución, todos estos estudiantes se consideran como la población total objetivo, por lo que no se requirieron diseños muestrales.

Instrumentos y materiales

Los instrumentos utilizados fueron los datos acerca del desempeño académico en la materia de Matemáticas recolectados por los profesores respectivos durante el semestre 2013-I, y consolidados en tres variables: "nota definitiva", "nivel" y "porcentaje visto". Para el análisis de los datos se usaron los softwares de *Minitab*, y *Weka* con el algoritmo y tratamiento de datos multivariados para análisis clúster descrito por Pardo & del Campo (2007) y Lebart, Morineau, & Piron, (1995).

Procedimiento y diseño estadístico

Se recolectaron los datos de las variables mencionadas para todos los estudiantes activos en los grados 7, 8 y 9 de la institución educativa Tándem para la finalización del período lectivo 2013-I. Después de corregir y depurar los datos, se corrió en *Minitab* un análisis de conglomerados por observaciones a manera de análisis preliminar para identificar un número adecuado de clústeres que posteriormente se asignaron en el análisis "conglomerado de k-medias" que arrojó las asignaciones finales ilustradas en el dendograma. Debido al reducido número de datos, no fue necesario realizar un análisis mixto (Pardo & del Campo, 2007).

Tabla de definiciones y variables

Tabla 1. Definiciones y variables a utilizar.

Nombre	**Definición conceptual**
Nota definitiva	Variable cuantitativa continua. Es la asignación valorativa total obtenida de la ponderación de las notas obtenidas por el estudiante en sus distintos momentos a lo largo del período lectivo. Su valor es un número decimal entre 0 y 10, siendo 10 la máxima valoración posible y 7.0 la mínima aprobatoria.
Porcentaje visto	Variable cuantitativa continua. Representa una valoración de la cantidad de temáticas abarcadas con respecto al total de temáticas planteadas en el plan de estudios contemplado para ese grado en esa materia particular. Su valor es un número decimal entre 0 y 1.
Nivel	Variable cualitativa ordinal. Es un juicio valorativo del docente acerca del grado de profundidad y complejidad de los contenidos a la que el estudiante está en capacidad de responder. Las categorías son: Alto, Medio o Bajo.

Resumen descriptivo de las variables

Variable nota definitiva:

Puede observarse que su media es de 7.3 con una desviación estándar de 0.6.

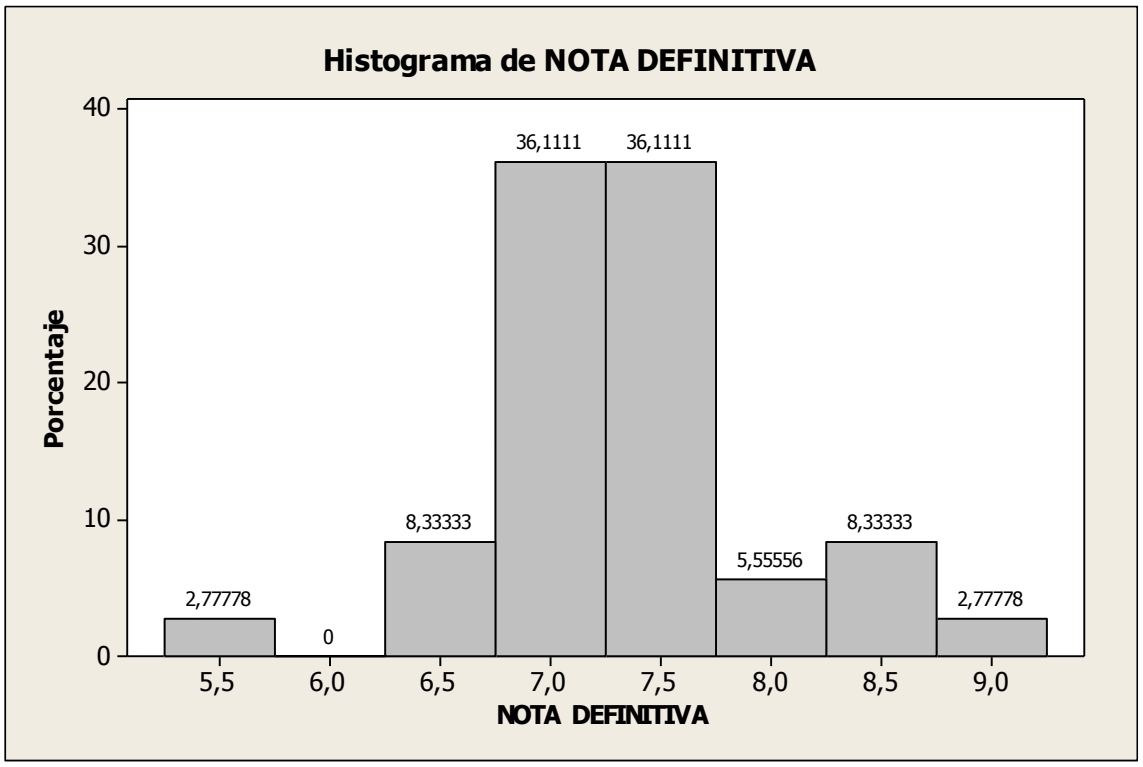

Figura 1. Histograma de la variable "nota definitiva".

Variable porcentaje del programa visto:

Puede afirmarse que en la mayor parte de los casos se alcanzó a abarcar por lo menos el 90% del programa de Matemáticas planeado para el período en cada uno de los grados.

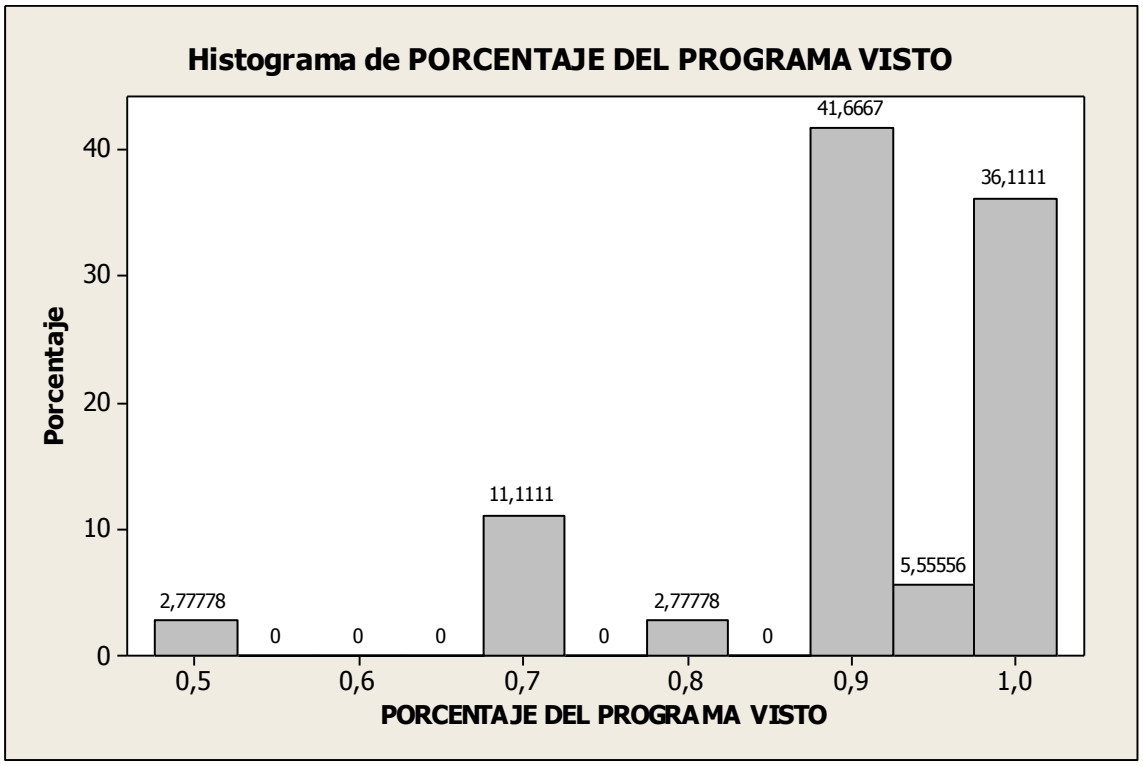

Figura 2. Histograma de la variable "porcentaje del programa visto".

Variable nivel:

Puede observarse que la mayoría de estudiantes se encuentran clasificados en un nivel Medio o Alto dentro de los contenidos programáticos de la materia de Matemáticas.

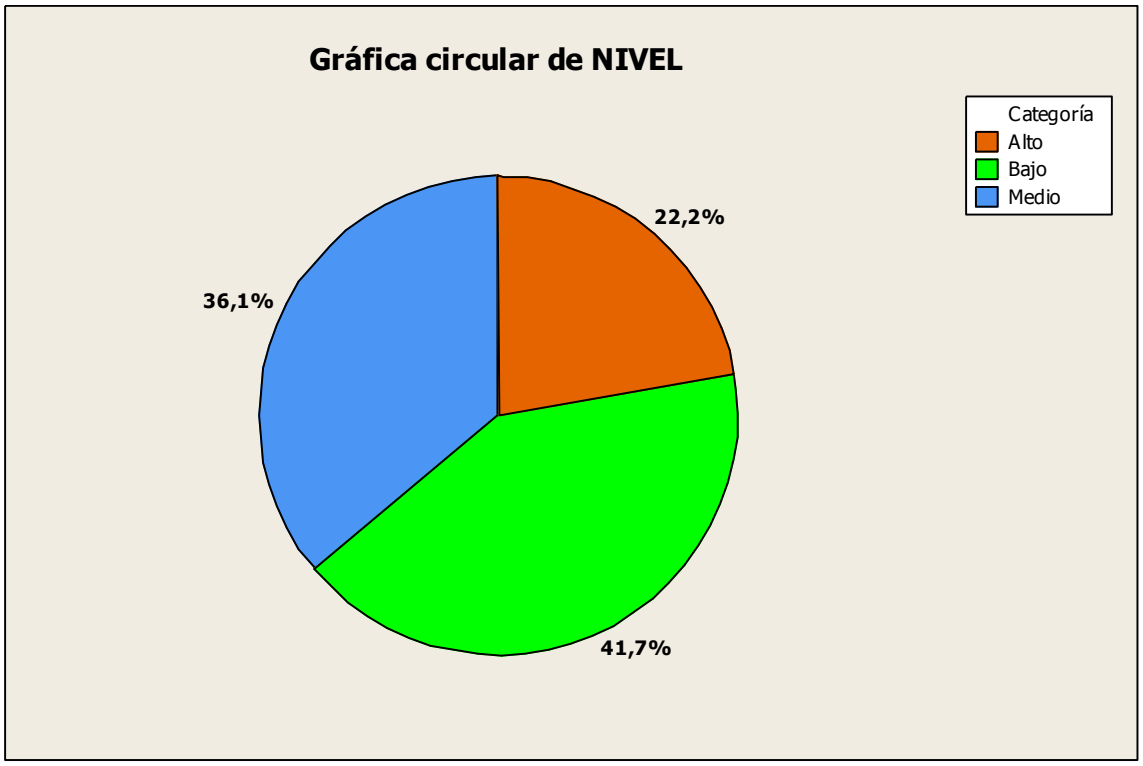

Figura 3. Diagrama de sectores de la variable "nivel".

Búsqueda de correlaciones y asociaciones entre las variables:

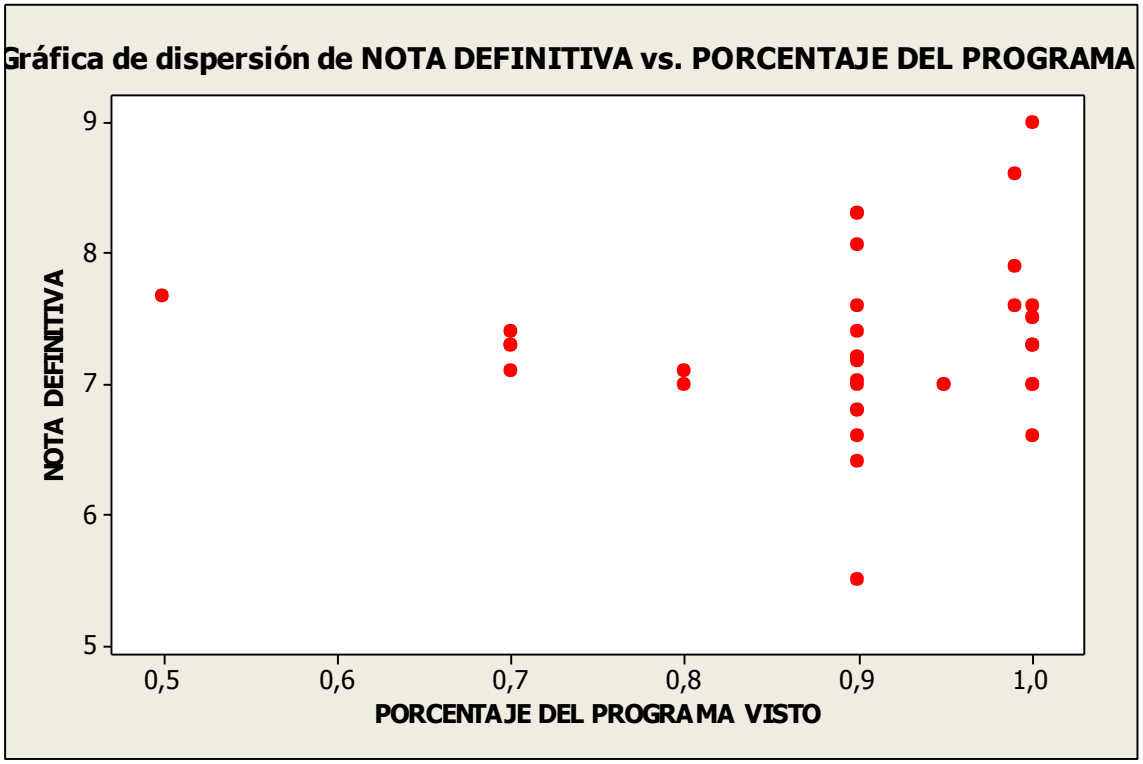

Figura 4. Diagrama de dispersión entre las variables "nota definitiva" y "porcentaje del programa visto".

Estadísticas tabuladas: NIVEL. GENERO

Filas: NIVEL Columnas: GENERO

	F	M	Todo
ALTO	2	5	7
BAJO	10	6	16
MEDIO	3	10	13
Todo	15	21	36

El coeficiente de correlación de Pearson obtenido fue de 0.076 con un valor p de 0.656, por lo que puede verse que no hay una relación estadísticamente significativa entre las variables “Nota definitiva” y “Porcentaje del programa visto”. Esto señala que en realidad el hecho de haber podido desarrollar poca o gran parte del programa no depende directamente

del desempeño obtenido por el estudiante. Por otro lado, se encontró un leve nivel de asociación entre la variable “nivel” y el género de los estudiantes; sin embargo no es una asociación estadísticamente significativa ($p=0.059$), por lo que no es posible concluir que el nivel de desempeño de los estudiantes en la materia de Matemáticas esté relacionado con su género.

Desarrollo del análisis clúster

Para la asignación de los grupos debe tenerse en cuenta que se van a formar de acuerdo al grado al que van a ingresar. Para esto, obsérvese que hay un estudiante de 5 que pasa a 6 y un estudiante de 6 que pasa a 7. Estos dos elementos no entran en el análisis clúster, al igual que tres estudiantes que ingresan a 11. Hay que tener en cuenta además que hay 6 estudiantes de noveno grado que reprobaron la materia, por lo que se incluyen en los que ingresan a la misma para ese grado.

A continuación se muestra el desarrollo del análisis para los estudiantes que aprobaron 7 e ingresan a 8:

La salida de Minitab para el análisis de conglomerado por observaciones es la siguiente:

Análisis de observaciones de conglomerado: NOTA DEFINIT. PORCENTAJE. ...

Distancia euclediana, Enlace de Ward
Pasos de amalgamación

Paso	Número de grupos	Nivel de semejanzal	Nivel de distancia	Grupos incorporados		Nuevo grupo	Número de obs. en el grupo nuevo
1	7	100,000	0,00000	3	4	3	2
2	6	95,457	0,10000	2	5	2	2
3	5	90,915	0,20000	6	8	6	2
4	4	87,886	0,26667	3	7	3	3
5	3	60,554	0,86837	3	6	3	5
6	2	30,143	1,53783	1	2	1	3
7	1	-71,666	3,77902	1	3	1	8

Puede observarse que el cambio más abrupto en el nivel de semejanza ocurre al seleccionar 3 grupos. Lo que, de acuerdo con Lebart et al. (1995), es un buen indicador del número de clústeres a seleccionar. El análisis de k-medias es el siguiente:

Análisis de grupos de K-medias: NOTA DEFINITIVA. PORCENTAJE. NIVEL CODIFICADO

Partición final

Número de grupos: 3

	Número de observaciones	Dentro de la suma de cuadrados del grupo	Distancia promedio desde el centroide	Distancia máxima desde centroide
Grupo1	1	0,000	0,000	0,000
Grupo2	2	0,005	0,050	0,050
Grupo3	5	0,284	0,222	0,305

Centroides de grupo

Variable	Grupo1	Grupo2	Grupo3	Centroide principal
NOTA DEFINITIVA	7,9000	7,3500	7,2600	7,3625
PORCENTAJE	0,9900	0,7000	0,9600	0,8987

Las distancias entre los centroides de grupos

	Grupo1	Grupo2	Grupo3
Grupo1	0,0000	1,1775	2,1001
Grupo2	1,1775	0,0000	1,0372
Grupo3	2,1001	1,0372	0,0000

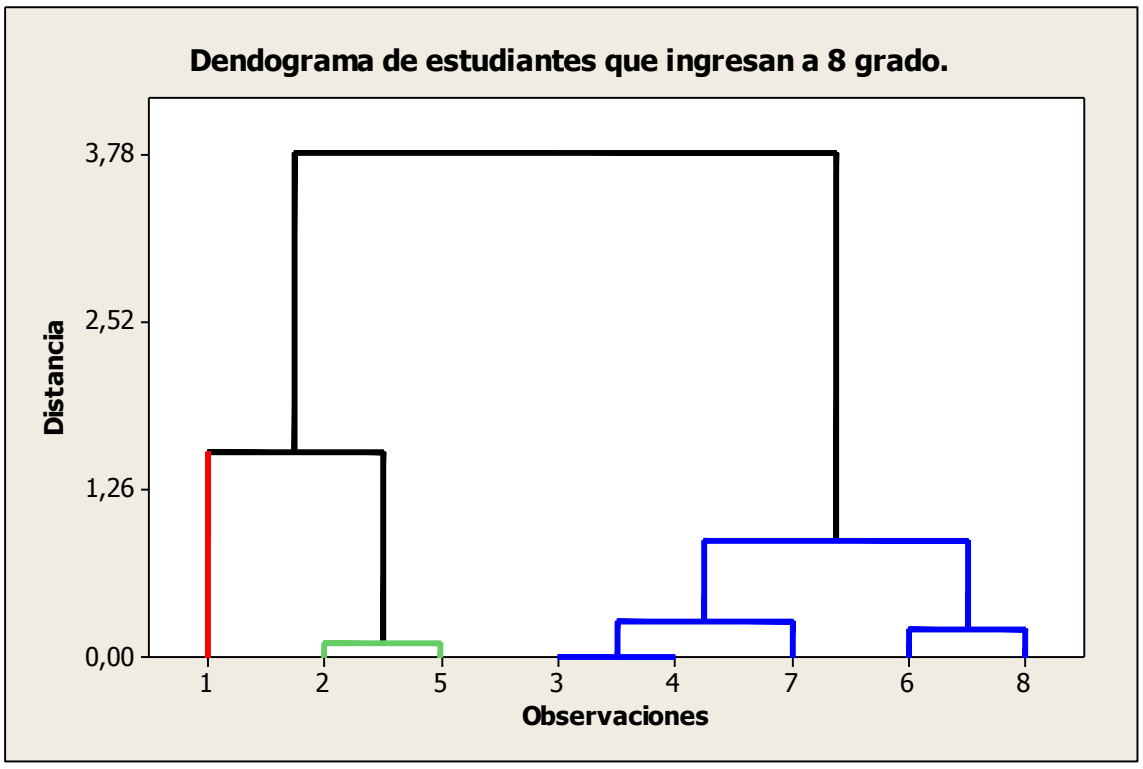

Figura 5. Dendograma de los estudiantes que ingresan a 8.

Con respecto a los estudiantes que ingresan a 9 grado:

El análisis de conglomerados por observaciones arroja los siguientes resultados:

Análisis de observaciones de conglomerado: NOTA DEFINIT. PORCENTAJE V. ...

Distancia euclediana, Enlace de Ward
Pasos de amalgamación

Paso	Número de grupos	Nivel de semejanzal	Nivel de distancia	Grupos incorporados		Nuevo grupo	Número de obs. en el grupo nuevo
1	13	100,000	0,00000	10	12	10	2
2	12	98,305	0,05831	2	7	2	2
3	11	97,094	0,10000	9	14	9	2
4	10	95,890	0,14142	3	5	3	2
5	9	92,762	0,24907	9	13	9	3
6	8	91,225	0,30193	3	4	3	3
7	7	88,172	0,40700	2	10	2	4
8	6	82,072	0,61688	8	9	8	4
9	5	70,289	1,02233	1	2	1	5
10	4	60,721	1,35156	3	6	3	4
11	3	52,078	1,64896	8	11	8	5
12	2	-9,682	3,77409	1	3	1	9
13	1	-107,202	7,12968	1	8	1	14

Que sugiere nuevamente tres clústeres para el análisis. El correspondiente resultado de clasificación por k-medias tiene el siguiente informe:

Análisis de grupos de K-medias: NOTA DEFINIT. PORCENTAJE V. NIVEL CODIFI

Partición final

Número de grupos: 3

	Número de observaciones	Dentro de la suma de cuadrados del grupo	Distancia promedio desde el centroide	Distancia máxima desde centroide
Grupo1	5	0,437	0,227	0,557
Grupo2	5	1,256	0,383	0,920
Grupo3	4	0,755	0,383	0,725

Centroides de grupo

Variable	Grupo1	Grupo2	Grupo3	Centroide principal
NOTA DEFINITIVA	7,0460	6,4200	7,5750	6,9736

```
PORCENTAJE VISTO  0,9280  0,9300  0,8750     0,9136

Las distancias entre los centroides de grupos

        Grupo1  Grupo2  Grupo3
Grupo1  0,0000  1,1798  1,1325
Grupo2  1,1798  0,0000  2,3102
Grupo3  1,1325  2,3102  0,0000
```

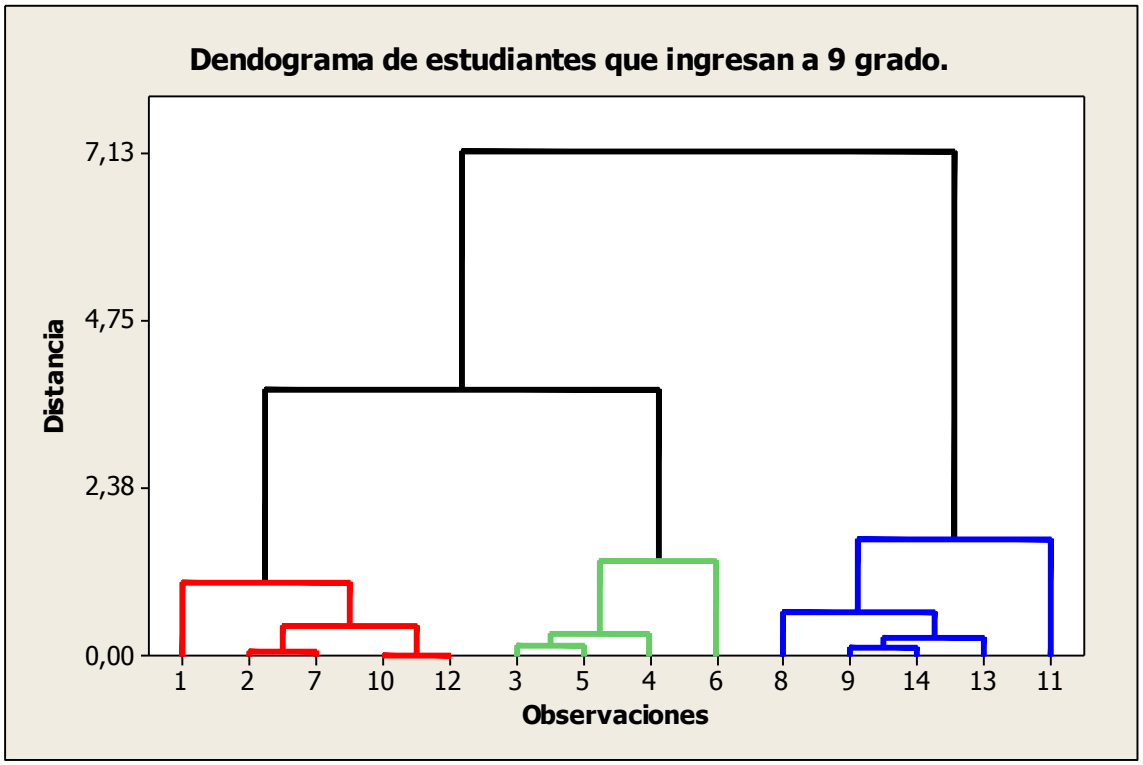

Figura 6. Dendograma de la distribución de estudiantes que ingresan a 9 grado.

Estudiantes que ingresan a 10 grado:

El análisis inicial de conglomerados es el siguiente:

Análisis de observaciones de conglomerado: NOTA DEFINIT. PORCENTAJE V. ...

Distancia euclediana, Enlace de Ward
Pasos de amalgamación

Paso	Número de grupos	Nivel de semejanzal	Nivel de distancia	Grupos incorporados		Nuevo grupo	Número de obs. en el grupo nuevo
1	13	100,000	0,00000	10	12	10	2
2	12	98,305	0,05831	2	7	2	2
3	11	97,094	0,10000	9	14	9	2
4	10	95,890	0,14142	3	5	3	2
5	9	92,762	0,24907	9	13	9	3
6	8	91,225	0,30193	3	4	3	3
7	7	88,172	0,40700	2	10	2	4
8	6	82,072	0,61688	8	9	8	4
9	5	70,289	1,02233	1	2	1	5
10	4	60,721	1,35156	3	6	3	4
11	3	52,078	1,64896	8	11	8	5
12	2	-9,682	3,77409	1	3	1	9
13	1	-107,202	7,12968	1	8	1	14

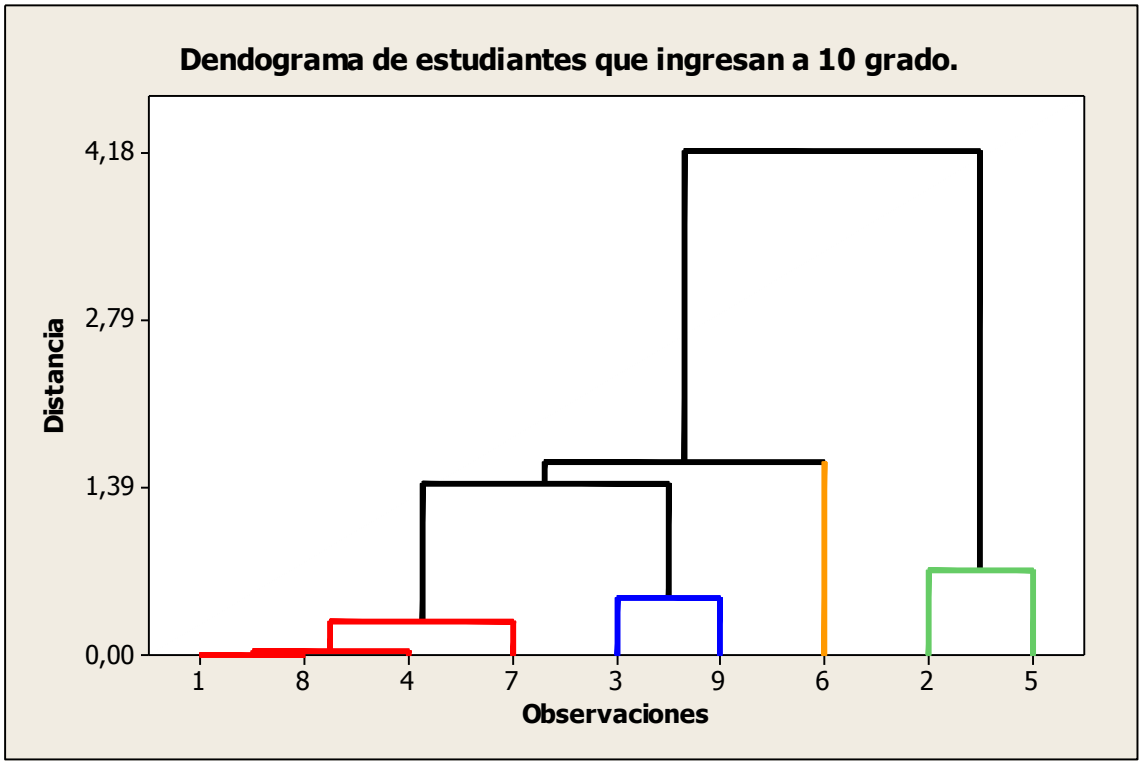

Figura 7. Dendograma de la distribución de estudiantes que ingresan a 10 grado.

Que sugiere 4 clústeres que reportan el siguiente resultado en el análisis de k-medias:

Análisis de grupos de K-medias: NOTA DEFINIT. PORCENTAJE V. NIVEL CODIFI

Partición final

Número de grupos: 4

	Número de observaciones	Dentro de la suma de cuadrados del grupo	Distancia promedio desde el centroide	Distancia máxima desde centroide
Grupo1	5	0,194	0,146	0,364
Grupo2	2	0,250	0,354	0,354
Grupo3	1	0,000	0,000	0,000
Grupo4	1	0,000	0,000	0,000

Centroides de grupo

Variable	Grupo1	Grupo2	Grupo3	Grupo4	Centroide principal
NOTA DEFINITIVA	7,2360	8,6500	8,0700	7,0000	7,6167
PORCENTAJE VISTO	0,9000	0,9500	0,9000	1,0000	0,9222

Las distancias entre los centroides de grupos

	Grupo1	Grupo2	Grupo3	Grupo4
Grupo1	0,0000	1,7326	0,8340	1,0323
Grupo2	1,7326	0,0000	1,1571	2,5933
Grupo3	0,8340	1,1571	0,0000	1,4680
Grupo4	1,0323	2,5933	1,4680	0,0000

Análisis de resultados y conclusiones

El objetivo principal del trabajo era dar solución al problema de clasificar estudiantes de educación personalizada en los grupos de 3 estudiantes más idóneos posibles de acuerdo a su desempeño académico y su capacidad para asimilar contenidos de la materia de Matemáticas, potenciando así sus habilidades en esta materia y optimizando el trabajo docente al homogenizar el nivel de desempeño de los estudiantes en cada grupo, permitiendo así la posibilidad de focalizar una estrategia puntual hacia cada grupo objetivo, ya sea para potenciar sus fortalezas o para superar sus dificultades. Se encontró que la población de estudiantes bajo estudio tenía una nota definitiva promedio de 7.3 con una baja dispersión (coeficiente de variación: 8.21%) en donde la mayor parte de ellos lograron abarcar al menos el 90% de las temáticas propuestas para la materia de Matemáticas. El 41.7% del total de estudiantes fueron clasificados en la categoría de "nivel bajo" por parte de sus profesores, mientras que el 36.1% y el 22.2% fueron clasificados en las categorías de "nivel medio" y "nivel alto" respectivamente. Con esta población de estudio, se empleó un análisis clúster con base en las variables "nota definitiva" y "porcentaje visto". El análisis clúster permitió encontrar los siguientes resultados:

- Con respecto a los estudiantes que ingresan a 8 grado se identificaron tres grupos: El primer grupo (con un estudiante) se caracteriza por estudiantes con notas definitivas cercanas a 7.9 y con un porcentaje visto del 99%. El segundo grupo (donde clasificaron dos estudiantes) se caracteriza por tener notas definitivas cercanas a 7.35 y haber completado el programa en un 70%. El tercer grupo (con 5 estudiantes) se caracteriza por estudiantes que obtuvieron notas definitivas cercanas a 7.26 y haber completado el programa en 96%.

- Con respecto a los estudiantes que ingresan a 9 grado se identificaron tres grupos: El primero (que consta de 5 estudiantes) se caracteriza por notas definitivas cercanas a 7.04 y un porcentaje visto del programa de 92%. El segundo (también con 5 estudiantes) se caracteriza por notas cercanas a 6.4 y un porcentaje visto del programa de 93% (que podría decirse, es el grupo de los repitentes). Finalmente el tercero (con 4 estudiantes) se caracterizan por haber obtenido una nota definitiva cercana a 7.57 y haber completado el programa en un 87%.
- Finalmente para los estudiantes que ingresan a 10 grado se identificaron cuatro grupos: El primero (con 5 estudiantes) se caracteriza por aquellos estudiantes que obtuvieron nota definitiva cercana a 7.23 y un porcentaje visto de 90%. El segundo (con 2 estudiantes) se caracteriza por aquellos estudiantes que obtuvieron nota definitiva cercana a 8.65 y alcanzaron un porcentaje visto de 95%. El tercero (de un estudiante) se caracteriza por estudiantes que obtuvieron una nota definitiva cercana a 8.07 y alcanzaron un porcentaje de programa visto de 90%. Finalmente el cuarto (también de un estudiante) lo caracterizan estudiantes con una nota definitiva cercana a 7.0 y tienen un porcentaje visto de 100%.

Como una última etapa del estudio, se realizó un análisis para determinar si la clasificación de los estudiantes en los distintos grupos estaba relacionada con el "nivel" de cada estudiante. Para esto se realizó una prueba de Chi-cuadrado para las variables "grupo" y "nivel". Lamentablemente el número de observaciones de aquellos que entraban a 8º y a 10º era insuficiente para dar algún resultado concluyente. Sin embargo, el análisis arrojó una asociación significativa entre el nivel y el grupo asignado para aquellos estudiantes que ingresaban a 9º ($p<0.00$). Esto representa un indicio de que la variable "nivel" puede ser usada en futuras investigaciones como factor de estudio para la clasificación de estudiantes, a pesar de que se necesiten más datos y más información para determinar cualquier influencia que

pueda tener esta variable en este tipo de problemáticas. Sumado a esto, puede decirse además que la clasificación realizada por el análisis clúster es coherente con el "nivel" al cual cada profesor percibe a sus estudiantes; en donde, en este caso, el análisis clasificatorio se ha anticipado a la categorización del nivel de los estudiantes, dando una evidencia más de que ha sido, al menos en el caso de los estudiantes que ingresan a 9º, una clasificación consistente con el contexto de los datos.

Como se mencionó en el marco teórico, la principal variable utilizada por la mayoría de los autores en este tipo de estudios clasificatorios es el del desempeño numérico del período inmediatamente anterior (Repáraz et al., 1990). Sin embargo, en el contexto de la educación personalizada, no es información suficiente un puntaje cuantitativo para evaluar el desempeño de un estudiante; ya que, por las mismas circunstancias de la educación personalizada, no hay un avance homogéneo en las temáticas para todos los estudiantes que cursan un mismo grado (como sí sucede en los colegios tradicionales). De aquí la necesidad de incluir en el estudio una variable que midiera esas diferencias en las temáticas vistas en términos porcentuales (variable "porcentaje del programa visto") para dar mayor coherencia al estudio y enriquecer los insumos del análisis de clasificación.

De esta manera se configura una propuesta de distribución de la población de estudiantes de instituciones de educación personalizada o semipersonalizada con base en las características de desempeño propuestas dando plenamente respuesta a los objetivos del estudio planteados.

Debido a que las características que permitieron la ejecución del estudio se retroalimentan cada semestre, es posible replicar el estudio cada período brindando una valiosa herramienta de planeación educativa a la institución, evidenciando la aplicación de las

técnicas estadísticas a la solución de un problema encontrado en el ámbito de desarrollo profesional del autor del trabajo.

Referencias

Antonenko, P. D., Toy, S., & Niederhauser, D. S. (2012). Using cluster analysis for data mining in educational technology research. *Educational Technology Research and Development*, *60*(3), 383–398. http://doi.org/10.1007/s11423-012-9235-8

Feldman, L., Goncalves, L., Chacón-Puignau, G., Zaragoza, N., & Pablo, J. De. (2008). Relaciones entre estrés académico, apoyo social, salud mental y rendimiento académico en estudiantes universitarios venezolanos. *Univ. Psychol.*, *7*, 739–752.

Lebart, L., Morineau, a, & Piron, M. (1995). Statistique exploratoire multidimensionnelle. *Statistique Exploratoire Multidimensionnelle.*, 439.

Luan, J. (2002). Data Mining and Its Applications in Higher Education, (113), 17–36.

Organista-Sandoval, J., & Henríquez-Ritchie, P. (2012). Revista Electrónica de Investigación Educativa Clasificación de estudiantes de nuevo ingreso a una universidad pública , con base a variables de desempeño académico , uso de tecnología digital y escolaridad de los padres Classification of Incoming Freshma, *14*, 34–55.

Pardo, C. E., & del Campo, P. C. (2007). Combinación de métodos factoriales y de análisis de conglomerados en R: El paquete factoclass. *Revista Colombiana de Estadistica*, *30*(2), 231–245.

Plazas, E., Penso, R., & López, S. (2006). Relación entre estatus sociométrico, género y rendimiento académico. *Psicología Desde El Caribe*, *17*, 176–195.

Renninger, K. A., & Wozniak, R. H. (1985). Effect of interest on attentional shift, recognition, and recall in young children. *Developmental Psychology*, *21*(4), 624–632. http://doi.org/10.1037/0012-1649.21.4.624

Repáraz, C., Tourón, J., & Villanueva, C. (1990). Estudio de algunos factores relacionados con el rendimiento académico en 8ª de EGB. *Actualidades En Psicología*.

Rico, J. J. H., & Habana, L. (2012). Ordinal Para La Predicción Del Rendimiento Académico, *33*(3), 252–267.

Samper, M., & Olarte, N. (2009). *TANDEM, Nuestra intención en la comunidad El Proyecto Educativo Institucional*. Universidad de la Sabana.

Steinley, D. (2003). Local Optima in K-Means Clustering: What You Don't Know May Hurt You. *Psychological Methods*, *8*(3), 294–304. http://doi.org/10.1037/1082-989X.8.3.294

Ullmer, J. (2012). Student Characteristics , Peer Effects And Success In Introductory Economics. *Journal of Economics and Economic Education Research*, *13*(1), 79–87.